Customer Churn Prediction Using Machine Learning

Noli A. Angeles[1], Hali Schoultz[1], and Oskar Montesdeoca[1]

[1]Arizona State University, Tempe, AZ 85281, USA

May 1, 2025

Abstract

Customer churn poses a significant challenge for subscription-based businesses, as retaining existing users is often more cost-effective than acquiring new ones. This project explores how machine learning models can be leveraged to predict customer churn and identify key behavioral and demographic indicators of churn risk. Using a dataset of approximately 37,000 users containing demographic details, engagement metrics, and customer feedback, we performed extensive exploratory data analysis and developed several predictive models.

We implemented and evaluated three classification algorithms: Naive Bayes, K-Nearest Neighbors (KNN), and Decision Trees. Each model underwent tailored preprocessing and encoding strategies to ensure fair comparison and robust evaluation. Initial findings revealed that the Decision Tree model outperformed others in terms of precision, recall, and AUC, and also provided interpretable insights through feature importance. In contrast, the Naive Bayes classifier underperformed, likely due to its assumption of feature independence.

Our analysis identified membership type, engagement behavior, and customer complaints as strong indicators of churn risk. The results demonstrate the value of predictive modeling in proactive churn management and support the use of interpretable machine learning approaches for strategic customer retention. Future work will involve exploring ensemble models and conducting hyperparameter optimization to further improve predictive accuracy.

1 Introduction

Customer experience has become a key factor for businesses. There is increasing pressure to not only attract new customers but also retain existing ones. This is where Customer Churn becomes a significant factor. Customer churn, also referred to as customer attrition or defection, describes the process in which customers cease the relationship with the business or stop using a product or service over a set period [13]. For service-based or subscription models, having high churn rates can be a threat to long-term business viability. Retaining existing customers is not only more cost-effective than acquiring new ones [19, 1, 16] but can also contribute to profitability, customer loyalty, and competitive edge [21, 11].

There has been interest in studying customer churn throughout recent decades. Some early studies in the 1990s were focused on customer loyalty, satisfaction, and defection, with some putting emphasis on service failures, pricing, or inconvenience as some leading causes for customer churn [6, 14]. As more data-driven technologies started to emerge in the 2000s, there was more focus on predicting customer churn. Some of the techniques that were starting to be used were CRM systems and data mining techniques, which allowed for better insights into customer behavior and churn [18, 24]. From the 2010s onward, researchers employ methods that range from logistic regression to hybrid neural networks [15, 20, 22].

As much as customer churn studies have continued to evolve, and the advancement of technologies has improved research, there remain some gaps and challenges, such as feature imbalance and the dynamic nature of consumer behavior [15, 26], as well as ethical concerns around data privacy [24].

To guide our analysis, we focused on the following research questions:

- **Main Question:** How can machine learning models be used to effectively predict customer churn in subscription-based services?
- **Subquestion 1:** What are the key predictors of customer churn in subscription-based models?
- **Subquestion 2:** What role do pricing strategies and service quality play in influencing customer retention?
- **Subquestion 3:** How can machine learning models be utilized effectively to forecast customer churn?

These questions shaped the direction of our exploratory analysis, modeling approach, and evaluation of results.

2 Background

2.1 Early Perspectives on Customer Churn (Pre-2000s)

In early literature, "customer churn" was labeled under different names. Customer switching [14], customer defection [5, 21], customer retention [7] and customer loyalty [6] are some of the alternative ways to call churn. However, they encompass the same idea of minimizing the rate at which customers stop using a product/service or maximizing the length of consumption. Customer retention can increase a company's profit versus obtaining new ones [5] and has a deeper impact than other factors that give them a competitive advantage [21]. Initial views on customer retention were highly specific to each industry (banking, telecoms, insurance, retail, etc.), thus it was hard to make generalizations and apply strategies to more than one industry [14]. Despite this, businesses can still get a better grasp of why customers choose to drop their service and go to another one.

Reichheld (1990) and Duffy (1998) both emphasized customer satisfaction as a primary driver of loyalty, though their perspectives on its underlying factors varied. [21, 6]. But what exactly contributed to satisfaction remained a key question. Keaveney (1995) defines eight main categories for a customer's reason to change or leave a service. In order of largest reason of a customer switching: core service failures, service failure encounters, pricing, inconvenience, employee responses to service failures, attraction by competitors, ethical problems, and involuntary switching [14]. In the banking sector, Ennew and Binks (1996) attribute quality of service and investment in customer relationship as a critical factor for retention [7]. Colgate et al. (1996) also found that price sensitivity, customer service, and easy accessibility were the main reasons for students leaving financial institutions[5]. On the other hand, Reichheld (1990) claims the main way for companies to keep customers from defecting is to outmatch competitors by continually improving their products and processes. Overall, it is the combination of core quality of the product and the customer service that contributes the most to customer loyalty and retention.

Businesses employed various strategies to approach customer retention during this era. Firstly, Colgate et al. (1996) suggested identifying and measuring the cost of the defection rate, as well as creating incentives to entice customers to stick with the service [5]. Loyalty programs, such as frequent flyer and membership reward systems, emerged as a widely used retention strategy to incentivize repeated business [6]. Duffy (1998) noted that certain approaches - such as free services and special member events - were less effective in reducing churn [6]. Although there is no way to completely keep all current customers, Reichheld (1990) suggested learning from the valuable information that defectors provide [21]. Efforts to create insight from these individuals can help businesses to find early warning signs of customer churn which in turn helps improve the business [21].

2.2 The Emergence of Data-Driven Churn Analysis (2000s-2010s)

While traditional retention strategies provided some insights, they were often reactive rather than predictive, relying on customer feedback rather than anticipating churn before it happened. As businesses grew and

consumer preferences became more dynamic, traditional customer satisfaction surveys and qualitative feedback became insufficient for understanding churn patterns at scale. Early literature showed that anecdotal feedback (surveys and experiences) was the main way to gain insights from current customers and those that have defected [21]. A 2002 study by Garland leveraged the widely used Juster's scale (scale of 1-11 to determine the probability that population will do something by a future date) to identify which customers were only LIKELY to defect [9]. The shift to data-backed decision-making became more prominent as the rate of data acquisition and datasets for businesses continued to grow in the early 90s and 2000s. Initially, the overwhelming amount of data made decision-making difficult. However, data mining techniques helped researchers derive meaningful insights, such as better gauging customer loyalty [18], identifying key attributes of customer churn and reducing data dimensionality [3], improving churn prediction [18, 3, 10], as well as developing customer relationship management (CRM) systems [24].

The rise of CRMs made it easier to track and understand customer behaviors and have highly customizable products that fit the needs of both customers and businesses alike [24]. In order for this system to work, it is mandatory to create a customer database that collects information such as transaction history and response to marketing initiatives/contacts [24]. Having such a valuable tool can improve customer retention and, in turn, increase revenue. Essentially, insights from customer preferences, behavioral patterns, and historical data fueled the widespread adoption of predictive modeling in business applications [18]. Cao and Shao (2008) utilized Support Vector Machine-Recursive Feature Elimination (SVM-RFE) for feature selection [3], a process Hadden et al. (2007) emphasize as both vital and often neglected in predictive modeling [10]. Other studies, such as Xie et al. (2009), explored alternative approaches, including the Improved Balanced Random Forests (IBRF) method [26]. This shift towards predictive analytics laid the foundation for more advanced churn prediction methodologies, which will be explored in the next section.

2.3 Modern Perspectives on Customer Churn (2010s-Present)

Even with the ever-growing amount of data being generated and machine learning techniques being refined, one idea has stayed the same: it is much more cost effective to keep current customers. Saran (2016) notes how the cost of customer acquisition is expensive and Mishra and Reddy (2017) mentions the difficulty of gaining new ones, thus retaining what customers a business already have is essential [1, 16]. To achieve this, we must find the key factors that drive churn and cause customers to switch. Chang and Chiu (2023) make strong claims that the value of service, customer satisfaction, and competitors are the main reasons for defection [4]. While Saran (2016) also supports dissatisfaction of a service as a key variable, they also suggest the cost and the subscription plan itself as a driving point [1]. Another reason that causes defection is customer engagement - the amount of consumption and generation of content - which is the most important factor for Wu et al. (2024) [25].

