



WPI

Development of Predictive and Prescriptive Analytical Models Using Customer, Revenue, and Usage Data

Project Team

Abigael Kihu awkihu@wpi.edu

Michael O'Connor moconnor@wpi.edu

Shiyu Wu swu4@wpi.edu

William Bazakas-Chamberlain wpbazakaschamber@wpi.edu

Project Advisors

Professor Wilson Wong
Department of Computer Science

Professor Robert Sarnie
WPI Business School

Project Co-Advisor

Professor Marcel Blais
Department of Mathematical Sciences

This report represents the work of WPI undergraduate students submitted to the faculty as evidence of completion of a degree requirement. WPI routinely publishes these reports on its website without editorial or peer review. For more information about the project's program at WPI, please see <http://www.wpi.edu/academics/ugradstudies/project-learning.htmls>

Abstract

In collaboration with the WPI MQP team, SaaSWorks is pursuing the development of a predictive model that integrates client data into classifications that enable the identification of key performance indicators correlated with a customer's lifetime value. The model pinpoints which customers display patterns indicative of a high churn risk cancellation and distinguishes them from those periodically churning and reactivating, re-purchasing or re-subscribing. These analytics will be productized into a configurable product feature set that drives strategic decisions within SaaSWorks clients' companies and assists in the automatic discovery of previously unknown key performance indicators.

Acknowledgments

First, we would like to thank our sponsor at SaaSWorks for the amazing opportunity to experience working closely within the Data Analytics field and for welcoming us into your company culture. We are grateful for the correspondence with Eva Shah and Jim O’Neill that helped guide us through our development and business analysis process.

We would like to thank our WPI advisors Professor Robert Sarnie, Professor Wilson Wong, and Professor Marcel Blais, for their academic support and guidance throughout the entirety of this project. Our regular meetings with them helped shape our project's direction and prepared us to be a highly functioning Agile-Scrum team.

Thank you,

Abigael Kihu

Michael O’Connor

Shiyu Wu

William Bazakas-Chamberlain

Executive Summary

Modern service and subscription-based businesses generate tremendous amounts of data throughout the continuous execution of their services and customer intake of those services. The success of those business models is based on the long-term retention of existing customers and recruitment of new customers. The value of a customer's experience with a company is a proportional relationship known as customer lifetime value (CLTV). A company's ability to predict an individual's CLTV and then take steps to increase that metric would enable a business to systematically increase its overall revenue and cement its growth. The largest impact to average CLTV for subscription-based businesses—which typically provide flat-rate or tiered-rate products—is not how much a customer spends at one time but the length of time a customer interacts with the business.

SaaSWorks—a software-as-a-service company that performs data analytics and knowledge discovery for other businesses—is in the development phase of optimizing its modeling and analytical tools to create services that can be generalized enough to fit various recurring revenue business models while also providing actionable context-dependent information to said businesses. They have enlisted the WPI MQP team to develop a model that generates predictive insight for SaaSWorks clients by analyzing their historical data to identify the Key Performance Indicators (KPIs) most correlated with CLTV.

The production of a predictive tool that automatically outputs the specific KPIs unique to individual business models has the potential to yield monumental impact to the growth of all business industries that enlist this service from SaaSWorks. Upon the completion of extensive explorative research and testing under time constrictions, the model feature set will be productized into a configurable and repeatable SaaS service. By providing a tool that enables near-automated identification of the KPIs and metrics most correlated with CLTV, the model can be applied to the business data of any generic client regardless of their industry or target customers. This service will objectively hold immeasurable value to all stakeholders – SaaSWorks clients and SaaSWorks – as it will have a direct and real-time impact on company performance and bottom-line profitability for all parties

The MQP team aimed to develop a system capable of identifying customers' predicted lifetime and the likelihood of a given customer terminating its relationship with a business. The system for determining these values must be robust enough to capture the relationships for various businesses and perform the necessary analysis to accurately select a viable model.

The team elected to employ Agile Scrum as the framework for software development and design. It enables continuous development, collaboration, and reorientation of goals as needed throughout the software creation process. For the development of models capable of leveraging historical data to forecast future behavior and metrics, the team chose to evaluate twelve different machine learning classification models and three models for regression analysis of lifetime. Additionally, principal component analysis (PCA) was also examined as a possible way to improve model accuracy through data dimensionality reduction. The models were repeatedly trained and tested using K-Folds cross-validation ($k=5$) to determine the average accuracy and Matthew's Correlation Coefficient—a special variation of the phi coefficient—to select optimal models. The data utilized in the training and testing was a subscription-based company's user-account per month data from January 2018 - October 2022.

The models were tested using six different data formats:

- First 3 Active Months
- First 3 Active Months with PCA
- Last 3 Active Months
- Last 3 Active Months with PCA
- Last 3 Active Months with Derived Fields
- Last 3 Active Months with Derived Fields and PCA

The formats with principal component analysis (PCA) consisted of six principal components that captured at least 95% of the variance that the true feature space contained. The derived fields that some formats provided were an attempt to capture a nonlinear relationship between the recurring revenue of customers and their usage of the services. For the formats with both PCA and the derived fields, the derived fields were generated before PCA, so they have weights in the components—see Appendix B for full results.

The testing results showed that random forest (estimators = 200) is a sufficient model for determining the classification of a customer as a terminated or active account. The team believes this data to indicate the ability to determine if currently-active customers exhibit behavior closer

to a previously terminated account or an active account. This will allow a given business to determine which incentive programs or other measures they should prioritize in order to encourage the customer to return. The regression attempts to predict customer lifetime were not as successful. Between the two models of linear regression and an artificial neural net with five hidden layers that uses ReLu as the activation function, the mean standard error was 16.6 and 10.7 months, respectively.

The team's work here indicates there are underlying relationships between the dataset and the likelihood of a given customer terminating their account. Future work on this topic is still needed to better identify the exact nature of the said relationship, as the nonlinear relationships between the features clouded the team's ability to identify them in the given timeframe. Work towards developing a more transparent classification model would also allow greater insight into the key performance indicators (KPIs) of all businesses and assist with prescriptive measures.

The team believes the program's efficiency could be greatly increased by adding a specifically formatted schema that automatically manipulates the data as it is made available instead of reformatting at model training time. Finally, a long-term study of the software's true accuracy rate is the recommended approach for future tests, as it currently is only compared to historical data.

Table of Contents

Abstract	ii
Acknowledgments	iii
Executive Summary	iv
Table of Contents	vii
List of Figures	xi
List Of Equations	xii
List of Tables	xiii
Authorship	xiv
1. Introduction	1
1.1 SaaSWorks	1
1.2 Problem Description	2
1.3 Goals	3
2. Research	4
2.1 Software-as-a-Service	4
2.2 Inactivity vs Permanent Churn	7
2.3 Customer Lifetime Value	8
2.4 SaaSWorks Approach to Agile	9
2.4.1 Kanban	10
2.4.2 Agile Scrum Sprint Planning	10
2.4.3 Jira Board Configuration	11
2.5 Competitors	12
2.6 Machine Learning	13
2.7 Regression Models	14
2.7.1 Multivariable Linear Regression (MLR)	14
2.7.2 Multilayer Neural Network	14
2.8 Classification Models	16
2.8.1 Decision Tree	16
2.8.2 Random Forest	18

2.8.3 Support Vector Machine (SVM)	19
2.8.4 Stochastic Gradient Descent (SGD)	20
3. Methodology	21
3.1 What is Agile Scrum?	21
3.1.1 Workflow of Agile Scrum	21
3.1.2 Daily Stand-up	22
3.2 Machine Learning	24
3.2.1 Cross Validation	24
3.2.2 Classification Metrics	24
3.2.3 Regression Metrics	25
4. Software Development Environment	27
4.1 Project Management Software	27
4.1.1 Slack	27
4.1.2 Jira	27
4.1.3 Discord	27
4.2 Source Code Management Software	27
4.2.1 GitHub	27
4.3 Integrated Development Environment Software	28
4.3.1 Python	28
4.3.2 Jet Brain's PyCharm	28
4.4 Data Sources and Database	28
4.4.1 Amazon Workspace	28
4.4.2 DataGrip	29
4.5 Software Tools	29
4.5.1 Visual Paradigm	29
5. Software Requirements	30
5.1 Software Requirements Gathering Strategy	30
5.2 Functional and Non-Functional Requirements	32
5.2.1 Functional Requirements	32
5.2.2 Nonfunctional Requirements	32
5.3 User Stories and Epics	34

5.3.1 Epics	34
6. Software Design	39
6.1 Primary Processes	39
6.2 UML Class Diagram	41
7. Software Development	42
7.1 User Story Formatting	42
7.2 Sprint 0: 10/24 - 10/26	43
7.2.1 Sprint Retrospective	43
7.2.2 Weekly Summary	44
7.3 Sprint 1: 10/27 - 11/3	44
7.3.1 Sprint Retrospective	46
7.3.2 Weekly Summary	47
7.4 Sprint 2: 11/4 - 11/10	48
7.4.1 Sprint Retrospective	50
7.4.2 Weekly Summary	51
7.5 Sprint 3: 11/11 - 11/17	53
7.5.1 Sprint Retrospective	56
7.5.2 Weekly Summary	57
7.6 Sprint 4: 11/18 - 12/1	60
7.6.1 Sprint Retrospective	62
7.6.2 Weekly Summary	63
7.7 Sprint 5: 12/2 - 12/8	69
7.7.1 Sprint Retrospective	70
7.7.2 Weekly Summary	71
7.8 Product Burndown Chart	72
7.9 Software Testing	73
8. Business and Project Risk Management	75
8.1 Risk vs Reward	76
8.2 Risk Culture	77
8.3 Additional Risks	78
8.3.1 Operational Risks	78

8.3.2 Security Risks	78
8.3.3 Market Risks	79
8.3.4 Accuracy Risks	79
8.3.5 Financial Risk	80
9. Assessment	81
9.1 Business Learnings	81
9.2 Technical Learnings	82
9.2.1 Data Management in Python	82
9.2.2 Machine Learning	83
10. Future Work	84
10.1 Increase Data Granularity	84
10.2 Implement a Shared Schema	85
10.3 Project Management	85
10.4 Increasing Interpretability	86
10.5 Forecast Testing	86
11. Conclusion	88
11.1 Method Applied	88
11.2 Result	88
11.2.1 Segmentation of Customer Base	89
11.2.2 Putting Data in Context of Time	90
11.2.3 Relationship with Customer Lifetime Revenue	92
11.3 Learning	94
References	95
Appendix A: PCA Weights	99
Appendix B: Results of Repeated Training and Testing of 12 Classification Models	107
Appendix C: Feature Importance	109
Appendix D: State Segmented Confusion Matrix Model Evaluations	116
Appendix E: Daily Stand-ups	126

List of Figures

Figure 1: Diagram of SaaSWorks Services	6
Figure 2: The Artificial Neural Network Model	15
Figure 3: Multilayer Neural Network Architecture Diagram	16
Figure 4: Example Decision Tree Visualization Created with Graphviz Using Last 3 Months of Active Customer Data (N=100)	17
Figure 5: Support Vector Machine	19
Figure 6: User Persona of the SaaSWorks Analyst Who Will Use the Model	30
Figure 7: User Persona of A WPI Professor Who Will Evaluate Work Using This Paper	31
Figure 8: Use Case Diagram for the Two Primary User Types	31
Figure 9: Process Diagram of Initial Model Selection	40
Figure 10: Process Diagram of Customer Incentive Recommendation	40
Figure 11: UML Class Diagram	41
Figure 12: User Story Formatting	43
Figure 13: Aggregate Confusion Matrices of Classification of Account Status Given First 3 Active Periods (N=9770)	58
Figure 14: Aggregate Confusion Matrices for Classification Models (N=49992)	65
Figure 15: Covid Pandemic Impact on Active Accounts in Different Locations	68
Figure 16: Confusion Matrix for the Stochastic Gradient Model	68
Figure 17: Product Burndown Chart	73
Figure 18: The Heatmap of Correlation Analysis	90
Figure 19: Time Periods Against the Number of Active Accounts	91
Figure 20: Time Periods Against the Churn Rate	91

List of Equations

Equation 1: Customer Churn Rate Calculation	7
Equation 2: Average Revenue Per Customer	8
Equation 3: Lifetime Estimation	9
Equation 4: Customer Lifetime Value	9
Equation 5: Multivariable Linear Regression	14
Equation 6: Calculating A Single Label for A Neural Network	15
Equation 7: Gradient Descent	20
Equation 8: Stochastic Gradient Descent	20
Equation 9: Accuracy Calculation	24
Equation 10: Matthew's Correlation Coefficient	25
Equation 11: Mean Squared Error	25

List of Tables

Table 1: Summary of Software Requirements	33
Table 2: User stories and Epics	34
Table 3: Sprint 1 Completed Story Points	44
Table 4: Sprint 1 Incomplete Story Points	45
Table 5: Sprint 1 Scope Change Story Points	46
Table 6: Sprint 2 Completed Story Points	48
Table 7: Sprint 2 Incomplete Story Points	48
Table 8: Sprint 2 Scope Change Story Points	50
Table 9: Sprint 3 Completed Story Points	52
Table 10: Sprint 3 Incomplete Story Points	52
Table 11: Sprint 3 Scope Change Story Points	54
Table 12: Sprint 4 Complete Story Points	60
Table 13: Sprint 4 Incomplete Story Points	61
Table 14: Sprint 4 Scope Change Story Points	62
Table 15: Model Accuracies	67
Table 16: Sprint 5 Complete Stories	69
Table 17: Sprint 5 Incomplete Stories	70
Table 18: Most Effective Models as Determined by Average MCC and Accuracy	93

Authorship

Chapter	Section	Subsection	Primary Author(s)	Primary Editor(s)
Abstract			William Bazakas-Chamberlain	Abigael Kihu
Executive Summary			Michael O'Connor	Abigael Kihu
Introduction	SaaSWorks		William Bazakas-Chamberlain	Abigael Kihu, Shiyu Wu
	Problem Description		Abigael Kihu	William Bazakas-Chamberlain, Shiyu Wu
	Goals			
Research	Software-as-a-Service		Abigael Kihu	Shiyu Wu
	Inactivity vs Permanent Churn			
	CLTV and CLR			
	SaaSWorks Approach to Agile	Kanban		
		Agile Scrum Sprint Planning		
		Jira Board Configuration		
	Competitors			
	Machine Learning	Overview	Michael O'Connor	Abigael Kihu
MLR				

		SGD	Shiyu Wu	William Bazakas-Chamberlain
		Multilayer Neural Network		
		Decision Tree	Michael O'Connor	Shiyu Wu
		Random Forest		
		SVM		
Methodology	What is Agile Scrum?	Workflow of Agile Scrum	Michael O'Connor	Abigael Kihu
		Daily stand-up		
	Machine Learning Model Testing	Cross Validation	Michael O'Connor	Abigael Kihu
		Regression Metrics		
		Classification Metrics		
	Software Development Environment	Project Management Software	Slack	Michael O'Connor, Shiyu Wu
Jira				
Discord				
Source Code Management Software		Github		
Integrated Development Environment Software		Amazon Workspace		
		DataGrip		
Software Tools		Visual Paradigm	Abigael Kihu	William Bazakas-Chamberlain

Software Requirements	Software Requirements Gathering Strategy		Michael O'Connor	Abigael Kihu
	Functional and Non-Functional Requirements	Functional Requirements		
		Non-Functional Requirements		
	User Stories and Epics	Epics	William Bazakas-Chamberlain	Abigael Kihu
Software Design	Primary Processes		Michael O'Connor	Abigael Kihu
	UML Class Diagram			
Software Development	User Story Formatting		Abigael Kihu	Shiyu Wu
	Sprint 0: 10/24 - 10/26	Sprint Retrospective	Abigael Kihu	Shiyu Wu
		Weekly Summary		
	Sprint 1: 10/27 - 11/3	Sprint Retrospective		
		Weekly Summary		
	Sprint 2: 11/4 - 11/10	Sprint Retrospective		
		Weekly Summary		
	Sprint 3: 11/1 - 11/17	Sprint Retrospective	Abigael Kihu	Shiyu Wu
		Weekly Summary		
	Sprint 4: 11/18 - 12/1	Sprint Retrospective		
		Weekly Summary		

	Sprint 5: 12/2 - 12/8	Sprint Retrospective			
		Weekly Summary			
	Product Burndown Chart		Abigael Kihu	William Bazakas-Chamberlain	
	Software Testing		Michael O'Connor	Abigael Kihu	
Business and Project Risk Management	Risk vs Reward		Abigael Kihu	William Bazakas-Chamberlain	
	Risk Culture				
	Additional Risks	Operational Risks			
		Security Risks			
		Market Risks			
		Accuracy Risks			
		Financial Risk			
Results	Business Learnings		Abigael Kihu, Shiyu Wu,	Abigael Kihu	
	Technical Learnings	Segmentation of Customer Base	William Bazakas-Chamberlain Michael O'Connor		
		Relationship with CLTV			

Future Work	Increase Data Granularity	William Bazakas-Chamberlain	Abigael Kihu
	Implement a Shared Schema		
	Project Management		
	Increasing Interpretability		
	A/B Testing		
Conclusion	Method Applied	William Bazakas-Chamberlain Shiyu Wu	Abigael Kihu, Shiyu
	Results	Michael O'Connor, Shiyu Wu	
	Learnings	Michael O'Connor	

1. Introduction

1.1 SaaSWorks

SaaSWorks is on a mission to democratize the data-driven insights and expertise required to build and manage high-growth, high-value businesses. To provide analytics insights to businesses in various industries, SaaSWorks must have a wide range of data specified algorithms suitable to analyze any company's data. As the start-up expands and grows their business, SaaSWorks is constantly developing and optimizing their modeling and analytical tools to create services that can be generalized enough to fit various recurring revenue business models while simultaneously having enough specifications to precisely perform in-depth analysis and return accurate reports.

SaaSWorks is an analytics service provider to recurring revenue businesses in various industries. Recurring revenue business is any company which makes money on a recurring basis from products or services. These companies subscribe to SaaSWorks to extract, cleanse, unify and enrich their billing, CRM and usage data to, combine, and enrich their invoice, billing, and CRM data to generate a single source of truth. The value of SaaSWorks services comes from their ability to integrate knowledge derived from their clients' multiple streams of context-dependent data analytics into succinct and actionable information.

SaaSWorks is actively developing solutions that maximize their clients' recurring, subscription, and membership revenue growth and Lifetime Value. These solutions are designed to enable their client's company executives to make informed decisions regarding the direction of their company. Ultimately, SaaSWorks will use a client's own historical data to forecast outcomes and address areas within the business which could be improved for optimal customer retention. In pursuit of the growth of the business, SaaSWorks is seeking to develop a Predictive Model that will alert the SaaSWorks client when a customer is at risk of churning — meaning canceling their subscription — so that necessary actions can be taken at an impactful time to prevent churn and retain the recurring revenue of that customer.

1.2 Problem Description

A subscription company's churn rate is the percentage of its customers who cancel their service over a specific period. “The churn rate is calculated by dividing the number of customers whose subscriptions have been canceled over a certain period by the total number of customers at the beginning of that period” (Zoho, 2022). This metric is vital for any business to measure its long-term success since the number of customers churned is directly proportional to the amount of revenue lost. It is a significant factor when determining a company's Customer Lifetime Value (CLTV) which is calculated by dividing Average Revenue per Customer (ARPC) by the churn rate.

With the knowledge that it is more cost-effective to invest in retaining current customers than to acquire new ones, SaaSWorks clients seek services to help them increase their CLTV. The ARPC is the average revenue generated from customers per month, which is found by dividing total revenue by the total number of customers in that given period. When addressing CLTV, a client's ARPC cannot be easily manipulated; thus, they must focus on the churn rate. One of the most important factors a subscription company can consider when it comes to improving its customer retention rate is reducing its churn rate. This value can be used to highlight patterns related to location, seasons, quality of services, subscription plans, and any other underlying factors impacting the retention rates of their customers.

SaaSWorks clients want to track their customer churn over every period, detect which variables directly affect churn, and use the customer behavior patterns associated with those variables to predict which customers have a high probability of churning before they do so. They need this information to identify which segments of a SaaSWorks client's customers should be sent outreach messages called Nudges, and pinpoint the exact time to send them. These Nudges are designed to influence customer behavior by prompting customers to continue paying for and using a company's services. The goal is to yield the greatest impact to topline revenue by reducing the chance of customer churn, boosting CLTV, and increasing the return on customer acquisitions.

Applying analytic tools tailored to a specific client's business involves a tremendous amount of man hours to create. Each client's business model, industry, or target audience affects the algorithms necessary to determine the KPIs related to their CLTV. Traditional software lacks

the ability to automatically create predictive models on a wide variety of data sets regardless of these factors. The team aims to develop such a model that will predict the probability and period of a customer churning which then prompts the SaaSWorks Analyst to send Nudges that will drastically decrease chances of permanent churn.

1.3 Goals

To address this untapped market of competent analytical software, SaaSWorks has collaborated with the WPI MQP team to develop a model that generates predictive insight for SaaSWorks clients. The development process was broken down into multiple steps, beginning with SaaSWorks client's customers being classified based on related demographics. The client's historical data was used to identify factors that affect a segment's CLTV. The plan is to identify the Key Performance Indicators (KPI) most correlated with CLTV and CLR after analyzing a client's dataset then to pinpoint which customer behaviors will drive the most improvement when impacted. These factors determine which customer behaviors are most impactful to CLTV and then generates them into a near-automated process that can be used on a generic client.

Not only will this process involve creating software, but also human intervention. The process will be used on varying clients meaning that the software developed will have different use cases, requiring some human intervention to generate optimal predictive insights. The SaaSWorks Analyst will use the information from the model to develop data-driven playbooks individually designed for each SaaSWorks client to impact customer behaviors, reduce churn rates and increase CLTV.

2. Research

2.1 Software-as-a-Service

Software-as-a-Service (SaaS) companies sell cloud-based software to their customers in a subscription-based approach that generates recurring revenue. They provide and maintain servers, databases, and software that allow their applications to be accessed by their customers over the internet from almost any device. SaaS customers often pay a subscription fee based on the number of features and resources required for the client's application (Levinson, 2007).

A SaaS company's major selling point is the efficiency and accessibility of its centrally managed applications that allow users to utilize their software online rather than installing it on-site. SaaS customers seek the benefits of scalability, saving costs, IT expertise, customer loyalty, increased market potential, and growth in their recurring revenue which are provided through SaaS services (Brook & Zhang, 2020).

SaaSWorks, Inc is a startup data analytics service provider that works with subscription, membership, and recurring revenue businesses to offer SaaS that optimizes their client's revenue, performance, and customer satisfaction. Based in Norwell, Massachusetts, SaaSWorks was founded on May 1st, 2019, by Jim O'Neill and Vipul Shah. The private startup received seed funding on Feb 6, 2020, from Conversion Venture Capital. Currently, SaaSWorks is comprised of 33 employees made up of US based full-time employees, consultants, and 3rd parties that provide flexible resources in South America and Eastern Europe (O'Neill & Shah, 2022). Their goal is to provide a service that combines people, processes, and a proprietary data platform to make sense of their customers' revenue data. Their team combines, cleans, and enriches invoicing, billing, and CRM data to provide straightforward KPIs that help their clients run better businesses. They identify the data field within the client's historical data which yields the most impact on revenue and customer retention to help them understand what the data is telling them and how they can use it to improve business practices. SaaSWorks mission is to democratize the data-driven insights and expertise required to build and manage high-growth, high-value businesses (O'Neill & Shah, 2019).

In the process of analyzing a client's data and identifying areas requiring improvement, SaaSWorks applies a variety of tools to identify, analyze, and organize a client's data to help them build data-driven, durable businesses. These steps complement each other and are used either separately or concurrently. Based on the needs of each client, SaaSWorks will individualize their services to fit their unique business models. These may include but are not limited to RevWorks, AccountWorks, and LeadWorks. For SaaSWorks clients, these services are a critical driver for trusting their data, which all begins with gaining an understanding of what the data means. With RevWorks, SaaSWorks analyzes a client's revenue data to access and stream insights that drive retention and growth. This allows them to build a data pipeline that continually produces tailored KPIs for daily, weekly, and monthly review. A SaaSWorks team uses this data to monitor performance, and reveal revenue drivers, opportunities, and risks. In an attempt to improve relationships and revenue outcomes with a company's customers while operationalizing data-driven engagement, SaaSWorks experts perform extensive research and deliver them to their client's customer-facing teams to help them identify target customers and improve services to reduce chance of churn. They use AccountWorks to access revenue data to power customer focus and priorities. Using LeadWorks SaaSWorks teams examine the client's conversion data by company, lead, and funnel stage to deliver key findings and underpinning data elements that will help them engage in the next phase of growth and provide a complete understanding of the drivers of the conversion and revenue (O'Neill & Shah, 2019). A diagram of the SaaSWorks Services is listed below in Figure 1.

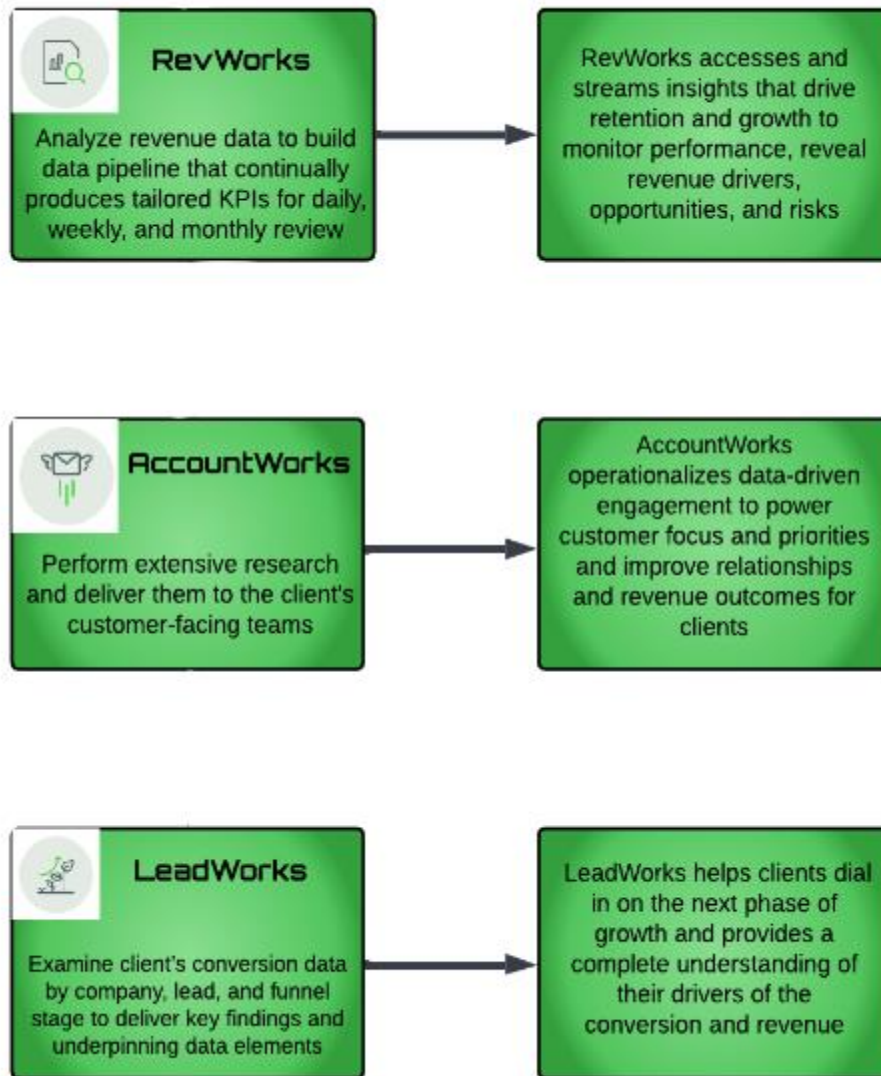


Figure 1: Diagram of SaaSWorks Services

SaaSWorks provides Subscription Companies interested in growing and understanding their business operations using their revenue data. They have built a portfolio of clients with a variety of companies such as Parkville, which operates a SaaS-based customer loyalty program platform to help businesses like gyms, salons, yoga studios, and retail stores create a customer reward program. They have also worked with Sentieo, a financial and corporate research platform that unifies analytics and modeling, market data, and document searches into a single

research management system. Fresh Technology is a SaaSWorks client that builds back-office and kitchen production systems to enable restaurants to process and control costs.

2.2 Inactivity vs Permanent Churn

Subscription based businesses – such as health and fitness studios – experience cycles of inactivity and reactivation from their customers that are documented in their usage and revenue data for each period. These companies record and analyze customer usage data and categorize customers as inactive when they have not attended any classes for one month and their recurring payments have stopped. They will be considered inactive if they surpass the grace period of a full month of consecutive inactivity. Many customers gain a status of inactivity following seasonal patterns, where customers are more likely to be actively attending classes during the summer and inactive during colder, busier months. These patterns also vary depending on the geographic region that the classes are held from various gyms, studios, etc in multiple locations in the US, and the social and economic circumstances in a given period. Certain circumstances such as the Covid-19 Pandemic have the potential to largely interfere with customer activity status and cause unprecedented churn rates. If the customer surpasses the grace period of one month without reactivation, they will be considered churned, as calculated in equation 1. Churned customers equate to revenue loss for subscription companies. These companies turn to SaaSWorks to compute and analyze their churn rates then propose impactful changes to reduce customer churn (Oliveira, 2019).

$$\text{customer churn rate} = \text{customers lost in period} / \text{customers at beginning of period}$$

Equation 1: Customer Churn Rate Calculation

There are various reasons why a customer may churn, involving both voluntary and involuntary churn. When a customer purposefully cancels their subscription, either by skipping paid classes or canceling future classes, they have voluntarily churned. This type of churn is indicative of an underlying problem in the company and customer dissatisfaction of services, unreasonable pricing strategies or simply the procurement of the wrong customers that do not align with the company's values or community. A customer has involuntarily churned if their

subscription is canceled due to a payment failure or decline. This type of churn does not provide useful information to SaaSWorks in terms of user experience of CLTV other than possible overpricing, so the team did not focus on this classification (I.R. Team, 2022).

In subscription-based health & fitness service providers that offer classes with various levels of progression – such as fitness classes that offer beginner, intermediate, and advanced levels of instruction – churn rates are most influenced by customers not being able to graduate to the next learning level. This may be due to certain courses not being engaging enough or customers lacking interest enough to stick through the program and earn graduate status. The beginner class levels exhibit higher churn rates than higher level classes, indicating these classes are not adequately capturing the interest of their customers or providing a service that the younger customers and beginners can deem valuable enough to continually attend.

2.3 Customer Lifetime Value

SaaSWorks has done extensive research to develop methods that can grow their clients' businesses. They utilize Customer Lifetime Value (CLTV) and Customer Lifetime Revenue (CLR) as growth indicators of their client's customer base to understand their client's needs and help them retain long-term customers. The importance of CLTV is that it reflects customer retention and drives customer acquisition cost (CAC), meaning that the higher a company's CLTV is, the lower their CAC becomes which enables them to grow their business quicker. With CAC being comparably more expensive than the costs of customer retention, most companies seek to improve their CLTV so that customers can be active and incurring revenue for longer periods. That being said, there are two ways to increase CLTV; reducing overall churn rates through improvement of Customer Relationship Management (CRM), and increasing the Average Revenue Per Customer (ARPC). ARPC which is the average revenue generated from customers per month, tracked and analyzed in customer segments. The two main ways of increasing ARPC are raising prices, and customer expansion through subscription upgrades and add-ons.

$$\text{average revenue per customer} = \text{total revenue} / \text{total customers}$$

Equation 2: Average Revenue Per Customer

When a customer churns, they cease to be a customer or client of the company, meaning they are no longer bringing in recurring revenue. Estimated lifetime duration (LTE) is the period of time that a customer is estimated to stay with the company prior to churning. This metric can also be used to calculate CLTV using the following equations:

$$\textit{lifetime estimation} = \frac{1}{\textit{customer churn rate}}$$

Equation 3: Lifetime Estimation

$$\textit{customer lifetime value} = \textit{average revenue per customer} * \textit{lifetime estimation}$$

Equation 4: Customer Lifetime Value

CLTV is most applicable to recurring revenue businesses and is most meaningful when tracked and analyzed at a segment level. The CLTV metric is likely to underestimate the lifetime value in recurring revenue businesses, so CLR is used to account for segments of customers that belong to the same weekly, monthly or annual period who churn and reactivate their subscriptions multiple times over their lifetime. CLR measures the average total lifetime revenue generated by each customer in a segment and specific period without leveraging churn rates or count of churn and reactivation. SaaSWorks uses these metrics to analyze their clients' historical data and identify those that most correlate with CLR and CLTV so they can predict areas of improvement to decrease churn rate and increase ARPC (O'Neill & Shah, 2022; Jackson, 2022).

2.4 SaaSWorks Approach to Agile

SaaSWorks has been continuously evolving their Agile methodology as the company and, more specifically, the product development team grows. Through past experience developing start-ups prior to his current role as Co-founder and Chief Technology Officer (CTO)

of SaaSWorks, Jim O'Neill (O'Neill) has applied extensive background knowledge and research into the Agile methodology that SaaSWorks currently practices.

It is understood that the process needed for a team of two developers is different from a team of 5, 10, 25, 50, etc., so SaaSWorks is constantly adjusting their methods to account for the growth that they have experienced in anticipation that it will continue to multiply. More structure is needed once a team exceeds five or more developers where multiple work streams are being developed concurrently. SaaSWorks agrees that with distributed teams, the investment in process and supporting tools and documentation is worth the formal engineering management to avoid miscommunications within the team. To account for the fast-learning cycles of early stage startups, SaaSWorks founders optimize for throughput vs. documentation and prioritize hiring active developers. Early-stage startups also tend to hire senior engineers who have the experience to lend themselves to more self-guided work, but SaaSWorks is inclusive to developers of differing skills ranging from recent graduates to engineers with twenty years of experience (O'Neill, 2022).

2.4.1 Kanban

Agile and DevOps software development are commonly implemented using the Kanban methodology. In addition to complete transparency of work, it mandates real-time capacity communication. Team members can always observe the status of each piece of work thanks to the visual representation of work items on a Kanban board. For an early-stage start-up, Kanban is a fitting framework to facilitate the identification of goals, timelines, and deliverables. When a product is being developed from the ground up, senior early-stage engineers are often not particularly fond of creating tickets and estimates, but they do value them when the product has already been built and new developers are joining who require more protective measures. SaaSWorks started using Kanban in their infancy and then evolved their sprints to build from tasks to detailed Epics in the last 6 months. Some areas of SaaSWorks operations still prefer the structure of Kanban such as the Sec/DevOps team which still currently use this framework.

2.4.2 Agile Scrum Sprint Planning

[Co-founder] is a big believer in using the right parts of the Agile framework that correspond to the size, scale, and maturity of the team. “My opinion is if you are a large

organization going from Waterfall, or Lean, or other processes and adopting Agile, you should go all in as you likely have dozens or hundreds of engineers in the organization” (O'Neill, 2022).

Although this is expected to change over time, SaaSWorks is currently practicing weekly Sprints with periodic backlog grooming. [Head of Product] will start to build more formal backlogs guiding the company to adopt more formal backlog grooming. SaaSWorks plans to begin performing more formal retrospectives once a formal Scrum owner is established in the upcoming weeks. Their present Pod concepts mirror some parts of Scrum but are not enforced. Namely, Pod owners would be Product Owners and there are no assigned Scrum Masters for each sprint. SaaSWorks has not yet adopted the formal “Scrum of Scrums” framework into their methodology, although they expect it to be introduced in the upcoming months once cross-workstream collaboration and dependencies are established (O'Neill, 2022).

2.4.3 Jira Board Configuration

SaaSWorks has made several modifications and custom configurations to different aspects of their Jira Board to accommodate their Agile practices. Story points are not considered appropriate for a SaaSWorks team until a development team knows the product codebase and learns how to accurately estimate them. They are expected to be introduced in the next 6-12 months once the team knows how to estimate and work volumes have increased to a point that requires scheduling beyond what is important and urgent. Senior developers at SaaSWorks have the capability to produce rough estimates and due dates for tasks which are created as Jira tickets and placed into epics to be assigned in each sprint. Developers can push back or move up a ticket between sprints as needed with confidence that they are working as fast as possible (O'Neill, 2022)

The replacement of due-dates has proven favorable to early-stage development where work hours are more intense as deadlines take priority when launching a product and collective work schedules are secondary. This is also partially to avoid story points becoming a gamed system where developers continue to increase the number and reduce their throughput that result in perfect burn-down charts with very little throughput. Used in limited situations and progressively more frequent as SaaSWorks commits to certain client deliverables, due dates are used by the teams (O'Neill, 2022).

SaaSWorks operates in pods which are equivalent to Scrum teams. “The Pod Owner and Primary Developer look over a 4-week timeframe and identify the work to be done, create Jira Epics, and Tasks, and then roughly lay out the work. A Pod currently only has a single Primary Developer and an optional Secondary developer if the work requires additional help. [Head of Product] and I generally partner and oversee to make sure the right work is in the right priority across the right teams at a more macro level” (O’Neill, 2022).

2.5 Competitors

SaaSWorks competitors operate in the market of driving core KPIs like CLTV using predictive analytics. Pecan.AI is a competitor of SaaSWorks that is best known for providing Predictive Churn Rates. Founded in 2016 by Zohar Bronfman CEO and Noam Brezis CTO, Pecan.AI provides services in the form of, “Predictive Analytics Software that Solves Critical Business Problems for Revenue-Driving and Operations Teams” (Bronfman & Brezis, 2016). They drive customer loyalty and retention initiatives for their clients by predicting high risk customers and revenue, detecting 85% of customer churn, and lowering customer churn by 20%. They reported an annual recovery rate growth of 300% in 2021 and 2020 with \$117 Million venture capital raised. Their workforce is made up of 90+ employees who operate in 12 countries worldwide to provide on average 30 million daily predictions to their clients (Bronfman & Brezis, 2016). Their strategy is to use machine learning to predict churn at every stage in the customer journey including Acquisition, Optimization, Customer Engagement, Customer Retention, and the Cross-sell/Upsell stage. Like SaaSWorks, Pecan operates using a four step strategy which consists of ingesting raw data & automated data preparation, model training, business predictions & audience creation, and push to ad platforms. Their clients enjoy the benefits of no long-term contracts, tech integration, simple pricing, time-to-value, and client experience. Pecan has worked with a variety of clients such as ABInBev, DELL, Ideal Image, Johnson & Jonson, Nestle, and Pepsico.

