Churn Analysis Using Classification and Clustering Techniques

William J. Owens

National University

ANA 630: Advanced Analytic Applications

Dr. Henry Mendoza Rivera

October 21, 2023

Abstract

In business, churn is a measure that represents a customer cancelling their subscription or choosing to no longer do business with a service provider. Ultimately, churned customers represent lost revenue and profit. Understanding and reducing churn can therefore help businesses retain more customers and increase their revenue and profitability. Furthermore, it generally costs more to acquire a new customer than to retain an existing one. Reducing churn can allow businesses to avoid the higher costs associated with customer acquisition, such as marketing and advertising expenditures. Lastly, the preemptive identification of customers who are at-risk for churn can enable businesses to engage in more effective retention measures. In this study, we perform an extensive churn analysis which includes statistical modeling with logistic regression. We then implement various classification methods and compare their predictive performance. In addition, we utilize clustering techniques for exploratory purposes and for feature engineering within our classification models.

1 Introduction

The phenomenon known as customer churn has been studied extensively within the field of telecommunications. This industry is one that is particularly sensitive to customer churn due to its primarily subscription-based revenue model. Customers often leave their current service provider for an alternative provider. This results in a significant loss of revenue for the organization, which is why efforts have been made to study this phenomenon, why it occurs, and how to prevent it.

Previous research has determined that factors such as use frequency, call quality, and customer service are all related to the likelihood of a customer churning (Ahn, Han, & Lee, 2006).

Other authors have posited that churn is heavily mediated by user activity status (Keramati & Ardabili, 2011). In terms of call quality, reception issues and network coverage are important considerations, and customers may leave their current provider for a competitor if that competitor can provide better reception and broader coverage. With respect to customer service, if complaints and billing errors are not handled promptly or appropriately, customers may churn as a result. Technological considerations are also relevant as churn determinants. Customers may leave a provider that does not adapt to newer technology and better features. Lastly, many customers consider pricing and special pricing deals when deciding whether to stay or leave their current provider (Almana, Aksoy, & Alzahran, 2014).

In prior research literature, the costs of churn on telecommunications providers have been well-documented. The global average churn rate of 2% translates to a total annual loss of over $100 billion (Almana, Aksoy, & Alzahran, 2014). It is also important to consider the cost of attracting new customers versus the cost of retaining existing ones. Various estimates suggest that there is between a five-fold and ten-fold difference in the cost of attracting new customers versus retaining existing customers (Celik & Osmanoglu, 2019). There is no doubt that attracting new customers is more costly. Researchers have estimated that telecommunications service providers can increase profits by 25 to 85% by simply reducing their customer churn rate by 5%. It appears that focusing on customer retention can produce a sizable return on investment (Almana, Aksoy, & Alzahran, 2014).

In order to prioritize retention, it would be advantageous to know which customers are likely to churn. Otherwise, untargeted retention efforts will be spread thin or lead to ballooning costs. Classification methods from machine learning can be used to identify who is at-risk of churning, and then service providers can subsequently target their retention efforts on these

individuals. This allows providers to reduce churn and minimize their retention campaign overhead (Keramati & Ardabili, 2011). The next section will focus on our research methodology. There are three main components of our work: statistical modeling and analysis, cluster analysis, and predictive modeling and analysis. Clustering techniques will be used for both exploratory data analysis and in the feature engineering process for our classification models.

2 Methods

2.1 Data

The data we utilized is from the Iranian Churn Dataset in the UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/563/iranian+churn+dataset). It is a random subsample of data collected from an Iranian telecommunications company over a period of 12 months. There are 3,150 observations within the dataset. Each record represents information pertaining to a single customer. There are 13 variables included in the dataset, including: number of call failures, whether a complaint was made (binary; "no" or "yes"), subscription length in months, charge amount from lowest to highest (ordinal – 10 levels), call seconds, number of calls, number of text messages, number of distinct phone calls, age group from younger to older (ordinal – 5 groups), tariff plan (binary; "pay-as-you-go" or "contractual"), activity status (binary; "inactive" or "active"), the calculated value of the customer, and churn status (binary – dependent variable; "no" or "yes"). There were 495 customers who churned, which is 15.71% of the sample. It should be noted that the ordinal variables were assumed to have an equal amount of separation between their factor levels and were thus treated as numeric variables when a decision had to be made for data processing operations.