To address customer churn we also have to pay attention to the main reasons why customers leave and tend to these behaviors. In order to increase customer loyalty, Mishra and Reddy (2017) suggest high quality customer service and offer reward points [16]. Wu et al. (2024) states to focus on increasing customer consumption for those that are less engaged than other subscribers [25]. However, the best way to prevent defection seems to be predicting churn early in the customer's experience with the service. Studies like Prabadevi and Kavitha (2023) use basic machine learning techniques like K-nearest neighbors, random forest, and logistic regression [20]. Momin et al. (2019) take these traditional techniques and add artificial neural networks to predict customer churn [17]. Mishra and Reddy (2017) take a deep learning route using convolutional neural networks. The previously mentioned machine learning and deep learning techniques seemed insufficient for Khattak et al. (2023), so they turned towards a hybrid deep learning model that combined bidirectional long/short-term memory and convolutional neural networks [15].

2.4 Controversies and Limitations in Customer Churn Research

Despite the many leaps and bounds in the customer retention field, there are still many concerns for both businesses and customers alike. CRM systems helped with data acquisition on a huge scale, but how much personal information is stored in these databases and how is it being used by these companies remains a

critical privacy question [24]. Biases in predictive algorithms make ML algorithms less accurate and could lead to poor decisions. There are also limitations to previous studies that we could try to address. Khattak et al. (2023) notes that their study only used binary classification that strictly dealt numerical features [15]. For Xie et al. (2009), the imbalance and noise in data sets have greatly affected previous studies [26]. Finally, Chang and Chiu (2023) claim that customers using different means of a service (computer and smartphone) could have very different behaviors and preferences [4].

2.5 Methodologies

Prabadevi et al. (2023) used four machine learning algorithms to predict customer churn: stochastic gradient booster, random forest, K-Nearest Neighbor(KNN), and logistic regression. They determined that stochastic gradient booster had the highest accuracy, and each method was investigated using Receiver-Operating-Characteristics (ROC) and Area-Under-Curve (AUC) [20]. An advantage of using stochastic gradient booster is the adaptability; however, due to the high adaptability, this method requires a large framework search during tuning, which can be a challenge [20]. Their data was from a telecommunications dataset.

2.6 Project Plan

Using Python, the plan for this project will begin with cleaning and data munging. After creating a cleaned dataframe, descriptive statistics like the correlation value between variables, as well as the covariance value between variables will be found. Graphs and visualizations will be created to visualize trends and patterns that may become apparent after the exploratory data analysis.

For this project, machine learning will be used to predict customer churn. In this dataset, the customer_churn variable is binary, so the methods that will be used for this project are Naive Bayes, K-Nearest Neighbors, and classification tree.

3 Methods

3.1 Data Sources, Preparation, and Cleaning

The primary dataset used for this analysis was sourced from a publicly available customer churn dataset [23], containing information on approximately 37,000 users across 23 features. These include demographic attributes, account activity, behavioral engagement metrics, and service feedback. The target variable, `churn_risk_score`, is a binary classification label, where a value of 1 indicates a high risk of churn and 0 indicates a low risk.

Initial inspection revealed several data quality issues requiring resolution before analysis. Numerous features contained missing or inconsistent values, including placeholder entries such as ‘’ ?‘, ‘’xxxxxxx‘‘, and ‘’Error‘‘, which were uniformly replaced with `NaN`. Missing categorical values (e.g., in `region_category`, `medium_of_operation`, and `preferred_offer_types`) were imputed using each column's mode, while missing numerical values (e.g., `avg_frequency_login_days`, `points_in_wallet`, and `avg_time_spent`) were filled using the median to mitigate the influence of outliers. Invalid negative values—particularly in features like time spent and wallet points—were also flagged as missing and imputed accordingly.

Feature engineering was applied to improve the dataset's predictive capacity. A `tenure_days` variable was derived from the joining date to capture how long each user had been with the platform. A composite `engagement_score` was constructed using time spent, login frequency, and wallet points to reflect user involvement. We also created a binary indicator, `had_complaint`, from the original `complaint_status` variable, and grouped membership tiers into a simplified `membership_grouped` variable.

In preparation for modeling, we dropped columns deemed irrelevant or potentially leading to data leakage, including `security_no`, `referral_id`, `joining_date`, and `last_visit_time`. Features (`X`) were separated from the target variable (`y`), and the latter was encoded using `LabelEncoder()` to convert class labels into binary numeric form.

The dataset was then split into training and testing sets using an 80/20 split, with a fixed random seed for reproducibility. We verified that class distributions between churners and non-churners were consistent across both sets to avoid introducing bias during model evaluation. Features were classified as either categorical or numerical based on their data types.

Preprocessing strategies were tailored to the modeling techniques. For models sensitive to feature scaling and categorical separation (e.g., KNN and Naive Bayes), categorical features were one-hot encoded using `OneHotEncoder()`. For the Decision Tree model, `OrdinalEncoder()` was used to convert categorical variables into ordered numeric values compatible with tree-based learning.

All preprocessing steps were encapsulated in `Pipeline` objects for each model. This modular approach ensured consistent data transformation, reduced risk of data leakage, and streamlined the training and evaluation processes.

3.2 Methods: Correlation Analysis

To explore the relationships among numerical features and the churn outcome, a correlation heatmap was generated. This analysis revealed that the most meaningful correlations were negative in direction. Specifically, points_in_wallet and avg_transaction_value demonstrated the strongest negative correlations with churn_risk_score, indicating that users who spend more or have accumulated more points are generally less likely to churn. Conversely, features such as avg_frequency_login_days showed a weak positive correlation, suggesting that frequent login behavior does not necessarily imply user retention and may, in some cases, signal dissatisfaction.

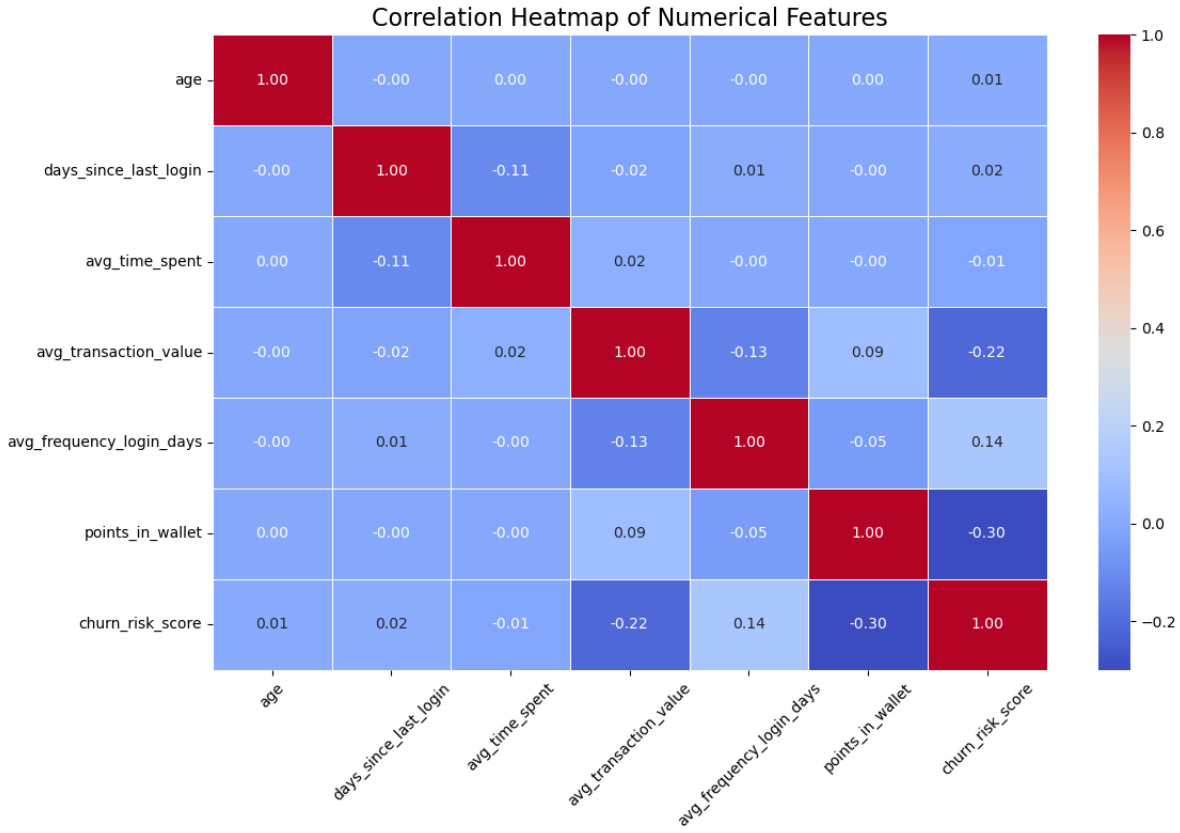

Figure 1: Correlation analysis of variables before feature engineering. The "strongest" relationship with churn would be 'points_in_wallet'