2.6 Machine Learning

High dimensional datasets have an inherent difficulty of being uninterpretable to humans and naive algorithms. Identifying useful patterns and intelligent insights on these large datasets was limited prior to the practical application of machine learning.

Machine learning (ML) is a branch of artificial intelligence that “focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy” (IBM Cloud Education, 2020). In general, machine learning algorithms work through the use of a statistical model that modifies itself during run-time to minimize error in its output as new data is provided to the algorithm, i.e. it *learns* with experience. Based upon its experience, the machine learning algorithm is able to make inferences about data it has not previously examined, and therefore, forecast future behaviors. This enables the generation of predictive models as long as sufficient and meaningful data exists for the model to learn from prior to prediction.

There are two primary forms of machine learning applications: classification and regression. Classification problems consist of dividing data sets into one of a discrete set of groups or categories (Mitchell, p. 54, 1997). Facial and speech recognition, email spam filtering, and object detection are all classic examples of ML classification applications. Regression problems attempt to model the relationships between a feature set and a continuous, numeric output (Dasgupta, 2012). Stock price prediction, market value of a house based on size and location, and predicting the next day’s temperature are examples of regression problems that can be approached with machine learning.

This form of artificial intelligence has seen a significant uptake in adoption across multiple industries—especially business analytics—due to its ability to detect and deliver insights from complex data sets and then accurately forecast future behavior based upon historic data. From a survey of businesses with \geq \$100 million in revenue (n=403), 83% of respondents reported increasing their budgets for AI/ML initiatives in FY 2019-2020 and 50% reported using or developing ML for “generating customer insight and intelligence” (Algorithmia, 2020).

2.7 Regression Models

2.7.1 Multivariable Linear Regression (MLR)

Multivariable linear regression (MLR) is a statistical model building the linear relationship between independent variables and labels (Maulud & Abdulazeez, 2020). Mathematically, a linear regression model can be represented as the relationship between a linear combination of independent variables and a resulting dependent value—see equation 5.

$$y = \beta_0 + \beta_1x_1 + \dots + \beta_mx_m + e$$

Where x_n is an independent variable, β_n is the regression coefficient for x_n , e is the noise, and β_0 is the intercept value

Equation 5: Multivariable Linear Regression (Maulud & Abdulazeez, 2020)

The MLR machine learning model works to estimate the β values—or the weights of the independent variables—by initially setting all of the weights to random values and determining the difference between the predicted dependent variable and the true value, this difference is the model’s error. The model will then modify the weights of each independent variable to minimize the error—this is typically done using gradient descent.

Models are modified repeatedly and as the error decreases, the values that are more closely related to the dependent variable have a higher magnitude weight assigned to them. A hypothetical example of this is using current barometric pressure, the current temperature, and the price of an ice cream sandwich to predict tomorrow’s temperature. Two of those variables are more closely related to tomorrow’s temperature; therefore, will have a greater weight—or coefficient—in the linear regression.

2.7.2 Multilayer Neural Network

The neural network is a popular model in areas like “medical diagnosis, geological survey for oil, and financial market indicator prediction” (Chaudhary & Sharma, 2021). From

Chaudhary and Sharma (2021), a neural network model works similarly as a collection of communicating simple processing units or neurons. To predict a single label, every independent variable is weighted, summed up, and applied to the activation function, and it is expressed in math as equation 6 and figure 2.

$$y_j = \text{activation}(b + \sum_{i=1}^m w_{ij}x_{ij})$$

Where w_{ij} is an element in a weight matrix, x_{ij} is an independent variable, m represents the number of rows in matrix X , n represents the number of columns in X , b represents the error.

Equation 6: Calculating a Single Label for a Neural Network (Chaudhary & Sharma, 2021)

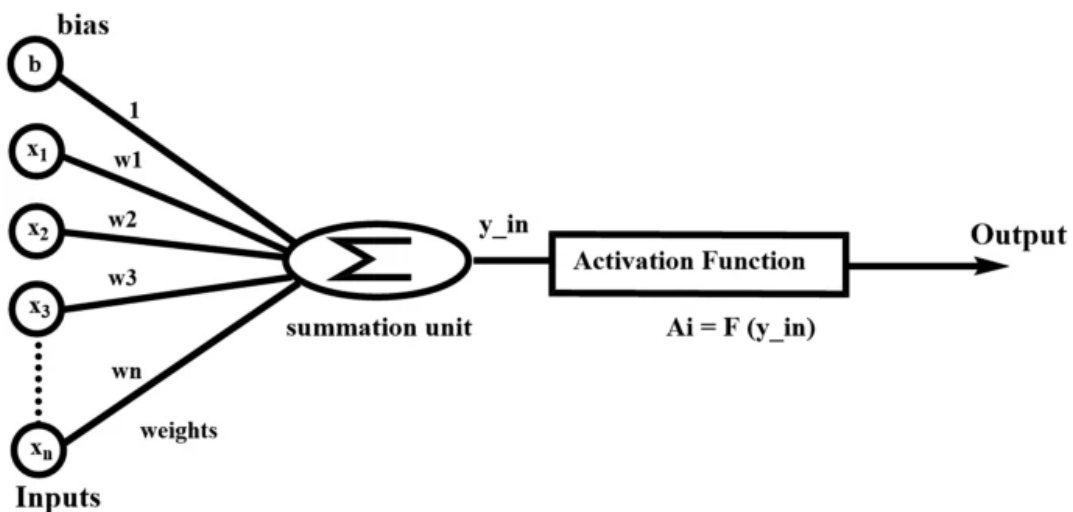


Figure 2: The Artificial Neural Network Model

Note. This picture shows the architecture of the simple Neural Network model. From “Multilayer Neural Network Design for the Calculation of Risk Factor Associated with COVID-19,” by A. Chaudhary & M. Sharma, 2021, *Augmented Human Research*, 6(1), <https://doi.org/10.1007/s41133-021-00044-4>.

Copyright 2021 by Springer Nature.

In this simple neural network model, the x 's are the “input layer”, and the y 's are the “output layer”, and usually a naive neural network model does not contain any hidden layers in its model (Bottou, 2012).

The multilayer neural network is a model with at least one hidden layer between output and input layers (Chaudhary & Sharma, 2021). As shown in Figure 3, a neural network with a hidden layer can be treated as two naive neural networks models combined: Matrix X produces the hidden layer Y as its output after processing the first weight matrix, and Y predicts output layer Z by processing the second weight matrix. For the model with multiple hidden layers, this process will be repeated.

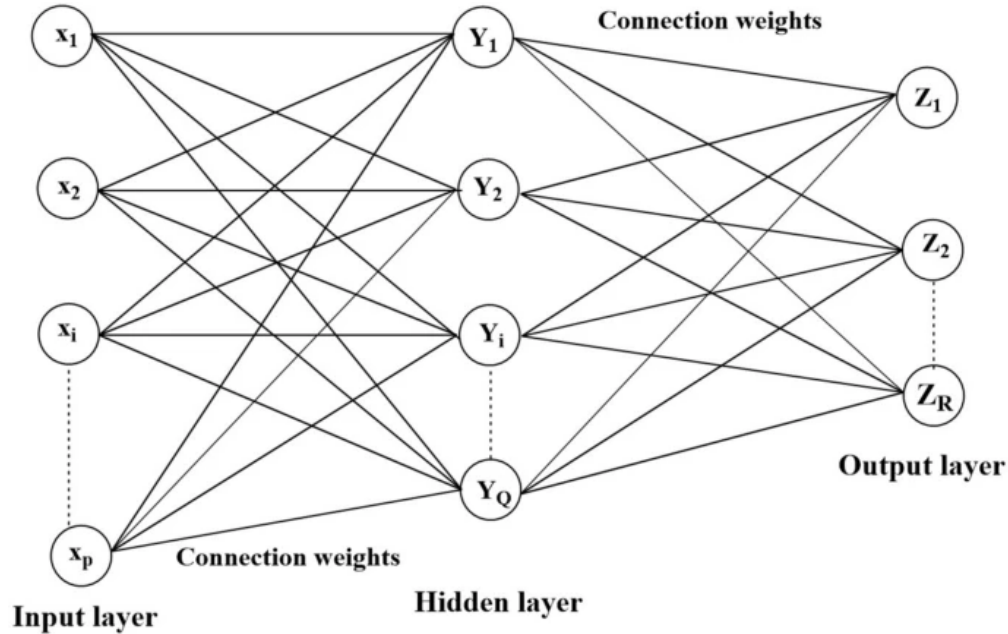


Figure 3: Multilayer Neural Network Architecture Diagram

Note. This picture shows the architecture of the Multilayer Neural Network model. From “Multilayer Neural Network Design for the Calculation of Risk Factor Associated with COVID-19,” by A. Chaudhary & M. Sharma, 2021, *Augmented Human Research*, 6(1), <https://doi.org/10.1007/s41133-021-00044-4>.

Copyright 2021 by Springer Nature.

2.8 Classification Models

2.8.1 Decision Tree

Decision tree classification models—see example in Figure 4—take the shape of directed tree graphs made of a series of nodes and edges. At each node, there is a question with a Boolean statement that directs the flow from the root node to the next level. At the leaf node, the classification is assigned.

The decision tree determines the questions using a criterion function. While there is technically a near infinite number of possible functions, the most common are Gini function, entropy, and the natural log of loss.

The Gini function is the measure of purity for a class after splitting a given attribute. The function is maximized to provide the greatest number of correct classifications at each level of the decision tree (Tangirala, 2020). Levels are added until a minimum level of purity is reached for the classes.

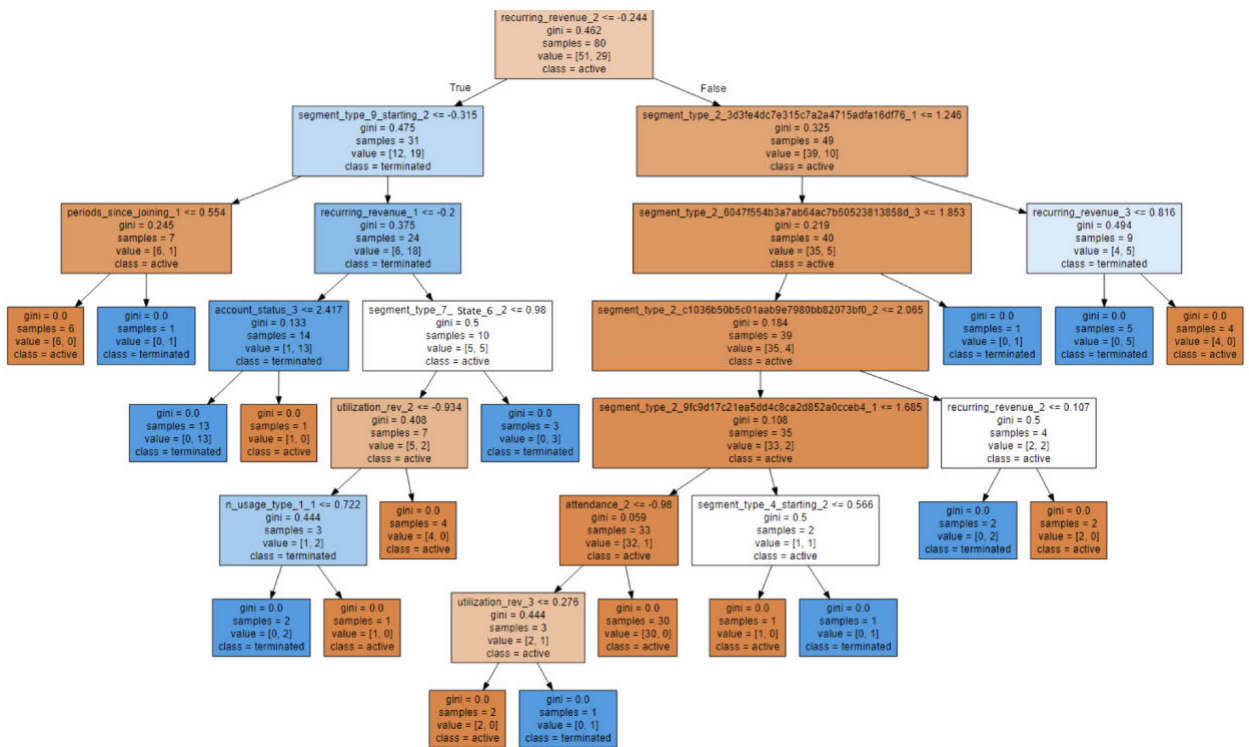


Figure 4: Example Decision Tree Visualization Created with GraphViz using Last 3 Months of Active Customer Data (n=100)

In contrast to the Gini coefficient score, entropy measures the level of disorder after the selection of an attribute. It measures the overall chaos caused by the selection of that attribute and the metric is therefore minimized to improve model effectiveness (Aning & Przybyła-Kasperek, 2022).

Finally, log loss is described as “the negative average of the log of corrected predicted probabilities for each instance” (Seita, 2022). Essentially, log loss works to determine the

probability of a datapoint belonging to a particular classification. It then compares the probability to the actual classification and chooses the attribute split to maximize the probability of the true class.

Decision trees are models that have the capability to handle large amounts of both numerical and categorical data to determine the predicted classification (Abdulazeez, 2021). Additionally, it is the easiest for humans to interpret as it can be processed as a simple flow chart that clearly identifies its decisions. However, decision trees perform poorly on corner cases and special populations as it generalizes classes to promote overall accuracy. An example of this is a decision tree that classifies animals as either dogs or ostriches. If the model chooses to base its decision on the number of legs, it will be overall accurate, but two-legged dogs will be misclassified.

2.8.2 Random Forest

Random forest models build upon the idea of decision trees, but instead of having a single, robust, and highly accurate tree, random forests use a collection of small and shallow trees. It determines its classification based upon a majority vote system.

Random forests also employ the same criterion as decision trees to train and improve the effectiveness of the overall classifier. However, instead of applying the criterion to all of the possible features and selecting the overall best, each tree is provided with a random subset of features and optimizes those. This system ensures that a more diverse feature selection is evaluated during testing and is less likely to misclassify the special population (Schonlau & Jou, 2020).

The main drawback of the random tree model is that it loses its interpretability in comparison to the decision tree model as instead of one set of decisions being made, a hundred or more decisions are being made. However, it often provides better predictions than the decision tree model in exchange. Also, when working with large data sets where the number of independent variables is more than samples, random forests have a better performance than models like logistic regression (Schonlau & Jou, 2020).

2.8.3 Support Vector Machine (SVM)

A support vector machine model produces a hyperplane that precisely separates data based on their different features (Mechelli et al., 2020, pp. 101). The hyperplane's accuracy is shown as having the largest possible distance between the classes and the boundary line—see Figure 5 below. This total distance is also known as the “margin” of the support vector machine model.

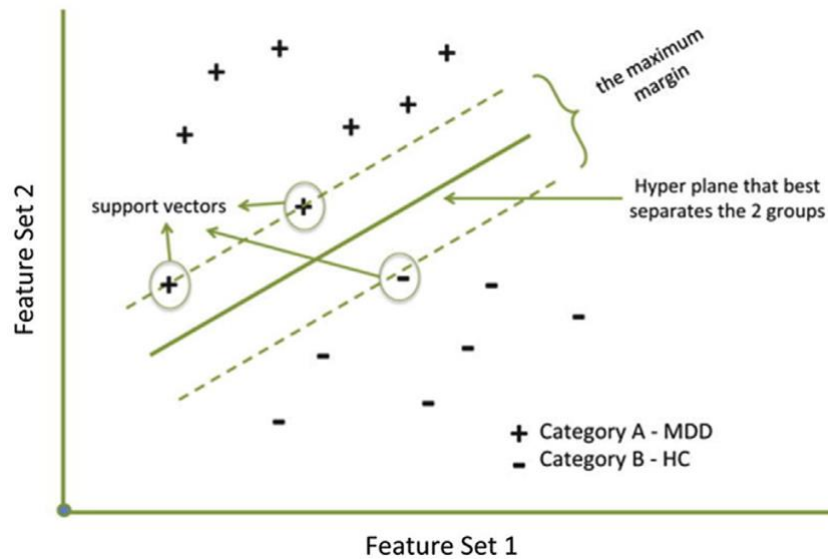


Figure 5: Support Vector Machine

Note. This picture shows how SVM separates data by using a hyperplane.

From *Machine Learning:*

Methods and Applications to Brain Disorders (p. 102),

by Mechelli, A., Pisner, D. A., & Schnyer, D. M. 2020, ScienceDirect. Copyright 2020 by ScienceDirect.

The separating plane can be generated in different forms that vary with effectiveness depending on the application. Typically, the boundary utilizes either a linear model, a polynomial model, or a radial-basis model. A linear model, as the name suggests, is a linear combination of the features such that the space is effectively divided between the classes. A polynomial model will generate a combination of the features from degree of 1-to-n to divide the feature space into the proper classifications. Radial-basis form of the SVM looks at the similarity between the features to order their importance and their overall effect on the classifier.

In comparison to random forest and decision trees this model is substantially harder to interpret and describe its decision-making process. Additionally, SVMs are typically less accurate in data sets with significant numbers of categorical variables. However, they are shown to perform well in high dimensional feature spaces (Purnami et al, 2015).

2.8.4 Stochastic Gradient Descent (SGD)

Stochastic Gradient Descent seeks the function f in order to minimize the total values of loss function, the total distance between the predicted and actual value (Bottou, 2012). To achieve the goal, analysts applied the gradient descent method, which can approach the minimum point by steps based on the value of the slope by using a weight matrix trained by different samples for each element inside, and its iteration follows equation 7 (Bottou, 2012).

$$\omega_{t+1} = \omega_t - \gamma \frac{1}{n} \sum_{i=1}^n \nabla_{\omega} l(z_i, \omega_t)$$

where l is the loss function, γ represents the learning rate, n is the sample size, z represents the coefficient randomly picking different samples with unknown distribution, and ω represents the weight matrix.

Equation 7: Gradient Descent

Following this idea, with a simplification, the new process, the Stochastic Gradient Descent algorithm, randomly selects pairs of independent variables and labels for each iteration instead of every prediction for x_i (Bottou, 2012). The iteration works similarly to the gradient descent, but it will use the same samples to train all weights matrix elements (Bottou, 2012).

$$\omega_{t+1} = \omega_t - \gamma_t \nabla_{\omega} l(z_t, \omega_t)$$

where the meanings of all variables are the same as the Gradient Descent algorithm.

Equation 8: Stochastic Gradient Descent

It then can be processed in the system with a large dataset (Bottou, 2012). In real life, SGD can be used in recognizing and separating the different types of clothes by calculating the weights for every pixel of their photos.

3. Methodology

3.1 What is Agile Scrum?

Pre-2000s, software development hinged primarily on a linear set of steps referred to as the “Waterfall Methodology.” The method consisted of steps that “flowed” from one into another with few loop-backs or checks on the progress. This system led to many products failing due to extended development timelines or changes in customer requirements that the teams could not meet. In response to the continuously changing nature of software development, the Agile Software Methodology was created to assist software teams in their responsiveness and integrate checks in the process. Agile Software Development—or just Agile—is an encompassing term for a collection of frameworks and practices deployed in a software development environment to enable small teams of self-organized, cross-functional peers to work quickly and collaboratively on customer-requirement-driven projects.

The system restructures teams outside the corporate archetype of a manager, assistant manager, and other subordinates. Instead, Agile encourages organizations where team members choose the task they have the skill to complete, i.e., not one person has to take all of the network communication tasks because of a job title; any team member can assist if the need is justified. The flattening of traditional corporate structure prioritizes “individuals and interactions over processes and tools” (Beck et al., 2001).

As these teams are self-organized, the practices they follow must be driven by a team's need to do so. Therefore, there is no “one” Agile methodology. The methods and processes used by any group must be generated, used, and then modified by the methodology's users. Some methodologies have been generalized as frameworks that provide a foundation for teams and then adapt to their needs.

3.1.1 Workflow of Agile Scrum

Frameworks such as Agile Scrum—an implementation of Agile—have teams work in short development cycles called sprints. Scrum also distinguishes itself from other Agile

frameworks by the usage of two unique team member roles, the Scrum Master that oversees Agile-related meetings and overall execution of the Agile process, and the Product Manager (or Product Owner) that works to represent customers and stakeholders' interests in the product.

Before the sprint, a meeting led by the Scrum Master is held to organize all user requirements that have been provided and to prioritize the work. The customer-defined requirements are divided into high-level features or epics. These epics contain significantly more work than the team can accomplish in one or even multiple sprints. Therefore, they are broken down further into small chunks called “user stories” that provide value to the product on their own and can also be developed within a single sprint. These stories are given “story points” that measure their overall difficulty in completely implementing. Typically, Scrum Masters will make note of any individual stories that are rated as larger than a single sprint and have the team break them into smaller parts again later in the meeting. All of the user stories are stored in a prioritized list known as a product backlog.

At the beginning of the sprint, developers meet to establish the work to be completed during the development period. During this planning session—called sprint planning, the Product Manager generates a list of related, high-priority tasks from the product backlog that the team has the capacity to complete within the given period. The team analyzes the list, makes suggestions for additional tasks or removal of specific stories, and develops an initial plan for accomplishing the agreed-upon stories. These stories are ordered into a sprint backlog and assigned to members. During this meeting, the Scrum Master works to ensure that the entire team agrees with the initial plan for each major sprint objective and that all appropriate backlog items are included in the sprint backlog (Schwaber & Sutherland, 2020).

3.1.2 Daily Stand-up

As development progresses through the sprint, the team meets once a day for an extremely brief meeting called the “daily stand-up” or “daily Scrum.” This meeting entails the Scrum Master asking each member of the team three questions: what they have accomplished since the last meeting, what will they accomplish before the next meeting, and is there anything that will prevent them from accomplishing their goal. The Scrum Master makes notes of anything that will hinder performance—known as blockers—and will communicate directly with

the necessary team member(s) to remove the obstacle. Additionally, this meeting communicates the status of each team member's progress to the rest of the team to ensure everyone is aware of the project's position. The Product Owner may also take time to do an overview of the sprint backlog or Kanban board to ensure it is up to date with current tasks.

After the sprint, the Scrum Master leads the Sprint Review meeting with the entire Scrum team. At this time, the Product Owner communicates the total amount of progress completed toward the final product during the sprint. This is typically done through a sprint burn-down chart that displays the number of user-story points that were completed relative to the total number of points remaining in the backlog. The developers analyze their overall progress towards completion of the software, communicate with stakeholders to ensure work is moving towards their requirements, and allow an opportunity for the team to modify the product backlog based on new information discovered during the sprint or provided by the stakeholders (Schwaber & Sutherland, 2020).

The final event of any sprint is the sprint retrospective meeting. The team members review the highs and lows of the last sprint and discuss the conditions that caused them to occur. The team identifies ways to increase the effectiveness in regard to processes, individuals, tools, or even acceptance criteria of stories. This may result in an update of the team's workflow or even additions to the next sprint's backlog. The cycle then continues for the next sprint until the project is brought to completion.

3.2 Machine Learning

3.2.1 Cross Validation

Machine learning leverages a phenomenal amount of mathematical power to unearth previously undetected patterns and relational systems in data that can inform decision-making and planning. However, no system is 100% effective in modeling consumer phenomenon. To determine the overall effectiveness of the system and have a basis for comparison from one model to another, the team assessed the models using K-Fold Cross Validation. Generally, cross-validation is seen as a “technique for evaluating ML models by training several ML models on subsets of the available input data and evaluating them on the complementary subset of the data” (Alpaydin, 2021).

Each row of the dataset was divided into one of five “folds” or groups. From the five groups, one was withheld as a test set, and the models were trained on the remaining. The models were then fed to the test dataset, and their accuracies were recorded. This process was repeated four additional times by withholding a different fold each time. The overall success of a model was described using the average accuracy from the five cycles.

3.2.2 Classification Metrics

For the classification models generated, two primary metrics were used to determine the effectiveness of the predicted classes: accuracy and phi correlation coefficient. Accuracy of the models is defined as the ratio of the sum of true positive, true negative assignments, and the total number of assignments as shown in equation 9.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where $TP = True\ Positive$, $TN = True\ Negative$, $FP = False\ Positive$, and $FN = False\ Negative$

Equation 9: Accuracy Calculation

Accuracy is a typically accepted measure of success that is simple to interpret and present. However, a secondary measure was utilized as accuracy is less effective in imbalanced class distribution scenarios as it tends to favor the majority class (Branco et al, 2015). An example of this phenomenon is found in medical image classification. If a model is choosing which X-ray contains a tumor and only 3% –an arbitrarily small number–of X-rays are positive for tumors, a model can just assign all images as “tumor free” and have an accuracy of 97% but it is a fundamentally flawed model.

When performing segmentation of the customer data set, several of the feature space subsets became imbalanced and therefore accuracy was an ironically inaccurate metric. To supplement accuracy in some distributions with significantly imbalanced distributions, Matthew’s Correlation Coefficient was used to measure model success in capturing the relationship between the feature space and the true label as it equally prioritizes both classes of the binary classification (Chicco et al, 2021).

$$MCC = \frac{Tp * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where $TP = True\ Positive$, $TN = True\ Negative$, $FP = False\ Positive$, and $FN = False\ Negative$

Equation 10: Matthew’s Correlation Coefficient

3.2.3 Regression Metrics

The regressions models generated were evaluated and compared utilizing two different metrics: mean squared error (MSE) and Coefficient of Determination (r-squared). Mean-squared error is the average squared error between the predicted value and the true value. As MSE increases, the effectiveness of the model to capture the true behavior of the data decreases.

$$MSE = \frac{1}{n} \sum_1^n (V_{predicted} - V_{true})^2$$

Equation 11: Mean Squared Error

R-squared was used as an additional metric and considered more *sensitive* than MSE (Jierula et al, 2021). The combination of the two enables a system that is unlikely to result in a tie when ranking effective models.

4. Software Development Environment

4.1 Project Management Software

4.1.1 Slack

Slack and Jira were the main agile software used by the MQP team.

Slack enables teams to communicate with each other quickly and in a more professional setting than most instant messaging platforms. As it is used in more than 750,000 businesses, it is a well-established and vetted technology (Slack, n.d.).

4.1.2 Jira

Jira provides features to track sprint and product backlogs, assign tasks and stories to individual members, automatically generate sprint progress charts (such as the burndown chart), and allow for customization of story information to meet specific team needs. Jira helps teams to know the progress and plan for the future at all times.

4.1.3 Discord

Discord is an instant messaging and voice chat platform. It allows the team to quickly communicate about the specifics of what they are working on throughout the workday and easily share files and information. Messages stay organized and relevant with multiple chat channels for different topics.

4.2 Source Code Management Software

4.2.1 GitHub

Git and GitHub were the primary code-base management systems utilized. Git (version 2.37.3) is an open-source program used in industry, education, and hobby communities for version control software development projects. When combined with GitHub, an online platform

used to remotely store code repositories, Git allows for teams of any size to work asynchronously on the same project.

4.3 Integrated Development Environment Software

4.3.1 Python

With over 57% of machine learning developers and data scientists using python for their programs (Voskoglou, 2019) due to its readability and robust libraries for advanced data analytics, Python (3.10) is the primary tool that will be used for the scripts written for this MQP. Additionally, several Python libraries are being reviewed for application. Specifically, NumPy, Sklearn, and TensorFlow will be tested for effectiveness when developing models.

4.3.2 Jet Brain’s PyCharm

Jet Brain’s PyCharm IDE was used as the text editor and integrated development environment for the team due to its native debugging tools that allow users to quickly isolate and analyze program data, profile memory and compute resource usage, and the team’s familiarity with the software.

4.4 Data Sources and Database

4.4.1 Amazon Workspace

The project Sponsor, SaaSWorks, was providing an Amazon Web Service’s Postgres database that contains all information that they can release to the team from their client’s sources. This data includes revenue and usage data, organized by reporting period, and numbered by account. [REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]



4.4.2 DataGrip

The team was accessing and querying the database using SQL via JetBrains DataGrip IDE due to its well-rounded set of programming tools and on the suggestion of the project's sponsor. The database entries may also be analyzed and parsed using Microsoft Access for quick, one-off analysis.

4.5 Software Tools

4.5.1 Visual Paradigm

Visual Paradigm is a professional diagram software that offers Agile and Scrum project management visualization tools. The team utilized their process mapping and roadmap tools to create a Use Case Diagram for the two primary user types, the MQP professor and the SaaSWorks Analyst.

5. Software Requirements

5.1 Software Requirements Gathering Strategy

Software requirements were primarily gathered from informal interviews with SaaSWorks as they defined the gap in their capabilities. The interviews were conducted via virtual meetings as SaaSWorks illustrated the business need of detecting when clientele are likely to exit programs and the rewards of retaining the clients.

Secondary to the informal interviews, the MQP team worked internally to develop requirements by brainstorming the necessary steps for classification and regression tasks, and the systems that needed to be in place to accomplish the data. The team utilized user personas—demonstrated in Figure 6 and Figure 7—in the brainstorming process to think through the typical usages and necessary features to accomplish those tasks. From those personas, the standard use cases—Figure 8—of the two primary user types were developed and used to determine more specific requirements for the software.


 <p>Alyssa</p>	<p>Alyssa Cote is a Health and Wellness Consultant at SaaSWorks.</p> <p>She has a Bachelor of Science focused in Public Communication from University of Vermont. Alyssa is an experienced Sales and Operations Manager in the health, wellness and fitness industry. She is skilled in Sales, Management, Recruiting and Training/Development.</p> <p>She has 5 years of experience working at OrangeTheory Fitness managing brand compliance, performance, sales, and marketing successes. She was responsible for administering monthly performance audits to increase revenue, member quantity, and qualities of service in studios along the entire Greater Boston region.</p>	<p>At SaaSWorks, Alyssa is responsible for customer outreach and retention of existing customers through the coordination of reminder messages sent in strategic intervals to influence customer behavior.</p> <p>She needs a predictive model that will help her identify which pockets of a SaaSWorks client's customers should be sent outreach messages and pinpoint the exact time that will yield the greatest impact to topline revenue and reduce the chance of customer churn.</p> <p>She wants to use the information from the model to develop data-driven playbooks designed for SaaSWorks Clients to implement in efforts to impact customer behaviors, increase CLTV and reduce churn rates.</p>
--	---	--

Figure 6: User Persona of the SaaSWorks Analyst Who Will Use the Model

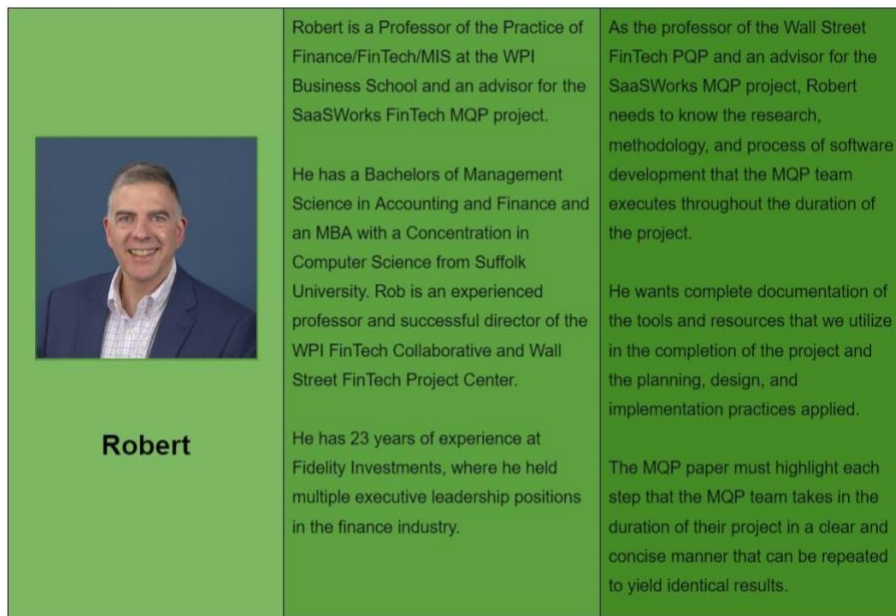


Figure 7: User Persona of a WPI Professor Who Will Evaluate Work Using This Paper

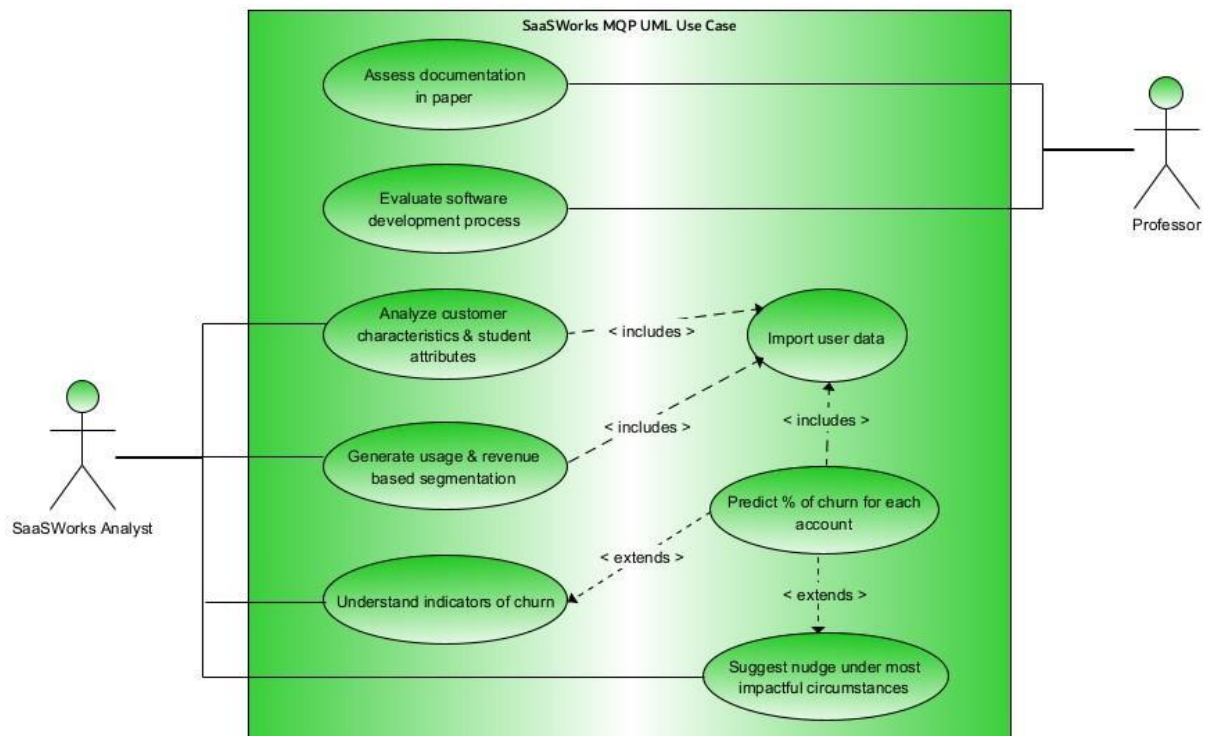


Figure 8: Use Case Diagram for the Two Primary User Types

5.2 Functional and Non-Functional Requirements

The requirements for the software were designated as one of two categories: function and nonfunction requirements. The team defines functional requirements as specifications given directly by the customer. Nonfunctional requirements include the necessary constraints placed on the product to ensure quality is maintained in the product. All requirements are summarized in Table 1.

5.2.1 Functional Requirements

SaaSWorks requested that the software utilize their custom data warehouse generated hosted on Amazon Web Services (AWS) PostgreSQL Database as our primary source of data. With the given data, the company needs the ability to be able to identify the likelihood of a customer ending their contract or relationship with their clients and then predict the outcome of retaining or failing to retain that group of customers. SaaSWorks also intends to integrate this product with pre-existing software used in their daily customer analysis process so it will need to have an output that can be easily assessed automatically by other software.

5.2.2 Nonfunctional Requirements

The MQP team agreed on designing the software using a class-based approach to ensure proper encapsulation of data and easier compatibility when other developers (both internal to the team and external) expand upon sections of the software.

With all large datasets, the data will need to be cleaned and normalized to a certain degree to allow for cross-distribution comparisons and to simplify machine learning development. Additionally, the true behavior and exact equation to determine the likelihood of a consumer ending their subscription is unknown and will vary from one business to another. Therefore, a series of models with varying parameters will need to be trained and assessed before an optimal model can be selected. Once a model is selected, infrastructure must be in place to use the model to predict what customers are at risk of ending their subscription.

Functional Requirements
Communicate with AWS PostgreSQL Database to pull data.
Identify customers at risk of leaving a client's business.
Determine the relationship between RFM and CLR on a customer's likelihood to continue their membership.
Prescribe interventions that positively impact a customer's interactions with the client.
Create a system that can be generalized beyond the singular dataset that is being used
Produce easily parsable files and media that can be used for external programs
Nonfunctional Requirements
Generate the software utilizing object-oriented programming
Normalize data for cross-distribution comparison
Handle irregularities in the data set
Automatically compare several models for best selection
Put in place architecture for both initial analysis and everyday reports

Table 1: Summary of Software Requirements

5.3 User Stories and Epics

5.3.1 Epics

The requirements of the final product were summarized into five primary epics based on the software requirements and requirements of the MQP. The epics provided a foundation for the team to further refine the goals into user stories that needed to be completed during development. The team developed user stories that identified steps and features that added value individually and collectively to the final product. Table 2 lists the user stories and their associated epics.