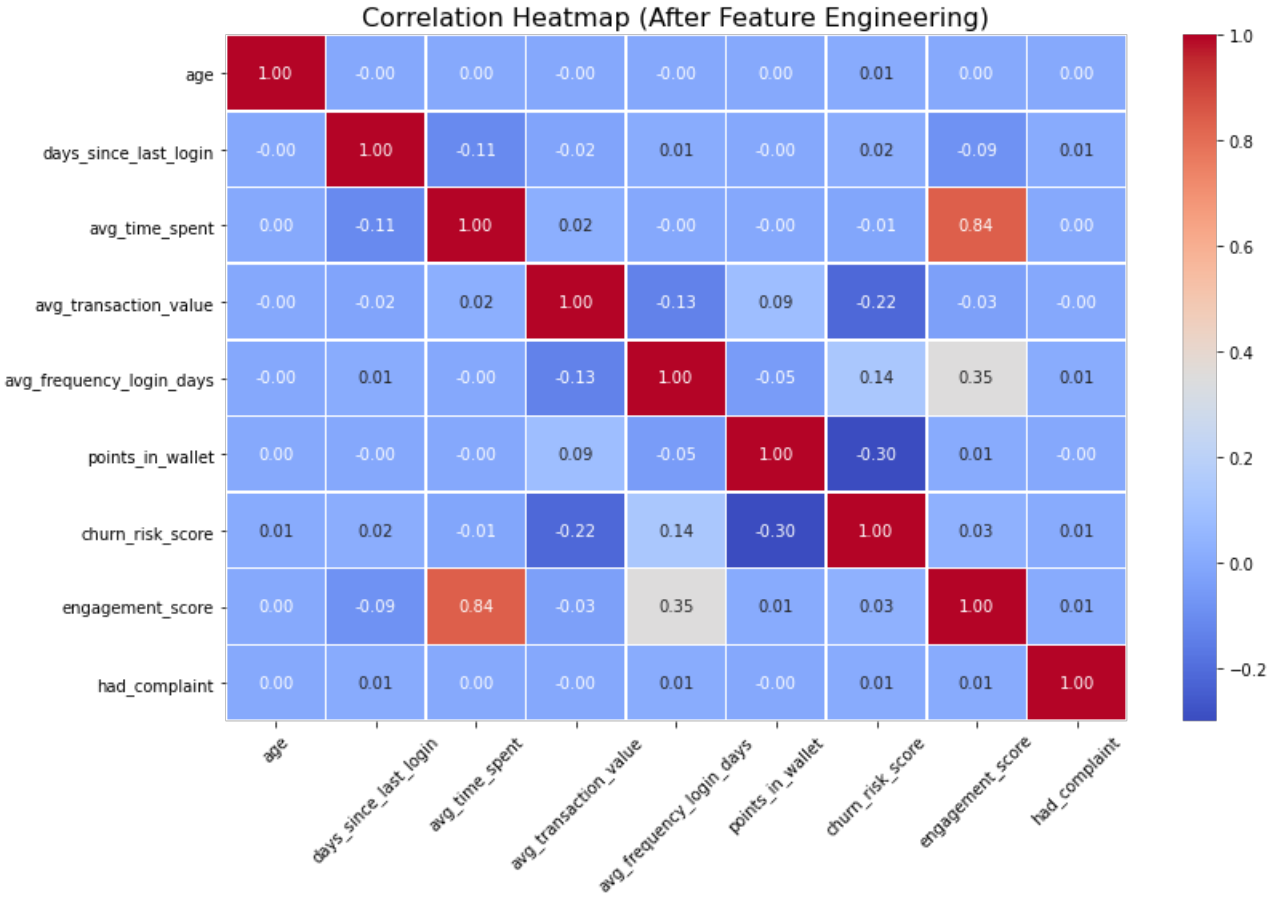

Figure 2: Correlation analysis of variables after feature engineering. No new relationships with churn are observed, indicating complex relationships between predictor variables and churn.

Engineered features such as engagement_score and had_complaint exhibited very weak linear correlations with the churn target, implying that their predictive contributions may arise through non-linear interactions rather than direct linear influence. Overall, the correlation heatmap highlighted the complexity of churn behavior and suggested that advanced modeling techniques would be necessary to capture the subtle patterns within the dataset.

3.3 Methods: Descriptive Statistics

Using the pandas function *.describe()*, we can see the mean, median, mode, and quartiles of every feature. Since it has been established that points_in_wallet and avg_transaction_value have a strong negative correlation, these will be explored first. The mean for points_in_wallet is 691 and for avg_transaction_value it is 29,271. The interquartile range for points_in_wallet and avg_transaction_value are shown below.

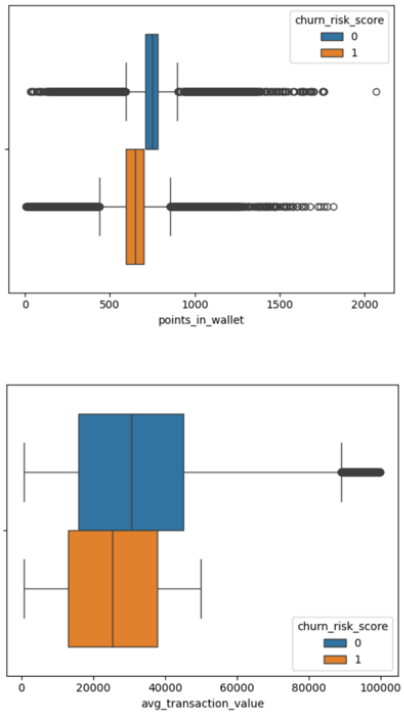

Figure 3: These box plots show the interquartile range of their respective features and the churn_risk_score

These figures help to visualize the negative correlation that points_in_wallet and avg_transaction_score have with churn_risk_score, showing that the higher the points in a customer's wallet and the higher the average transaction value, the less at risk a customer is for churn.

The standard deviation of points_in_wallet is 177.2, and the standard deviation of avg_transaction_value is 19444.8. The range of these variables is quite large, with the range of points_in_wallet being 2,062.6 and the range of avg_transaction_value being 99,113.6.

Covariance differs from correlation in that covariance can take any value, instead of just -1 to +1. From the covariance matrix using churn_risk_score, avg_transaction_value has a very low score of -2112.5, and points_in_wallet also has a low score of -26.5. Avg_frequency_login_days has a weak positive covariance score of 0.537.

3.4 Methods: Distributions

In this section, we analyzed how the average churn risk score varies across several categorical variables. We believe this is an important step in our data analysis as we are trying to identify what are some factors related to high or low churn that we can see, which is linked to one of our main questions on this project. The visualizations were created using `pandas` for grouping and aggregation, and `matplotlib.pyplot` for bar plots. Each figure includes value annotations and a dynamic y-axis range for better visibility of subtle differences.