Sprint Completed	User Story	Story Points
Epic 1: Data Preprocessing		
1	As an analyst, I need to be able to perform sophisticated analysis on the data in Python to discover insights.	1
1	As an analyst, I need to clean the data so I can report useful metrics to my customers.	5
1	As an analyst, I need to be able to run models which only take input of numerical data on categorical data.	1
1	As an analyst, I need the lifetime of a customer for calculations and predictions.	2
1	As an analyst, I need the revenue of a customer for calculations and predictions.	1
3	As an Analyst, I need to analyze time-dependent data to foresee trends.	2
4	As an analyst, I need to target and label customers the period right before the period they churn.	1

1	As an analyst, I need to group the customers based on a wide variety of characteristics and behaviors so I can identify the most impactful customer categories.	1
4	As an analyst, I need an interface to interact with my predictive modeling scripts.	3
N/A	As an analyst, I need to identify customer attributes that will segment the data to remove noise.	3
Epic 2: Data Analytics and Trend Identification		
2	As an analyst, I need to find the potential relationship between time period and other variables.	3
2	As an analyst, I need the ability to manually identify trends in data.	2
N/A	As an analyst, I need to visualize the relationship between variables on a graph.	1
N/A	As an analyst, I need to visualize the relationship between variables on a graph based on segmented data.	2
1	As an analyst, I want to be able to view long term customer behavior patterns.	3
1	As an analyst, I need to identify the range of a given column of data.	1
1	As an analyst, I need to be aware of outliers in my data set so I can identify special cases and behavioral anomalies.	2
1	As an analyst, generating summary statistics allows me to get a better understanding of the data.	2
2	As an analyst, I need to compare lifetime revenue between different segments of customers.	2
2	As an analyst, I need to visualize the lifetime revenue for different segments of customers.	2

4	As an analyst, I need to determine if a relationship exists between attendance rate and account termination.	3
5	As a developer, I need to linearize data to apply it to machine learning models.	3
3	As an analyst, I need to categorize customers as higher and lower value customers.	3
3	As a developer, I want to reduce the dimensionality of the data to reduce compounding error in machine learning models.	1
Epic 3: Predictive Modeling		
2	As a developer, I need to reformat the database output to be interpretable by machine learning models.	1
2	As an analyst, I want to filter out predictors that do not correlate with customer lifetime revenue.	2
2	As an analyst, I need to understand the deciding power of different features in my classification model.	2
N/A	As an analyst, I want to identify correlations with customer lifetime revenue and other numerical metrics.	5
4	As an analyst, I need to determine the effectiveness of decision trees as a classification model so I can choose an optimal model.	3
4	As an analyst, I need to determine the effectiveness of random forests as a classification model so I can choose an optimal model.	2
3	As an analyst, I need a way to assess the results of the machine learning training and testing.	1
4	As an analyst, I need a way to visualize the results of the machine learning training and testing.	1

2	As an analyst, I need to determine if there is a significant relationship between categorical data and churn.	3
4	As an analyst, I need to compare trends of a customer with accounts of similar value.	3
4	As an analyst, I need to analyze significant differences between active and terminated accounts.	3
3	As a developer, I need to convert segmented customer data into machine learning model ready data.	1
3	As an Analyst, I need to be able to predict if an account will take an extended period of inactivity from the company.	3
4	As an analyst, I need a model to predict the lifespan of each account version.	5
4	As an analyst, I need to analyze the last or most recent months of a customer's lifespan for churn prediction.	3
3	As an analyst, I need to check for correlation between variables using linear regression.	3
3	As an analyst, I need models that can accurately determine which customers are likely to step away from the client so I can identify which customers need Nudges.	1
3	As an analyst, I need to determine if a random forest model is optimal for determining which customers are likely to step away from the client so I can identify which customers need Nudges.	1
3	As an analyst, I need to remove collinear variables, so the models are more interpretable.	2
4	As an analyst, I need to train and examine models as quickly as possible so I can determine the best classification system.	2

Epic 4: Prescriptive Modeling

N/A	As an analyst, I need to relate real world events and customer behavior with key metrics to determine what course of action the business should take.	8
N/A	As an analyst, I need to simulate the impact of improving metrics and report a dollar value return on improving metrics so I can generate suggestions that primarily affect those metrics.	5
Epic 5: Documentation		
3	As a professor, I need to understand how the team operated and progressed as an Agile Scrum Team in their MQP to assess their performance.	6
4	As a professor, I need a good understanding of the background and context of the project.	2
4	As a professor, I need to learn the context and research for a better understanding of the project.	2
4	As a professor, I need to evaluate the methods used during the project's completion.	3
4	As a professor, I need to validate the software used during the project's completion.	5
4	As a professor, I need to understand the stakeholders' project goals to ensure they are completed.	2
5	As a professor, I need to evaluate the project group's approach to the problem.	2
5	As a professor, I need to understand the project group's software development workflow.	8
5	As a professor, I need to understand the impacts and risks the project has on the business stakeholder.	4
5	As a professor, I need to evaluate the project group's reflection on their project.	2

5	As a professor, I need to ensure the project group has long term and out of scope ideas.	2
5	As a professor, I need to evaluate the overall result of the project as seen by the project group.	2

Table 2: User stories and Epics

6. Software Design

6.1 Primary Processes

The team has identified two primary processes that the software undertakes during its usage. Firstly, it intakes data and derives a statistical model to determine the likelihood of a customer discontinuing their services with a SaaSWorks client– as pictured in Figure 9. Secondly, on a daily basis, SaaSWorks will deploy the software to determine which of their customers exhibit characteristics of a “likely to terminate” customer and would therefore benefit from an incentive to return to the business– as pictured in Figure 10.

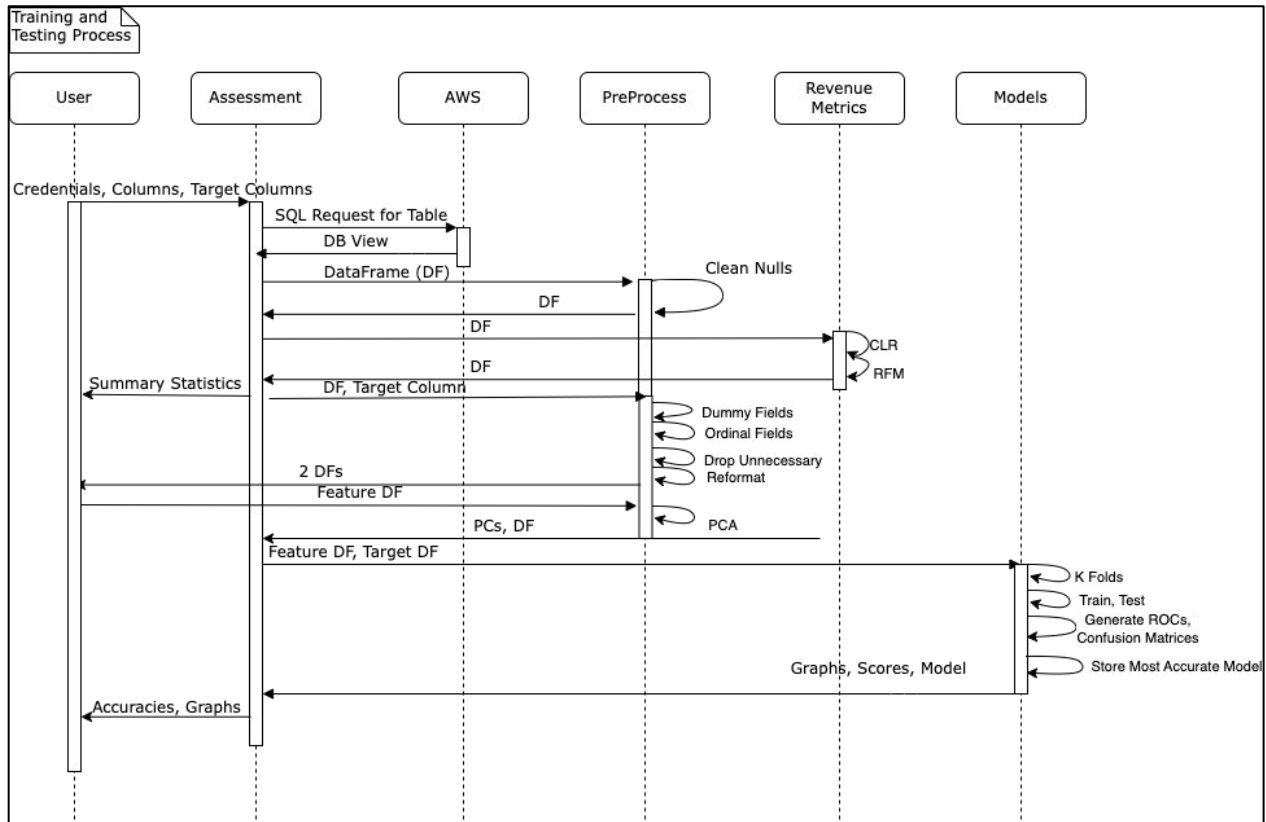


Figure 9: Process Diagram of Initial Model Selection

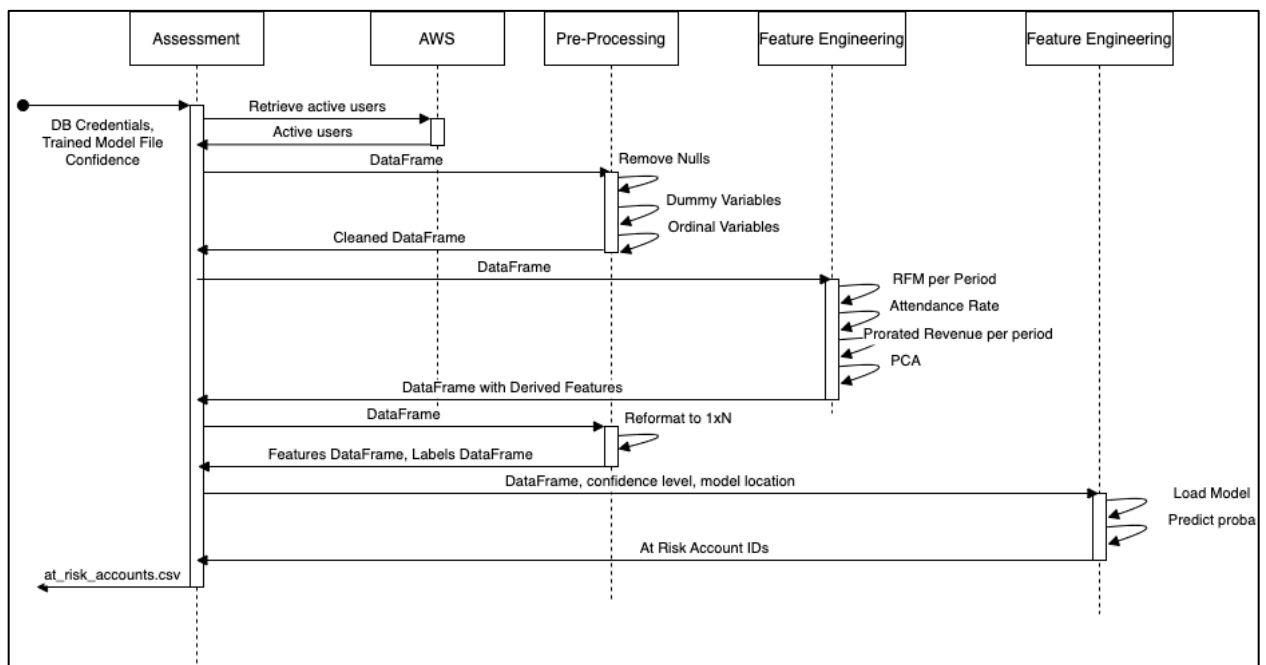


Figure 10: Process Diagram of Customer Incentive Recommendation

6.2 UML Class Diagram

From a software architecture perspective, the primarily deliverable of this project is uninteresting. The overall series of classes– depicted in Figure 11– are essentially independent of one another except for the main class that acts as the pipeline for pushing data from one class to another. The system does not require substantial flexibility or program branching so there was no rationale for a complex class structure.

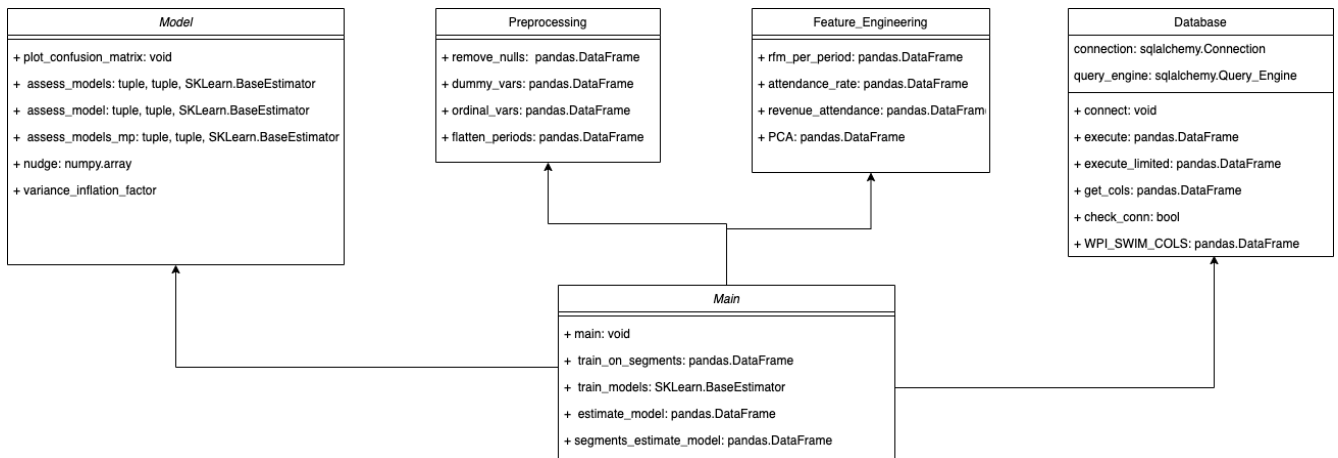


Figure 11: UML Class Diagram

7. Software Development

The team developed a routine of meeting in person on Thursdays and Mondays when we had an agenda that required collaboration and in-depth discussions. Daily stand-ups took place at 10:15 am on every working day. The team ran on Thursday-to-Thursday Sprints and worked daily from 10 am to 4 pm. Often we worked individually outside of our typical working hours to reach goals and deadlines outlined in the Jira Board. On Thursdays, we performed our sprint retrospective following the completion of a sprint and met with our MQP advisors for a weekly check-in. On Mondays, the team joined the SaaSWorks weekly kickoff meeting on zoom and some team members attended weekly individual meetings with their department advisors. On Tuesdays, Wednesdays, and Fridays, the team worked remotely, remaining on-call and available to meet from 10 am to 4 pm. Our regular meetings with SaaSWorks include 10 am Weekly Kick-offs on Mondays, 11 am Weekly Check-in meetings with Eva and Jim, 10:30 am Weekly Development Planning on Thursdays, 12 pm Friday Fun Demonstrations, and one on one meetings with Eva and Alyssa at 10:30 am on Fridays to discuss business and stakeholder analysis. The team occasionally scheduled meetings with advisors for guidance and strategy advice.

7.1 User Story Formatting

Our user stories are formatted on Jira using the configuration illustrated in Figure 12. A story can be either part of an epic or its own separate issue, but both typically include a title, description, status, story point, assignee, and child issue(s) if any.

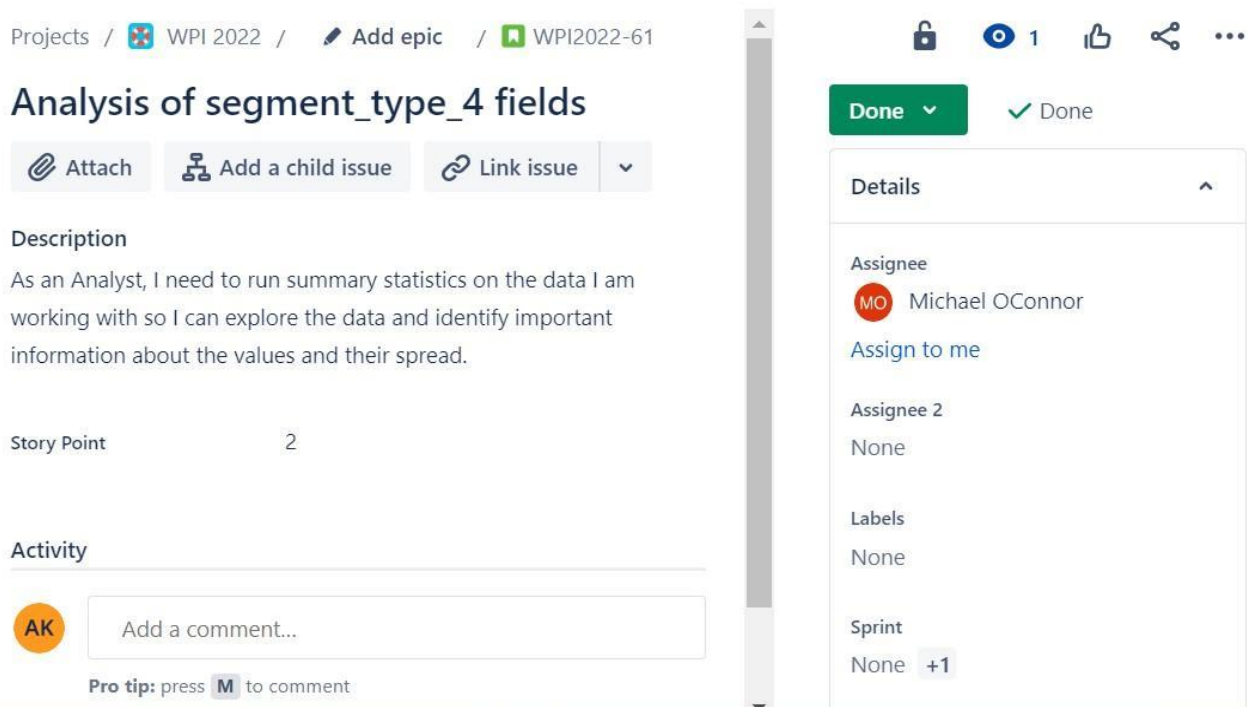


Figure 12: User Story Formatting

7.2 Sprint 0: 10/24 - 10/26

7.2.1 Sprint Retrospective

During our first sprint, the installation of software packages went smoothly including the setting up of Jira and the creation of epics and user stories. We had multiple interactions with our sponsor through planned meetings, emails, and Slack correspondence to discuss the dataset and project goals which have all been very insightful. All our meetings with our professors and advisors have equally been successful and helpful toward the progression of our project. We put in extensive time and effort towards initial approach research to gain a deeper understanding of our data and the techniques we can use to garner the most success.

Improvements could be made in the areas of individualized work which were undefined at the moment due to the uncertainties in the project as we were still in the planning phase. Some team members still needed to be familiarized with tools such as Git Hub, SQL, and Python and its classes. A system needed to be created to send out status updates for SaaSWorks, and we

needed to create a more efficient weekly meeting schedule with SaaSWorks to ensure that all meetings we attended were relevant to the progression of the project.

We utilized online resources and peer assistance to educate teammates on the usage of platforms and tools like GitHub, SQL, and Python. Online examples and practice repositories in GitHub were examples of learning techniques we implemented. To address the status update issue, we sent a snippet of our implementation sheet at regular intervals to communicate our progress with SaaSWorks. Additionally, we met with SaaSWorks to review which meetings require our attendance and input.

7.2.2 Weekly Summary

As a team, we utilized this first week of our project to conduct research about the software, platforms, and tools that we will be using such as Agile, Jira, GitHub, SQL, and Python. We began pre-processing with database exploration and knowledge discovery, as well as extensive planning through Epics and user stories on Jira to the capacity that is currently possible. We conducted multiple meetings with SaaSWorks to inquire about the data sample and discuss project goals and met with our advisors for further guidance when needed. We established spreadsheets and modeling tools to support documentation and implementation efforts. Individually, we were assigned user stories which we began working on at the end of the week.

7.3 Sprint 1: 10/27 - 11/3

Planned Story Points: 29

Completed Story Points: 20

Table 3 lists the completed user stories, the story owner, and story Points from Sprint 1.

Sprint 1: 10/17 - 11/3			
User Story	Story Owner	Story Points	Key
Link correct database to Python	Michael O'Connor	1	WPI2022-18

Remove or fill the holes on the working dataset	Shiyu Wu	5	WPI2022-19
Check for outliers	Shiyu Wu	2	WPI2022-20
Find a lifetime of a customer	William Bazakas-Chamberlain	2	WPI2022-34
Find total revenue of a customer.	Abigael Kihu	1	WPI2022-33
Convert categorical data to discrete numerical.	Shiyu Wu	1	WPI2022-31
Apply More Advanced Query Patterns to DB	Michael O'Connor	3	WPI2022-21
Generate summary statistics for each field	Shiyu Wu	2	WPI2022-54
Find range (spread) for each column.	Shiyu Wu	1	WPI2022-36
Finalize Use Case Diagram	Abigael Kihu	Task	WPI2022-58
Implementation Spreadsheet	Abigael Kihu	Task	WPI2022-60
Analysis of segment_type_4 fields	Michael O'Connor	2	WPI2022-61

Table 3: Sprint 1 Completed Story Points

Incomplete Story Points: 13

Incomplete stories and tasks were automatically added to the Sprint 2's backlog. Table 4 lists the incomplete user stories, the story owner, and story Points from Sprint 1.

Sprint 1: 10/17 - 11/3			
User Story	Story Owner	Story Points	Key
Find average CLR for segments.	William Bazakas-Chamberlain	2	WPI2022-35
Research what kind of normalization we need for different models	N/A	Task	WPI2022-30
UI for selecting desired fields	Michael O'Connor	3	WPI2022-55

Generate an RFM score for individual customers, avg for segments	William Bazakas-Chamberlain	3	WPI2022-56
Identify factors to segment customers by.	N/A	3	WPI2022-23
Generate Segments	William Bazakas-Chamberlain	1	WPI2022-10
Find mean, median, and mode for each numerical feature.	Shiyu Wu	1	WPI2022-37

Table 4: Sprint 1 Incomplete Story Points

Scope Changes - Story Points Added During Sprint: 4

Table 5 lists the additional story points, the story owners, and their story Points from Sprint 1.

Sprint 1: 10/17 - 11/3			
User Story	Story Owner	Story Points	Key
Find range (spread) for each column.	Shiyu Wu	1	WPI2022-36
Finalize Use Case Diagram	Abigael Kihu	Task	WPI2022-58
Implementation Spreadsheet	Abigael Kihu	Task	WPI2022-60
Analysis of segment_type_4 fields	Michael O'Connor	2	WPI2022-61
Find mean, median, and mode for each numerical feature.	Shiyu Wu	1	WPI2022-37

Table 5: Sprint 1 Scope Change Story Points

7.3.1 Sprint Retrospective

During Sprint 2, the team handled working remotely very well in response to our IDs losing access to 50 Prescott for most of the week. We acknowledged that we split up the work in

terms of Jira tasks and user stories evenly, and with additional peer assistance and VC calls to have group discussions, we were able to complete most tasks on Jira on-time, despite the lost day of work due to database access issues. We developed many well-structured business and software requirement questions to ask SaaSWorks which led to many clarifications in our project objectives. The team did a good job advocating for us when communicating with our project sponsors and advisors.

The team could improve to the extent that we broke down tasks on Jira for maximum efficiency. We ran into some issues with the database being down on Friday, which we could not avoid, but database connectivity could definitely be improved. Our translation of Jira tasks to code could also use some work to improve understanding of desired outcomes for that specific code.

The team planned to improve how we broke down tasks on Jira during the planning phase by analyzing and discussing user requirements for Jira stories. This change also helped to improve the issue of translating tasks to code, alongside improvements in terms of the use of more pseudocode, development planning, and VC calls to talk through unknowns and areas of difficulty while coding.

7.3.2 Weekly Summary

Our first sprint of the term and first Thursday to Thursday sprint design started well after completing sprint 1 with good timing. We experienced a lost working day when the database was down, meaning we could not access the data or run queries, and our ID access issues for 50 Prescott which forced us to work remotely from Monday to Friday. From this, we learned that the hybrid approach worked best for the team, with in-person meetings on Mondays and Thursdays to attend sponsor and advisor meetings and conduct sprint planning together, and the rest of the week working remotely with everyone being on-call from 10am-4pm. Our meetings with SaaSWorks and our advisors provided a lot of clarity for our project goals and allowed us to work efficiently to progress toward these goals. The creation of Agile Personas and UML use case diagrams were the first steps in the business value and project risk analysis of SaaSWorks which would be further developed as the project progresses. Improvements were made to the capabilities of the Database class in Python, allowing for precise queries and data filtering using field definitions provided by Eva. Data Preprocessing and basic analysis classes were finalized

and pushed to GitHub. CLR for all account IDs were analyzed and refined using SQL queries, leading to the development of segmentation and data visualization plots used to compare customer fields to CLR.

7.4 Sprint 2: 11/4 - 11/10

Planned Story Points: 25

Completed Story Points: 8

Table 6 lists the completed user stories, the story owner, and story Points from Sprint 2.

Sprint 2: 11/4 - 11/10			
User Story	Story Owner	Story Points	Key
Evaluate relationship between categorical data and churn	Michael O'Connor	3	WPI2022-41
Scatter Plots of Data and CLR Split by Categorical Data	William Bazakas-Chamberlain	2	WPI2022-91
Find mean, median, and mode for each numerical feature.	Shiyu Wu	1	WPI2022-37
Find average CLR for segments.	William Bazakas-Chamberlain	2	WPI2022-35

Table 6: Sprint 2 Completed Story Points

Incomplete Story Points: 30

Table 7 lists the incomplete user stories, the story owner, and story Points from Sprint 2.

Sprint 2: 11/4 - 11/10			
User Story	Story Owner	Story Points	Key
Linearize data for lifetime regression	Michael O'Connor	3	WPI2022-25
Software Requirements (WIP)	Michael O'Connor	2	WPI2022-84

Find the correlations for numerical values using Python.	Shiyu Wu	3	WPI2022-22
Remove predictors with low correlation to CLR.	Shiyu Wu	2	WPI2022-42
Remove predictors with collinearity.	Michael O'Connor	2	WPI2022-43
Create data frame for each proposed model	N/A	1	WPI2022-29
Linear regression model	Shiyu Wu	3	WPI2022-44
Histogram of occurrences for numerical and categorical data	N/A	2	WPI2022-90
Week 1: Implementation Documentation	Abigael Kihu	2	WPI2022-95
Business Value and Risk Assessment	Abigael Kihu	Task	WPI2022-59
Identify factors to segment customers by.	N/A	3	WPI2022-23
Generate an RFM score for individual customers, avg for segments	William Bazakas-Chamberlain	3	WPI2022-56
Generate Segments	William Bazakas-Chamberlain	1	WPI2022-10
UI for selecting desired fields	Michael O'Connor	3	WPI2022-55
Research on what kind of normalization we need for different models.	N/A	Task	WPI2022-30

Table 7: Sprint 2 Incomplete Story Points

With Sprint 2 having only 2 full working days rather than our usual 5, the number of sprints and story points completed was significantly less than planned. This was partially due to the lost working day on Monday for WPI Wellness Day and mainly as a result of the pivot in our project approach following a discussion with Eva during our weekly Wednesday check-in with SaaSWorks that enlightened the group that we were working with time series data rather than an aggregate dataset as previously assumed.

This led the team to take a step back from our Jira board and tasks we were working on to dedicate a day to researching time series data so that we could have a basis of understanding necessary to plan our next steps, reducing another working day from our sprint.

Scope Changes - Story Points Added During Sprint: 13

Table 8 lists the additional story points, the story owner, and story Points from Sprint 2.

Sprint 2: 11/4 - 11/10			
User Story	Story Owner	Story Points	Key
Find mean, median, and mode for each numerical feature.	Shiyu Wu	1	WPI2022-37
Identify factors to segment customers by.	N/A	3	WPI2022-23
Generate an RFM score for individual customers, avg for segments	William Bazakas-Chamberlain	3	WPI2022-56
Find average CLR for segments.	William Bazakas-Chamberlain	2	WPI2022-35
Generate Segments	William Bazakas-Chamberlain	1	WPI2022-10
UI for selecting desired fields	Michael O'Connor	3	WPI2022-55
Research on what kind of normalization we need for different models.	N/A	Task	WPI2022-30

Table 8: Sprint 2 Scope Change Story Points

7.4.1 Sprint Retrospective

Throughout Sprint 2, the team worked with a number of data visualization tools to graph RFM and correlation between customer revenue and usage data. Similarly, the creation of a UML Use Case Diagram and Agile Personas were helpful to the team to distinguish the specific functions of the model we should prioritize to fulfill SaaSWorks goals for this project. The team's transition to a hybrid approach had gone smoothly as the team continued to be productive and efficient with a Monday and Thursday in-person schedule and the remaining weekdays

remote. Our communication remained strong within the team and with SaaSWorks and our Advisors through the use of Discord, Slack, Email, and text messaging. With that, our resource sharing had been effective in keeping all members of the team in the loop with the progress of the project and the necessary background knowledge.

With the completion of Sprint 2 yielding fewer user stories completed than started, we realized that during our past sprint planning sessions, we heavily overestimated the capacity of user stories that we could realistically complete in a single sprint. A meeting with Eva and Jim from SaaSWorks revealed to the team that the dataset we were working with was a time series rather than aggregate data like we had approached the project believing. This led us to the realization that we were spending too much time on less important areas of the project and needed to restructure our prioritization of the user stories on Jira.

We planned to revisit our Jira board to closely assess and prioritize user stories to better plan out our sprints. Using UML diagrams to plan the functions of our model, we were able to plan more thoroughly with a capacity of 20 story points per sprint split evenly between the 4 of us. Along with this, we planned to execute a complete Jira backlog refinement as a result of our goals being pivoted due to the discovery of our data being a time series.

7.4.2 Weekly Summary

The team began our sprint with an in-depth sprint planning where we later realized that we had overestimated our capacity to complete the excessive amount of user stories we assigned to this sprint. We worked on analyzing categorical data using correlation analysis tools, RFM, and SQL queries with the goal of identifying relationships between customers and their categorical data. There were not any statistically significant relationships found, and on Wednesday, Eva pointed out that we were working with a time series dataset rather than aggregate data. With this realization, we shifted our focus to researching methods of time series analysis and reconsidering our approach to our development. On Friday we prepared a Demonstration presentation of our current work to SaaSWorks which we presented during the 12pm Friday Fun Demo meeting. It was a successful presentation. During our sprint retrospective, we discussed this pivot in our goals and created a new plan for how we would address the time series data as opposed to how we had been working under the assumption that it

was an aggregate dataset. Due to the WPI Wellness Day being on Monday 11/7, the sprint had one less working day than usual.

7.5 Sprint 3: 11/11 - 11/17

Planned Story Points: 17

Completed Story Points: 12

Table 9 lists the completed user stories, the story owner, and story Points from Sprint 3.

Sprint 3: 11/11 - 11/17			
User Story	Story Owner	Story Points	Key
Arrange Data into 1xN Vectors for Time Series	Michael O'Connor	2	WPI2022-96
Evaluate if SVM will work for gen pop	Michael O'Connor	3	WPI2022-97
Convert Segmented data to ML model ready data	William Bazakas-Chamberlain	1	WPI2022-102
Principal component analysis	Michael O'Connor	1	WPI2022-26
Linear Regression Model	Shiyu Wu	3	WPI2022-44
Generate Segments	William Bazakas-Chamberlain	1	WPI2022-10
Analyze usage of Random Forest on Inactivity Prediction	Michael O'Connor	1	WPI2022-104

Table 9: Sprint 3 Completed Story Points

Incomplete Story Points: 38 (- 24 from paper) = 14 incomplete story points

Table 10 lists the incomplete user stories, the story owner, and story Points from Sprint 3.

Sprint 3: 11/11 - 11/17			
User Story	Story Owner	Story Points	Key
Identify factors to segment customers by.	N/A	3	WPI2022-23

Software Requirements (WIP)	Michael O'Connor	2	WPI2022-84
Analyze Usage (RFM) Based Segmentation	William Bazakas-Chamberlain	3	WPI2022-98
Create new account statuses (Ready to Churn and First Usage)	Michael O'Connor	1	WPI2022-99
Analyze Accounts based on Activity Status	Shiyu Wu	3	WPI2022-100
Analyze Usage Period and Churn Rate Over Time	Shiyu Wu	3	WPI2022-101
Implementation Documentation Spreadsheet	Abigael Kihu	6	WPI2022-86
Introduction	William Bazakas-Chamberlain	2	WPI2022-78
Software Development	Abigael Kihu	8	WPI2022-87
Business and Project Risk Management	Abigael Kihu	4	WPI2022-80
Research	Abigael Kihu	2	WPI2022-79
Analyze usage of Decision Trees on Inactivity Prediction	Michael O'Connor	1	WPI2022-103

Table 10: Sprint 3 Incomplete Story Points

Scope Changes - Story Points Added During Sprint: 33

Table 11 lists the additional user stories, the story owner, and story Points from Sprint 3.

Sprint 3: 11/11 - 11/17			
User Story	Story Owner	Story Points	Key
Convert Segmented data to ML model ready data	William Bazakas-Chamberlain	1	WPI2022-102
Generate an RFM score for individual customers, avg for segments	William Bazakas-Chamberlain	3	WPI2022-56

Introduction	William Bazakas-Chamberlain	2	WPI2022-78
Principal component analysis	Michael O'Connor	1	WPI2022-26
Software Development	Abigael Kihu	8	WPI2022-87
Business and Project Risk Management	Abigael Kihu	4	WPI2022-80
Research	Abigael Kihu	2	WPI2022-79
Linear regression model	Shiyu Wu	3	WPI2022-44
Analyze usage of Decision Trees on Inactivity Prediction	Michael O'Connor	1	WPI2022-103
Generate Segments	William Bazakas-Chamberlain	1	WPI2022-10
Analyze usage of Random Forest on Inactivity Prediction	Michael O'Connor	1	WPI2022-104
Remove predictors with collinearity.	Michael O'Connor	2	WPI2022-43
SGD Classifier	Shiyu Wu	3	WPI2022-105

Table 11: Sprint 3 Scope Change Story Points

Many significant user stories were planned, executed, and completed in this Sprint which greatly impacted the direction of our project as we approached the half-way mark of the term. To account for many incomplete issues, many of the issues added mid-sprint as listed in the figure of scope changes were under the documentation of development process epic, totaling 24 added story points from this epic alone. This was the result of the team placing more emphasis on accurate progress tracking on Jira as we proceeded with the completion of the paper. A few issues related to this epic, namely the Software Development chapter of the paper, will continue to be updated throughout the term up until the end of the final sprint, on 12/1/22.

7.5.1 Sprint Retrospective

During sprint 3, the team was able to quickly and efficiently pivot the direction of the project after receiving new information about the dataset we were working with. After dedicating Monday to researching time series data and conducting multiple meetings with Eva and our advisors throughout the week, we were able to collect enough information to plan the next steps in our development process. On Friday, we presented a demonstration of our progress so far to the SaaSWorks team and received insightful feedback. On Saturday, the team met up to have an extra brainstorming session outside of our usual working hours, making up for any time lost during our short period of uncertainty. The brainstorming session consisted of every team member writing/illustrating their understanding of the project so far and developing a few drafts for the process map of our project on the whiteboard, which helped us visualize the requirements and necessary steps to achieve the goals set in place. By the end of the sprint, the team had successfully created a number of SVM and SGD classification models, decision trees, random forests, and first-generation linear regression models. The MQP paper was heavily reformatted and revised and we began the editing process alongside the writing of new sections and subsections.

We noticed that our Jira board is still not very reflective of the teammates' individual work, meaning that outside of our implementation documentation spreadsheet updated daily, we were not sufficiently tracking our work and progress on Jira. We also needed to work on our prioritization of the tasks assigned on Jira to ensure that the most impactful tasks and stories were completed before we attempted the less significant work. Regarding our Saturday meeting, although it was a good brainstorming session, we were not as productive as planned since we did not complete the sprint planning session in a timely manner. Following our pivot in the project, we realized that we could improve our communication with professors when seeking guidance in order to have more in-depth discussions.

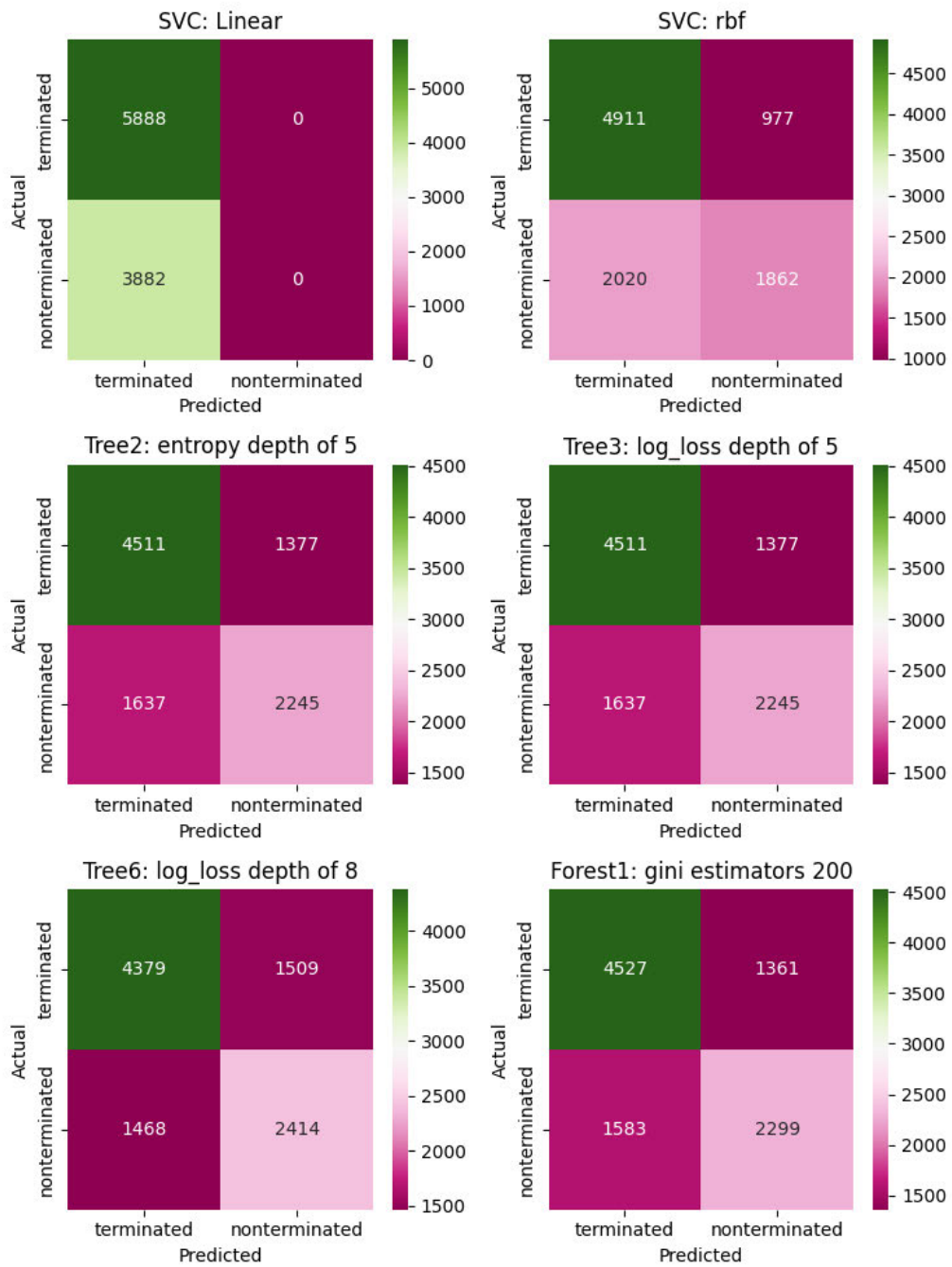
To address the issue with Jira tracking, we would begin updating the Jira board more frequently during the sprint, ideally after every daily stand-up, to ensure that our current and future actions aligned with the bigger picture planned out in Jira. Additionally, we planned to place more emphasis on task and story prioritization during our Jira sprint planning sessions moving forward.