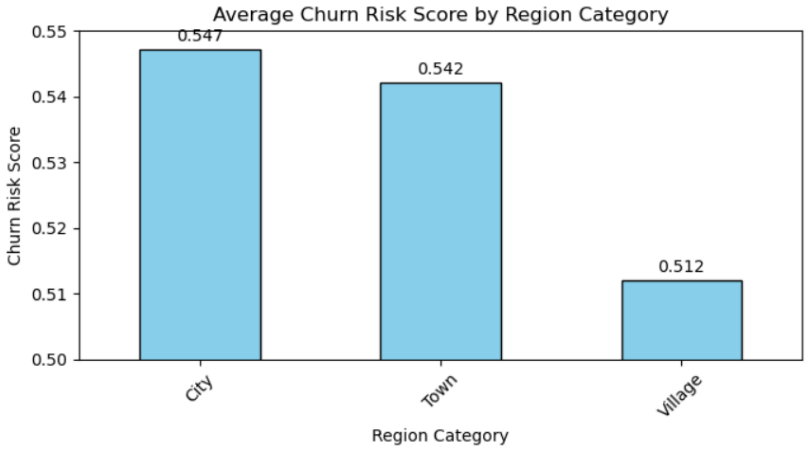

Figure 4: Average Churn Risk Score by Region Category

City residents show the highest churn risk (0.547), followed by Town (0.542). Village users have the lowest churn (0.512), suggesting lower service alternatives or greater stability in rural areas.

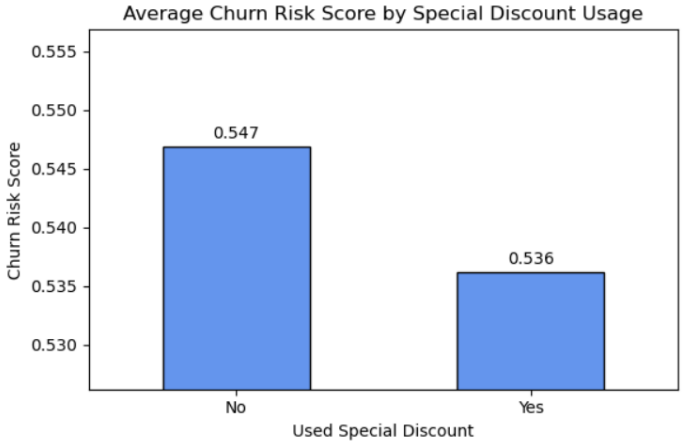

Figure 5: Average Churn Risk Score by Special Discount Usage

Customers who did not use discounts are more likely to churn (0.547) compared to those who used them (0.536). This indicates that targeted discounts can enhance retention.

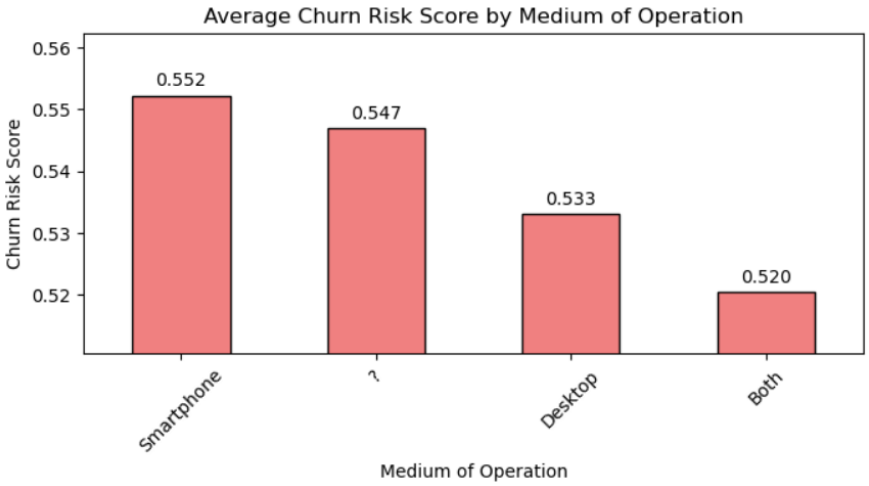

Figure 6: Average Churn Risk Score by Medium of Operation

Smartphone-only users have the highest churn (0.552), while customers using both desktop and mobile platforms have the lowest (0.520). Multi-platform engagement appears to reduce churn likelihood.

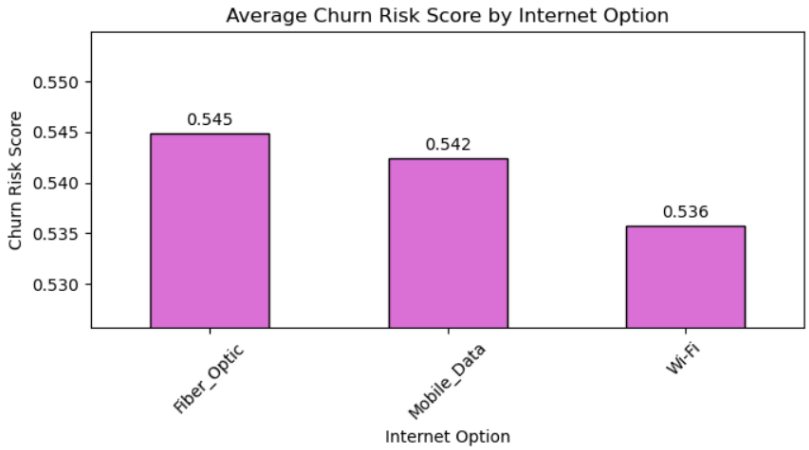

Figure 7: Average Churn Risk Score by Internet Option

Fiber Optic (0.545) and Mobile Data (0.542) users churn more than Wi-Fi users (0.536). Wi-Fi use may reflect more stable or home-based customer behavior.

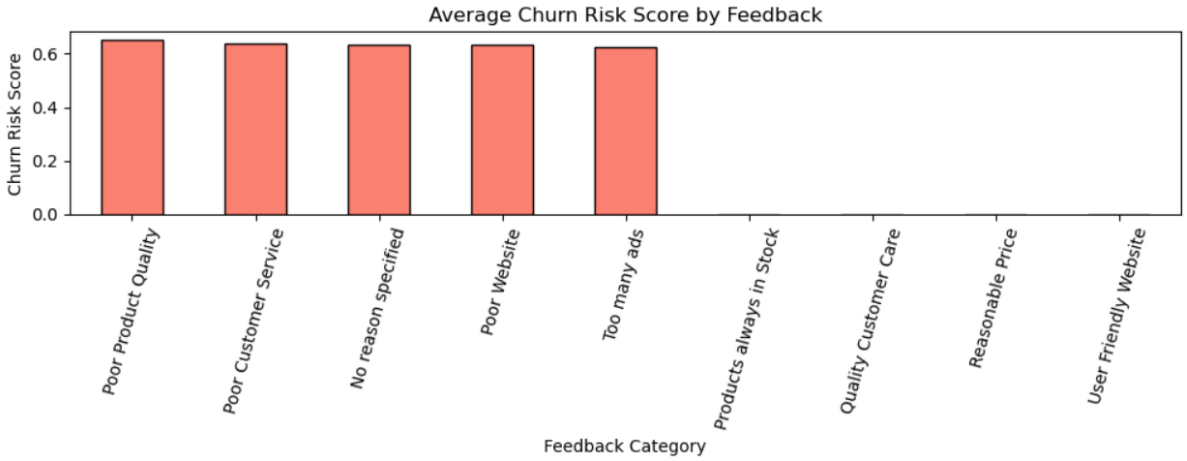

Figure 8: Average Churn Risk Score by Feedback

High churn is strongly correlated with negative feedback such as poor product quality, customer service, or website experience. Positive feedback categories correlate with near-zero churn.