7.5.2 Weekly Summary

This week, the team worked on the MQP paper, adding sections and subheadings that detailed the techniques and research we have implemented thus far. We started a thorough editing process over the already written sections and made changes to improve the paper's accuracy and flow. We created the initial models for determining whether a consumer would engage in prolonged inactivity. This required data reformatting, cross-validation, and model accuracy evaluation. SVMs, decision trees, and random forests were all tested. From each of these models an accuracy score and MCC score was determined based on the confusion matrix results. The confusion matrix shows the distribution of classification compared to the true classification of the data.

The matrices—as shown in Figure 13—show the number of real terminated classes and real non-terminated (or currently active) accounts used to test the model by the sum of the true positive predictions (top left sector) and the false negative predictions (top right sector), and the sum of the false positive predictions (bottom left sector) and the true negative predictions (bottom right sector) respectively. The predicted classifications—shown on the X-axis—show the total number of feature vectors categorized as terminated (sum of true positive and false positive) and non-terminated (sum of true negative and false negative). An ideal confusion matrix has a low proportion of false positives and false negative predictions.

Each model generated an average accuracy of 68-70% with three periods' worth of data per customer ($n=9770$). We developed a stochastic gradient descent classifier model, SGD, that predicts whether an account will churn in the next month given the historical data and a linear regression model that predicts the life span for each account version ID. The team was able to create a function that generates segments of customers based on any unique attribute, which can be standardized and fed into various Machine Learning models. This breakthrough allowed us to connect the data containing customer segments to the ML model by breaking down the data frame of customer information into their respective segments and then converting the data frame to the required form for the ML model.



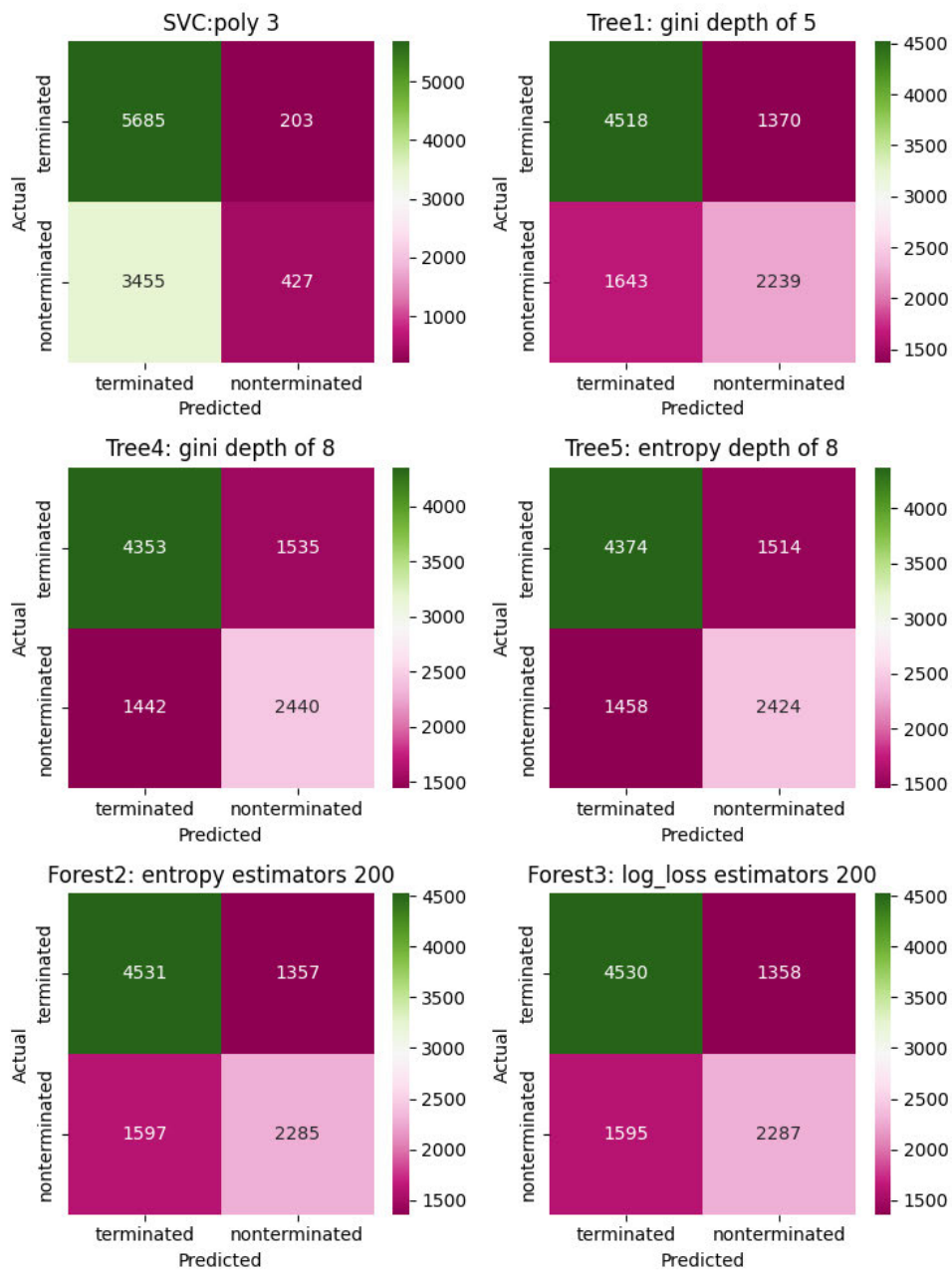


Figure 13: Aggregate Confusion Matrices of Classification of Account Status Given First 3 Active Periods (n=9770)

Note: The color and number in the confusion matrix sector indicates the number of data points that follow into that category.

7.6 Sprint 4: 11/18 - 12/1

Planned Story Points: 43

Completed Story Points: 27

Table 12 lists the completed user stories, the story owner, and story Points from Sprint 4.

Sprint 4: 11/18 - 12/1			
User Story	Story Owner	Story Points	Key
Introduction	William Bazakas-Chamberlain	2	WPI2022-78
Research	Abigael Kihu	2	WPI2022-79
Analyze Usage Period and Churn Rate Over Time	Shiyu Wu	3	WPI2022-101
Analyze Accounts based on Activity Status	Shiyu Wu	3	WPI2022-100
Analyze Usage (RFM) Based Segmentation	William Bazakas-Chamberlain	3	WPI2022-98
Create new account statuses (Ready to Churn and First Usage)	Michael O'Connor	1	WPI2022-99
Create a deep learning model.	Shiyu Wu	5	WPI2022-28
Analyze last 2 months of data for customer status prediction	Michael O'Connor	3	WPI2022-180
Abstract	Abigael Kihu	Task	WPI2022-73
Oversampling and Undersampling	Shiyu Wu	1	WPI2022-187
Create a Main script to demo software	Michael O'Connor	2	WPI2022-182
Nudge on Probability	Michael O'Connor	1	WPI2022-183

Table 12: Sprint 4 Complete Story Points

Incomplete Story Points: 31

Table 13 lists the incomplete user stories, the story owner, and story Points from Sprint 4.

Sprint 4: 11/18 - 12/1			
User Story	Story Owner	Story Points	Key
Linearize data for lifetime regression	Michael O'Connor	3	WPI2022-25
Software Requirements (WIP)	Michael O'Connor	2	WPI2022-84
Software Development	Abigael Kihu	8	WPI2022-87
Business and Project Risk Management	Abigael Kihu	4	WPI2022-80
Train Models Concurrently	Michael O'Connor	2	WPI2022-181
Create a Main script to demo software	N/A	1	WPI2022-182
Software Development Environment	N/A	5	WPI2022-82
Nudge on Probability	Michael O'Connor	1	WPI2022-183
Save Models	Michael O'Connor	1	WPI2022-184
Feature Importance Visualization	William Bazakas-Chamberlain	1	WPI2022-185
Methodology	N/A	3	WPI2022-81
Label and generate CLR graphs	William Bazakas-Chamberlain	Task	WPI2022-186

Table 13: Sprint 4 Incomplete Story Points

Scope Changes - Story Points Added during Sprint: 15

Table 14 lists the additional user stories, the story owner, and story Points from Sprint 4.

Sprint 4: 11/18 - 12/1			
User Story	Story Owner	Story Points	Key
Abstract	Abigael Kihu	Task	WPI2022-73
Train Models Concurrently	Michael O'Connor	2	WPI2022-181
Create a Main script to demo software	N/A	1	WPI2022-182
Software Development Environment	N/A	5	WPI2022-82
Nudge on Probability	Michael O'Connor	1	WPI2022-183
Save Models	Michael O'Connor	1	WPI2022-184
Feature Importance Visualization	William Bazakas-Chamberlain	1	WPI2022-185
Methodology	N/A	3	WPI2022-81
Label and generate CLR graphs	William Bazakas-Chamberlain	Task	WPI2022-186
Oversampling and Undersampling	Shiyu Wu	1	WPI2022-187

Table 14: Sprint 4 Scope Change Story Points

7.6.1 Sprint Retrospective

During this sprint the team recognized that we had substantially improved in the areas that we had pointed out as needing improvement during our last sprint retrospective. We collectively began updating our Jira board more frequently to assure that we are always on task and working on something that will contribute to the success of the project. As the sprint progressed, we faced more instances of project pivoting to our objectives, and we handled them as well if not better than in previous stances of pivoting with quick recovery and redirecting to the next best objective. Our team discussions and brainstorming sessions continued to fuel our

productivity and generate new ideas or solutions to blockers. The team has developed a high-level understanding of our project goals and process that has allowed us to integrate the feedback of our stakeholders into our methodology. Overall, the team has shown proficiency in identifying which paths may work and which are not viable and having the capability to cut our losses when we realize that something is not working to restructure rather than to waste time and resources on unsuccessful paths, especially with the limited time remaining in the term.

The team could improve on reducing the number of points we assign to a single sprint as we have noticed that we often overfill our sprints which often leads to a number of stories remaining incomplete at the end of the sprint as we lack the time to get to all of them. With Sprint 4 being longer than usual with the addition of 2 days from the previous week, we were able to complete a majority of our sprints, but we still recognized how overfilled our sprint was. We could also improve our prioritization and focus of specific tasks/stories to see them through to the end in a timely manner. We noticed that we had been planning and discussing completing a sample code pack to be sent to SaaSWorks at the beginning of this sprint, but we did not get to it until the final day of the sprint.

The team will focus on learning and accurately estimating our work capacity, especially as our final sprint approaches, we will carefully plan our Jira board to ensure that we can complete all necessary sprints before our deadline. We plan to prioritize tasks and stories in order of importance rather than picking and choosing which stories to approach first as previously done. We may assign due dates for particular stories rather than allowing the full length of the sprint to complete important issues.

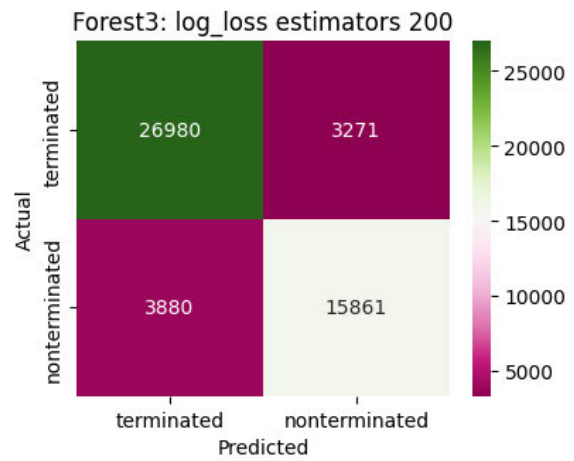
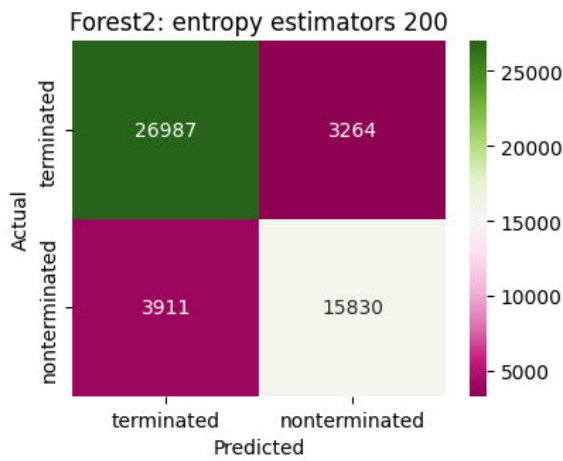
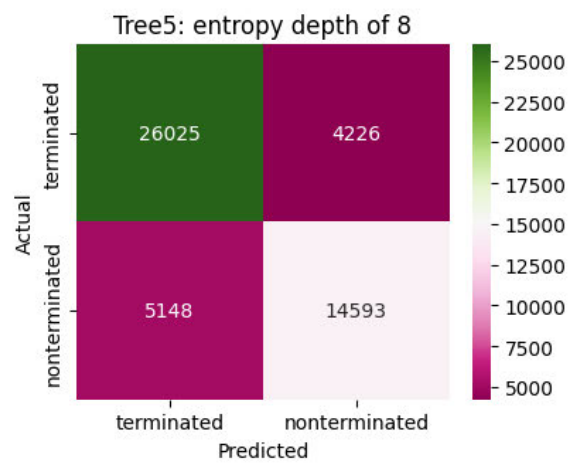
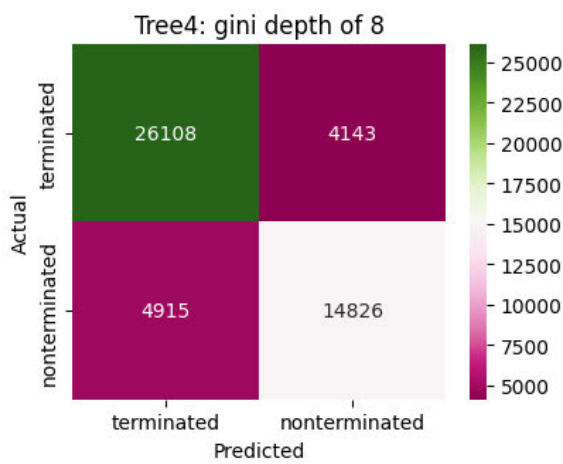
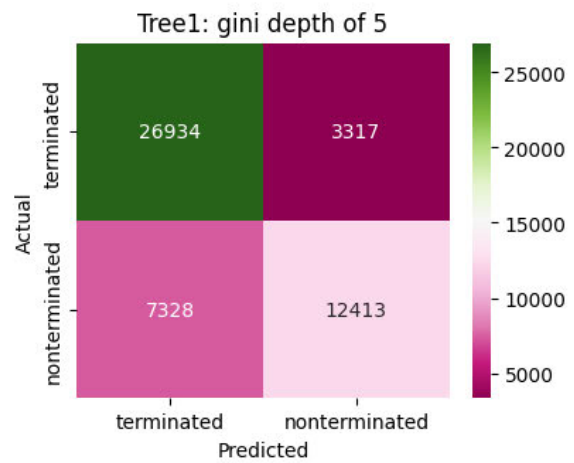
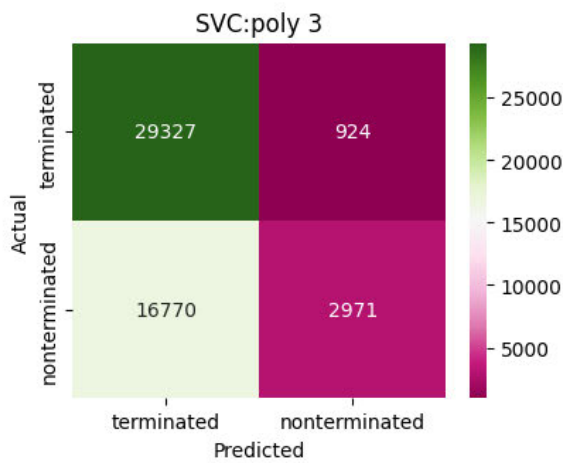
7.6.2 Weekly Summary

During sprint 4, the team prioritized technical and documentation finalization efforts as deadlines approached with the end of the term being a few days away. We worked on writing the remaining chapters in the paper and making final edits before our 12/2 deadline to submit the first draft. We wrote and edited the Abstract, Introduction, Research, Business Value, and Risk Assessment and parts of the Methodology Software Development Environment chapters. We continuously updated the Software Development chapter, Jira board, and Implementation Documentation spreadsheet as well. On the technical side, the sprint initially focused on cleaning up some of the spaghetti code and making sure the code is flexible enough for the other team

members to run their analysis without having to reinvent the machine learning wheel. For the latter half of the sprint, we worked on a new angle of the process by changing the question that the model was geared towards. Initially, it was "based on the first n months, does a given account demonstrate behavior closer to a terminated or active account". We changed it to "given the LAST n months...". The change of context resulted in better accuracy—see Table 15 and Figure 14—and a model that should be better suited for the goal of the project. The confusion matrices in figure 15 Additionally, this should allow for SaaSWorks to essentially use the data in a "rolling window" setting where the data for each account can update every month as the prior context was static and would not change with time. We worked on analyzing the effects that PCA had on the new data. It appears that the dimensionality reduction is also reducing the accuracy in comparison with just using the full feature set (which is of a significantly higher dimensionality). The team also worked to resolve the imbalanced training for churn prediction at the account version level—i.e., when accounts churn and return at an irregular, periodic interval instead of when accounts churn and are not reactivated—by applying an oversampling technique and creating a deep learning model. The deep learning model (ANN with 5 hidden layers) resulted in a MSE of 11. In context, this indicates that on average the model was off by +/-3.3 months. As the average account will only be active for approximately 8 months before it will churn, the MSE was deemed too high by the team and SaaSWorks. The maximum MSE for acceptance was defined as approximately 4 (or +/- 2 months).

We developed a better way of segmenting customer data to fit the ML model, which involved creating functions for this and worked towards saving the output of ML models for each segment and saving their confusion matrices. Additional work was done by the developers to complete the writing of comments and documentation for existing classes, and then uploaded to Github.

Overall, it was a successful week with most of our planned stories being completed. We sent a Current Status and Final Milestones update and a Code Snapshot to SaaSWorks for them to evaluate our progress before we submit our final deliverables on 12/8.



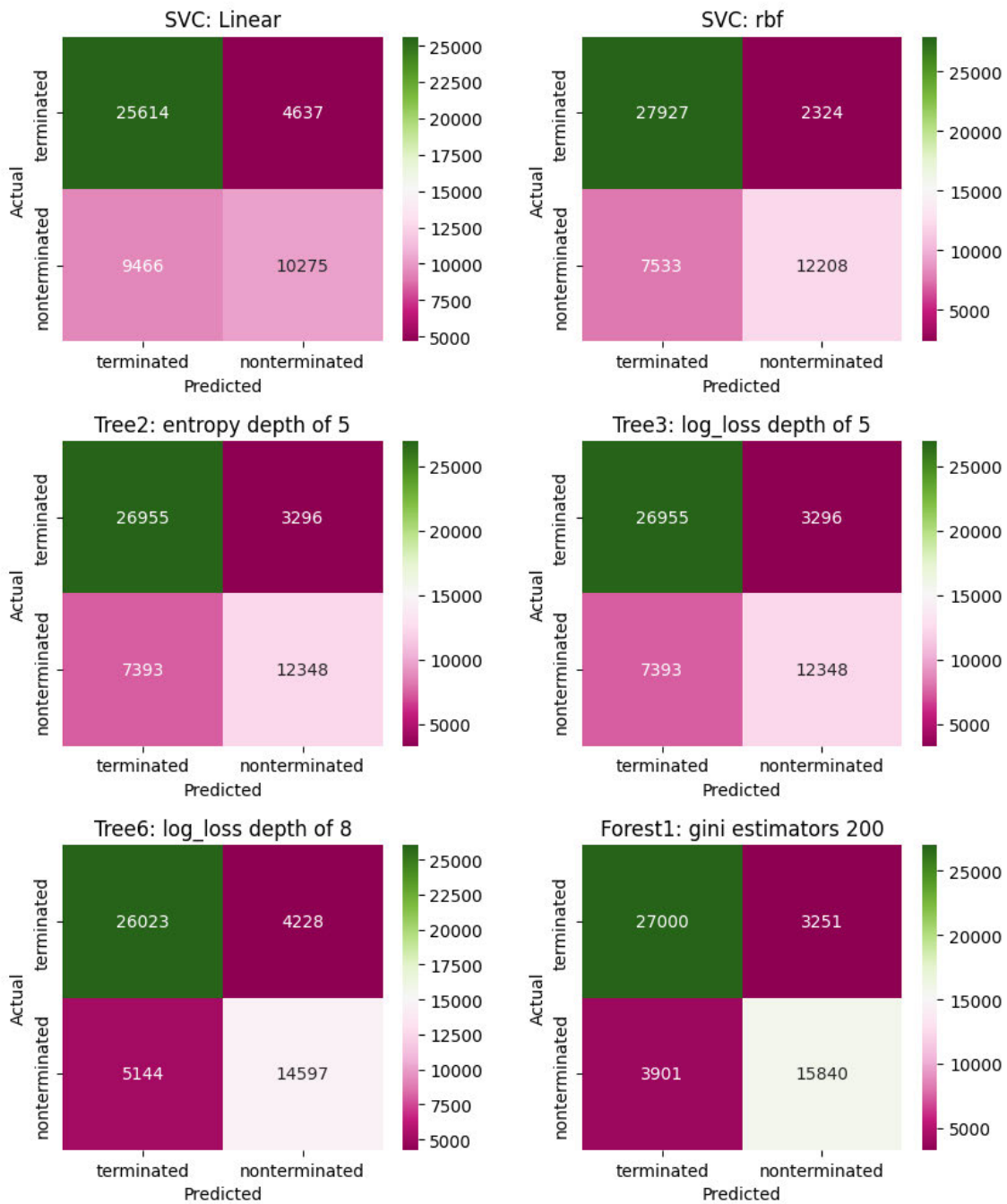


Figure 14: Aggregate Confusion Matrices for Classification Models (n=49992)

Note: The color and number in the confusion matrix sector indicates the number of data points that follow into that category.

A number of models were developed during this sprint to visualize our findings and outcomes of the project thus far. As shown in Figure 15 - the Aggregate Confusion Matrices use

49992 samples in various classification models to depict ML model predictions vs actual data. A Confusion matrix for the Stochastic Gradient Model is displayed in figure 16.

Table 15 represents the model accuracies of the decision trees and random forests used. In figure 15, the Covid Pandemic Impact on active accounts from different locations is illustrated in the multivariable line graph below.

Model	Accuracy Score	Matthew's Correlation Coefficient
SVM: Linear	0.7179	0.3924
SVM: RBF	0.8028	0.5831
SVM: Polynomial 3rd Degree	0.6461	0.2188
Decision Tree: Gini (Max depth 5)	0.7871	0.5471
Decision Tree: Entropy (Max depth 5)	0.7862	0.5452
Decision Tree: Log Loss (Max depth 5)	0.7862	0.5452
Decision Tree: Gini (Max depth 8)	0.8188	0.6190
Decision Tree: Entropy (Max depth 8)	0.8125	0.6050
Decision Tree: Log Loss (Max depth 8)	0.8125	0.6051
Random Forest: Gini (200 Estimators)	0.8569	0.6992
Random Forest: Entropy (200 Estimators)	0.8565	0.6983
Random Forest: Log Loss (200 Estimators)	0.8570	0.6994

Table 15: Model Accuracies

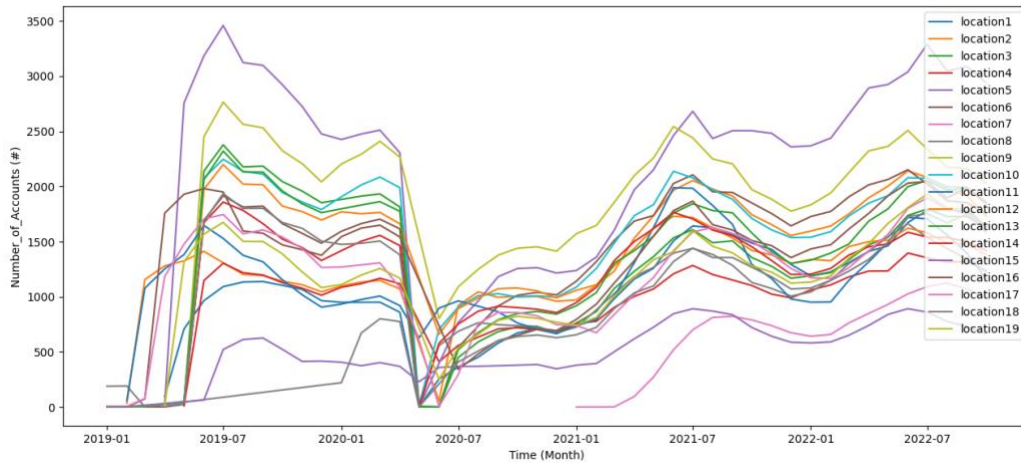


Figure 15: Covid Pandemic Impact on Active Accounts in Different Locations

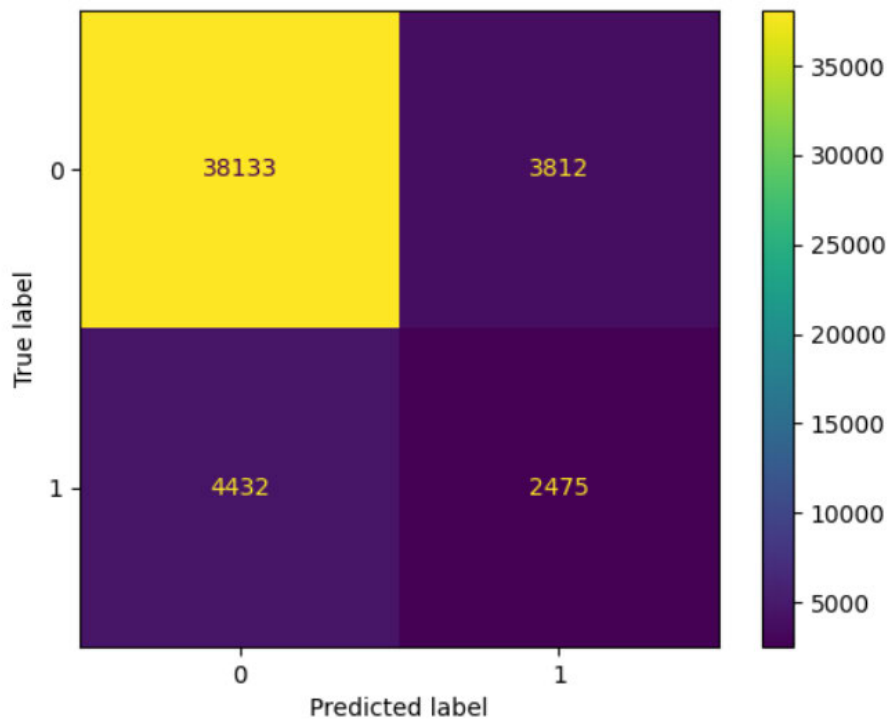


Figure 16: Confusion Matrix for the Stochastic Gradient Model

7.7 Sprint 5: 12/2 - 12/8

Planned Story Points: 51

Completed Story Points: 47

Table 16 lists the completed user stories, the story owner, and story Points from Sprint 5.

Sprint 5: 12/2 - 12/8			
User Story	Story Owner	Story Points	Key
Clean Project Structure to Conform to Process Diagram	Michael O'Connor	3	WPI2022-192
Address Professor Comments	William Bazakaschamberlain	Task	WPI2022-191
Future Work	William Bazakaschamberlain	2	WPI2022-188
References	ALL	Task	WPI2022-178
Assessment	William Bazakaschamberlain	2	WPI2022-88
Software Development	Abigael Kihu	8	WPI 2022-87
Source Code Management Software	Abigael Kihu	Task	WPI2022-83
Visualize and Evaluate Model Results	Michael O'Connor	1	WPI2022-53
Creating descriptions for each machine learning model	Shiyu Wu	3	WPI2022-194
Conclusion	Shiyu Wu	2	WPI2022-189
Feature Importance Visualization	Michael O'Connor	1	WPI2022-185
Table of Contents	Abigael Kihu	Task	WPI2022-106
Implementation Documentation Spreadsheet	Abigael Kihu	6	WPI2022-86
Software Design	Michael O'Connor	2	WPI2022-85
Software Requirements (WIP)	Michael O'Connor	2	WPI2022-84

Software Development Environment	Michael O'Connor	5	WPI2022-82
Methodology	Michael O'Connor	3	WPI2022-81
Business and Project Risk Management	Abigael Kihu	6	WPI2022-80
List of Tables	Abigael Kihu	Task	WPI2022-77
List of Figures	Abigael Kihu	Task	WPI2022-76
Acknowledgement	Abigael Kihu	Task	WPI2022-73
Executive Summary	Michael O'Connor	1	WPI2022-74

Table 16: Sprint 5 Complete Stories

Incomplete Story Points: 4

Table 17 lists the incomplete user stories, the story owner, and story Points from Sprint 5.

Sprint 5: 12/2 - 12/8			
User Story	Story Owner	Story Points	Key
Label and generate CLR graphs	William Bazakaschamberlain	Task	WPI2022-186
Save Models	Michael O'Connor	1	WPI2022-184
Linearize data for lifetime regression	Unassigned	3	WPI2022-25

Table 17: Sprint 5 Incomplete Stories

Scope Changes - Story Points Added during Sprint: 0

7.7.1 Sprint Retrospective

As the team made final edits to our development and documentation efforts during our last sprint cycle, we made significant progress in the restricted time frame available. The team

made great strides to make the necessary adjustments and additions to the code and paper as advised by our sponsors and advisors with a quick turnaround time. Regarding prioritization, we made notable improvement in time management of the remaining issues on Jira to approach deadlines with ease. This was made possible through proficient story point estimations and equal division of work. The Code Snapshot delivery to SaaSWorks went smoothly as we were able to engage in an open dialogue about the data feature set and logic behind model selections. By the end of the final week, the team was able to finalize the paper, complete the debugging process, and push the final code to GitHub. The areas of improvement that the team could touch on include maintaining a consistent format throughout the paper to maintain uniformity for comprehensive reading. During the editing process, the team could have expressed better feedback internally for more concise and precise edits. To counter the issue of inconsistent formatting from individual teammates compiling their work into one document, we were able to establish a style guide that all could follow.

7.7.2 Weekly Summary

This week the team's priority was to make finalization efforts to the project in terms of writing, editing, and formatting the paper to accurately reflect our development process. We focused on the writing, editing, and updating of the table of contents, the Abstract, Executive Summary, Methodology, Software Development, Business and Risk Analysis, Future Work, and Conclusion chapters. Additionally, the team rewrote the classification model descriptions in the Research section and added significant information to the ML, Assessment, Software Testing, and a few other areas of the paper. We finalized the program features to be delivered to SaaSWorks. On Wednesday, there were a couple last requests that had not yet been implemented, so work was done to bring those into existence – specifically, decision tree visualization, formatting outputs to be readable, addition of more CLI arguments, adding additional documentation both in the code and as a README. A description was written for every model used and tables were created within the Software Development section to list every user story worked on during every sprint. The team updated the visuals for the number of active accounts vs churn rates graphs by replacing the location names with a simplified format of “location n”. We spent some time editing the output of a function that potentially may be pushed

to production as well, but overall, the main focus of the team was the completion of the project in its entirety.

7.8 Product Burndown Chart

Total Completed Story Points: 114

Burndown charts are visual information radiators that display the work that is projected to be completed in a sprint in handwritten, drawn, printed, or digital display that enables users to monitor the status of a project. Agile teams use burndown charts to visualize how much work has already been completed, how much work still needs to be done, and how much time is remaining to finish it. The graph "burns down" to zero on or before the last day of the time period as tasks are completed (Lucidchart, 2020).

The Product Burndown chart in figure 17 represents the team's progress throughout the sprint cycles in completion of the development and documentation of the predictive model. The chart highlights the incremental increase of user stories assigned as the team gained further insights and capabilities to contribute to all aspects of the project. This allowed the team to more accurately estimate story points and complete more stories in a given sprint toward the last two sprints as compared to the initial sprints.

There were 51 remaining story points that were not completed. These are from abandoned user stories due to the pivoting the team experienced throughout the project as time progressed, and specifications were defined. These user stories were concluded to be no longer relevant or progressive towards the stakeholders' goals, so the team ceased working on them and the story points were disregarded in pursuit of more impactful issues to complete. The 114 story points completed were crucial to the development of the project.

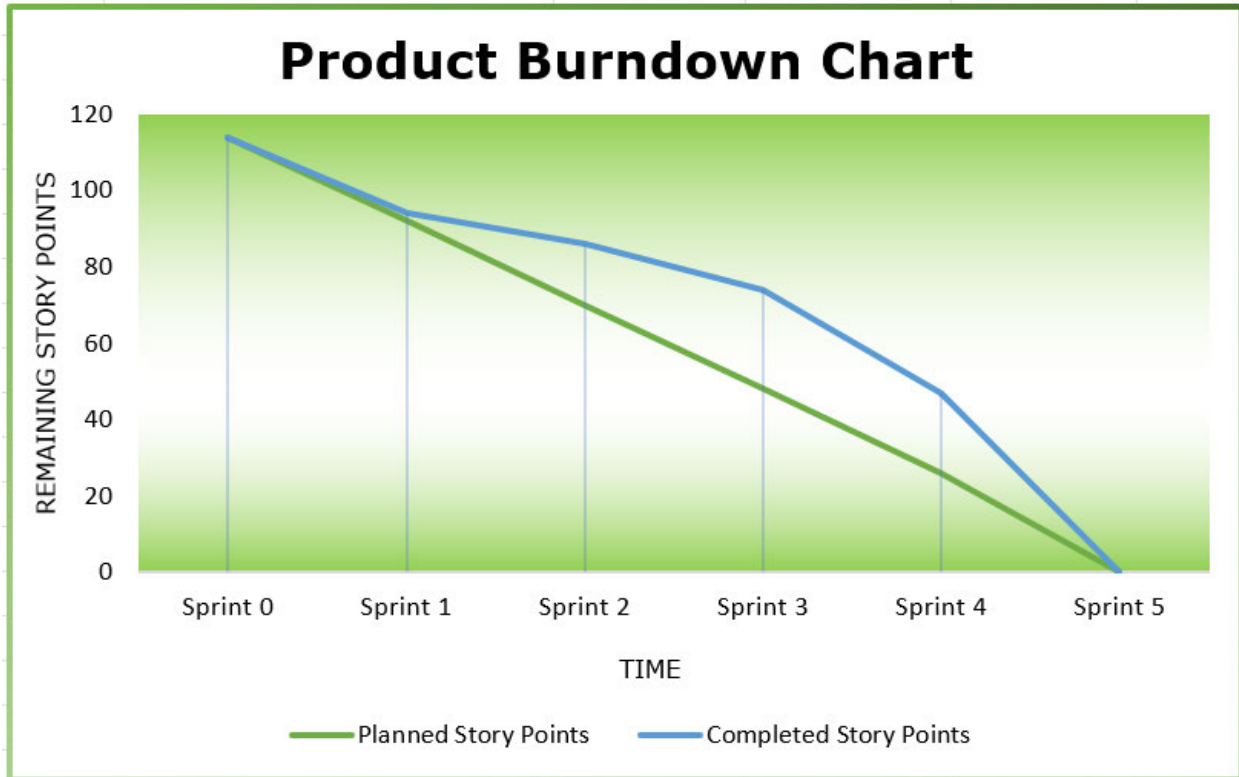


Figure 17: Product Burndown Chart

7.9 Software Testing

The machine learning models were tested against models of similar type (either regression or classification). Each grouping of models was trained and tested using the same randomly chosen to split five times utilizing K-Folds Cross Validation ($k=5$). After each sequence, the accuracy and MCC scores were recorded. After all five sequences, the metrics were averaged for each model and then compared to select the best model as determined by the MCC score with accuracy as a tiebreaker. The results for Cross-Validation comparisons can be found in Appendix B.

The best model was then cloned as a cleaned, unfitted model and then provided 75% of the feature set and then tested with the remaining 25% of the data set using permutation feature importance to determine the impact of each of the features. Each feature was withheld from the dataset 30 times and the delta in the MCC was recorded and averaged to determine the impact

the feature had on the predictive power of the best model. The full results of the feature importance for the overall most effective model can be found in Appendix C.