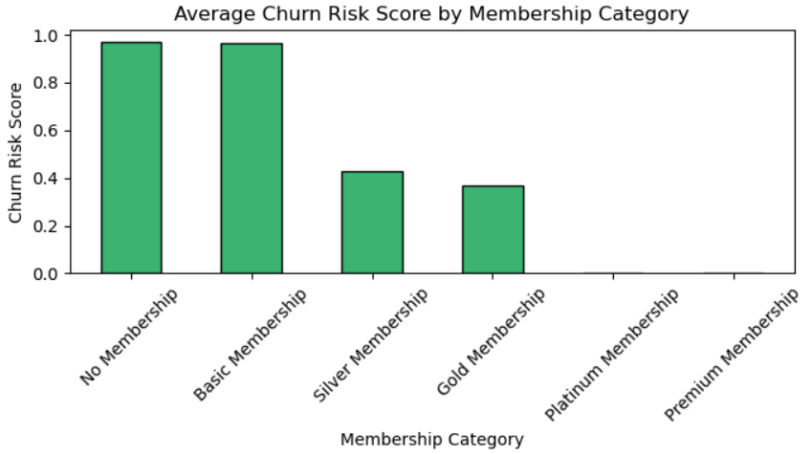

Figure 9: Average Churn Risk Score by Membership Category

Churn risk is highest for users with No or Basic Membership (0.95), while it drops substantially for Silver and Gold members. Premium and Platinum members have negligible churn, reinforcing the value of membership engagement.

3.5 Methods: Naive Bayes Classifier

The first model we implemented was a Multinomial Naive Bayes classifier. This algorithm was selected due to its simplicity, efficiency with high-dimensional data, and suitability for categorical feature spaces. It operates under the assumption of conditional independence among features, which, while rarely true in practice, still often yields competitive results in classification tasks. As described in the data preparation section, categorical variables were encoded using `OneHotEncoder()` to produce binary indicator variables for each category. Numerical features were passed through unchanged. This transformation was wrapped inside a scikit-learn Pipeline to ensure consistency and avoid data leakage.

After training the model on the training set, we evaluated its performance using multiple metrics. The classification report showed an accuracy of 55%, with a precision of 0.57, recall of 0.47, and F1-score of 0.52 for the churn class (class 1). The confusion matrix revealed that the model correctly predicted 1880 churners and 2090 non-churners, but also misclassified 2155 actual churners and 1274 non-churners. We further evaluated the model's ability to rank churn risk using the ROC curve. The area under the curve (AUC) was 0.56, indicating that the model's discrimination between churners and non-churners was only slightly better than random guessing. Finally, 5-fold cross-validation was performed on the full dataset, producing an average cross-validation accuracy of 53.3%. These results suggest that while Naive Bayes was efficient and interpretable, it was not the most effective model for predicting churn in our dataset.

3.6 Methods: K-Nearest Neighbors

K-Nearest Neighbors can be used for both classification and regression tasks. It works by using proximity to find out information or predictions about a point of data. It is a simple and effective algorithm. Similar to Naive Bayes Classifier, `OneHotEncoder()` was used on the categorical features to produce binary indicator variables for each category. After training and testing the model on the train and test data, a classification report was ran on the model, as well as other descriptive statistics. The F1-score of K-Nearest Neighbors model was 0.59, and the accuracy was 0.65. The recall of predicting non-churners (0) was 0.56, and the recall of predicting churners (1) was 0.71. The confusion matrix showed that it correctly predicted 1891 non-churners and 2912 churners. The area-under-curve (AUC) was 0.70 for this model. This number is closer to 1 than the Naive Bayes Classifier, meaning it is a better performing classifier. The root mean squared error (RSME) indicates how well the predictions of a model match the actual data, and the RSME was 0.592. This number could be lower, but it is a good start.

3.7 Methods: Classification Tree

Classification tree is a supervised learning algorithm that is used to predict categorical outcomes, and in this case we will be using it to predict customer churn. It uses a model of decision rules based on input features. We decided on this model because of its interpretability, non-linearity and ability to capture feature interactions. The decision tree utilized an Original Encoder within a dedicated pipeline to ensure data integrity and prevent leakage. The model seems to have performed very well across key metrics, with an accuracy score of 91.34 precision of 0.92 and F1 score of 0.92. The confusion matrix distribution showed a strong classification boundary with low misclassification rates in both classes. ROC curve showed strong separation between the two classes, with AUC of 0.91 which indicates the model is performing very well in ranking customers by likelihood of churn. 5-fold cross-validation was also performed with a high accuracy level of 90.71 confirming that the performance is consistent across different splits of the data. Feature importance was also extracted from the trained model, and results show that membership tier and wallet balance were the most significant drivers of churn. This suggests that customers in lower tiers or with low wallet balances may be more likely to churn, potentially due to a lower perceived value of incentives. While the Classification Tree achieved strong performance across all metrics, it's important to consider that some factors could impact on performance, such as potential overfitting and data leakage.

4 Results

4.1 Results: Exploratory Data Analysis

The exploratory data analysis made with distributions showed some important patterns related to customer churn. We can see that users located in urban regions are more likely to churn, likely due to having more access to competing services. Membership tier is a strong predictor of churn on initial look, and could be an important factor on retaining customers as well as customer feedback, with negative feedback correlating with high churn while positive one can be linked to less churn. We can also see that customer that take

advantage of discounts show lower risk, indicating that promotional strategies could be an effective tool to reduce churn risk.

4.2 Results: ROC Curves

To evaluate the models' ability to distinguish between churners and non-churners, we plotted the ROC curves for the Naive Bayes, K-Nearest Neighbors (KNN), and Decision Tree classifiers. The area under the curve (AUC) was calculated for each model to summarize performance across all classification thresholds.

The Decision Tree model achieved the highest AUC of 0.91, indicating strong discrimination between classes. The KNN model obtained an AUC of 0.70, suggesting moderate classification performance, while the Naive Bayes model achieved an AUC of 0.56, indicating slightly better than random performance.

The ROC curves for all three models are shown in Figure 4. As shown, the Decision Tree's curve rises sharply toward the top-left corner of the plot, reflecting its superior ranking ability compared to Naive Bayes and KNN.

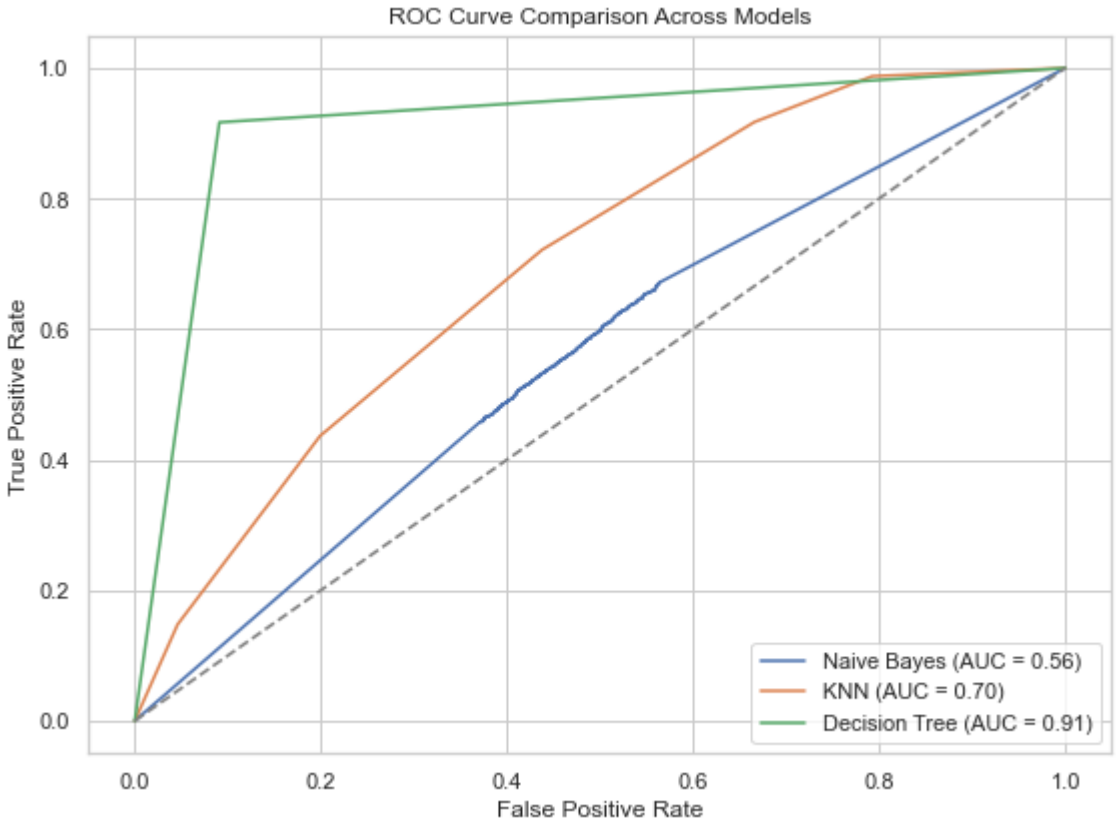

Figure 10: Comparing ROC curves of Naive Bayes, KNN, and Classification Tree models. The Classification Tree had the highest area under the curve (AUC) demonstrating stronger performance in distinguishing between the two classes.

4.3 Results: Naive Bayes Classifier

```
Naive Bayes CV Accuracy: 0.5330
```

Figure 11: Naive Bayes 5-Fold Cross-Validation

Naive Bayes (Multinomial) Report:

	precision	recall	f1-score	support
0	0.49	0.62	0.55	3364
1	0.60	0.47	0.52	4035
accuracy			0.54	7399
macro avg	0.54	0.54	0.54	7399
weighted avg	0.55	0.54	0.53	7399

Figure 12: Naive Bayes Classification Report

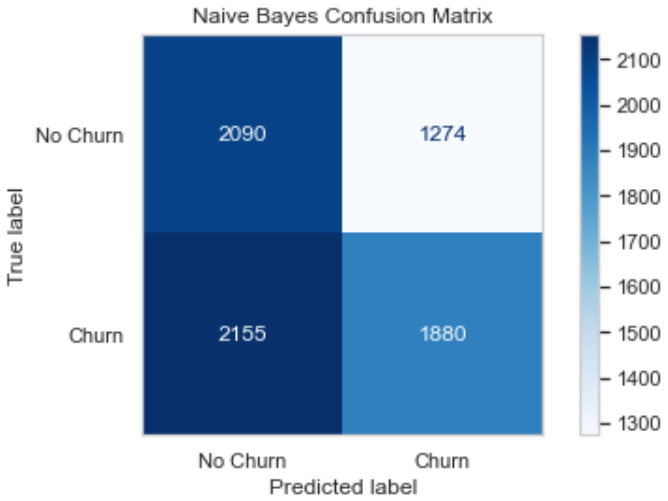

Figure 13: Naive Bayes Confusion Matrix

Naive Bayes achieved an accuracy of 54% on the test set and a cross-validation (CV) accuracy of 53.3%. Its classification report revealed precision and recall values around 0.49–0.60. Meanwhile, the confusion matrix for the same model, incorrectly predicted more churn values than it correctly predicted no churn and churn values. The Naive Bayes model struggled with correctly predicting if a customer has churned. With an accuracy of 0.56, the Naive Bayes had an accuracy that was slightly better than random.

4.4 Results: K-Nearest Neighbors

```
KNN CV Accuracy: 0.6526
```

Figure 14: K-Nearest Neighbors 5-Fold Cross-Validation

KNN Report:

	precision	recall	f1-score	support
0	0.63	0.56	0.59	3364
1	0.66	0.72	0.69	4035
accuracy			0.65	7399
macro avg	0.65	0.64	0.64	7399
weighted avg	0.65	0.65	0.65	7399

Figure 15: K-Nearest Neighbors Classification Report

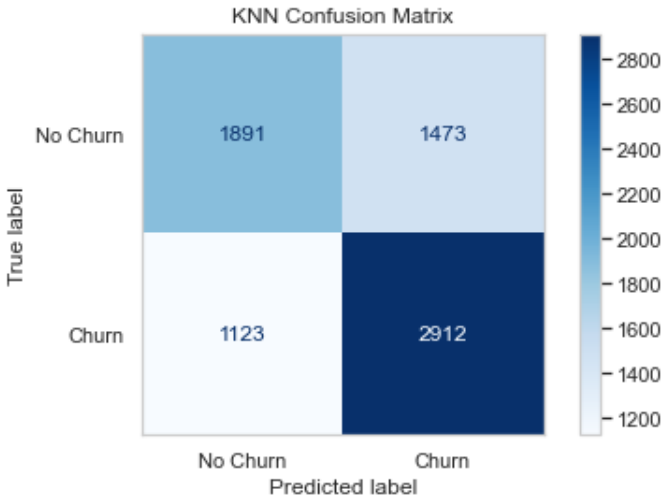

Figure 16: K-Nearest Neighbors Confusion Matrix

From the classification report, the K-Nearest Neighbors model achieved a precision of 0.66, recall of 0.72, and F-1 score of 0.69 for the churn variable. From the 5 fold cross-validation, this model had an accuracy of 65.26%, and KNN had an AUC of 0.70. This suggests that the model performed moderately. The confusion matrix and ROC curve confirm this. The KNN correctly predicted 2912 churn values, while wrongly predicting 1123 no churn values as churn values.

4.5 Results: Classification Tree

CV Accuracy: 0.90714

Figure 17: Classification Tree 5-Fold Cross-Validation

```
Accuracy: 0.9133666711717799

Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.91      0.91      3364
           1       0.92      0.92      0.92      4035

    accuracy                           0.91      7399
   macro avg       0.91      0.91      0.91      7399
weighted avg       0.91      0.91      0.91      7399


Confusion Matrix:
 [[3058  306]
 [ 335 3700]]
```

Figure 18: Classification Tree Report

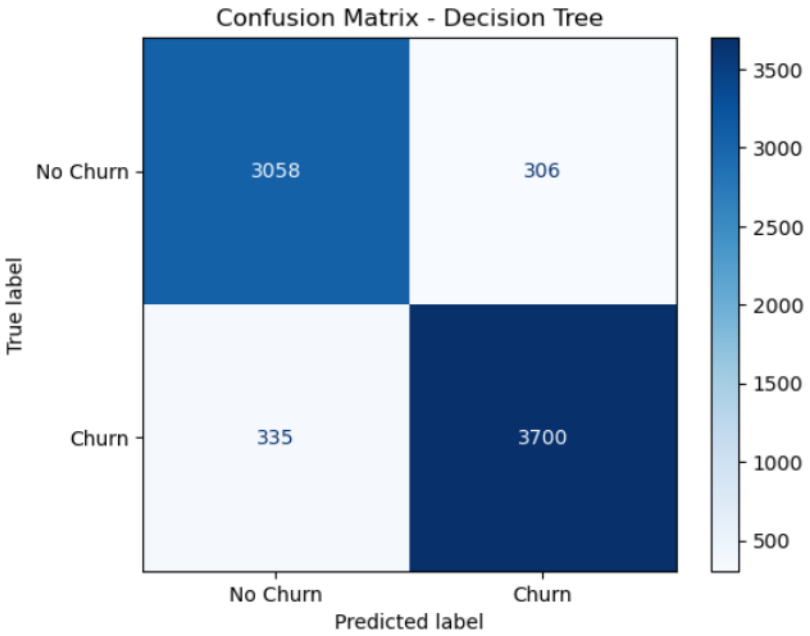

Figure 19: Classification Tree Confusion Matrix

The classification tree model showed strong predictive performance in identifying customer churn, with an accuracy of 91.34%, a precision of 0.92, and an AUC score of 0.91. These results showcase how effective decision trees can be in handling categorical churn predictions problems and being capable of capturing non-linear relationships and complex feature interactions [2]. However, decision trees can be prone to overfitting if model depth and complexity are not controlled [12]. For this project, using pipelines to ensure proper preprocessing and separation between training and testing sets was done to mitigate the risk.

Observing the mean accuracy of 90.71% obtained with a 5-fold cross-validation, we can see that the model has strong generalization capability within the current dataset. These results are consistent with previous research [20] in which decision tree models were also found to be effective for customer churn prediction.

The Classification Tree model also provided insight into the top factors influencing churn through feature importance analysis. Among the top ten most important features, `membership_grouped` and `points_in_wallet` emerged as the strongest predictors of customer behavior. Customers with lower membership tiers and fewer

wallet points showed a higher likelihood of churn, suggesting that membership level and loyalty program engagement are critical to retention.