8. Business and Project Risk Management

SaaSWorks is working to strengthen and maximize the growth of their clients' businesses by identifying which of their KPIs most correlate with their CLTV and CLR. With these metrics that they clean, organize, and concisely report to their clients' CFOs, SaaSWorks wants to have the capability to make accurate predictions of which customers are at a high risk of churning and in what period they are expected to churn. SaaSWorks will use this information to make suggestions to their clients about which customers should be sent Nudges at which specific times in order to influence them to continue doing business with the company and reduce their risk of churn. SaaSWorks needs these Nudges to be backed up by comprehensive data analysis, modeling tools, and logic rather than a black box that makes suggestions to no explanation. With the development of a Predictive model designed to do just this, the goal is that these Nudges will be sent with accuracy that will yield a significant impact on the churn rates of a SaaSWorks client and thus increase the CLTV and CLR of that company. The value of this project lies within the Predictive model producing viable results for a SaaSWorks client in the form of increased CLTV and CLR which will encourage the growth of the company. These capabilities are expected to increase retention and willingness to spend of a SaaSWorks client.

The success of this project will be beneficial to all stakeholders, the SaaSWorks clients whose data is applied to the model, and SaaSWorks itself. As their clients experience company growth through increased CLTV and decreased customer churn rates, SaaSWorks will sequentially increase their own customer lifetime value as clients realize that their services are beneficial and worth long-term customer retention. SaaSWorks clients who may have been at risk to churn will disregard the idea of dropping their services once they experience revenue growth unattainable without the intervention of SaaSWorks. Other companies and competitors of clients who are unfamiliar with SaaSWorks services will recognize the growth of their clients through market research and seek similar outcomes. Upon the realization of their inability to produce the same results without the expertise and specialized tools that SaaSWorks provides, they too will inquire about their services, contributing to the growth of the startup.

8.1 Risk vs Reward

Although there are high-value stakes associated with this project, there are risks that should be considered as well. With SaaSWorks being in its early stages as a startup, the risk of failure is always present as the company works overtime to establish itself as a high-value SaaS company. As explained by a SaaSWorks co-founder and the CTO, “There is always another company willing to outwork and outsmart you, so focus on time to value and time to market rather than a work-life-balance which you will have plenty of time if your company fails” (O’Neill, 2022). SaaSWorks is constantly trying to develop services that will differentiate itself from other SaaS providers in terms of accuracy, accessibility, and results. There are many opportunities for failure that SaaSWorks must avoid or overcome, varying based on the outcomes of this project.

The development of this predictive model will increase the value of the services that SaaSWorks offers by providing a tool that enables near-automated identification of the KPIs and metrics most correlated with CLTV and CLR that can be applied to the business data of any generic client regardless of their industry or target customers. Doing so will progressively increase the revenue of SaaSWorks as they diversify their offered services with the introduction of this model specifically designed to drive business growth and improvement. This service will objectively hold more value to all stakeholders – SaaSWorks clients and SaaSWorks – as it will have a direct and real-time impact on company performance and bottom-line profitability for both parties. “The bottom line refers to a company’s net profits after deducting its cost of goods sold, fixed overhead and administrative expenses, interest charges on loans and other debts, depreciation and amortization charges, and federal, state, and local income taxes” (Woodruff, 2022). Companies generate this value for shareholders who find importance in its ability to “emphasize the actions that influence a company’s net income or net profits. It is the concept that makes companies consider their social and environmental factors besides their financial performance” (WallStreetMojo, 2022). SaaSWorks ability to positively impact the bottom line will make it a valuable asset to businesses in all industries.

Upon the successful creation of this Predictive Model, it will be “productized into a configurable and repeatable product feature-set” (O’Neill, 2022) that will impact clients’ willingness to pay for their software services as a direct result of the increase in revenue that

clients experience by using the software. Willingness to pay is the maximum amount of money a client would sacrifice in exchange for the software service. The best way to improve willingness to pay is to improve their service quality and/or variety. The success of this project will increase the effectiveness and demand of the SaaSWorks services, fulfilling both the quality and variety aspects of their Willingness to Pay variables. As their clients experience the benefits of their services, SaaSWorks will gain new customers through referrals and marketing. This influx of new customers will drive sales and increase revenue for SaaSWorks alongside their clients.

To counter the risk of possible project failure if undesirable outcomes are reached, it is understood that in consideration to the limited time allotted for development, setting the model up correctly to be generalized enough to fit various data sets while concisely identifying the unique KPIs associated with CLTV and CLR for each company is far more important than the outputs that are produced.

8.2 Risk Culture

“Risk culture involves values, beliefs, knowledge, attitudes, and understanding of risk shared by stakeholders associated with a business. It encapsulates the amount and type of risk a business is willing to accept in pursuit of key objectives” (Adamson, 2022). SaaSWorks approaches risks rationally and highly calculated, with every possible outcome extensively researched and evaluated to plan for optimal results. Through taking risks moderately, SaaSWorks is dedicated to doing the research, analysis, and work that many companies know they need to do, but simply cannot manage financially or through labor hours. They evaluate each prospect for their data compatibility and completion before embarking on the project and devoting their time and resources. The SaaSWorks team leverages the risks of their work on their meticulous solution platform, whose value will ideally speak for itself through the generation of high ROI (O’Neill, 2022). Their risk culture is healthy and within reason, meaning that SaaSWorks has laid a good foundation of company discipline and decisiveness that will benefit them in the long run.

8.3 Additional Risks

8.3.1 Operational Risks

Operational risks are those associated with “the risk of losses caused by flawed or failed processes, policies, systems or events that disrupt business operations” (Morgan, 2021). The operational risks of this project are associated with the potential loss of the investment of time and resources dedicated to testing and customizing the model to suit SaaSWorks service specifications, merging the model with SaaSWorks software, and introducing it to the catalog of SaaSWorks services if the model fails or produces inaccurate outputs.

If the project does not achieve the success that is expected, meaning that the model does not accurately identify key KPIs most impactful to CLTV and CLR, or if the Nudges it suggests do not deliver significant improvements to company growth, SaaSWorks will be at risk of their own CLTV being lower than anticipated. This may potentially result in revenue loss for SaaSWorks due to client churn if they are not able to deliver suggestions that result in client improvement and growth. SaaSWorks already provides services that analyze and organize their clients’ data into a more concise and easily accessible format for increased company understanding. However, in pursuit of growth, SaaSWorks must bring their business to the next level by introducing a service that applies their data analytics to real-time business performance to make recommendations that will positively influence future revenue and growth. As a start-up SaaS business, SaaSWorks recognizes that having the capability to impact a client’s bottom-line income is necessary to retain long-term customers and establish themselves as a high-value SaaS company. It is especially important that SaaSWorks assists their clients in generating substantial growth and establishing business playbooks that align with healthy KPIs during the first year of receiving their services.

8.3.2 Security Risks

The security risks pertain to the sensitive client data that SaaSWorks has access to, which they cannot allow to be breached or shared. They have taken extensive precautions to mitigate this risk during the development cycle in the form of NDAs, which all team members have signed. The use of virtual desktops has been put in place to ensure that the data will be secured

and, reduce the chance of confidential information being transferred or manipulated between non-SaaSWorks servers. All team members and advisors privy to the project plan have agreed to uphold the security measures established and to never take information from the SaaSWorks systems or discuss it with others not directly working on this project without clear permission.

8.3.3 Market Risks

The current market for analytical data lacks the tools to automatically create predictive models on a variety of data sets. If the project is successful as planned, SaaSWorks will possess a powerful data analytics and predictive modeling tool considered to be the holy grail of business analysis. It will attract numerous subscription companies to SaaSWorks seeking their services for the results they produce. This will bring market awareness to their unique services and inspire potential competitors of SaaSWorks to replicate their services and challenge their market share. This could potentially risk SaaSWorks losing current or potential clients if their competitors are successful in recreating their services and providing them at either a greater value or lower cost. SaaSWorks can overcome this risk by competitively pricing their services while maintaining high-value results with consistent accuracy and satisfactory delivery to their clients.

8.3.4 Accuracy Risks

In this industry of data analytics, accuracy is an extremely important component of a company's services that contribute to the value of the company. Any Nudges sent to SaaSWorks clients which are based on inaccurate analytics will be useless to the client as they cannot generate any meaningful results or contribute to the growth of the company. Since they do not precisely represent the dataset, inaccurate analytics will identify insignificant KPIs or make unnecessary oversights that will yield no value to a SaaSWorks client. To avoid such miscalculations, SaaSWorks stresses the importance of utilizing modeling and Machine Learning (ML) tools designed and tested specifically to find trends inconspicuous to the human eye to avoid human error. ML is constructed on a self-sufficient operational model and updated as needed to facilitate software programs and systems that enable the analysis of substantial datasets to present to clients through faultless experiences.

Within the umbrella of accuracy risks, this project is subject to risks of overfitting, the precision of time, and customer identification. In terms of overfitting, like SaaSWorks services, this model must be generalized enough to be applicable to any variety of business models if they fit certain criteria for the data being observed. The model must simultaneously have the capability to adjust to the specificities of certain businesses while maintaining a level of precision and accuracy in the outputs and Nudges suggested. It is important that the customers who are at high risk of churn are identified correctly and as soon as possible so that the proposed Nudges can be sent out in a timely manner to allow clients to prepare impactful outreach messages before the customers are lost. The precision of the time sent is a large factor in the success of these Nudges; because if the Nudges are sent too late after a customer has been identified as being at risk of churn, the customers will more than likely have churned already. If they are sent out too early before a customer is officially at risk of churn, then the Nudges will have no effect on customer retention (O'Neill, 2022).

8.3.5 Financial Risk

The financial risk may arise if SaaSWorks makes a large investment into this project in anticipation of company growth that does not live up to its potential. However, “as an early-stage startup, there are various investments SaaSWorks will make knowingly at a financial loss with the assumption the technology we build will provide future benefits and profitability.” (O'Neill, 2022). All previously mentioned risks have connections to financial risks. This risk can be mitigated by SaaSWorks precisely and extensively producing, testing, protection and distribution of the model and data involved.

9. Assessment

9.1 Business Learnings

As the WPI MQP team was fully assimilated into the SaaSWorks working system, the team was able to dive into an invaluable learning experience that provided us with insights into how Data Analytics companies operate in the FinTech Industry. The importance of data analytics was soon revealed to the team as having the capability to gain valuable insights into business operations that help businesses uncover hidden opportunities and threats, support informed data-driven decisions, and provide a competitive edge to companies utilizing this tool in the market. As the business world reaches new levels of growth and advancement in today's fast-paced and data-driven market, data analytics will continue to be a vital and powerful tool to help businesses uphold their relevance.

The team learned how a well-functioning team should operate through our use of the Agile Scrum framework which proved to be useful on the platform Jira during our explorative development process. We utilized the configuration and personalization functionalities to thoroughly plan our project using epics, user stories, and child issues. These issues were assigned to one or more team members to ensure that a team member is responsible for every detail of the project. The team embraced the ability to create and remove stories as needed and track progress using story points within each sprint cycle. We found that consistently updating the Jira board is crucial to ensure that everyone is on track and making meaningful progress while upholding the priority commitments that impact the direction of the project the most.

The significance of effective stakeholder analysis and engagement in the form of systematic identification, analysis, planning and implementation of actions designed to influence stakeholders was recognized as the project specifications and goals were more clearly defined. Our consistent communication with our sponsors played a vital role in ensuring that their business needs actualized within the constrictions and scope granted. By continuously updating sponsors with a detailed overview of current progress and goals, the team was able to achieve a comprehensive understanding of the project that enabled increased efficiency through the precise and relevant feedback from those updates (APM, 2022).

The implementation of a hybrid workflow suited the team's work style by involving open collaboration and promoting synergy. The balance of in-person and remote meetings allowed the team to work on individual tasks from respective offices then reconvene to combine ideas and participate in sponsor and advisor meetings together.

9.2 Technical Learnings

Throughout the project term, the MQP team had to develop new and improve upon current technical skills to accomplish the project goals set. The team had to improve their Python and SQL skills in order to adapt to the large amount of data being manipulated. There was also much for the team to learn in the ways of machine learning. This initially included research into which models would be ideal for our use case but extended into research model evaluation metrics, visualizing models, and understanding their reasoning.

9.2.1 Data Management in Python

The first thing the team was prompted to learn was PostgreSQL, the relational database system which SaaSWorks used to contain all of the data relevant to the project. Initially the team had to form a basic understanding of the PostgreSQL system to set up a database connection in Python. Once this was learned the team could begin the next steps of pulling data from the database and working with it in Python. To accomplish these next steps, the team needed to be well versed in SQL. Code for filtering data in SQL was much easier to write than filtering in Python. Because of this, the team realized that investing time into improving their SQL skills would save working time and computation time from filtering data in Python. Another factor pushing the team to improve on their SQL was that the data set was so large the queries had to be detailed enough to filter for need-specific data, which pushed the team to learn how to write more detailed queries.

Once the data was stored in Python, the team used Pandas to manipulate the data for analysis. Pandas is a data management Python library that has functions for cleaning, manipulating, analyzing, and exploring data. Pandas is the same library used by SaaSWorks themselves. Because varying analysis methods and use cases required data in different formats,

being knowledgeable on the different methods included in Pandas allowed the team to complete different types of analysis.

The team also had some key takeaways from working with a large dataset. Given that the working dataset was so large, there was not enough random-access memory on the virtual machines being used to load the dataset all at once. Additionally, computations done on larger sets of data take far too long to run. For these reasons the team had to find a way to extract a portion of the data that is still representative of the majority. Any sample taken from the dataset may also need to maintain the same distribution as the original dataset to understand the true value of the data. One method the team learned for filtering data was to use random sampling. Random sampling pulls an unbiased subset from the data set to run computations on. To test the validity of each random sample multiple random samples are taken and the results are compared against one another. When the model or analysis being done onto the random set changes, recreating tests with the same input would not be possible without random seeding. Random seeding saves the seed used to generate the random data set meaning the same random set can be generated multiple times. This can be used to test the same dataset on multiple models with the dataset remaining random.

9.2.2 Machine Learning

Prior to the project term, the team had minimal knowledge in the field of machine learning but were knowledgeable enough to dictate what further knowledge to pursue to achieve the project's goals. To begin, the team had to learn about various machine learning models to identify which would be ideal for their use case. The two model types chosen for analysis were classification and regression. Classification models classified customers as exhibiting behavior similar to churning or not churning customers. Regression models could detect relationships between variables, identifying key performance indicators in our case.

Through further knowledge discovery and meetings with professors the team picked four classification models, decision trees, random forests, support vector machines, and stochastic gradient descent as well as two regression models, multivariable linear regression, and a multilayer neural network. After learning how to best fit the data for these models they experimented to optimize the results. Initially the team used accuracy as an evaluation metric and assumed that models created with high accuracy were always useful. The team knew there were

further steps to be taken to validate the results of the trained models. Accuracy accounts for overall correct predictions but fails to consider how the model is guessing. As, on a normally distributed dataset, a classification model could be correct half the time by solely choosing one classification for every piece of data. It was only after generating confusion matrices, showing both sides of the binary classification, that the team learned this fault with accuracy. After researching this problem, the team learned to use another metric, Matthew's coefficient, to evaluate the models. Matthew's coefficient prioritizes both sides of a binary classification and gives insight into if a trained model has solid reasoning behind its classifications.

Through further testing and experimentation with the models, the team learned the effects of formatting how the data was imputed to the models. This was a key learning moment as the team realized that the way they formatted the data changed the questions they were looking to answer with machine learning. The models went from answering the question of "will this account ever churn?" to "given the *last* n months of data, does the customer exhibit behavior more like an active customer or a terminated customer?" simply by discovering new ways to present the data to the models. The team also learned that a model's performance will change depending on how much data is inputted, meaning the team had to find the optimal amount of data. Too little might give low accuracies but requiring too much customer data may limit the potential customers that fit that criterion.

10. Future Work

Given that this project contained a large amount of research and experimentation, there is a significant amount of future work that could be done to improve and build upon current project structure to increase the efficiency of the aforementioned processes and further determine the true effectiveness of the program when implemented.

10.1 Increase Data Granularity

One of the simplest changes to the project that could result in better findings is changing the granularity of the data from monthly to weekly. The customer data given to us by SaaSWorks was split by month, but the team believes weekly data would have allowed for better and more

accurate trend identification as a higher level of detail could bring to light new behavioral patterns. To summarize, more useful data is more useful.

10.2 Implement a Shared Schema

One change essential to future work on this project is an editable database. The main benefits of an editable database are saving data tables, eliminating the need to run SQL queries multiple times, and creating columns or derived fields in the database.

The current process requires users to import data to Python using a SQL query. The data is then cleaned, analyzed, combined with other query results, reformatted, and then tested in Python. Depending on the complexity and size of the query, it could take from a few seconds to ten-plus minutes. As developers needed to test the system repeatedly to verify results and test functionality, this consumed significant time.

The solution used by our developers was to limit the queries to fewer data points. This solution is unideal for multiple reasons. First, a reduction in the data set could accidentally exclude significant patterns or data points to which the models are never exposed. When limiting the queries before segmenting the customers, it is often the case that the resulting data frames would not have any or enough of either churned or active customers, which is not enough data to train the model. Additionally, it became harder to replicate bugs as processing larger and larger datasets included significant downtime while the program waited on results. When a bug does not occur with a small subset but only with a sufficiently large set, a day of development may be wasted due to waiting on SQL queries and manipulation. Overall, our developers found that losing time to running queries was inevitable and mitigated as best as possible. With an editable database, queries would still need to be run, but the query would no longer have to filter through any data. It would just need to get the information from a saved table which would greatly decrease the queries run times.

10.3 Project Management

The project would also benefit from future work in certain project management aspects. When conducting research or testing, often, group members would be far too focused on their

individual tasks and lose sight of how their work connects to the project goals. It was not until later in the term, when creating a current state project summary for stakeholders that the team recognized the importance of collaborating to create write-ups summarizing our understanding of the project and what the team had accomplished. The team had initially avoided these as it was assumed to be more working time lost to planning, but the activity refocused the team and enabled more effective collaboration that typically saved more time than it cost. The project summaries improved the quality of feedback the team received from stakeholders. By providing SaaSWorks with a detailed overview of current progress and goals, they gained a more encompassing understanding of the project and, in return, provided more relevant recommendations.

10.4 Increasing Interpretability

Future efforts should be directed toward developing more interpretable outputs for the model's reasoning. There will always be a tradeoff in machine learning between precision and accuracy and interpretability. Typically, if a model is more generic and its reasoning can clearly be understood, it will likely suffer from being less accurate than a model which has a more complex way of making a classification. Choosing the ideal model depends on a client's needs and improving the way a model is evaluated will help find those ideal models. This effort falls into the field of transparent AI, which aims to generate robust models capable of near-white-box reasoning.

10.5 Forecast Testing

The most important piece of future work would be the real-time testing of predictive results. Although the team can test the accuracy of models on past data, the only way to determine the prescriptive power of the models and fine-tune the system is to apply the model to real-time data and then conduct the true predictive accuracy on a customer set. The team would verify that the customers identified as at risk of leaving would leave without intervention. It is important to note that many churning customers could just be seasonal. To verify a permanent churn, SaaSWorks would wait thirteen months as this was the time defined that a customer must

be inactive before being considered terminated. If the model was verified as accurate in determining customers likely to churn on this imaginary future set, it could then be applied to identify future behavior.

11. Conclusion

11.1 Method Applied

Over the course of this project, the WPI MQP team worked in collaboration with SaaSWorks to analyze the relationship between customer account data and customer retention, and then leverage the analysis to generate systems capable of detecting customer loss before it occurs. To achieve this goal, the team applied several tools—recency, frequency, monetary analysis on customer lifetime revenue, aggregate statistical summary, correlation analysis based on a timeframe, Principal Components Analysis, and plotting graphs of churn rate and number of active accounts over a given time period. The team also built up several machine learning models—Linear Regression, Support Vector Machine, Decision Tree, Stochastic Gradient Descent, Random Forest, and Multi-layer Neural Network—to predict the lifespan of an account version id and whether it tends to churn in the nth month.

11.2 Result

Overall, the team was able to generate a process and software for generating models capable of classifying accounts as “exhibiting behaviors of a terminated account or not” based on the provided historical data from SaaSWorks. The software can then identify currently active accounts with a specific probability – as defined by SaaSWorks for their particular need – of being more like a terminated account than active. From there, additional incentives can be applied to bring the customer back to the SaaSWorks client. Additionally, the application can identify the significant factors corresponding to that prediction.

However, the exact relationship of the features to the final prediction of the most accurate model cannot be determined at this time due to an apparent nonlinear relationship between the most significant features, such as recurring revenue, and the rest of the feature space.

The data does not always provide the answers analysts expect. [REDACTED]

[REDACTED]
connection to the churn rate. However, in correlation analysis-based aggregate by period start

date, no variable has a significant correlation with churn rate. [REDACTED]

[REDACTED]

[REDACTED]

The team gained insights from graphs plotted for analysis. Based on the graph of the churn rate and number of active accounts, the Covid-19 pandemic impacted the business of the subscription-based classes as in about March 2020 churn rates spiked suddenly, and the number of accounts dropped like a valley. The development team also performed a recency, frequency, and monetary analysis. From the result, most customers had activities in the last 100 days, and the majority paid less than \$2,000 in total as of the final period in the dataset.

To predict the life span for each account version ID, developers applied linear regression and multi-layer neural network models and got mean squared error metrics of about 16 and 11 for the test. The team also used random forest, decision tree, support vector machine, and stochastic gradient descent models to analyze whether an account version ID would churn in its nth month. Based on the confusion matrices, those models present good accuracy at predicting true negative but not true positive labels.

11.2.1 Segmentation of Customer Base

Identification of impactful customer segments enables businesses to focus on advertisement, recruitment, and incentive to encourage the growth of that customer subset. One metric useful to subscription businesses is a customer's recency, frequency, monetary value score (RFM). This metric is used to group customers based on their value to the business, a customer with the highest RFM score had most recently interacted with the business, interacted with the business the most frequently, and provided the most revenue to the business. This project defines Recency as the customers' last month with the business, frequency as the customers attendance rate multiplied by their lifetime, and monetary value as a customer's total amount spent with the business.

Customer segmentation, by definition, brings together very similar customers, which in this project caused several issues. When segmenting by RFM, it was noticed that segments with lower RFM scores consisted of only terminated customers while segments with higher RFM scores consisted of only active customers and, therefore, could not be analyzed by the machine learning models as it was a single classification. Another issue with RFM segmentation is that

the initial calculation used for RFM included aggregate data from a customer's whole lifespan. This variable could inflate the model's accuracy if the model was only supposed to have data from a certain set of periods. The customers were also segmented by various categorical variables, with models being trained, run, and evaluated on the segmented data (see Appendix D). For other segmentation types, it caused drastic imbalance issues that required additional metrics to properly evaluate the results. Additionally, oversampling was utilized to combat this but added computational expense and possible overfitting on the minority class as it repeated the same data points.

11.2.2 Putting Data in Context of Time

To set up a new column as churn rate where the WPI MQP team segmented the data by different time periods and formed a timeframe. The developers hoped to see the correlation between churn rate and variables, and they hypothesized that the attendance rate, average recurring revenue, and periods could have a strong connection with churn rate. However, the heatmap of correlation analysis in figure 18 did not provide the expected insight of data. The columns with high correlation were the results instead of reasons of churn rate.

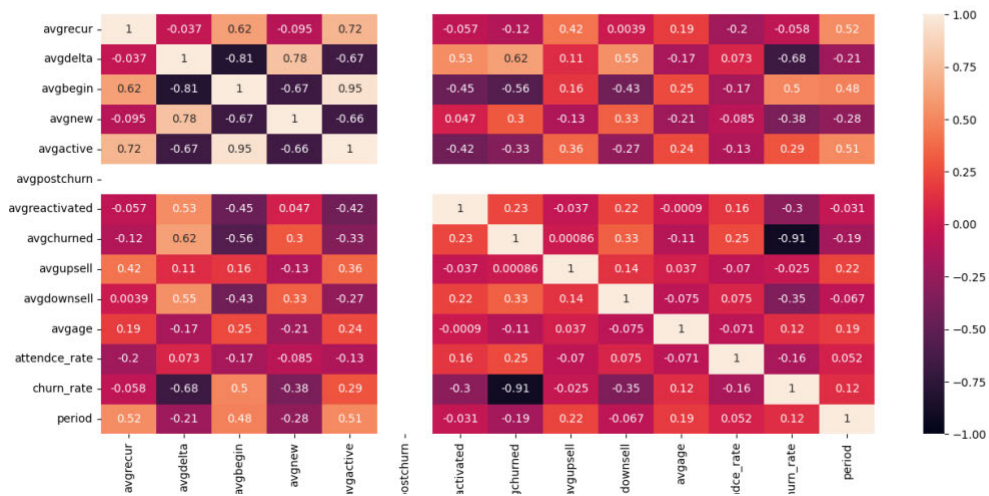


Figure 18: The Heatmap of Correlation Analysis

Enlightened by the meeting with sponsors, the developers recognized that the COVID pandemic might affect data. To prove this assumption, they plotted the graphs illustrated in

figure 19 and 10 with time periods against the churn rate and number of active accounts for different locations.

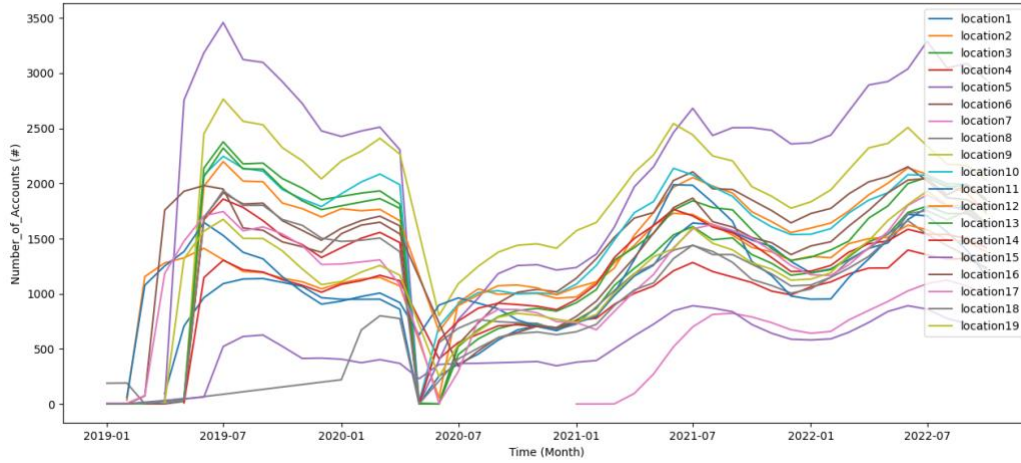


Figure 19: Time Periods Against the Number of Active Accounts

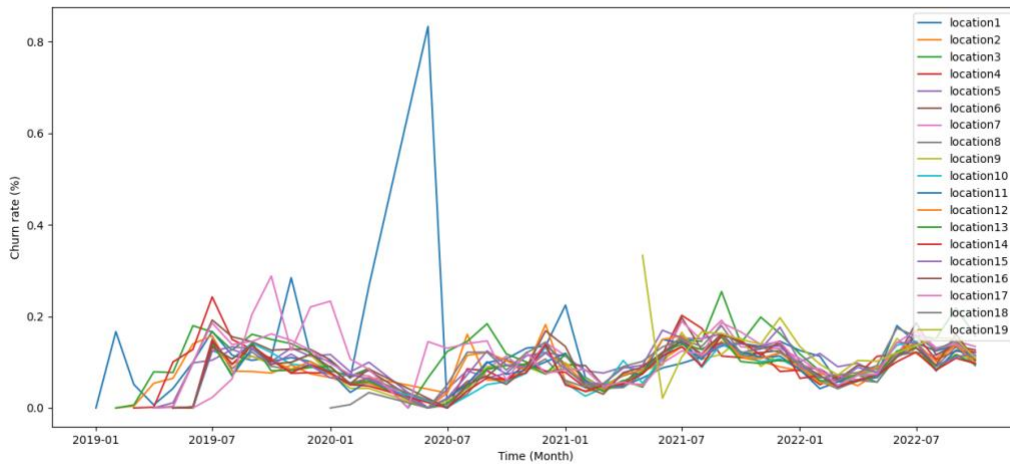


Figure 20: Time Periods Against the Churn Rate

As the graphs shown, COVID-19 impacted the churn rate and number of active accounts for classes between Jan and Jul in 2020, and the data was thus affected.

11.2.3 Relationship with Customer Lifetime Revenue

As the data provided to us from SaaSWorks was a subscription-based company that primarily charged flat fees (based on locality), it was identified early on that the greatest impact on an individual's lifetime revenue was their lifetime and not the amount or expense of the product they purchased as it could be in traditional business models. The focus in descriptive and predictive analytics gravitated towards identifying factors and relationships influencing customer lifetime. The team approached this from both a regression and classification perspective.

On the regression side, developers worked to accurately predict how long a customer would maintain their subscription before permanently terminating. To achieve the goal, the developers aggregated data by account version id, set the maximum of the period since joining as labels, and applied Multi-Linear Regression (MLR) and Multi-Layer Neural Network (MLNN). Mean Squared Error (MSE) was the chosen metrics to test these two models. The average MSE 5 MLNN models was 10.7140, and the mean MSE for 5 MLR models was 16.6334. A good MSE value should be closed to 0. Compared to the range, from 0 to 29 months of lifespan, these predictions indicated success for answering this question.

For the classification approach, the question was initially phrased as “given the first n months of data, does the customer exhibit behavior more like an active customer or a terminated customer.” This showed significant success—see Table 18 for comparison to other queries and see Appendix B for full table of results—and led to further analysis of the data to determine what the underlying behavior was that produced this relationship. However, after the results of that analysis, the question was rephrased to “given the *last* n months of data, does the customer exhibit behavior more like an active customer or a terminated customer.” This reformatting of the data resulted in a significant increase in accuracy and MCC score; see Appendix B for specific results of all classification models. Logically, this makes sense, as one would expect the most recently captured behavior of accounts to be more variance between active and terminated accounts than in the beginning when accounts appear to act in similar patterns.

The model training and testing results indicate that random forest is typically the most effective model for the dataset provided as it consistently scored the highest accuracy and MCC—see Table 18 for specific scores. When used in combination with the derived features of attendance rate and utilization revenue, random forests with entropy as the criterion marginally

outperformed the other random forest. The reason behind that is possibly due to the significant amount of categorical variables within the feature set as three nominal categorical variables were encoded to 39 Boolean features for ML processing. As previously discussed, random forests have an advantage over SVMs in feature sets containing significant categorical features. Additionally, random forests typically have higher test accuracy than decision trees as their wider selection of features reduces likelihood of overfitting (Schonlau & Jou, 2020).

Model Type	Last 3 Months with PCA		Last 3 Months no PCA		Last 3 months with PCA and Derived Fields		Last 3 months with Derived Fields	
	Accuracies	Matthews	Accuracies	Matthews	Accuracies	Matthews	Accuracies	Matthews:
SVM: Linear	0.70327	0.36080	0.78758	0.55242	0.71137	0.37922	0.78628	0.54971
SVM: RBF	0.79278	0.56244	0.79668	0.57015	0.78408	0.54251	0.80238	0.58239
SVM: Polynomial 3rd Degree	0.64146	0.20826	0.74147	0.44687	0.63376	0.19343	0.75338	0.47364
Decision Tree: Gini (Max depth 5)	0.78198	0.53788	0.83938	0.66296	0.74997	0.46698	0.83758	0.65866
Decision Tree: Entropy (Max depth 5)	0.78058	0.53425	0.82788	0.64163	0.74957	0.46657	0.82648	0.63781
Decision Tree: Log Loss (Max depth 5)	0.78058	0.53425	0.82778	0.64140	0.74977	0.46702	0.82638	0.63753
Decision Tree: Gini (Max depth 8)	0.79678	0.57289	0.87849	0.75092	0.76667	0.50436	0.87749	0.74965
Decision Tree: Entropy (Max depth 8)	0.79658	0.57040	0.87639	0.74783	0.76658	0.50396	0.87559	0.74523

Decision Tree: Log Loss (Max depth 8)	0.79618	0.56959	0.87619	0.74704	0.76678	0.50435	0.87579	0.74555
Random Forest: Gini (200 Estimators)	0.83188	0.64579	0.88349	0.75618	0.79928	0.57513	0.88489	0.75937
Random Forest: Entropy (200 Estimators)	0.82968	0.64104	0.88279	0.75472	0.80048	0.57777	0.88759	0.76531
Random Forest: Log Loss (200 Estimators)	0.83348	0.64928	0.88269	0.75486	0.79608	0.56834	0.88679	0.76343

Table 18: Most Effective Models as Determined by Average MCC and Accuracy.

Note: the highest score in each category is indicated in a darker green highlight.

11.3 Learning

As a learning experience, the team strongly believes that this has been a positive one. The team was able to experience and be integrated into a modern hybrid working environment and experience the pros and cons of the space. The entire team successfully navigated an Agile-Scrum workflow for the full lifecycle of a product. Developers had the opportunity to deepen their understanding of machine learning approaches, modern applications, and their ability to explain said approaches to external entities. The entire team gained a fuller understanding of the complexity interwoven into business analytics and the necessity for said complexity.

References

- Adamson, D. (2022, July 6). What's essential for a great risk culture within a business? GetRiskManager. Retrieved December 13, 2022, from <https://getriskmanager.com/risk-culture/>
- Algorithmia. (2020, December 10). 2021 enterprise trends in ML - Algorithmia. Algorithmia. Retrieved December 3, 2022, from https://info.algorithmia.com/hubfs/2020/Reports/2021-Trends-in-ML/Algorithmia_2021_enterprise_ML_trends.pdf?hsLang=en-us
- Alpaydin, E. (2021). Machine learning. Amazon. Retrieved November 17, 2022, from <https://docs.aws.amazon.com/machine-learning/latest/dg/cross-validation.html>
- Aning, S., & Przybyła-Kasperek, M. (2022). Comparative study of twoling and entropy criterion for decision tree classification of Dispersed Data. *Procedia Computer Science*, 207, 2434–2443. <https://doi.org/10.1016/j.procs.2022.09.301>
- Beck, K., Beedle, M., van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., Martin, R. C., Mellor, S., Thomas, D., Grenning, J., Highsmith, J., Hunt, A., Jeffries, R., Kern, J., Marick, B., Schwaber, K., & Sutherland, J. (2001, February). Principles behind the Agile Manifesto. Retrieved September 18, 2022, from <https://agilemanifesto.org/>
- Branco, P., Torgo, L., & Ribeiro, R. (2015, May 13). A survey of predictive modeling under imbalanced distributions. *arXiv.org*. Retrieved December 2, 2022, from <https://arxiv.org/abs/1505.01658>
- Brook, C., & Zhang, E. (2020, April 30). What is a SAAS company? Digital Guardian. Retrieved December 12, 2022, from <https://digitalguardian.com/blog/what-saas-company>
- Charbuty, B., & Abdulazeez, A. (2021). Classification based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, 2(01), 20–28. <https://doi.org/10.38094/jastt20165>
- Chaudhary, A., & Sharma, M. (2021). Multilayer neural network design for the calculation of risk factor associated with covid-19. *Augmented Human Research*, 6(1). <https://doi.org/10.1007/s41133-021-00044-4>
- Chicco, D., Tötsch, N., & Jurman, G. (2021, February 4). The Matthews Correlation Coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and

- markedness in two-class confusion matrix evaluation - biodata mining. BioMed Central. Retrieved December 2, 2022, from <https://biodatamining.biomedcentral.com/articles/10.1186/s13040-021-00244-z>
- Dasgupta, A., Sun, Y. V., König, I. R., Bailey-Wilson, J. E., & Malley, J. D. (2012, May 7). Brief review of regression-based and machine learning methods in genetic epidemiology: The Genetic Analysis Workshop 17 experience. *Genetic epidemiology*. Retrieved December 2, 2022, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3345521/#:~:text=Regression%2Dbased%20supervised%20methods%20attempt,are%20estimated%20from%20the%20data.>
- Gierula, A., Wang, S., OH, T.-M., & Wang, P. (2021, March 5). Study on accuracy metrics for evaluating the predictions of damage locations in deep piles using artificial neural networks with acoustic emission data. *MDPI*. Retrieved December 2, 2022, from <https://www.mdpi.com/2076-3417/11/5/2314/htm>
- How to create a Scrum Burndown Chart. *Lucidchart*. (2020, October 8). Retrieved December 9, 2022, from <https://www.lucidchart.com/blog/how-to-create-a-scrum-burndown-chart>
- IBM Cloud Education. (2020, July 15). What is machine learning? *IBM*. Retrieved December 2, 2022, from <https://www.ibm.com/cloud/learn/machine-learning>
- I.R. Team, I. R. (2022). Strategies for avoiding involuntary churn. *IR*. Retrieved December 5, 2022, from <https://www.ir.com/guides/involuntary-churn>
- Jackson, D. (2022, October 12). How to calculate customer lifetime value - the LTV formula. *Baremetrics*. Retrieved December 2, 2022, from <https://baremetrics.com/academy/saas-calculating-ltv>
- James Woodruff has been a ...more. (2022, May 31). Bottom line vs. top line: What's the difference for small business owners? *The Bottom Line*. Retrieved December 13, 2022, from <https://www.nationalfunding.com/blog/business-bottom-line/>
- Know your revenue. *SaaSWorks*. (2019). Retrieved December 9, 2022, from <https://www.saasworks.com/>
- Krstajic, D., Buturovic, L. J., Leahy, D. E., & Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics*, 6(1). <https://doi.org/10.1186/1758-2946-6-10>
- Levinson, M. (2007, May 15). Software as a service (SAAS) definition and solutions. *CIO*.