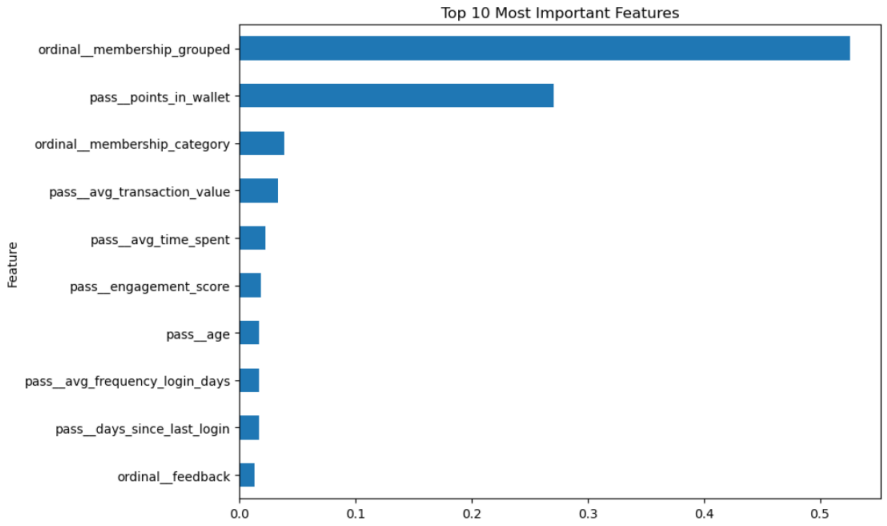

Figure 20: Top 10 Significant features influencing churn

5 Discussion

5.1 Discussion: Exploratory Data Analysis

After cleaning our data, the exploratory data analysis that was completed on our dataset revealed some correlations, both positive and negative.

The second sub-question is: *What role do pricing strategies and service quality play in influencing customer retention?* From our exploratory data analysis, the customers that were most likely to churn had negative feedback.

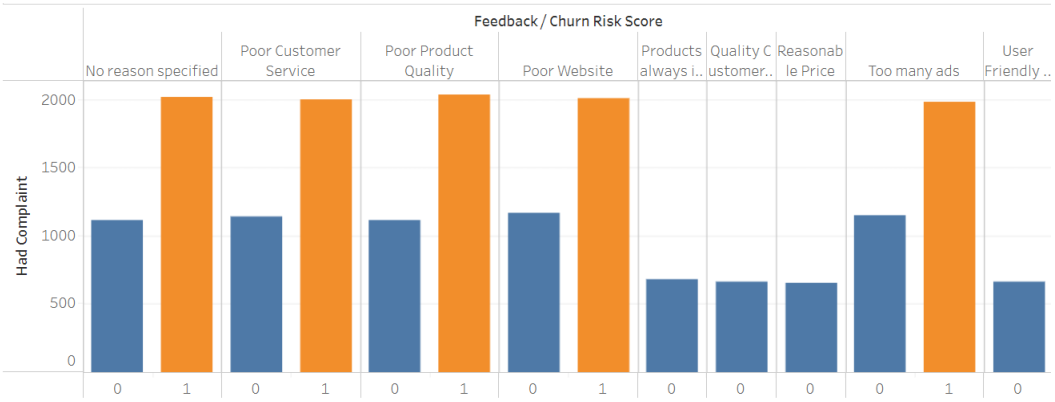

Figure 21: Feedback vs Churn Risk Score

This figure shows that when a company has a poor product, customer service, or website, customers are more likely to churn. Yet, when a customer's feedback is "Product is always in stock", "Quality customer

care", "Reasonable price", or "User friendly website", the likelihood of a customer to churn is much lower. In this case of this data, those categories had no churn at all.

To answer the second part of sub-question two: membership category is the most important feature when determining customer churn. Looking at Figure 8, customers with no membership and the basic membership are the most likely to churn. Customers with the silver or gold membership are half as likely to churn as customers with basic or no membership. Customers with platinum or premium membership are very unlikely to churn; in this dataset, they did not churn at all.

5.2 Discussion: Naive Bayes

The Naive Bayes classifier demonstrated modest performance in predicting customer churn. Based on the classification report, the model achieved a precision of 0.60, recall of 0.47, and F1-score of 0.52 for the churn class (class 1). These results suggest that while the model was somewhat able to identify churners, it struggled significantly with recall, indicating that many actual churners were missed. This is further supported by the confusion matrix, which revealed a large number of false negatives — cases where churners were misclassified as non-churners. Such misclassifications could have serious business consequences, as failing to identify at-risk customers would lead to lost revenue opportunities.

The ROC curve for Naive Bayes yielded an AUC of 0.56, indicating that the model's ability to distinguish between churners and non-churners was only slightly better than random guessing. These results highlight a key limitation of the Naive Bayes approach: its assumption of feature independence. In complex behavioral datasets like customer churn, features often interact in non-trivial ways, and the independence assumption leads to suboptimal predictive performance. Our findings are consistent with previous work by Prabadevi et al. [20] and Hadden et al. [10], who observed that Naive Bayes tends to underperform compared to tree-based or ensemble methods in customer churn prediction contexts.

From an ethical standpoint, deploying a weak model like Naive Bayes could lead to unfair or ineffective business practices. Misclassifying churn-prone customers as loyal could result in neglecting customers who actually need retention efforts, while misallocating marketing resources toward low-risk customers. Therefore, it is critical to prioritize models with stronger predictive performance in order to support fair, data-driven decision-making and efficient customer relationship management strategies.

5.3 Discussion: K-Nearest Neighbors

The K-Nearest Neighbors (KNN) model demonstrated moderate effectiveness in predicting customer churn. While it did not outperform the tree-based model, KNN showed stronger classification performance than Naive Bayes, particularly in identifying churners with reasonable balance between precision and recall. The ROC curve indicated that the model had a fair ability to distinguish between churn and non-churn cases. However, its reliance on distance-based calculations in a high-dimensional feature space—due to one-hot encoding—may have limited its overall effectiveness. Additionally, the model exhibited some tendency to misclassify non-churners as churners, which could lead to unnecessary retention efforts in a real-world application.

KNN is very simple to execute because the model's main thing to be determined is the distance between points based on various information.[20] KNN is usually chosen as a machine learning method due to its' simplicity. It is one of the easiest models to implement and understand. It is also very adaptable to new data. KNN is also appealing to use because it only has 2 main parameters: the value of K and the distance metric, which is a low number when compared to something like AI.[20]

The findings of this model are in line with what others have found. Prabadevi et al. found that K-Nearest Neighbors gives a decent value.[20] This model would not be the worst for a business to employ, but there certainly are better machine learning algorithms to use like decision trees. Ethically, all models are subject to bias and K-Nearest Neighbors is no exception.

5.4 Discussion: Classification Tree

The Classification Tree model produced the strongest predictive performance in this project, with a precision of 91. 34% and an AUC score of 0.91. These metrics, along with a cross-validated mean accuracy of 90. 71%, confirms the high ability of the model to distinguish churners from nonchurners. The strength of decision trees lies in their ability to handle non-linear relationships and categorical variables without requiring intensive preprocessing [2]. In our implementation, the model not only provided accurate predictions, but also contributed valuable information with feature importance for churn.

The feature importance analysis addressed one of the core project questions: *What are some key predictors in churn?* Among the top-ranked features were `membership_grouped` and `points_in_wallet`, indicating that loyalty status and customer engagement play critical roles in predicting churn. These findings suggest that customers with higher membership tiers and greater wallet balances are more likely to remain with the company and avoid churn, while those with basic memberships or lower balances are more at risk of leaving.

These findings have significant business implications. The interpretability of the Classification Tree allows marketing and customer retention teams to design targeted strategies based on the model's results, such as offering loyalty upgrades or rewards to customers with high risk of churn. The tree structure provides a clear rationale for why a customer might churn, making it easier to act on the insights in a transparent and ethical manner. In this way, the model serves as both a predictive and a decision support tool, offering practical value to real-world churn management initiatives.

To answer the third sub-question: *How can machine learning models be utilized effectively to forecast customer churn?*, we can compare our descriptive statistics from all three machine learning algorithms that have been implemented on this dataset. The classification tree algorithm had the best performance with an accuracy of 91.34%, while the KNN and Naive Bayes algorithms had accuracies of 65.26% and 53.3%, respectfully.

5.5 Discussion: Comparison of Models

When comparing the three models side by side, the Classification Tree emerged as the most effective approach for predicting customer churn. It achieved the highest precision, recall, and AUC scores, and offered interpretable outputs through feature importance rankings. The K-Nearest Neighbors model demonstrated moderate success, particularly in recall, but its performance was hindered by the high-dimensional, sparse feature space created by one-hot encoding. While it captured some non-linear relationships, it lacked the robustness and clarity of the tree-based approach. In contrast, the Naive Bayes classifier underperformed across all metrics. Its simplifying assumption of feature independence likely contributed to its inability to model complex customer behavior accurately. Overall, these findings suggest that more flexible, structure-aware models like Classification Trees are better suited to the churn prediction task than simpler, assumption-heavy algorithms.

6 Challenges and Limitations

Several challenges emerged while implementing and evaluating the Naive Bayes classifier for customer churn prediction. First, Naive Bayes assumes that all features are conditionally independent given the target variable. This assumption is rarely true in real-world behavioral datasets such as churn prediction, where features such as engagement, transaction value, and membership type are often interdependent. As a result, the model struggled to accurately capture complex relationships between features, leading to lower predictive performance.

Another challenge involved the nature of the features themselves. Although categorical features were one-hot encoded to fit the Naive Bayes model's requirements, the resulting high-dimensional and sparse feature space may have diluted the model's ability to learn meaningful patterns. Additionally, the weak correlations observed during exploratory data analysis suggested that no single feature strongly predicted churn, further limiting the effectiveness of a model like Naive Bayes that relies heavily on individual feature contributions.

The challenges presented from K-Nearest Neighbors are that prediction may be slow if a large dataset is used. KNN can also require a lot of memory and storage, which can be costly for a business. Additionally, it can be difficult for KNN to set a suitable K value for a given set of training data, and KNN can also be sensitive to irrelevant parameters, which can skew data.

Despite the strong performance of the Classification Tree model, several challenges were encountered during its development. One of the main issues is that decision trees are highly prone to overfitting, especially when the tree is grows too deep or complex [12]. Another important factor to consider is that small changes in the dataset can lead to large changes in the tree structure, making the model less stable and harder to interpret over time. A major challenge to consider as well is the threat of concept drift, where customer behavior patterns may evolve, potentially reducing the model's accuracy if not properly monitored [8].

7 Conclusions and Recommendations

This project set out to explore how machine learning models can be used to effectively predict customer churn in subscription-based services. By applying three different classification models—Naive Bayes, K-Nearest Neighbors (KNN), and Classification Trees—we aimed to assess model performance and identify key predictors that influence customer attrition.

Our results indicate that the Classification Tree model significantly outperformed both Naive Bayes and KNN across all major evaluation metrics, including accuracy, precision, recall, and AUC. This model also provided interpretability through feature importance analysis, which identified membership level and points in wallet as two of the most influential variables in predicting churn. In contrast, Naive Bayes struggled due to its simplifying assumption of feature independence, and KNN showed moderate success but was affected by high-dimensional feature encoding and sensitivity to irrelevant inputs.

The exploratory analysis further supported the importance of service quality and pricing strategies in retention. Negative feedback and lower membership tiers were associated with higher churn risk, reinforcing the value of customer satisfaction and loyalty programs. These insights directly address our subquestions and highlight the effectiveness of machine learning in guiding churn mitigation strategies.

Recommendations:

- **For future research**, we recommend exploring ensemble-based models such as Random Forests or Gradient Boosting, which can offer improved predictive power and handle feature interactions more robustly.
- **For businesses**, actionable strategies include targeting at-risk customers based on membership level and wallet engagement, and enhancing feedback loops to quickly address service dissatisfaction.
- **For model deployment**, it is critical to monitor for concept drift, as customer behavior patterns evolve over time. Regular retraining and validation should be built into the deployment pipeline.

In summary, this study demonstrates that machine learning can provide meaningful insights and decision support for customer churn prediction. With continued iteration and tuning, these models can play a central role in proactive customer retention strategies.

References

[1] S. A. and C. D. A survey on customer churn prediction using machine learning techniques. *International Journal of Computer Applications*, 154:13–16, 11 2016.

[2] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth International Group, Belmont, California, 1984.

[3] K. Cao and P.-j. Shao. Customer churn prediction based on svm-rfe. *2008 International Seminar on Business and Information Management*, 1:306–309, 2008.

[4] P.-C. Chang and Y.-P. Chiu. Factors influencing switching intention and customer retention of over-the-top (ott) viewing behavior in taiwan: The push–pull– mooring model. *Emerging Media*, 1(2):196–217, 2023.

[5] M. Colgate, K. Stewart, and R. Kinsella. Customer defection: a study of the student market in ireland. *The International Journal of Bank Marketing*, 14(3):23–29, 1996.

[6] D. Duffy. Customer loyalty strategies. *Journal of Consumer Marketing*, 15(5):435–448, 1998.

[7] C. Ennew and M. Binks. The impact of service quality and service characteristics on customer retention: Small businesses and their banks in the uk. *British Journal of Marketing*, 7:219–230, 1996.

[8] J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, and A. Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):1–37, 2014.

[9] R. Garland. Estimating customer defection in personal retail banking. *International Journal of Bank Marketing*, 20:317–324, 2002.

[10] J. Hadden, A. Tiwari, R. Roy, and D. Ruta. Computer assisted customer churn management: State-of-the-art and future trends. *Computers Operations Research*, 34(10):2902–2917, 2007.

[11] R. Iyengar et al. Return on loyalty: A strategic perspective on customer retention. *Journal of Marketing*, 86(4):1–23, 2022.

[12] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer, 2013.

[13] C. P. Johny and P. P. Mathai. Customer churn prediction: A survey. *International Journal of Advanced Research in Computer Science*, 11(4):1–5, 2020.

[14] S. M. Keaveney. Customer switching behavior in service industries: An exploratory study. *Journal of Marketing*, 59(2):71–82, 1995.

[15] A. Khattak, Z. Mehak, H. Ahmad, M. U. Asghar, M. Z. Asghar, and A. Khan. Customer churn prediction using composite deep learning technique. *Scientific Reports*, 13(1):17294, 2023.

[16] A. Mishra and U. S. Reddy. A novel approach for churn prediction using deep learning. *2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pages 1–4, 2017.

[17] S. Momin, T. Bohra, and P. Raut. Prediction of customer churn using machine learning. *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, pages 203–212, 2020.

[18] K. Ng and H. Liu. Customer retention via data mining. *Artificial Intelligence Review*, 14:569–590, 2000.

[19] P. E. Pfeifer and P. W. Farris. The elasticity of customer value to retention: The duration of a customer relationship. *Journal of Interactive Marketing*, 18(2):20–31, 2004.

[20] B. Prabadevi, R. Shalini, and B. Kavitha. Customer churning analysis using machine learning algorithms. *International Journal of Intelligent Networks*, 4:145–154, 2023.

[21] F. F. Reichheld and W. E. Sasser. Zero defeofions: Quoliiy comes to services. *Harvard business review*, 68(5):105–111, 1990.

[22] A. Rodan, H. Faris, J. Al-sakran, and O. Al-Kadi. A support vector machine approach for churn prediction in telecom industry. *International journal on information*, 17, 08 2014.

[23] P. Trivedi. Customer churn dataset. *Hugging Face*, 2022.

[24] R. S. Winer. A framework for customer relationship management. *California Management Review*, 43(4):89–105, 2001.

[25] B. Wu, G. Guo, and P. Luo. The effect of subscriptions on customer engagement. *Journal of Business Research*, 178:114638, 2024.

[26] Y. Xie, X. Li, E. Ngai, and W. Ying. Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, 36(3, Part 1):5445–5449, 2009.

Appendix A

Project Code: https://github.com/theoneandnoli/Capstone