- Retrieved December 12, 2022, from <https://www.cio.com/article/272086/web-services-software-as-a-service-saas-definition-and-solutions.html>
- Manasa. (2022, February 8). Ai and ML are transforming SAAS. read the downsides. IndustryWired. Retrieved December 2, 2022, from <https://industrywired.com/ai-and-ml-are-transforming-saas-read-the-downsides/>
- Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, 1(4), 140–147. <https://doi.org/10.38094/jastt1457>
- Mechelli, A., Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In *Machine learning: Methods and applications to brain disorders* (pp. 101–121). essay, Academic Press.
- Montavon Grégoire, Orr, G., Müller Klaus-Robert, & Bottou, L. (2012). Stochastic Gradient Descent Tricks. In *Neural networks: Tricks of the Trade* (Vol. 7700, pp. 421–436). essay, Springer.
- Morgan, L. (2021, October 12). What is operational risk? definition from searchcompliance. Security. Retrieved December 3, 2022, from <https://www.techtarget.com/searchsecurity/definition/operational-risk>
- Project Management, A. for. (2022). What is stakeholder engagement? APM. Retrieved December 8, 2022, from <https://www.apm.org.uk/resources/find-a-resource/stakeholder-engagement/#:~:text=Definition,those%20business%20needs%20are%20met.>
- Purnami, S. W., Andari, S., & Pertiwi, Y. D. (2015). High-dimensional data classification based on smooth support Vector Machines. *Procedia Computer Science*, 72, 477–484. <https://doi.org/10.1016/j.procs.2015.12.129>
- Reply. (2022, April 18). What is churn rate? - types, calculation & industry benchmarks. Essential Business Guides. Retrieved December 13, 2022, from <https://www.zoho.com/finance/essential-business-guides/subscriptions/what-is-churn-rate.html>
- Schonlau, M., & Zou, R. Y. (2020). The Random Forest Algorithm for Statistical Learning. *The Stata Journal: Promoting Communications on Statistics and Stata*, 20(1), 3–29. <https://doi.org/10.1177/1536867x20909688>
- Schwaber, K., Sutherland, J. (2020). *The Scrum Guide The Definitive Guide to Scrum: The*

- Rules of the Game. scrumguides.org. Scrum Alliance. Retrieved September 19, 2022, from <https://scrumguides.org/docs/scrumguide/v2020/2020-Scrum-Guide-US.pdf#zoom=100>.
- Setia, M. (2022, December 2). Log loss - logistic regression's cost function for beginners. Analytics Vidhya. Retrieved December 9, 2022, from <https://www.analyticsvidhya.com/blog/2020/11/binary-cross-entropy-aka-log-loss-the-cost-function-used-in-logistic-regression/>
- Slack. (n.d.). Slack for enterprises. Slack Technologies. Retrieved October 11, 2022, from <https://slack.com/enterprise#:~:text=More%20than%2075%2C000%20businesses%20use,apps%20just%20for%20your%20organization.>
- Tangirala, S. (2020). Evaluating the impact of gini index and information gain on classification using decision tree classifier algorithm*. International Journal of Advanced Computer Science and Applications, 11(2). <https://doi.org/10.14569/ijacsa.2020.0110277>
- Voskoglou, C. (2019, December 11). What is the best programming language for Machine Learning? Medium. Retrieved September 28, 2022, from <https://towardsdatascience.com/what-is-the-best-programming-language-for-machine-learning-a745c156d6b7>
- WallStreetMojo, W. E. (2022, July 22). Bottom line. WallStreetMojo. Retrieved December 2, 2022, from <https://www.wallstreetmojo.com/bottom-line/>

Appendix A: PCA Weights

The following table represents the linear combination of features for the six principal components that account for 95% of the variance captured by the true feature space when applied to the last 3 active months (n=9999)

PC 0	PC 1	PC 2	PC 3	PC 4	PC 5	Feature Name
-4.67E-05	-1.52E-04	-5.94E-04	8.23E-05	4.95E-04	5.91E-04	account_status_1
1.42E-02	1.26E-02	-2.48E-02	1.63E-02	3.33E-02	1.16E-01	periods_since_joining_1
4.58E-01	-4.51E-01	-1.77E-01	5.17E-01	-1.43E-02	-5.11E-01	recurring_revenue_1
7.21E-03	3.32E-03	7.92E-04	7.00E-03	2.10E-01	2.53E-02	segment_type_4_1
3.96E-03	2.27E-03	4.81E-03	-4.23E-03	2.79E-01	-1.40E-02	segment_type_4_starting_1
3.96E-03	2.27E-03	4.81E-03	-4.23E-03	2.79E-01	-1.40E-02	segment_type_4_since_first_active_1
5.36E-03	1.84E-03	4.95E-03	1.82E-03	2.86E-01	-1.07E-02	segment_type_4_since_last_active_1
1.70E-04	7.63E-05	-4.01E-03	8.79E-03	-9.90E-02	2.29E-02	segment_type_9_1
-1.49E-03	-3.59E-04	-1.53E-03	5.28E-03	-1.05E-01	1.16E-02	segment_type_9_starting_1
-1.49E-03	-3.59E-04	-1.53E-03	5.28E-03	-1.05E-01	1.16E-02	segment_type_9_since_first_active_1
-9.40E-04	-8.97E-04	-1.90E-03	7.42E-03	-1.01E-01	1.33E-02	segment_type_9_since_last_active_1
1.91E-03	-4.79E-03	-1.72E-03	-3.52E-03	1.11E-03	-7.17E-03	n_usage_type_1_1
-2.44E-03	5.27E-03	1.05E-03	5.08E-02	-1.28E-03	-4.62E-02	n_usage_type_2_1

3.21E-03	-6.71E-03	-4.66E-03	1.05E-02	-5.96E-03	-9.92E-03	n_usage_type_3_1
6.11E-04	-1.33E-03	-4.01E-04	-9.05E-03	1.36E-04	8.06E-03	attendance_1
4.79E-01	-5.37E-01	-1.95E-01	-4.82E-01	9.35E-03	4.32E-01	utilization_rev_1
1.22E-04	-1.87E-04	-1.59E-04	1.35E-03	-2.64E-04	-3.88E-04	segment_type_7_State_3_1
-5.69E-04	-8.52E-05	-5.65E-04	-1.65E-03	8.06E-04	-7.77E-04	segment_type_7_State_8_1
-9.22E-05	-1.42E-04	-9.49E-05	-3.42E-04	-8.42E-04	-6.28E-05	segment_type_7_State_9_1
2.03E-04	1.84E-04	2.18E-04	-4.18E-04	1.06E-04	5.31E-04	segment_type_7_State_7_1
-8.20E-04	-6.85E-04	-9.63E-04	-2.17E-04	2.42E-03	-1.51E-03	segment_type_7_State_6_1
2.28E-03	-2.25E-05	7.88E-04	4.30E-03	-1.98E-03	9.18E-04	segment_type_7_State_1_1
-9.74E-04	5.90E-04	2.21E-04	-2.57E-03	6.57E-05	1.20E-03	segment_type_7_State_2_1
-5.09E-04	4.92E-05	-5.87E-05	-5.24E-04	-1.40E-06	-8.28E-06	segment_type_7_State_5_1
-3.13E-05	6.41E-05	3.15E-05	1.07E-04	-2.08E-04	6.02E-05	segment_type_7_State_4_1
3.85E-04	2.36E-04	5.81E-04	-3.00E-05	-9.98E-05	3.51E-05	segment_type_2_28b9cfa57b9a96de198455231ef0ba70_1
-1.89E-04	-2.61E-04	-3.53E-04	-4.74E-05	3.95E-04	-4.40E-04	segment_type_2_2f1868fd96a641a332c6189ed8e53804_1
-1.96E-04	1.48E-04	4.78E-05	-5.18E-04	-5.31E-04	4.70E-04	segment_type_2_3c3d2a058982d064818dc7754080cf6f_1
1.69E-03	-2.84E-05	4.13E-04	3.03E-03	-1.25E-03	7.34E-04	segment_type_2_3d3fe4dc7e315c7a2a4715adfa16df76_1
-2.03E-04	-5.68E-05	-6.77E-05	-2.95E-04	9.00E-04	-2.78E-04	segment_type_2_41953e52c4e7a0477852699f673cae5c_1
7.93E-05	-5.31E-05	-3.06E-05	2.80E-04	-5.93E-07	-2.43E-04	segment_type_2_4b1f46c494f0ab295024a02cd6ce8970_1
-5.57E-05	-1.86E-04	-1.01E-04	-3.50E-04	-4.93E-04	-1.87E-04	segment_type_2_4edefdd1e38b6bbf184b31ed8c7559f5_1

-4.07E-04	1.61E-04	1.91E-04	-1.20E-03	8.52E-04	2.39E-04	segment_type_2_50fefb5efb085fd11b1a4fd2b6dda0aa_1
-2.89E-04	4.18E-05	-3.62E-04	-8.36E-04	1.09E-03	-2.11E-04	segment_type_2_5b2c2ddaf3de0e11cad8ccbe17e743aa_1
-5.09E-04	4.92E-05	-5.87E-05	-5.24E-04	-1.40E-06	-8.28E-06	segment_type_2_6047f554b3a7ab64ac7b50523813858d_1
-3.71E-04	2.81E-04	-1.82E-05	-8.53E-04	-2.55E-04	4.88E-04	segment_type_2_6257f8b0af3a10f6c496c6689b8f58f7_1
-2.57E-04	-5.39E-05	-1.97E-04	2.55E-04	7.20E-04	-3.95E-04	segment_type_2_81f0268e5bf69bc375a57f477832cea7_1
-2.24E-04	5.91E-05	-1.01E-04	-4.64E-04	2.11E-04	-3.79E-04	segment_type_2_9fc9d17c21ea5dd4c8ca2d852a0cceb4_1
-3.13E-05	6.41E-05	3.15E-05	1.07E-04	-2.08E-04	6.02E-05	segment_type_2_a49a2793abf714d8d4a7fe9af8a94c34_1
2.03E-04	1.84E-04	2.18E-04	-4.18E-04	1.06E-04	5.31E-04	segment_type_2_c1036b50b5c01aab9e7980bb82073bf0_1
-9.22E-05	-1.42E-04	-9.49E-05	-3.42E-04	-8.42E-04	-6.28E-05	segment_type_2_c9851c697a10db731b17662e01c479e1_1
-1.71E-04	-3.13E-04	-3.45E-04	-1.30E-04	4.05E-04	-3.91E-04	segment_type_2_d3a555e01f1e72fc98bbe5e63cd45d64_1
5.91E-04	5.85E-06	3.75E-04	1.27E-03	-7.33E-04	1.84E-04	segment_type_2_ea599d139e26a15813027695c97d18ff_1
-1.30E-04	-8.68E-04	-1.50E-03	2.60E-04	2.54E-04	-1.98E-03	account_status_2
1.43E-02	1.22E-02	-2.66E-02	1.63E-02	3.55E-02	1.18E-01	periods_since_joining_2
3.51E-01	1.34E-01	4.94E-01	4.18E-01	1.59E-02	2.85E-01	recurring_revenue_2
7.52E-03	4.86E-03	1.89E-03	4.38E-03	2.26E-01	2.25E-02	segment_type_4_2
3.96E-03	2.27E-03	4.81E-03	-4.23E-03	2.79E-01	-1.40E-02	segment_type_4_starting_2
3.96E-03	2.27E-03	4.81E-03	-4.23E-03	2.79E-01	-1.40E-02	segment_type_4_since_first_active_2
5.38E-03	1.78E-03	5.20E-03	1.71E-03	2.87E-01	-1.14E-02	segment_type_4_since_last_active_2
2.50E-04	1.45E-04	-3.99E-03	8.37E-03	-9.90E-02	2.32E-02	segment_type_9_2

-1.49E-03	-3.59E-04	-1.53E-03	5.28E-03	-1.05E-01	1.16E-02	segment_type_9_starting_2
-1.49E-03	-3.59E-04	-1.53E-03	5.28E-03	-1.05E-01	1.16E-02	segment_type_9_since_first_active_2
-8.79E-04	-7.26E-04	-1.48E-03	7.37E-03	-1.01E-01	1.31E-02	segment_type_9_since_last_active_2
6.34E-04	2.73E-03	2.96E-03	3.46E-03	2.21E-03	3.31E-03	n_usage_type_1_2
-2.51E-03	1.85E-03	-1.16E-02	4.15E-02	3.28E-03	2.62E-02	n_usage_type_2_2
2.66E-03	7.22E-06	7.66E-03	-1.42E-03	-8.99E-03	-1.16E-03	n_usage_type_3_2
4.39E-04	-1.77E-04	2.06E-03	-6.50E-03	-5.12E-04	-4.06E-03	attendance_2
3.74E-01	1.05E-01	6.83E-01	-3.39E-01	-4.02E-02	-2.05E-01	utilization_rev_2
1.22E-04	-1.87E-04	-1.59E-04	1.35E-03	-2.64E-04	-3.88E-04	segment_type_7_State_3_2
-5.69E-04	-8.52E-05	-5.65E-04	-1.65E-03	8.06E-04	-7.77E-04	segment_type_7_State_8_2
-9.22E-05	-1.42E-04	-9.49E-05	-3.42E-04	-8.42E-04	-6.28E-05	segment_type_7_State_9_2
2.03E-04	1.84E-04	2.18E-04	-4.18E-04	1.06E-04	5.31E-04	segment_type_7_State_7_2
-8.20E-04	-6.85E-04	-9.63E-04	-2.17E-04	2.42E-03	-1.51E-03	segment_type_7_State_6_2
2.28E-03	-2.25E-05	7.88E-04	4.30E-03	-1.98E-03	9.18E-04	segment_type_7_State_1_2
-9.74E-04	5.90E-04	2.21E-04	-2.57E-03	6.57E-05	1.20E-03	segment_type_7_State_2_2
-5.09E-04	4.92E-05	-5.87E-05	-5.24E-04	-1.40E-06	-8.28E-06	segment_type_7_State_5_2
-3.13E-05	6.41E-05	3.15E-05	1.07E-04	-2.08E-04	6.02E-05	segment_type_7_State_4_2
3.85E-04	2.36E-04	5.81E-04	-3.00E-05	-9.98E-05	3.51E-05	segment_type_2_28b9cfa57b9a96de198455231ef0ba70_2
-1.89E-04	-2.61E-04	-3.53E-04	-4.74E-05	3.95E-04	-4.40E-04	segment_type_2_2f1868fd96a641a332c6189ed8e53804_2

-1.96E-04	1.48E-04	4.78E-05	-5.18E-04	-5.31E-04	4.70E-04	segment_type_2_3c3d2a058982d064818dc7754080cf6f_2
1.69E-03	-2.84E-05	4.13E-04	3.03E-03	-1.25E-03	7.34E-04	segment_type_2_3d3fe4dc7e315c7a2a4715adfa16df76_2
-2.03E-04	-5.68E-05	-6.77E-05	-2.95E-04	9.00E-04	-2.78E-04	segment_type_2_41953e52c4e7a0477852699f673cae5c_2
7.93E-05	-5.31E-05	-3.06E-05	2.80E-04	-5.93E-07	-2.43E-04	segment_type_2_4b1f46c494f0ab295024a02cd6ce8970_2
-5.57E-05	-1.86E-04	-1.01E-04	-3.50E-04	-4.93E-04	-1.87E-04	segment_type_2_4edefdd1e38b6bbf184b31ed8c7559f5_2
-4.07E-04	1.61E-04	1.91E-04	-1.20E-03	8.52E-04	2.39E-04	segment_type_2_50fefb5efb085fd11b1a4fd2b6dda0aa_2
-2.89E-04	4.18E-05	-3.62E-04	-8.36E-04	1.09E-03	-2.11E-04	segment_type_2_5b2c2ddaf3de0e11cad8ccbe17e743aa_2
-5.09E-04	4.92E-05	-5.87E-05	-5.24E-04	-1.40E-06	-8.28E-06	segment_type_2_6047f554b3a7ab64ac7b50523813858d_2
-3.71E-04	2.81E-04	-1.82E-05	-8.53E-04	-2.55E-04	4.88E-04	segment_type_2_6257f8b0af3a10f6c496c6689b8f58f7_2
-2.57E-04	-5.39E-05	-1.97E-04	2.55E-04	7.20E-04	-3.95E-04	segment_type_2_81f0268e5bf69bc375a57f477832cea7_2
-2.24E-04	5.91E-05	-1.01E-04	-4.64E-04	2.11E-04	-3.79E-04	segment_type_2_9fc9d17c21ea5dd4c8ca2d852a0cceb4_2
-3.13E-05	6.41E-05	3.15E-05	1.07E-04	-2.08E-04	6.02E-05	segment_type_2_a49a2793abf714d8d4a7fe9af8a94c34_2
2.03E-04	1.84E-04	2.18E-04	-4.18E-04	1.06E-04	5.31E-04	segment_type_2_c1036b50b5c01aab9e7980bb82073bf0_2
-9.22E-05	-1.42E-04	-9.49E-05	-3.42E-04	-8.42E-04	-6.28E-05	segment_type_2_c9851c697a10db731b17662e01c479e1_2
-1.71E-04	-3.13E-04	-3.45E-04	-1.30E-04	4.05E-04	-3.91E-04	segment_type_2_d3a555e01f1e72fc98bbe5e63cd45d64_2
5.91E-04	5.85E-06	3.75E-04	1.27E-03	-7.33E-04	1.84E-04	segment_type_2_ea599d139e26a15813027695c97d18ff_2
-9.39E-04	-3.47E-03	5.06E-03	2.62E-04	-6.92E-04	-7.24E-03	account_status_3
1.42E-02	1.02E-02	-3.11E-02	1.74E-02	3.66E-02	1.12E-01	periods_since_joining_3
3.82E-01	4.62E-01	-3.22E-01	2.86E-01	-9.63E-03	4.38E-01	recurring_revenue_3

7.39E-03	4.70E-03	-2.39E-04	3.67E-03	2.40E-01	1.89E-02	segment_type_4_3
3.96E-03	2.27E-03	4.81E-03	-4.23E-03	2.79E-01	-1.40E-02	segment_type_4_starting_3
3.96E-03	2.27E-03	4.81E-03	-4.23E-03	2.79E-01	-1.40E-02	segment_type_4_since_first_active_3
5.38E-03	1.83E-03	5.88E-03	4.30E-04	2.87E-01	-1.10E-02	segment_type_4_since_last_active_3
2.84E-04	5.56E-04	-3.48E-03	8.32E-03	-9.88E-02	2.41E-02	segment_type_9_3
-1.49E-03	-3.59E-04	-1.53E-03	5.28E-03	-1.05E-01	1.16E-02	segment_type_9_starting_3
-1.49E-03	-3.59E-04	-1.53E-03	5.28E-03	-1.05E-01	1.16E-02	segment_type_9_since_first_active_3
-8.20E-04	-1.33E-04	-8.18E-04	7.02E-03	-1.02E-01	1.47E-02	segment_type_9_since_last_active_3
1.75E-03	3.66E-03	-6.03E-03	7.36E-03	3.15E-03	1.43E-02	n_usage_type_1_3
-1.26E-03	-3.39E-03	1.12E-03	3.32E-02	8.36E-06	4.56E-02	n_usage_type_2_3
3.42E-03	8.35E-03	-7.58E-03	-9.13E-03	-8.41E-03	-8.68E-03	n_usage_type_3_3
2.84E-04	6.73E-04	-3.21E-04	-5.55E-03	2.32E-05	-7.55E-03	attendance_3
3.89E-01	5.14E-01	-3.36E-01	-3.49E-01	-9.17E-03	-4.31E-01	utilization_rev_3
1.22E-04	-1.87E-04	-1.59E-04	1.35E-03	-2.64E-04	-3.88E-04	segment_type_7_State_3_3
-5.69E-04	-8.52E-05	-5.65E-04	-1.65E-03	8.06E-04	-7.77E-04	segment_type_7_State_8_3
-9.22E-05	-1.42E-04	-9.49E-05	-3.42E-04	-8.42E-04	-6.28E-05	segment_type_7_State_9_3
2.03E-04	1.84E-04	2.18E-04	-4.18E-04	1.06E-04	5.31E-04	segment_type_7_State_7_3
-8.20E-04	-6.85E-04	-9.63E-04	-2.17E-04	2.42E-03	-1.51E-03	segment_type_7_State_6_3
2.28E-03	-2.25E-05	7.88E-04	4.30E-03	-1.98E-03	9.18E-04	segment_type_7_State_1_3

-9.74E-04	5.90E-04	2.21E-04	-2.57E-03	6.57E-05	1.20E-03	segment_type_7_State_2_3
-5.09E-04	4.92E-05	-5.87E-05	-5.24E-04	-1.40E-06	-8.28E-06	segment_type_7_State_5_3
-3.13E-05	6.41E-05	3.15E-05	1.07E-04	-2.08E-04	6.02E-05	segment_type_7_State_4_3
3.85E-04	2.36E-04	5.81E-04	-3.00E-05	-9.98E-05	3.51E-05	segment_type_2_28b9cfa57b9a96de198455231ef0ba70_3
-1.89E-04	-2.61E-04	-3.53E-04	-4.74E-05	3.95E-04	-4.40E-04	segment_type_2_2f1868fd96a641a332c6189ed8e53804_3
-1.96E-04	1.48E-04	4.78E-05	-5.18E-04	-5.31E-04	4.70E-04	segment_type_2_3c3d2a058982d064818dc7754080cf6f_3
1.69E-03	-2.84E-05	4.13E-04	3.03E-03	-1.25E-03	7.34E-04	segment_type_2_3d3fe4dc7e315c7a2a4715adfa16df76_3
-2.03E-04	-5.68E-05	-6.77E-05	-2.95E-04	9.00E-04	-2.78E-04	segment_type_2_41953e52c4e7a0477852699f673cae5c_3
7.93E-05	-5.31E-05	-3.06E-05	2.80E-04	-5.93E-07	-2.43E-04	segment_type_2_4b1f46c494f0ab295024a02cd6ce8970_3
-5.57E-05	-1.86E-04	-1.01E-04	-3.50E-04	-4.93E-04	-1.87E-04	segment_type_2_4edefdd1e38b6bbf184b31ed8c7559f5_3
-4.07E-04	1.61E-04	1.91E-04	-1.20E-03	8.52E-04	2.39E-04	segment_type_2_50fefb5efb085fd11b1a4fd2b6dda0aa_3
-2.89E-04	4.18E-05	-3.62E-04	-8.36E-04	1.09E-03	-2.11E-04	segment_type_2_5b2c2dda3de0e11cad8ccbe17e743aa_3
-5.09E-04	4.92E-05	-5.87E-05	-5.24E-04	-1.40E-06	-8.28E-06	segment_type_2_6047f554b3a7ab64ac7b50523813858d_3
-3.71E-04	2.81E-04	-1.82E-05	-8.53E-04	-2.55E-04	4.88E-04	segment_type_2_6257f8b0af3a10f6c496c6689b8f58f7_3
-2.57E-04	-5.39E-05	-1.97E-04	2.55E-04	7.20E-04	-3.95E-04	segment_type_2_81f0268e5bf69bc375a57f477832cea7_3
-2.24E-04	5.91E-05	-1.01E-04	-4.64E-04	2.11E-04	-3.79E-04	segment_type_2_9fc9d17c21ea5dd4c8ca2d852a0cceb4_3
-3.13E-05	6.41E-05	3.15E-05	1.07E-04	-2.08E-04	6.02E-05	segment_type_2_a49a2793abf714d8d4a7fe9af8a94c34_3
2.03E-04	1.84E-04	2.18E-04	-4.18E-04	1.06E-04	5.31E-04	segment_type_2_c1036b50b5c01aab9e7980bb82073bf0_3
-9.22E-05	-1.42E-04	-9.49E-05	-3.42E-04	-8.42E-04	-6.28E-05	segment_type_2_c9851c697a10db731b17662e01c479e1_3

-1.71E-04	-3.13E-04	-3.45E-04	-1.30E-04	4.05E-04	-3.91E-04	segment_type_2_d3a555e01f1e72fc98bbe5e63cd45d64_3
5.91E-04	5.85E-06	3.75E-04	1.27E-03	-7.33E-04	1.84E-04	segment_type_2_ea599d139e26a15813027695c97d18ff_3

Appendix B: Results of Repeated Training and Testing of 12 Classification Models

- Accuracies and MCC for the six different analyses conducted on the customer data
- For models that implemented principal component analysis, PCA captured 95% of the variance.
- Models with derived fields included two engineered features:
 - Attendance: $(n_usage_1+n_usage_3)/(n_usage_1+n_usage_3+n_usage_3)$
 - Utilization: Attendance * Recurring Revenue
 - The derived features were calculated prior to PCA application in the applicable analysis.
- N = 9999, Terminated accounts = 3975, Active Accounts = 6024

Model Type	First 3 Months with PCA and Derived Fields		First 3 Months with Derived Fields		Last 3 Months with PCA		Last 3 Months no PCA		Last 3 months with PCA and Derived Fields		Last 3 months with Derived Fields	
	Accuracies	Matthews	Accuracies	Matthews	Accuracies	Matthews	Accuracies	Matthews	Accuracies	Matthews	Accuracies	Matthews:
SVM: Linear	0.60246	0.00000	0.70677	0.37230	0.70327	0.36080	0.78758	0.55242	0.71137	0.37922	0.78628	0.54971
SVM: RBF	0.68177	0.31077	0.71507	0.39040	0.79278	0.56244	0.79668	0.57015	0.78408	0.54251	0.80238	0.58239
SVM: Polynomial 3rd Degree	0.62236	0.14899	0.67107	0.27998	0.64146	0.20826	0.74147	0.44687	0.63376	0.19343	0.75338	0.47364
Decision Tree: Gini	0.67427	0.31787	0.75067	0.49548	0.78198	0.53788	0.83938	0.66296	0.74997	0.46698	0.83758	0.65866

(Max depth 5)												
Decision Tree: Entropy (Max depth 5)	0.67297	0.31395	0.75017	0.49869	0.78058	0.53425	0.82788	0.64163	0.74957	0.46657	0.82648	0.63781
Decision Tree: Log Loss (Max depth 5)	0.67307	0.31419	0.75017	0.49869	0.78058	0.53425	0.82778	0.64140	0.74977	0.46702	0.82638	0.63753
Decision Tree: Gini (Max depth 8)	0.67257	0.31695	0.74947	0.49693	0.79678	0.57289	0.87849	0.75092	0.76667	0.50436	0.87749	0.74965
Decision Tree: Entropy (Max depth 8)	0.67067	0.31705	0.74937	0.49804	0.79658	0.57040	0.87639	0.74783	0.76658	0.50396	0.87559	0.74523
Decision Tree: Log Loss (Max depth 8)	0.67127	0.31817	0.75057	0.50092	0.79618	0.56959	0.87619	0.74704	0.76678	0.50435	0.87579	0.74555
Random Forest: Gini (200 Estimators)	0.68787	0.33761	0.74987	0.47799	0.83188	0.64579	0.88349	0.75618	0.79928	0.57513	0.88489	0.75937
Random Forest: Entropy (200 Estimators)	0.68647	0.33364	0.74957	0.47791	0.82968	0.64104	0.88279	0.75472	0.80048	0.57777	0.88759	0.76531
Random Forest: Log Loss (200 Estimators)	0.69097	0.34419	0.75277	0.48524	0.83348	0.64928	0.88269	0.75486	0.79608	0.56834	0.88679	0.76343

Appendix C: Feature Importance

Feature importance is determined by the decrease in Matthew's Correlation Coefficient in a test set when feature is removed and is averaged over 30 trials. The following feature importance is based on the model with the overall highest MCC and accuracy: Random Forest with 200 estimators using entropy as its loss function with no PCA and with the two derived features.

Feature Name	Impact on MCC	STD
recurring_revenue_2	0.151	0.023
recurring_revenue_1	0.132	0.025
recurring_revenue_3	0.11	0.029
utilization_rev_2	0.04	0.018
periods_since_joining_3	0.025	0.019
utilization_rev_1	0.022	0.02
periods_since_joining_1	0.016	0.017
segment_type_7_State_1_2	0.015	0.006
segment_type_7_State_1_1	0.015	0.004
utilization_rev_3	0.014	0.017
n_usage_type_2_1	0.013	0.006
n_usage_type_3_1	0.011	0.006
periods_since_joining_2	0.011	0.016

n_usage_type_3_2	0.01	0.009
segment_type_2_3d3fe4dc7e315c7a2a4715adfa16df76_1	0.009	0.004
segment_type_2_50fefb5efb085fd11b1a4fd2b6dda0aa_3	0.008	0.002
segment_type_7_State_1_3	0.008	0.004
segment_type_2_ea599d139e26a15813027695c97d18ff_1	0.008	0.003
segment_type_4_2	0.007	0.01
account_status_3	0.006	0.007
segment_type_9_since_last_active_2	0.006	0.004
segment_type_7_State_2_1	0.005	0.006
segment_type_9_starting_3	0.005	0.01
segment_type_4_since_first_active_3	0.004	0.007
segment_type_4_starting_1	0.004	0.006
attendance_2	0.004	0.007
segment_type_9_since_first_active_1	0.004	0.009
attendance_1	0.003	0.006
n_usage_type_2_2	0.003	0.004
segment_type_7_State_5_1	0.002	0.004
segment_type_7_State_5_2	0.002	0.004
segment_type_7_State_2_3	0.001	0.003
segment_type_7_State_8_1	0.001	0.004

segment_type_9_3	0.001	0.005
segment_type_7_State_5_3	0.001	0.003
segment_type_2_81f0268e5bf69bc375a57f477832cea7_2	0.001	0.003
segment_type_2_6047f554b3a7ab64ac7b50523813858d_3	0.001	0.003
segment_type_2_81f0268e5bf69bc375a57f477832cea7_3	0.001	0.003
segment_type_2_4b1f46c494f0ab295024a02cd6ce8970_3	0.001	0.003
segment_type_2_3d3fe4dc7e315c7a2a4715adfa16df76_3	0.001	0.003
n_usage_type_1_1	0.001	0.006
segment_type_2_4b1f46c494f0ab295024a02cd6ce8970_1	0.001	0.003
segment_type_2_6047f554b3a7ab64ac7b50523813858d_1	0.001	0.003
segment_type_2_6047f554b3a7ab64ac7b50523813858d_2	0.001	0.003
account_status_1	0.001	0.003
attendance_3	0.001	0.004
segment_type_7_State_3_1	0.001	0.004
segment_type_2_28b9cfa57b9a96de198455231ef0ba70_1	0.001	0.002
segment_type_7_State_3_3	0.001	0.002
segment_type_2_6257f8b0af3a10f6c496c6689b8f58f7_2	0.001	0.002
segment_type_7_State_3_2	0.001	0.002
segment_type_2_3c3d2a058982d064818dc7754080cf6f_1	0.001	0.003
segment_type_2_3d3fe4dc7e315c7a2a4715adfa16df76_2	0.001	0.004

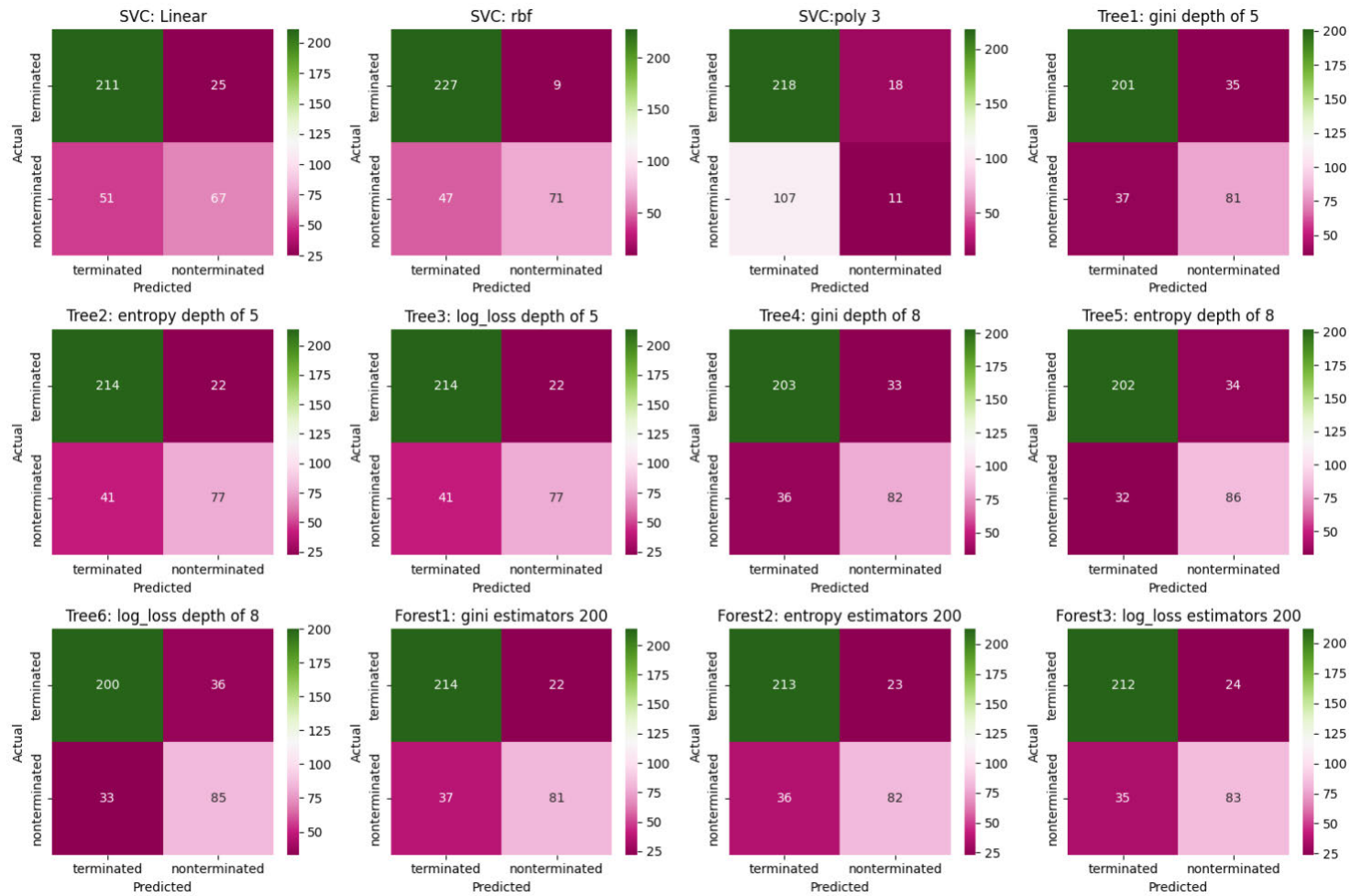
n_usage_type_3_3	0.001	0.007
segment_type_4_since_first_active_1	0.001	0.008
segment_type_9_starting_1	0.001	0.008
segment_type_2_2f1868fd96a641a332c6189ed8e53804_1	0	0.002
segment_type_2_4b1f46c494f0ab295024a02cd6ce8970_2	0	0.002
segment_type_2_41953e52c4e7a0477852699f673cae5c_2	0	0.002
segment_type_2_2f1868fd96a641a332c6189ed8e53804_3	0	0.002
segment_type_4_1	0	0.009
segment_type_9_since_first_active_2	0	0.009
segment_type_2_a49a2793abf714d8d4a7fe9af8a94c34_1	0	0
segment_type_2_50fefb5efb085fd11b1a4fd2b6dda0aa_1	0	0
m_1	0	0
segment_type_7_State_9_1	0	0
segment_type_7_State_7_1	0	0
segment_type_7_State_6_1	0	0
segment_type_2_41953e52c4e7a0477852699f673cae5c_1	0	0
segment_type_2_4edefdd1e38b6bbf184b31ed8c7559f5_1	0	0
segment_type_2_5b2c2ddaf3de0e11cad8ccbe17e743aa_1	0	0
segment_type_2_9fc9d17c21ea5dd4c8ca2d852a0cceb4_1	0	0
account_status_2	0	0

segment_type_2_6257f8b0af3a10f6c496c6689b8f58f7_1	0	0
segment_type_2_d3a555e01f1e72fc98bbe5e63cd45d64_1	0	0
segment_type_2_c9851c697a10db731b17662e01c479e1_1	0	0
segment_type_2_81f0268e5bf69bc375a57f477832cea7_1	0	0
segment_type_2_c1036b50b5c01aab9e7980bb82073bf0_1	0	0
segment_type_7_State_4_1	0	0
segment_type_2_ea599d139e26a15813027695c97d18ff_3	0	0
segment_type_7_State_6_2	0	0
segment_type_2_4edefdd1e38b6bbf184b31ed8c7559f5_2	0	0
segment_type_7_State_9_3	0	0
segment_type_2_3c3d2a058982d064818dc7754080cf6f_3	0	0
m_3	0	0
segment_type_2_41953e52c4e7a0477852699f673cae5c_3	0	0
segment_type_2_4edefdd1e38b6bbf184b31ed8c7559f5_3	0	0
segment_type_2_5b2c2ddaf3de0e11cad8ccbe17e743aa_3	0	0
segment_type_2_ea599d139e26a15813027695c97d18ff_2	0	0
segment_type_2_d3a555e01f1e72fc98bbe5e63cd45d64_2	0	0
segment_type_2_c9851c697a10db731b17662e01c479e1_2	0	0
segment_type_2_6257f8b0af3a10f6c496c6689b8f58f7_3	0	0
segment_type_2_a49a2793abf714d8d4a7fe9af8a94c34_2	0	0

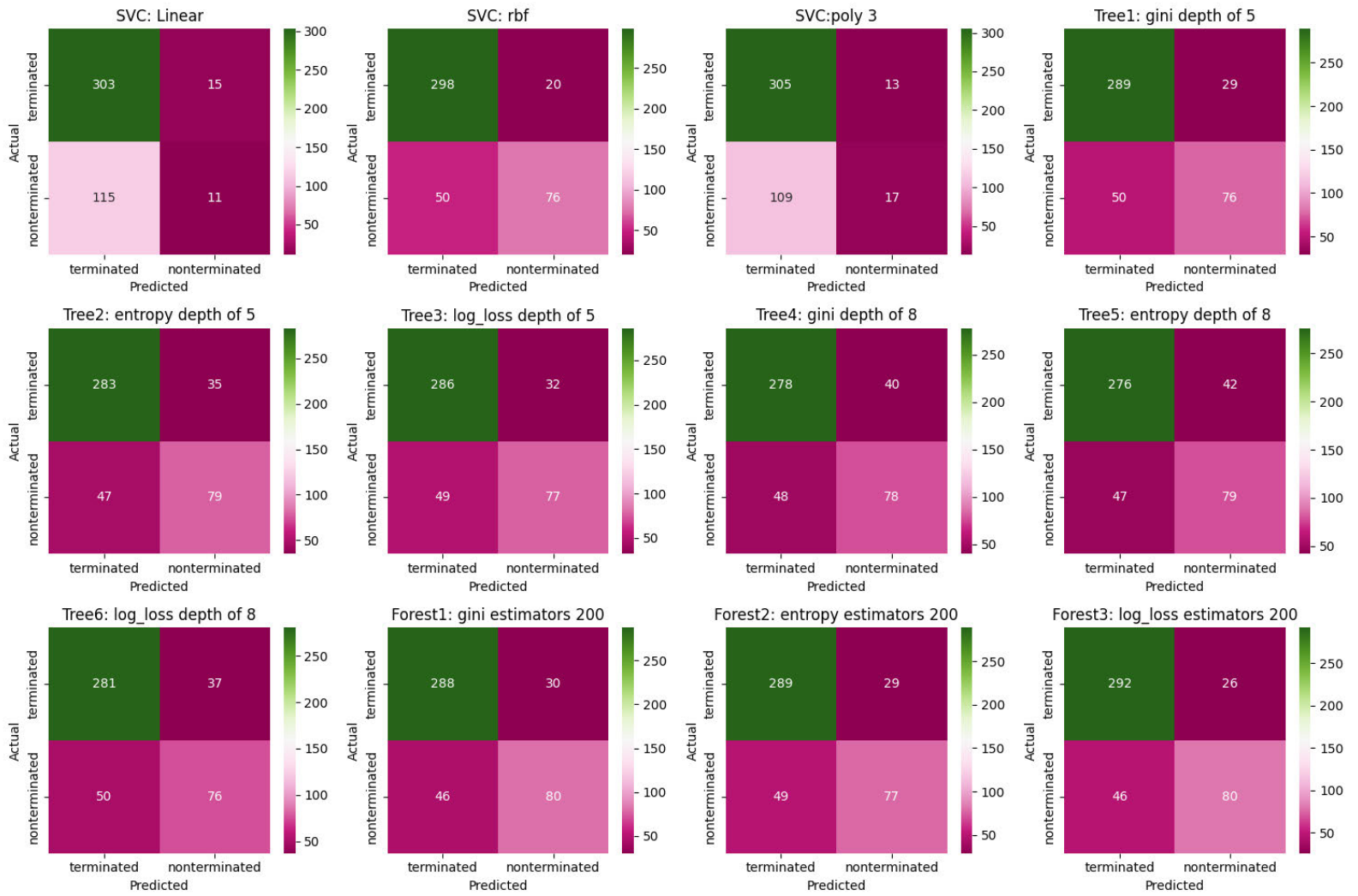
segment_type_2_9fc9d17c21ea5dd4c8ca2d852a0cceb4_2	0	0
segment_type_7_State_4_3	0	0
segment_type_2_50fe5fb5efb085fd11b1a4fd2b6dda0aa_2	0	0
segment_type_2_5b2c2ddaf3de0e11cad8ccbe17e743aa_2	0	0
segment_type_7_State_6_3	0	0
segment_type_7_State_4_2	0	0
segment_type_2_9fc9d17c21ea5dd4c8ca2d852a0cceb4_3	0	0
segment_type_2_d3a555e01f1e72fc98bbe5e63cd45d64_3	0	0
segment_type_2_a49a2793abf714d8d4a7fe9af8a94c34_3	0	0
segment_type_2_c9851c697a10db731b17662e01c479e1_3	0	0
segment_type_7_State_8_2	0	0
m_2	0	0
segment_type_2_2f1868fd96a641a332c6189ed8e53804_2	0	0
segment_type_2_c1036b50b5c01aab9e7980bb82073bf0_3	0	0
segment_type_2_28b9cfa57b9a96de198455231ef0ba70_3	0	0.004
segment_type_9_starting_2	0	0.006
segment_type_2_3c3d2a058982d064818dc7754080cf6f_2	0	0.002
segment_type_7_State_9_2	-0.001	0.002
N_usage_type_2_	-0.001	0.003
segment_type_2_28b9cfa57b9a96de198455231ef0ba70_2	-0.001	0.003

n_usage_type_1_2	-0.001	0.006
segment_type_9_since_last_active_1	-0.002	0.005
segment_type_4_since_last_active_3	-0.002	0.009
segment_type_9_since_last_active_3	-0.002	0.004
segment_type_9_2	-0.003	0.005
segment_type_4_3	-0.003	0.009
segment_type_9_1	-0.004	0.005
segment_type_7_State_8_3	-0.004	0.006
segment_type_4_since_first_active_2	-0.004	0.006
segment_type_9_since_first_active_3	-0.005	0.008
segment_type_4_since_last_active_1	-0.006	0.01
segment_type_7_State_2_2	-0.006	0.008
segment_type_7_State_7_3	-0.009	0.002
segment_type_2_c1036b50b5c01aab9e7980bb82073bf0_2	-0.009	0.002
segment_type_7_State_7_2	-0.009	0.002
segment_type_4_starting_3	-0.009	0.003
segment_type_4_starting_2	-0.009	0.005
segment_type_4_since_last_active_2	-0.01	0.009
n_usage_type_1_3	-0.01	0.006

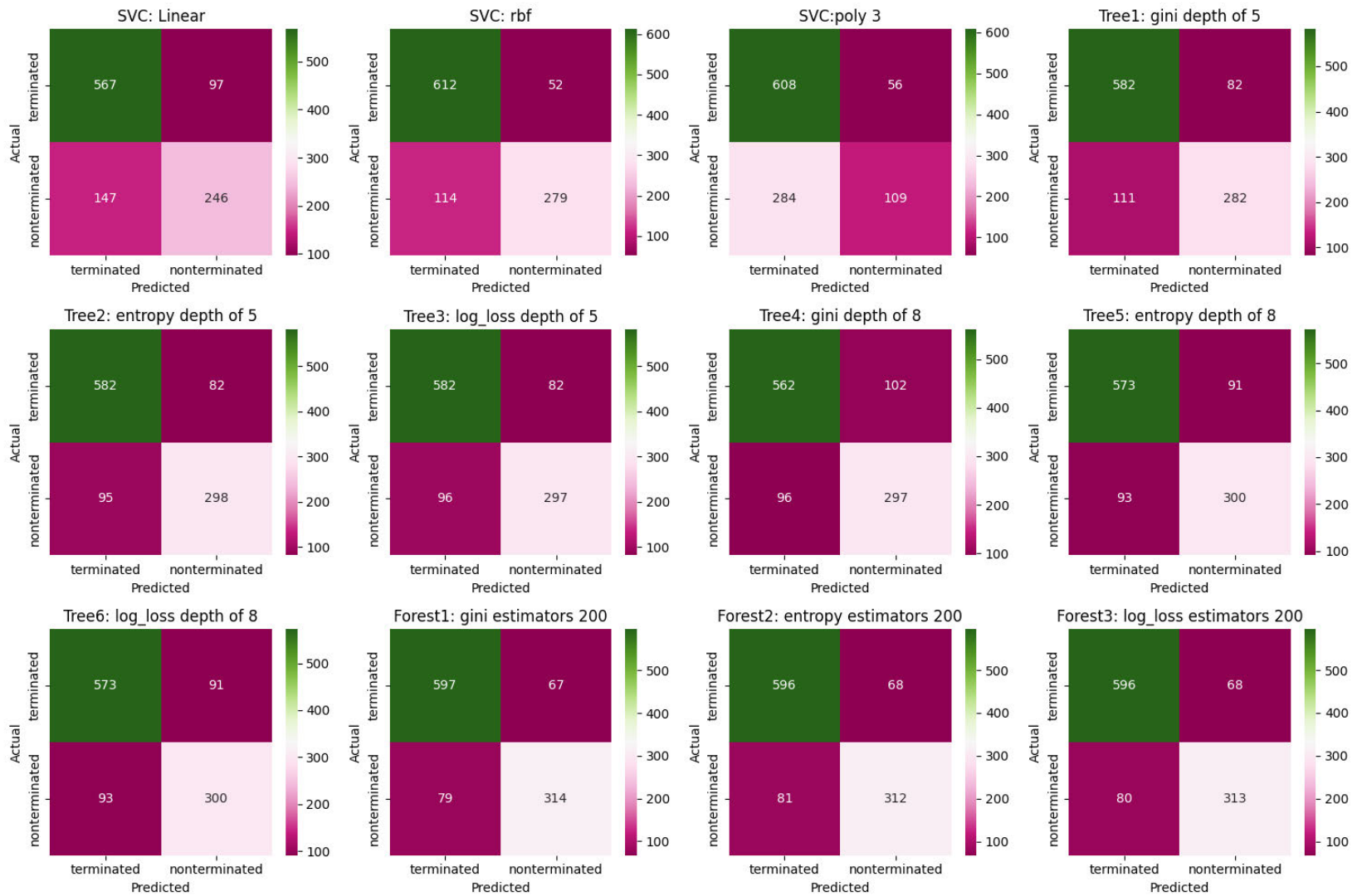
Appendix D: State Segmented Confusion Matrix Model Evaluations



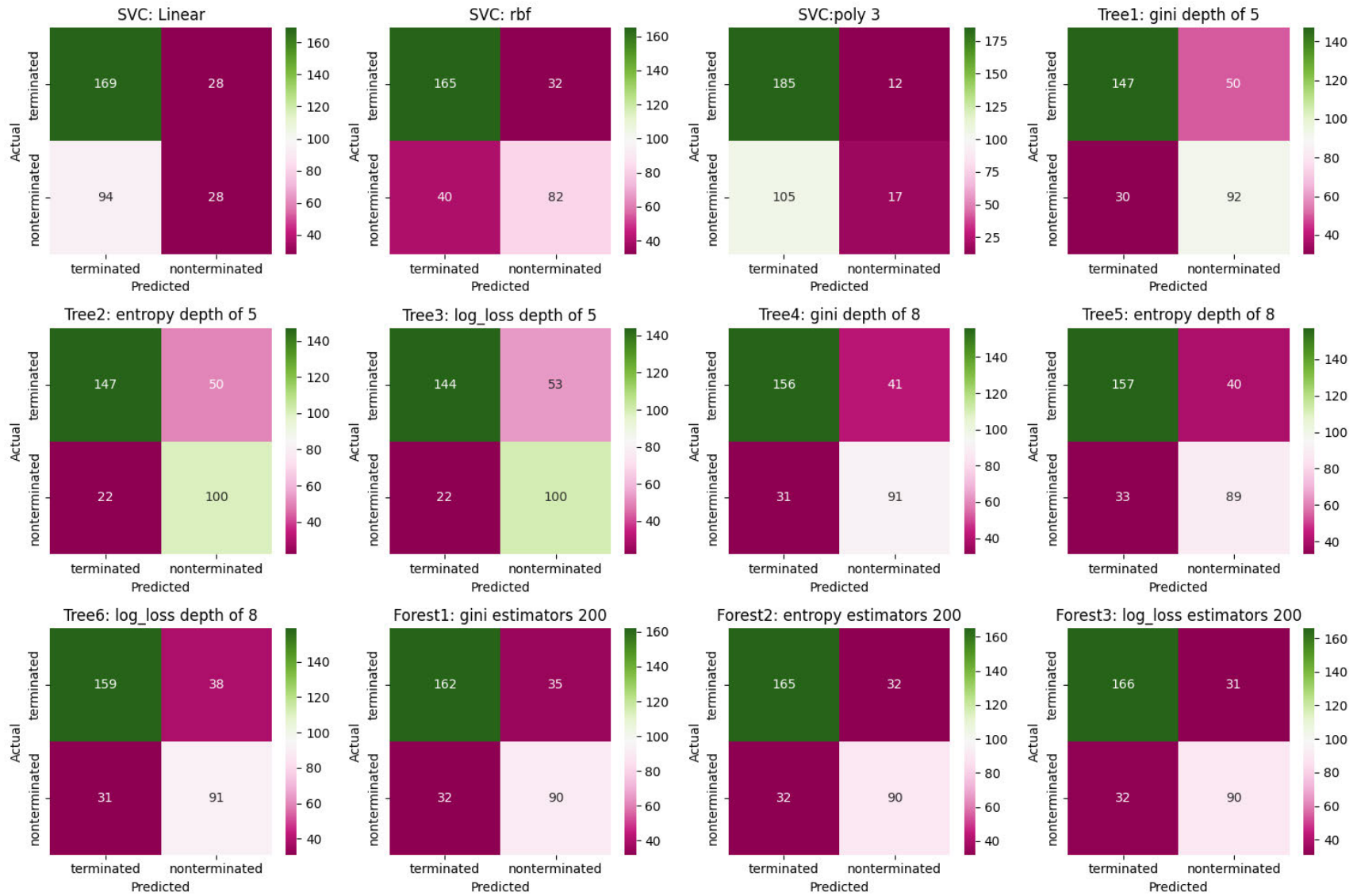
State_10 Segment Confusion matrices for classification models (n=354)



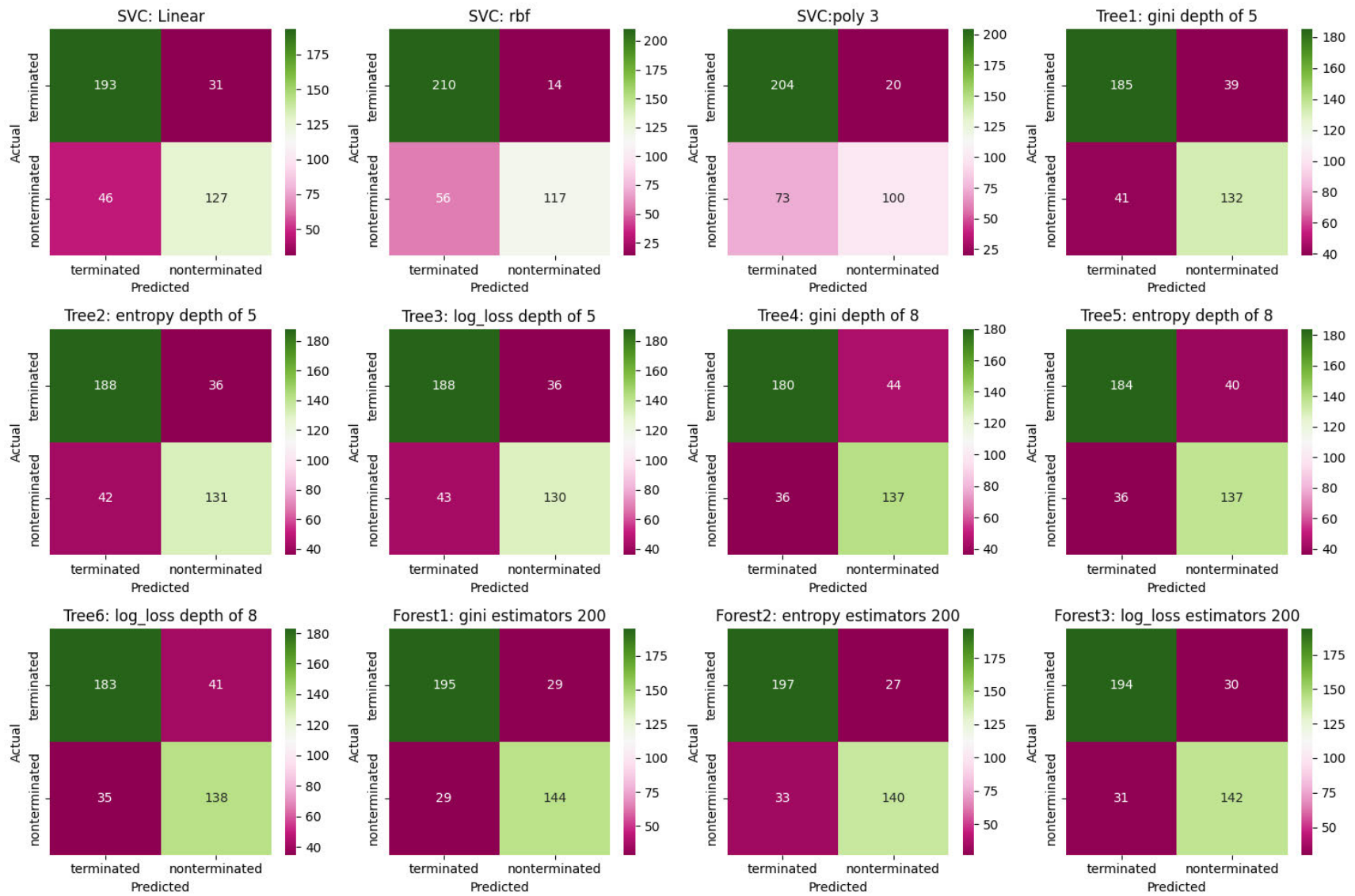
State_3 Segment Confusion matrices for classification models (n=444)



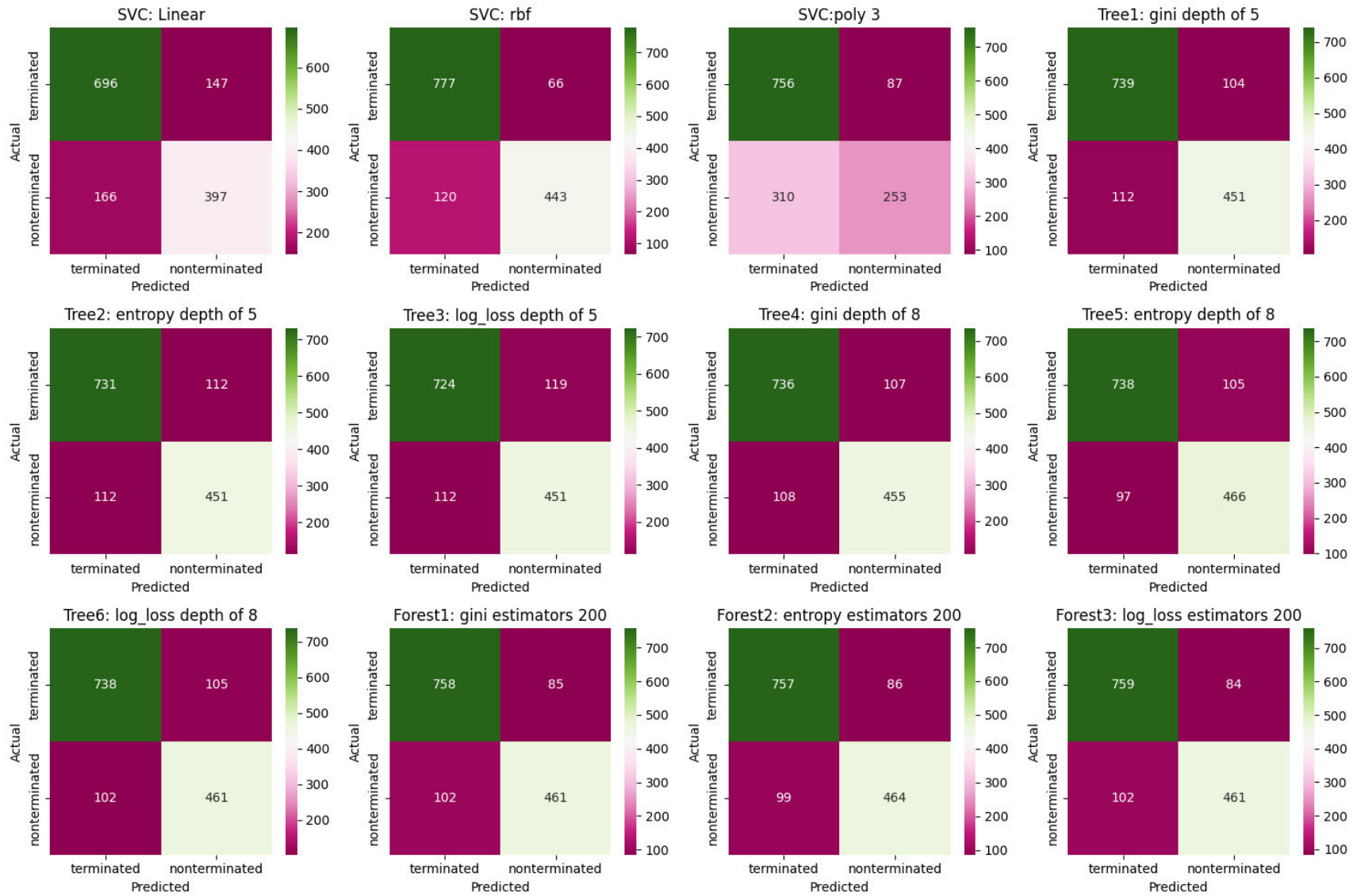
State_8 Segment Confusion matrices for classification models (n=1057)



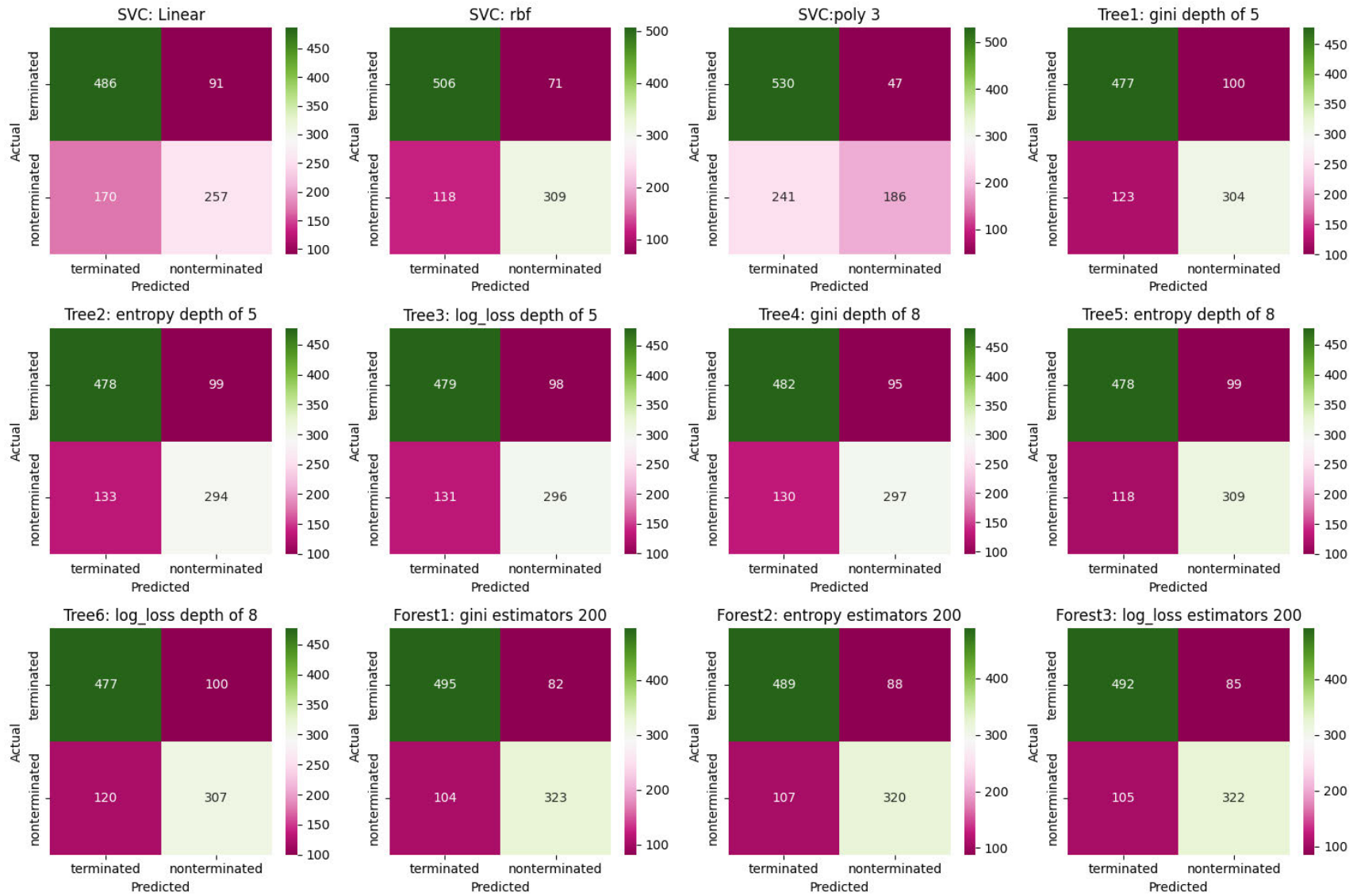
State_9 Segment Confusion matrices for classification models (n=319)



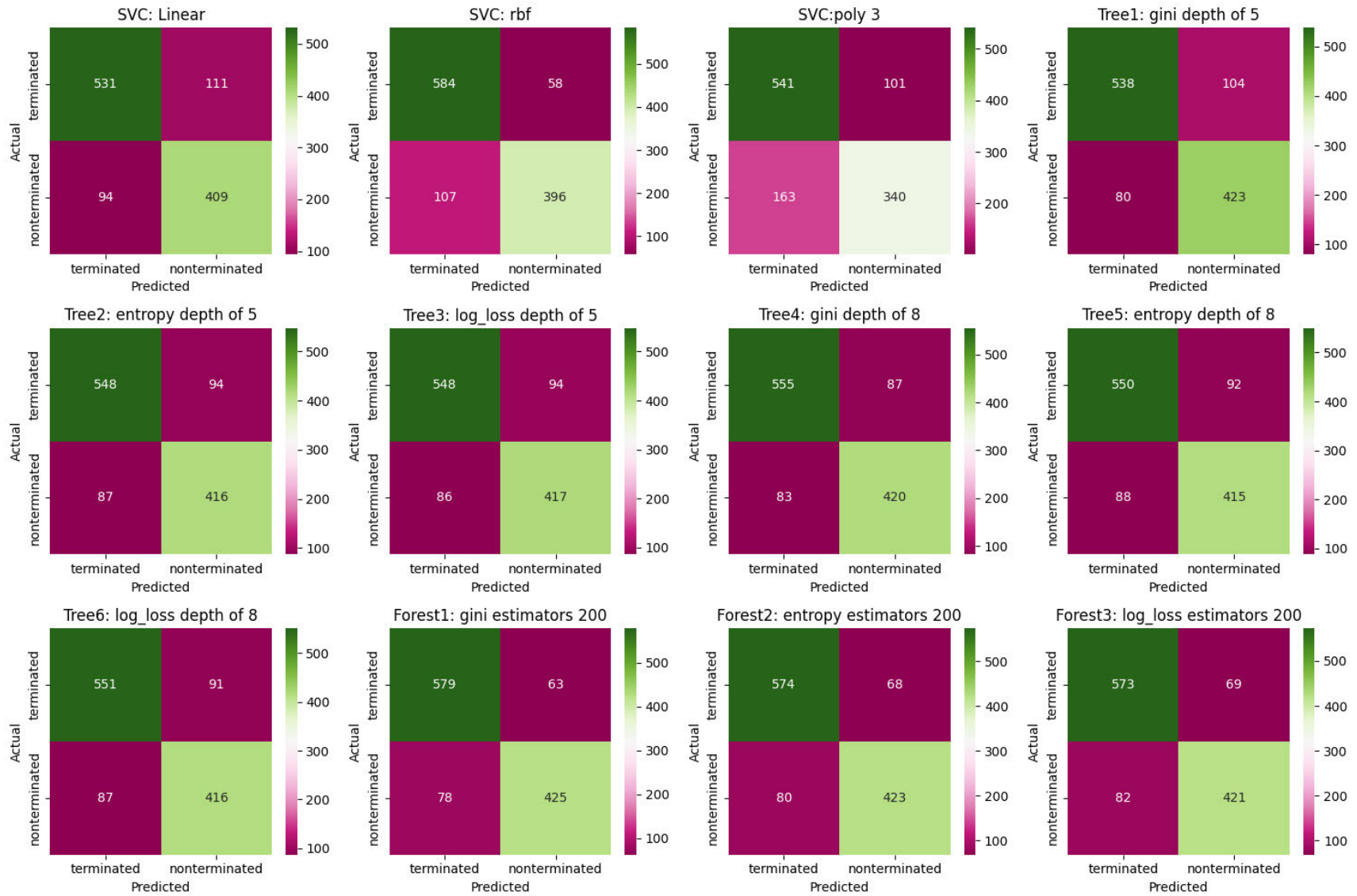
State_7 Segment Confusion matrices for classification models (n=397)



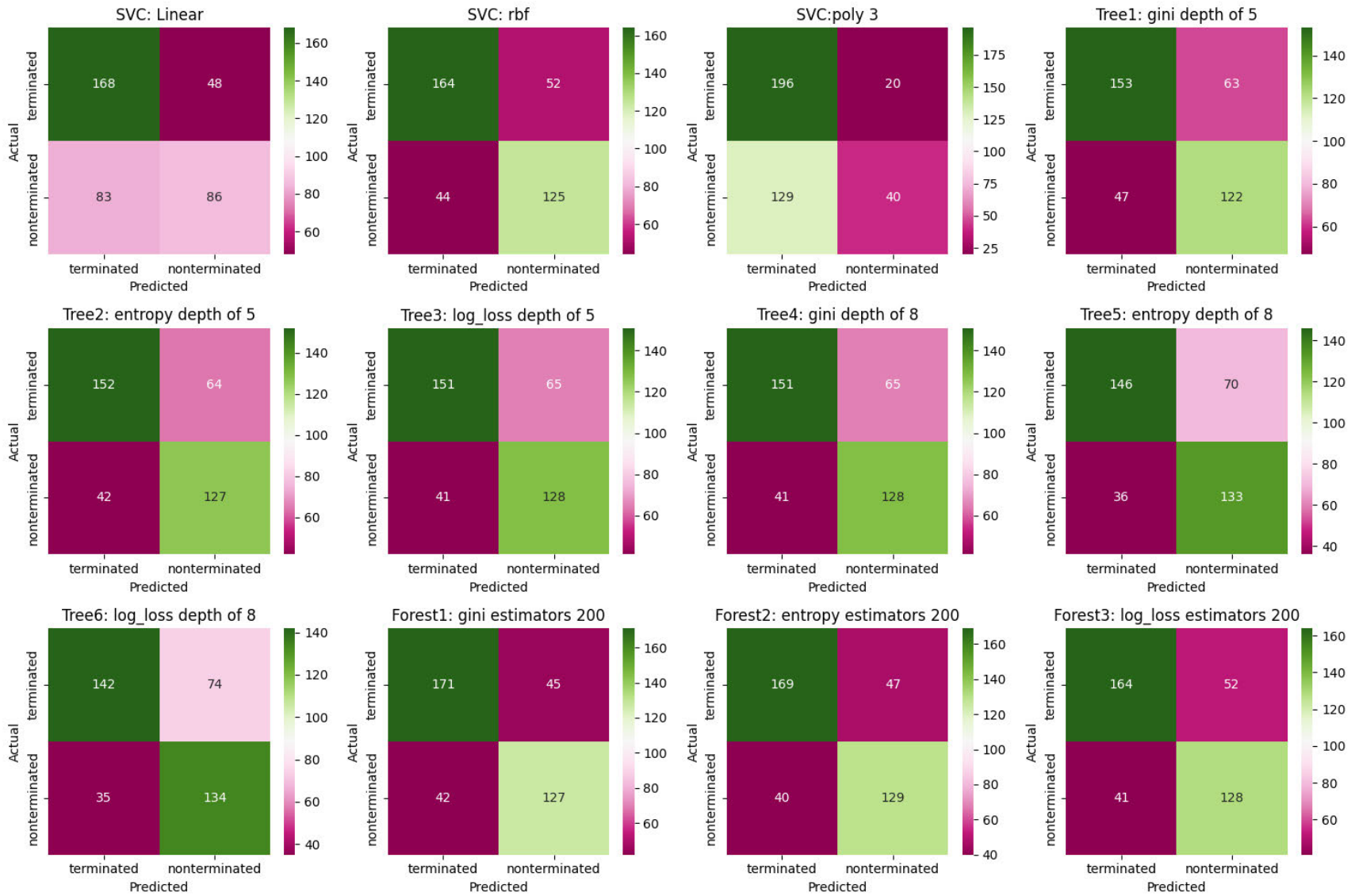
State_6 Segment Confusion matrices for classification models (n=1406)



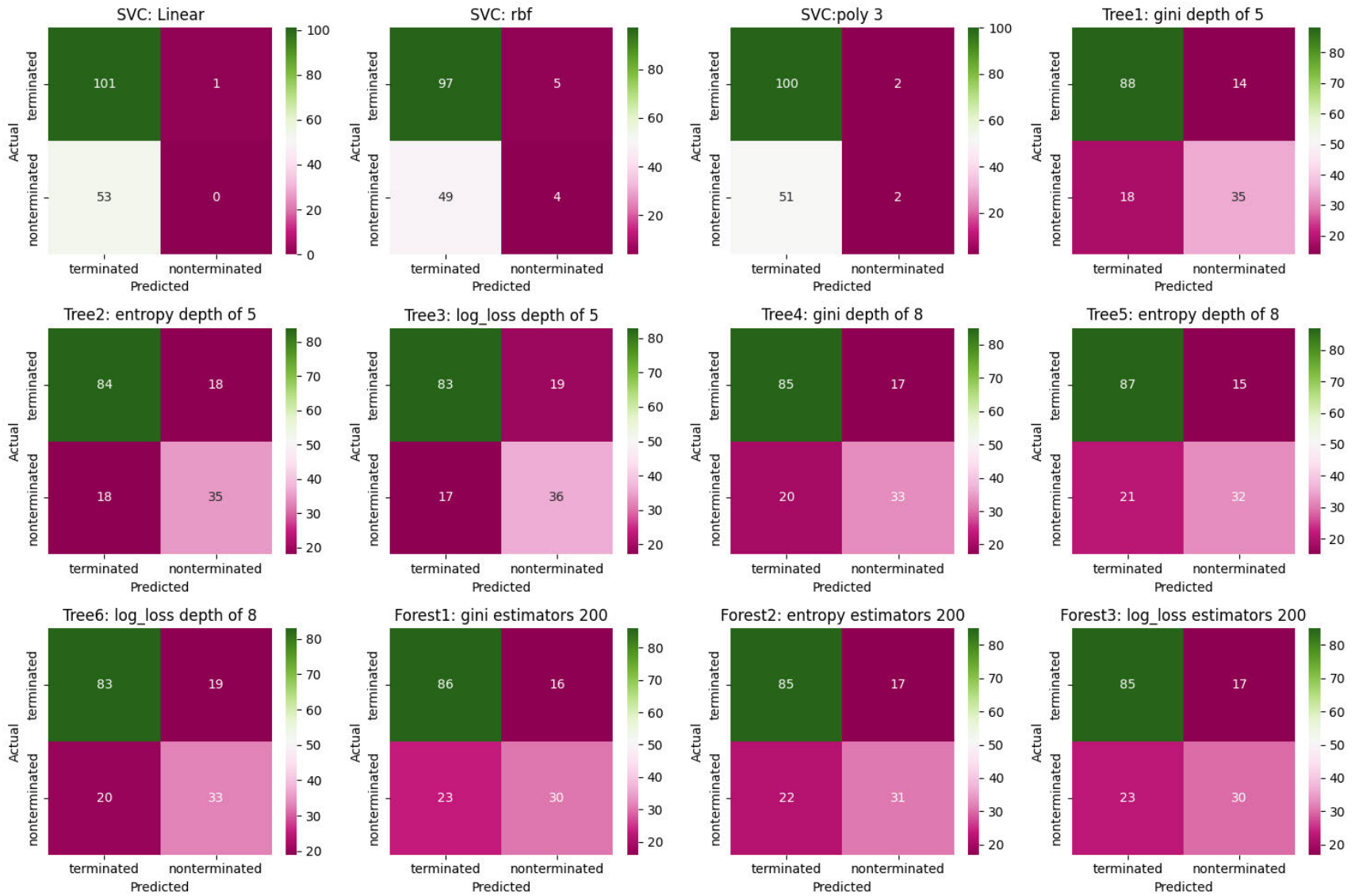
State_1 Segment Confusion matrices for classification models (n=1004)



State_2 Segment Confusion matrices for classification models (n=1145)



State_5 Segment Confusion matrices for classification models (n=385)



State_4 Segment Confusion matrices for classification models (n=155)

Appendix E: Daily Stand-ups

Question	Name	Mon 01/24	Tue 01/25	Wed 01/26	Thur 01/27	Fri 01/28	Sprint Retrospective
What have you worked on since our last meet?	Abigail	- researched and pre-planning of project scope - researched scheduling tools and format - read articles on prescriptive modeling	- researched Agile project management techniques - looked in articles about how to build productive and prescriptive models	- set up Jira and generate basic user stories on Jira - prepared questions for meeting with Iira - looked into different formats to document our daily stand-ups	- created a spreadsheet to track project implementation efforts and track progress in daily stand-ups - created spreadsheet to store all research materials for future reference	- prepared an agenda for meeting with Iira about business value and risk assessment - looked into agile personas	<p>What went well?</p> <p>During our Sprint Zero, the installation of software packages went smoothly including the setting up of Jira and the creation of Epics and user stories. We had multiple interactions with our sponsor through planned meetings, emails, and Slack correspondence to discuss the dataset and project goals which have all been very insightful. All our meetings with our professors and advisors have equally been successful and helpful toward the progression of our project. We put in extensive time and effort towards initial approach research to gain a deeper understanding of our data and the techniques we can use to garner the most success.</p>
	Michael	- set up software environment - got connected to database - started looking into possible challenges with prescriptive analytics - refreshed understanding of knowledge discovery and database	- started reviewing writing Epics in Jira - researched UML modeling such as use case diagrams	- set up Jira and generate basic user stories on Jira - looked into different UML diagrams we might use	- continued working on epic and stories in Jira to clearly define them - talked to Iira for further explanation on table data	- set up python connection to SQL database - set up a format documenting actions in the code	
	Shiya	- set up software environment - reviewed python, machine learning and data analytics skills	- concentrated on solving time conflict issues - reviewed SQL - looked over dataset	- set up Jira and generate basic user stories on Jira - review questions discussed with Prof Blake - read articles on prescriptive learning and how to remove unnecessary data	- discussed field descriptions for dataset with Iira - set up user stories - reviewed packages for python	- final range for each column and filter holes to data learned how to use SQL	
	William	- reviewed SQL - further reading of "Know Your Metrics" python article	- finished sections of the W3 Schools SQL tutorial - read "Data Driven Growth with Python" tutorial - watched Youtube videos about SaaS business models and data mining	- set up Jira and generate basic user stories on Jira - watched Youtube videos on data mining and identifying patterns to data	- explored and began applying dataset - set up Jira - watched Youtube video on SQL and data mining	- created queries for finding lifetime of customer - watched and took notes on a video introduction to database trees	
What will you work on today?	Abigail	- initial outlining and planning of Jira - coordinate a plan for maximum efficiency of seven weeks	- test out some Project Management platforms discussed in research - share prescriptive modeling article with team	- add more stories to Jira - read and take notes on article about z-score	- research agile personas - set up all spreadsheets for future organization and documentation	- start looking into SQL and how to pull data efficiently - secondary analysis for table for - complete a weekly summary to send to Professors - agile personas	<p>What could be improved?</p> <p>Improvements can be made in the areas of Individualized work which is undefined at the moment due to the uncertainties in the project as we are still in the planning phase. Some team members still need to be familiarized with tools such as Git, Hub, SQL, and Python and its classes. A system needs to be created to send out status updates for SaaSWorks and we need to create a more efficient weekly meeting schedule with SaaSWorks to ensure that all meetings we are attending are relevant to the progression of the project.</p>
	Michael	- watch Youtube videos on knowledge discovery refresh understanding of different UML charts	- label Epics in Jira - create software requirements user stories in Jira	- add more user stories to Jira - start building UML diagrams - discuss with Iira the fields in the database	- sprint planning for next week - meet with Prof Wong, Serrão and Blake	- set up technical documentation in record packages and look to use	
	Shiya	- review SQL - watch Youtube videos and read articles on prescriptive learning	- read article on prescriptive learning	- ask one question about the code - review data filtering - look for article related to coding	- sprint planning for next week - meet with Prof Wong, Serrão and Blake	- take another task on Jira such as z-score or analysis	
What if anything is blocking your progress?	Abigail	- no significant blockers	- issue with connecting to SaaSWorks Jira	- no significant blockers	- no significant blockers	- no significant blockers	<p>What do you plan to change?</p> <p>We will utilize online resources and peer assistance to educate ourselves on the usage of platforms and tools like Git Hub, SQL, and Python. Online examples and practice repositories in GitHub are examples of learning techniques we will implement. To address the status update issue we plan to send a snippet of our implementation sheet at regular intervals to communicate our progress with SaaSWorks. Additionally we will meet with SaaSWorks to review which meetings require our attendance and input.</p>
	Michael	- no significant blockers	- no significant blockers	- need a better understanding of the database - need to refresh on UML diagrams	- no significant blockers	- no significant blockers	
	Shiya	- no significant blockers	- issue with connecting to SaaSWorks Jira	- no significant blockers	- no significant blockers	- no significant blockers	
	William	- no significant blockers	- issue with connecting to SaaSWorks Jira	- no significant blockers - need a better understanding of the database	- uncertainty due to no specific tasks assigned to work currently	- no significant blockers	
Weekly Summary	Abigail	This week my objective was to develop a better understanding of our project through research, then use that understanding to begin planning out the scope and implementation. I familiarized myself with Agile practices and began utilizing Jira, Github, and SQL, which will be used heavily in our project. I created spreadsheets to document and track our progress and resources, and conducted and planned meetings with SaaSWorks to gather information about the value and flow of this project.					<p>As a team, we utilized this first week of our project to conduct research about the software, platforms and tools that we will be using such as Agile, Jira, Github, SQL, Python, and Machine Learning. We began pre-processing with database exploration and knowledge discovery, as well as extensive planning through Epics and user stories on Jira to the capacity that is currently possible. We conducted multiple meetings with SaaSWorks to inquire about the data sample and discuss project goals and meet with our advisors for further guidance when needed. We established spreadsheets and modeling tools to support documentation and implementation efforts. Individually we were assigned user stories which we began working on at the end of the week.</p>
	Michael	This week, I worked on configuring the software environment for myself and guided other team members in their setup since we were given all of the necessary permissions to access the data prior to Friday. After the environment was setup, I participated and facilitated the backlog planning sessions. I also worked on observing the data provided by SaaSWorks to better understand the scope of the data.					
	Shiya	In the past week, I worked mainly on installing software environments and virtual workspace and setting up the Jira board with my teammates. After that, we assigned our stories, and I have almost completed three stories before this week end.					
	William	This week, most of my efforts were focused on setting up the virtual workspace with all of the correct software and permissions. I also worked on sprint planning to create manageable tasks and user stories on the Jira.					

Sprint 0 Daily standups, Sprint Retrospectives, and weekly summary

Question	Name	Fri 10/28	Mon 10/31	Thu 11/1	Wed 11/2	Thu 11/3	Sprint Retrospective
What have you worked on since we last met?	Almgren	- prepared an agenda for meeting with Eva about business value and risk assessment - moved into agile personas	- met with with Eva & Alyssa - began planning an Agile persona -	- developed questions for next meeting with Eva and Alyssa - designed an agile persona for user - reviewed paper	- completed Agile persona - assessed remaining tasks for paper and added them to Jira - practical SQL queries of 70 records - downloaded Visual Foxpro and started making UML use case diagram	- worked on SQL query code to start customer revenues by account ID with help of Williams - worked on ordering items from Michael revenue to lowest - worked on UML use case diagram - reviewed agile persona	
	Michael	- set up python connection to SQL database - set up a format - documenting actions in the code	- connected database from Friday - and GitHub review session - sprint retrospective - started through next steps	- started working on generating a schema - developed questions for meeting with Eva - worked on Agile Persona - touch on SQL code for total revenue	- attempted to work on a schema but experienced crashing in virtual machine due to data overload	- re-ran queries on SQL to filter data after speaking with Eva - helped Shyu with GitHub and William with main function	
	Shyu	- find range for each column and find holes in data - learned how to use SQL	- prepared data filtering and basic analyze tools	- worked method on checking outliers - developed methods for data analysis	- packed preprocessing program to GitHub - noticed a small issue in function when pushed to database - finished data analysis	- started pre-processing and basic analysis classes to finish after speaking with Eva - fixed bug in basic analyze class	
	Williams	- created queries for finding lifetime of customer - worked and took notes on a video introduction to decision trees	- developed queries - wrote python code - explored power of SQL	- generated SQL queries for receiving revenue - spent time learning python classes and calling queries to Postgres database	- worked with Michael to generate graphs - worked on generating CLR of customers - started on queries for SQL	- worked on functions for generating tables segmentations using customer with state attributes	
What will you work on today?	Almgren	- start looking into SQL and how to pull data efficiently - summary analyze Eva asked for - complete a weekly summary to send to Professor's agile persona	- work on Agile Personas - explore SQL to pull revenue data	- add a new case type as professor - complete Agile Personas - work on finding total revenue of a customer on SQL - add new user stories on Jira for my tasks	- discuss story points to convert some tasks on Jira to stories - review Agile persona created - work on receiving revenue SQL queries - create UML use case diagram	- finish UML use case diagram - update sprint planning - start new task on Jira	
	Michael	- set up technical documentation to receive packages and code to use	- explore functions of segment 4 & segment 7 - prepare summary statistics for other fields in group	- continue working on a schema - identify cause of python crashing	- filter data to reduce crashing in virtual machine - finish compiling a schema	- lead sprint planning session - break up tasks on Jira - discuss some objectives	
	Shyu	- take another task on Jira such as a schema or analyze	- test program in database - ask Eva about definition of Null values	- test methods and push them to GitHub - begin another task on Jira	- test data analysis and push code to GitHub - begin another task on Jira	- add mode function to basic analyze class then push to GitHub - sprint planning - start new task on Jira	
What (if anything) is blocking your progress?	Almgren	- no significant blockers	- no significant blockers	- no significant blockers	- no significant blockers	- no significant blockers	
	Michael	- no significant blockers	- no significant blockers	- unknowns causing python to crash	- virtual machine crashing due to data overload - used sufficient amount of data after filtering	- time constraint	
	Shyu	- no significant blockers	- unknown about null values	- unable to download a package on python	- no significant blockers - bug found during data filtering	- no significant blockers	
	Williams	- no significant blockers	- Need better understanding of data	- unsure of what library to use to generate graphs	- unsure about use case for code because we don't have segments	- SQL queries take a very long time	
Weekly Summary	Almgren	This week I met with Eva and Alyssa to begin planning and creating agile personas for our project and team, the Professor and SasWorke Analyst. This is the first step in my business value and project risk analysis to identify the users and stakeholders of the model we are creating. I familiarized myself with SQL, while completing a first task to find the total revenue of customer accounts with the help of my teammates. I significantly increased my understanding of our project objectives and processes through meetings with my teammates, SasWorke, professors, and further research.					
	Michael	I primarily worked to improve the capability of the Database class in Python. It enabled a more precise ability to query and request data from PostGreSQL. Additionally, I worked to refine the data that is being pulled from the database as it initially crashed my Amazon WorkSpace by overfilling the available memory in the computer. After discussing the definition of valid data with Eva and the team, the system is properly filtering based on query inputs. I also assisted team members develop and deploy queries and subqueries for their tasks.					
	Shyu	In the past week, I have finished the Preprocessing and basicAnalyze classes and push them to GitHub. I also explored the database with my groupmates for better understanding on what we would need to do next.					
	Will	This past week I generated the queries required for getting data related to CLR. I also created python code for calculating CLR and generating plots relating to CLR. Started working on comparing customer categories to CLR.					
							<p>During Sprint 1, the team handled working remotely very well in response to our 100% leader access to 50 Percent for most of the week. We acknowledged that we split up the work in terms of iterations and user stories evenly and with additional peer assistance and VC calls to have group discussions, we were able to complete most tasks on Jira on-time, despite the last day of work due to database access issues. We developed many well-structured business and software requirement questions to ask SasWorke which led to many clarifications in our project objectives. The team did a great job advocating for our selves when communicating with our project sponsors and advisors.</p> <p>What went well?</p> <p>During Sprint 1, the team handled working remotely very well in response to our 100% leader access to 50 Percent for most of the week. We acknowledged that we split up the work in terms of iterations and user stories evenly and with additional peer assistance and VC calls to have group discussions, we were able to complete most tasks on Jira on-time, despite the last day of work due to database access issues. We developed many well-structured business and software requirement questions to ask SasWorke which led to many clarifications in our project objectives. The team did a great job advocating for our selves when communicating with our project sponsors and advisors.</p> <p>What could be improved?</p> <p>The team could improve on the extent that we break down tasks on Jira for maximum efficiency. We ran into some issues with the database being down on Friday, which we could not avoid, but database connectivity could definitely be improved. Our breakdown of tasks to code could also use some work to improve understanding of desired outcomes for that specific code.</p> <p>What can we change to make improvements?</p> <p>The team plans to improve how we break down tasks on Jira during the planning phase by analyzing and discussing user requirements for Jira stories. This change will also help to improve the logic of translating tasks to code, alongside improvements in terms of the use of more pseudocode, development planning, and VC calls to talk through unknowns and areas of difficulty while coding.</p>

Sprint 1 Daily standups, Sprint Retrospectives, and weekly summary

Owners	Team	Thu 11/4	Mon 11/7	Tue 11/8	Wed 11/9	Thu 11/10	Agenda Retrospective
What have you worked on since we last met?	Ahiguel	- participated in sprint planning and retrospective - check in meetings with Sasoworks and Professors - summarized sprint retrospective and weekly summary - finished up use case diagram	W P I W E L L N E S S D A Y	- met with Eva & Alysa to discuss UML use case diagrams and make appropriate changes - reviewed past MVP papers to get ideas for formatting - started looking into sprint burndown charts	- continued formatting paper - planned the figures and graphs to add to paper - researched sprint burndown charts and how to manually create them	- finished formatting table of contents and sections for paper - researched time series machine learning methods	Throughout Sprint 2, the team worked with a number of data visualization tools to graph RFM and correlation between customer revenue and usage data. The graphs are a valuable tool to conceptualize the relationships between customer segments and data fields which will be most useful for our predictive model. Similarly, the creation of a UML Use Case Diagram and Agile Personas were helpful to the team to distinguish the specific functions of the model we should prioritize to fulfill the SaaSWorks' goals for this project. The team's transition to a hybrid approach has gone smoothly as the team continues to be productive and efficient with a Monday and Thursday in-person schedule and the remaining weekdays remote. Our communication remains strong within the team and with SaaSWorks and our Advisors through the use of Discord, Slack, Email, and text messaging. With that, our resource sharing has been effective in keeping all members of the team in the loop with the progress of the project and the necessary background knowledge.
	Michael	- participated in sprint planning and retrospective - check in meetings with Sasoworks and Professors - continued working with database and plan to merge users		- started looking over ways to review categorical data and feature selection with Chi square to test independence by categorical variables - created new categorical variable to identify period right before a customer churn	- finished Chi squared analysis for churn rate - found no significant correlation between segments and account status - discovered statistic Cramer's V	- researched time series classification methods - found method that classifies likelihood of customers to go into default	
	Shinye	participated in sprint planning and retrospective - check in meetings with Sasoworks and Professors - set up and tested methods for hand analysis then passed the to Github		- completed correlation graph for numerical values - found bug in preprocessing - already tested entire database	- attended a session/class to learn how to use GitLab - half completed removing low correlation CLR calculation methods	- research time series dataframe	
	William	participated in sprint planning and retrospective - check in meetings with Sasoworks and Professors - finished function for graphing CLR by category for each user - generalized function to take in any segment type		- pulled recency data for RFM to determine amount of days since customer was last active	- created recency and frequency clusters for revenue - performed K means clustering for revenue	- watched videos and read articles on time series data - looked at data in sql to extract usage data for time series dataset	
What will you work on today?	Ahiguel	- add descriptions to user stories on Jira - look at paper to begin formatting and adding sprints to software development sections	- do more research for MVP paper - begin formatting and writing sections in paper - research sprint burndown charts and learn how to manually create them	- begin writing and editing sections in the paper - meet with SaaSWorks	- watch videos on time series - complete sprint retrospective and weekly summary - add sprint into paper	What could be improved? with the completion of Sprint 2 yielding less user stories completed than started, we realized that during our past sprint planning sessions we heavily over-estimated the capacity of user stories that we can realistically complete in a single sprint. A meeting with Eva and Jim from SaaSWorks revealed to the team that the dataset we are working with is a time series rather than aggregate data like we had approached the project believing. The team agreed that we could do a better job at unifying our tasks so that we are all working on similar areas of the project at the same time to encourage consistent progression aligned with all our work. This led us to the realization that we may be spending too much time on less important areas of the project and need to restructure our prioritization of the user stories on Jira.	
	Michael	- start looking at categorical correlation tests - look into linear regression	- look into categorical variable as a response variable to indicate chance of customer churning - finish Chi Square	- research Cramer's V in depth in relation to feature selection - evaluate statistical dependent against categorical independent variables	- attempt apply time series method to dataset using CSV - create vector of features for each account ID consisting their data over a 5 month period to look into likelihood of churn		
	Shinye	- work on new tasks on Jira and complete outstanding tasks	- begin finding categorical correlations - scatter xy graph on Jira board	- complete CLR calculation methods to remove variables with low correlation	- create table for time series data frame - perform correlation analysis on the table		
	William	- work on other graphing tasks on Jira such as scatterplots - work on RFM - finish outstanding tasks from Sprint 1	- look into K means clustering - generate segments based on value of customer - split up segments by categorical data - finish RFM and begin another graphing task on Jira	- generate RFM scores for segments - help Shinye with CLR for Use Case - work on scatterplots by CLR	- write queries to get average tractive time before churn for customers - be able to create graphs for customer based on varying categorical data		
What if anything is blocking your progress?	Ahiguel	- no significant blockers	- no significant blockers	- no significant blockers	- no significant blockers	- unknowns about time series data	What can we change to make improvements? We plan to revisit our Jira board to closely assess and prioritize user stories to better plan out our sprints. Using UML diagrams to plan the functions of our model, we will be able to plan more thoroughly with a capacity of 20 storypoints per sprint split evenly between the 4 of us. Along with this, we plan to execute a complete Jira backlog refinement as a result of our goals being pivoted due to the discovery of our data being a time series.
	Michael	- no significant blockers	- need to research Chi square libraries	- need to consider possible outcomes and action plans if there is found to be no correlation - unsure about how to test work	- assumptions needed to apply to model that could prevent accuracy		
	Shinye	- no significant blockers	- no significant blockers	- failed to run CLR calculation methods	- need to identify periods considered risky to churn for each customer for relevant data		
	William	- no significant blockers	- no significant blockers	- inquire into possibility of using pivot	- unknowns about how to analyze time series data		
Weekly Summary	Ahiguel	During this sprint, I focused on completing our Use Case Diagram and Agile personas and organizing the resources and meeting discoveries so that we can align our goals with those of SaaSWorks. Once I completed this, I moved on to formatting our paper after researching past MVP papers and documentation tools. I arranged the table of contents and organized the sections and subsections of the paper in a logical order to clearly and cohesively document the progress of our project.					
	Michael	The first half of the sprint revolved around analyzing categorical data for relationship with the status of a given account. I used chi-squared and Cramer's V analysis to look for indicators of a relationship, but I was unable to establish any statistically significant relationships within the given categorical data. However, I shifted focus to understanding time series classification techniques after Eva pointed out that the nature of the data is not conducive of aggregate analysis.					
	Shinye	At the beginning of this week, I worked on making correlation analysis tools. After meeting with Eva, I looked for a new approach doing data analysis and prediction. Now, I am working on making a timeframe to do correlation analysis.					
	William	This past week most of my work was done in Python and SQL. Firstly, I read articles on how a recency, frequency, monetary value (RFM) scores are calculated. I then created SQL queries to grab the required data to complete the calculations. Then in Python I wrote functions to find cluster customers based on their S, F, and M.					

Sprint 2 Daily standups, Sprint Retrospectives, and weekly summary

Question	Name	Fri 11/11	Mon 11/14	Tue 11/15	Wed 11/16	Thu 11/17	Sprint Retrospective
What have you worked on since we last met?	Abigail	- finished sprint retrospective and weekly summary - added sprint info to software Development section of paper - continued researching time series data	- worked on editing paper - refined research section and revised process map of SaaSWorks services	- calculated completed user story points from completed sprints on Jira - added sprint information to Software Development section of paper	- added diagrams to paper - began deep editing of introduction section moving down into rest of paper	- updated implementation documentation spreadsheet for sprint 3 - assessed progress of the MQP paper	<p>What went well?</p> <p>During sprint 3, the team was able to quickly and efficiently pivot the direction of the project after receiving new information about the dataset we are working with. After dedicating Monday to researching Time Series data and conducting multiple meetings with Eva and our advisors throughout the week, we were able to collect enough information to plan our next steps in our development process. On Friday we presented a demonstration of our progress so far to the SaaSWorks team and received insightful feedback. On Saturday, the team met up to have an extra brainstorming session outside of our usual working hours, making up for any time lost during our short period of uncertainty. The brainstorming session consisted of every team member writing/illustrating their understanding of the project so far and developing a few drafts for the process map of our project of the on the white board, which helped us visualize the requirements and necessary steps to achieve the goals set in place. By the end of the sprint, the team had successfully created a number of SVM and SGD classification models, decision trees, random forests, and first generation linear regression models. The MQP paper was heavily reformatted and revised and we began the editing process alongside the writing of new sections and subsections.</p>
	Michael	- team found a possible classification method using support vector classifier - tried to test that classifier on python	- finished iteration 1 of new data format - ran a support vector machine classifier with 70% accuracy	- implemented random forest and decision trees for current classification methods - yielded same accuracy results as SVM	- worked on paper - discussed paper formatting with Abby - worked on defining cross validation	- discovered and resolved UUID error in code - continued working on software architectural section - created a process diagram for training development models	
	Shiyu	- worked on SQL queries to pull significant fields related to churn rate - prepared powerpoint slides for demo	- completed correlation table - found that correlation table provides no significant insights regarding churn rate	- worked on creating a regression model on the account version ID lifespan level - completed model and encoder	- worked on regression model - attempted to filter data	- worked on machine learning model - finished the linear regression model and found that prediction model yields low score - worked on SGD model	
	William	- worked on SQL queries - finished up RFM - created graphing and clustering for RFM	- calculated RFM based on usage data	- created another function called "segment_by" to create list of dataframe for individual RFM scores	- analyzed account segments - developed queries for each individual account status then disregarded idea	- discovered and resolved UUID error in code - filtered machine learning data by segmentation of account attributes	
What will you work on today?	Abigail	- create and present a powerpoint for SaaSWorks demo - add graphs and visual data to paper	- continue revising Research section - edit remainder of paper then add more details	- write project proposal using documented information from meetings, emails, and teamwork	- add paper sections to Jira board - continue editing	- finalize sprint retrospective and weekly summary fro sprint 3 - continue editing and writing MQP paper - update Jira board	<p>What could be improved?</p> <p>We noticed that our Jira board is still not very reflective of the individual work that each teammate is doing, meaning that outside of our implementation documentation spreadsheet which is updated daily, we are not sufficiently tracking our work and progress on Jira. We also need to work on our prioritization of the tasks assigned on Jira to ensure that the most impactful tasks and stories are getting completed before we attempt the less significant work. Regarding our Saturday meeting, although it was a good brainstorming session, we were not as productive as planned since we did not complete the sprint planning session in a timely manner. Following our pivot in the project, we realized that we can improve our communication with professors when seeking guidance in order to have more in-depth discussions.</p>
	Michael	- create and present a powerpoint for SaaSWorks demo - continue testing machine learning to flatten array of periods and classifications	- attempt to segment data by region to improve accuracy and remove noise - help Abby with story point counting for sprints	- work on predicting lifetime of individual version IDs - figure out which approach is	- answer Eva's questions on email - continue working on software requirements section of paper - meet with Will and Shiyu to discuss models	- work on project suggestions that Eva discussed - put together a quick software snapshot for SaaSWorks -	
	Shiyu	- create and present a powerpoint for SaaSWorks demo - create graph of correlation - start working on deep learning	- create new account status called ready to churn to retest correlation to churn rate - work on segmentation of account status	- generate table for regression model on account version ID lifespan- modify Michael's table of accounts to test model	- finish debugging regression model after filtering - run model to collect data	- modify method of modeling to increase accuracy of results - continue with SGD and regression modeling	
	William	- create and present a powerpoint for SaaSWorks demo - produce users segmented by RFM scores and their respective graphs	- work on making function return a list of data frames based on segments - work on modeling analysis	- look into developing a query building function to call SQL queries in Python	- create function that will automatically segment accounts by each different attribute - start working on paper	- get the filtered data to run on machine learning model	
What (if anything) is blocking your progress?	Abigail	- no significant blockers	- no significant blockers	- no significant blockers	- no significant blockers	- no significant blockers	<p>What can we change to make improvements?</p> <p>To address the issue with Jira tracking, we will begin updating the Jira board more frequently during the sprint, ideally after every daily stand-up, to ensure that our current and future actions align with the bigger picture planned out in Jira. Additionally, we plan to place more emphasis on task and story prioritization during our Jira sprint planning sessions moving forward.</p>
	Michael	- issue with python compiling SQL queries	- no significant blockers	- no significant blockers	- syntax error issue	- no significant blockers	
	Shiyu	- no significant blockers	- unknowns about correlation related to churn	- no significant blockers	- filtering data takes a very long time using pandas	- no significant blockers	
	William	- SQL queries not working	- no significant blockers	- unknowns about what to do next	- no significant blockers	- unknown errors in filtered code	
Weekly Summary	Abigail	This week I focused my efforts on the MQP paper, creating sections and subsections that cover the methods and research that we have implemented so far. I began an immersive editing process over the sections already written and made revisions to adjust the flow and accuracy of the paper.					<p>This week, the team worked on the MQP paper, adding sections and subheadings that detail the techniques and research we have implemented thus far. We started a thorough editing process over the already written sections, and made changes to improve the paper's accuracy and flow. We created the initial models for determining whether a consumer will engage in prolonged inactivity. This required data reformatting, cross validation, and model accuracy evaluation. SVMs, binary trees, and random forests were all tested. Each generated an average accuracy of 68-70% with three periods worth of data per customers (n=9770). We developed a stochastic gradient descent classifier model, SGD, that predicts whether an account will churn in the next month given the historical data, and a linear regression model that predicts the life span for each account version ID. The team was able to create a function that generates segments of customers based on any unique attribute which can be standardized and fed into the Machine Learning models. This breakthrough allowed us to connect the data containing customer segments to the ML model by breaking down the dataframe of customer information into their respective segments then converting the dataframe to the required form for the ML model.</p>
	Michael	This week, I primarily worked on developing the first generation of models for predicting if a customer will take an extended period of inactivity. This involved data reformatting, cross validation, and simple assessment of models. I tested SVMs, binary trees, and random forests.					
	Shiyu	This week I focused on create a linear regression model and a sophisticated gradient descent classifier model. The linear regression model predicts the life span for each account version id, and SGD predict whether an account will churn in the next month given the historical data.					
	William	This week, I worked on connecting the data containing customer segments to the ML model. This involves breaking a dataframe of customer into their respective segments and converting it into the required form for the ML model. I also created a function to generate segments of customers based on any unique attribute, these segments can all be standardized and fed into the ML models.					

Sprint 3 Daily standups, Sprint Retrospectives, and weekly summary

Question	Name	Tue 11/16	Wed 11/17	Thu 11/22	Fri 11/23	Thu 11/24	Mon 11/28	Tue 11/29	Wed 11/30	Thu 12/1	Sprint Retrospective	
What have you worked on since we last met?	Abigail	- summarized sprint retrospectives and weekly summaries - updated Jira board to reflect time for greater effort - finished remaining code for paper - Jira - edited paper	- updated authorship table within Abstract/Intro/Introduction - began working on Research section in Jira - edited paper				- finished Abstract, Introduction and Research chapters of the paper - updated Software Development chapter	- continued editing paper as a whole - wrote software build section in the software requirements	added a few sections to Research Chapter - wrote and updated Software Development environment		- worked on finalizing paper - continued on a few under documentation Jira on Jira	
	Michael	- reorganized Jira board - helped Will set up (learning) issues and tags - retrospectives comments to classmate	- started running additional ML models using J and J search - wrote - created a new method that allows models to be trained concurrently - validated predictions based on last n number of frames				- finished organizing all final training and testing data to a class setting - class code files - made sample of code to send to GitHub - completed new method that allows models to be trained concurrently - validated predictions based on last n number of frames	- began applying and understanding of general budget prediction using classifiers ML - deep learning model and found that significant difference between linear regression	- discussed regression model with professor Shih - analyzed last 3 months for work around 10 to find optimal direction - and practice to find direction	- completed method creating database table for last 3 months - each account ID - finished connecting to code	- met with Fira to discuss high level goals and questions - continued to work on code refactoring and merging with Will	
	Shiba	- fixed bug by S2D modeling - began writing on PCA and normalization for linear regression model	- worked on creating graphs for churn rates on time period and number of active accounts via period of time to observe effects of COVID - created confusion matrix for S2D				- worked with Michael to make queries for getting last n months of data	- read and edited section 2 and 3 of paper - cleaned up C++ class - checked new algorithm, a convolution	- cleaned up C++ class for logic - began - worked with Michael to reduce branch - worked on printing and debugging them - read and edited paper		- worked with Michael to merge model loading class - organized confusion matrix by region and looked into it to get confusion matrix by state	OUT SCUL
	William	- got machine learning model to run in ML of existing customers - worked on generating code for making PCA, stability against	- took over research paper - tried to generate paper previously only used for Jira				- worked on methodology chapter about Software Development chapter	- wrote Database Value and Risk analysis chapter - edit paper at kitchen standing with Fira	- got results to add to Paper - document our development process Database Value and Risk analysis chapter		- got paper mostly finished by tonight to be sent to advisors tomorrow - complete sprint retrospective and weekly summary	
What will you work on today?	Abigail	- continue writing and editing paper - sprint planning and review Jira	- complete abstract and introductions - edit and write research section - update software development section				- work on methodology chapter about Software Development chapter - continue editing entire paper	- write Database Value and Risk analysis chapter - edit paper at kitchen standing with Fira		- get paper mostly finished by tonight to be sent to advisors tomorrow - complete sprint retrospective and weekly summary		
	Michael	- sprint planning - increase status coding - continue working on class systems	- finish class systems - create code package together to send to Software for review				- understand why PCA is identifying features based on most powerful classifiers - set up C++ and hardware - write method for model selection - merge and set up maps to prepare for final map	- answer Fira's questions with a more specific - create a package maps with details to send to Fira		- identify high level goals and distinguish what we can achieve when is unrealistic		
	Shiba	- work on deep learning model to get a better score - sprint planning on Jira	- modify database then continue S2D modeling - write comments for code and push to GitHub - look into S2D model for gradient descent over model				- add comments to code and update on Github - final merge for Shiba in production model	- continue writing comments to code - connect code to class - ready to upload code to class	- merge and review all code to GitHub		OUT SCUL	
	William	- fix bug, created what generating ML modeling - talk about next steps for learning algorithms	- come up with a solution to getting last n months of data for each account				- finalize code with comments and clean up class - read and edit to create one file for Software	- start reading and editing section 3 - merge with code - set up to create one file for Software	- work on finalizing last file		- capture abstract - continue code on GitHub	
What of anything is blocking your progress?	Abigail	- no significant blockers	- no significant blockers				- no significant blockers	- no significant blockers	- no significant blockers	- no significant blockers		
	Michael	- no significant blockers	- no significant blockers				- no significant blockers	- no significant blockers	- no significant blockers	- no significant blockers		
	Shiba	- no significant blockers	- no significant blockers				- no significant blockers	- no significant blockers	- no significant blockers	OUT SCUL		
	William	- 10M is a datapoint statistic	- no significant blockers				- no significant blockers	- no significant blockers	- no significant blockers	- no significant blockers		
Weekly Summary	Abigail	During this sprint I prioritized writing the remaining chapters in the paper and making final edits before our deadline to submit the draft tomorrow. I wrote and edited the Abstract, Introduction, Research, Database Value and Risk Assessment and parts of the Methodology/Software Development Environment chapters. I continuously updated the Software Development Chapter, Jira board, and Implementation/Documentation spreadsheet as well.										
	Michael	This sprint was heavily focused on cleaning up some of the spaghetti and making sure the code is flexible enough for the other team members to use their models without having to know the machine learning workflow. For the latter half of the sprint, I began working on a new angle of the process by changing the queries that the model now generated because they "based on the first 10 months, then a given account demonstrate behavior (down to a granular or active account) in "given the LAST n months." The change of context resulted in better accuracy and a model that should be better suited for the goal of the project. Additionally, this could allow for Software to eventually use this as a "rolling window" setting where the data for each account can update every month as the other context (the data) and model (and training) with the data for each account can update every month as the other context updates and would also change with time. We worked on analyzing the effects of our PCA based on the same data as the first S2D. Additionally, we focused on writing a new method for the ML models, and then updating it to be used in GitHub. We developed a better way of generating customer data to fit the ML models, which involved creating functions for the and various contexts using the output of ML models for each segment and testing their confusion matrices. Overall it was a successful week with most of our planned items being completed. We used a correct State and Final Metrics table and a Code Snippet to Software for them to evaluate our progress before we submit our final deliverables on 12/1.										
	Shiba	During this sprint, I worked with suboptimal training by applying crossentropy technique. I created a deep learning model and got to see final RMSE. Also, I finish writing all comments for all my method, and I even updated all items to GitHub.										
	William	During this sprint I first worked to develop a better way of segmenting customer data to fit the ML model, this involved creating functions for this. I also worked to use the output of ML models for each segment and use their confusion matrices.										

Sprint 4 Daily standups, Sprint Retrospectives, and weekly summary

Questions	Name	Fri 12/2	Mon 12/5	Thu 12/6	Wed 12/7	Thu 12/8	Sprint Retrospective
What have you worked on since we last met?	Abigail	- worked on finalizing paper for draft submission	- worked on writing, editing, and formatting paper - rework abstract to fit 900 character limit	- continually updated formatting of paper - updated sprint 4, in Software development section	- did an intensive rereadthrough of the paper - started thinking about project to SaaSWorks presentation	- finalized business learnings section - made updates and edits as needed	<p>What went well?</p> <p>As the team made final edits to our development and documentation efforts during our last sprint cycle, they made significant progress in the restricted time frame available. The team made great strides to make the necessary adjustments and additions to the code and paper as advised by our sponsors and advisors with a quick turnaround time. In regards to prioritization, they made notable improvement in time management of the remaining issues on Jira to approach deadlines with ease. This was made possible through proficient story points estimations and equal division of work. The Code Stashbox delivery to SaaSWorks went smoothly as they were able to engage in an open dialogue about the data feature set and logic behind model selections. By the end of the final week, the team was able to finalize the paper, complete the debugging process, and push the final code to GitHub.</p>
	Michael	- worked on finalizing paper for draft submission	- worked on a few sections of paper - ran new analysis to answer Eva's questions - rework feature importance - created new approaches and format to share with Eva	- added about 20 pages worth of data into appendices of the paper - reformat confusion matrices - respond to Eva's questions - added to sections 7, 9 and 10	- continued writing and editing paper - created new outputs and graphs for the paper	- continued editing and writing entire paper	
	Shyla	- worked on finalizing paper for draft submission	- worked on paper - worked on debugging code	- worked on editing paper - added data to paper	- worked on ML sections of the paper for further expansion	- continued editing and writing entire paper	
	William	- worked on finalizing paper for draft submission	- worked on paper - worked to bug fix older functions written on SQL and Pandas	- began writing future work section - made edits to paper	- continued writing and editing paper - wrote future work and technical learnings sections	- continued editing and writing entire paper	
What will you work on today?	Abigail	- prioritize the completion of paper to submit tonight for professor feedback	- heavy reformatting of paper - write executive summary - organize tables, equations and figures	- write business learnings section - do a full read through of paper	- do another intensive rereadthrough of the paper for final edits - write business learnings section	- summarize sprint retrospective and weekly summary - create product burndown chart - complete final edits and corrections from professor feedback to submit paper tonight	<p>What could be improved?</p> <p>The areas of improvement that the team could touch on include maintaining a consistent format throughout the paper to maintain uniformity for comprehensive reading. During the editing process, the team could have expressed better feedback internally for more concise and precise edits.</p>
	Michael	- prioritize the completion of paper to submit tonight for professor feedback	- meeting with professor Hads to discuss ML models - final out why data is behaving oddly	- add more graphics into the paper - write code testing section -	- piece code together and fix bugs that come up - continue writing and editing paper	- complete final edits and corrections from professor feedback to submit paper tonight	
	Shyla	- prioritize the completion of paper to submit tonight for professor feedback	- continue working on paper - finish debugging	- add to section 2, 9, and 11 to discuss Machine Learning	- continue adding to section 2 - push final code to GitHub	- complete final edits and corrections from professor feedback to submit paper tonight	
	William	- prioritize the completion of paper to submit tonight for professor feedback	- write section 9 and 11 in the paper - finish writing function	- continue writing future work section - work on Technical learnings section - add visuals to appendices and tables	- continue making fixes to paper based on professor feedback	- complete final edits and corrections from professor feedback to submit paper tonight	
What (if anything) is blocking your progress?	Abigail	- no significant blockers	- no significant blockers	- no significant blockers	- no significant blockers	- no significant blockers	<p>What can we change to make improvements?</p> <p>To counter the issue of inconsistent formatting from individual teammates completing their work into one document, they were able to establish a style guide that all could follow.</p>
	Michael	- no significant blockers	- no significant blockers	- no significant blockers	- no significant blockers	- no significant blockers	
	Shyla	- no significant blockers	- no significant blockers	- no significant blockers	- no significant blockers	- no significant blockers	
	William	- no significant blockers	- may finish writing functions to be able to generate graphs for the paper	- no significant blockers	- no significant blockers	- no significant blockers	
Weekly Summary	Abigail	This week the team's priority was to make finalization efforts to the project in terms of writing, editing, and reformatting the paper to accurately reflect our development process. I focused on the writing, editing and updating of the table of contents, the Abstract, Executive Summary, Software Development, Business and Risk Analysis, and Business Learnings and Conclusion.					<p>This week the team's priority was to make finalization efforts to the project in terms of writing, editing, and reformatting the paper to accurately reflect our development process. They focused on the writing, editing, and updating of the table of contents, the Abstract, Executive Summary, Methodology, Software Development, Business and Risk Analysis, Future Work, and Conclusion chapters. Additionally, the team reworked the classification model descriptions in the Research section and added significant information to the ML, Assessment, Software Testing, and a few other areas of the paper. They finalized the program features to be delivered to SaaSWorks. On Wednesday, there were a couple last requests that had not yet been implemented, so work was done to bring those into existence - specifically, decision tree visualization, formatting outputs to be readable, addition of more CL arguments, adding additional documentation both in the code and as a README. A description was written for every model used and tables were created within the Software Development section to list every user story worked on during every sprint. The team updated the visuals for the number of active accounts vs churn rates graphs by replacing the school names with a simplified format of "school n". They spent some time editing the output of a function that potentially may be pushed to production as well, but overall the main focus of the team was the completion of the project in its entirety.</p>
	Michael	This week, I finalized the features in the program that will be delivered to SaaSWorks. On Wednesday, there was a couple last requests that had not yet been implemented, so work was done to bring those into existence - specifically, decision tree visualization, formatting outputs to be readable, addition of more CL arguments, adding additional documentation both in the code and as a README. Additionally, I reworked the classification model descriptions in Section 2. Added significant amounts to the ML sections in the paper, the results section, the testing section, and touched a few other parts of the paper					
	Shyla	Write the description for every model and the conclusion and finding for whole project. Make the Jira table for each sprint. Make two graphs about the number of active accounts and churn rate with replacing school name as "school n"					
	William	Throughout this week I primarily worked on finalizing and editing the paper. I specifically worked on the technical learning, future work, and conclusion sections. I also briefly spent time editing the output of a function that potentially may be pushed to production.					

Sprint 5 Daily standups, Sprint Retrospectives, and weekly summary