2.2 Statistical Analysis

The purpose of this portion of our research was to better understand the relationships between the independent variables and the dependent variable, churn status. To this end, we observed the relationships between binary categorical independent variables and the binary dependent variable using contingency tables, chi-square tests of association, and mosaic plots. The potential associations between ordinal categorical variables and the binary dependent variable were examined with contingency tables, the Cochran-Armitage trend test, and spineplots. We also observed the relationships between continuous independent variables and the binary dependent variable using boxplots. After this exploratory analysis, we created an initial logistic regression model using the 12 available independent variables and the dependent variable "churn". This model can be summarized by the following:

$$logodds(churn)=\beta0 + \beta1(call\_failures) + \beta2(complaint) + \beta3(months\_subscribed) + \beta4 (charge\_amt) + \beta5(total\_call\_secs) + \beta6(total\_calls) + \beta7(total\_texts) + \beta8(distinct\_calls) + \beta9 (age\_group) + \beta10(tariff\_plan) + \beta11(active\_status) + \beta12(age) + \beta13(customer\_value)$$

Goodness of fit was subsequently evaluated based on AIC, BIC, and Deviance metrics. We also used the Hosmer-Lemeshow test to assess the fit of our logistic regression model and checked VIF measures to determine whether multicollinearity was present. After this, we performed a backward elimination procedure, eliminating statistically insignificant variables from the model in stepwise fashion, until the resulting model contained only statistically significant independent variables. This reduced model was then evaluated using the same metrics, tests, and checks as the initial model.

Lastly, an interaction model was created by adding the interaction term for “complaint” and “active_status” to the previously reduced model. We hypothesized that customers who made a complaint and were active users would be even more likely to churn. Therefore, this interaction term may provide additional explanatory power and improve the fit of our model. After fitting the interaction model, we evaluated it using the same methods as before. We also calculated odds ratios and 95% odds ratio confidence intervals and plotted the Deviance residuals and studentized Pearson residuals for our model.

2.3 Cluster Analysis

Our objective with the cluster analysis we performed was to investigate any latent patterns or trends that existed within the data using an array of clustering techniques. The first clustering method we implemented on the data was the K-Prototypes algorithm. We selected this technique because the popular K-Means algorithm is only suitable for continuous numeric features. Since we have mixed data, we needed to select an algorithm that accommodates both numeric and categorical features. With K-Prototypes, the Euclidean distance is calculated for numeric features, while a matching dissimilarity measure is computed for categorical features. To find the optimal number of clusters for the algorithm, we examined elbow and silhouette score plots (these have been adjusted for compatibility with K-Prototypes). We then used a number for the k parameter based on these visualizations. The cluster labels were then examined with the original data to discern any observable patterns that may be insightful for customer segmentation.

The second clustering method we explored was a combination of Factor Analysis of Mixed Data (FAMD) and K-Means. FAMD is a dimensionality reduction technique that can be used when there are both continuous and categorical features present. It is essentially a mixture of Principal

Component Analysis (PCA) and Multiple Correspondence Analysis (MCA). The continuous variables are standardized as in PCA and the categorical variables are one-hot encoded, cross-tabulated in contingency tables, and used to compute distances between levels via the chi-square statistic. The processed data is then reduced to the orthogonal linear combinations of the variables that explain the most variance. We used the results of performing FAMD as input for the K-Means algorithm. Since the mixed data was reduced to its principal components, K-Means was more viable than if we still had unprocessed categorical variables present. Silhouette plots with different values for k were examined to determine the optimal number of clusters. The resulting cluster labels were used to group the original data for pattern inspection. We also produced a FAMD plot with the clusters color-coded to examine overall cluster separation.

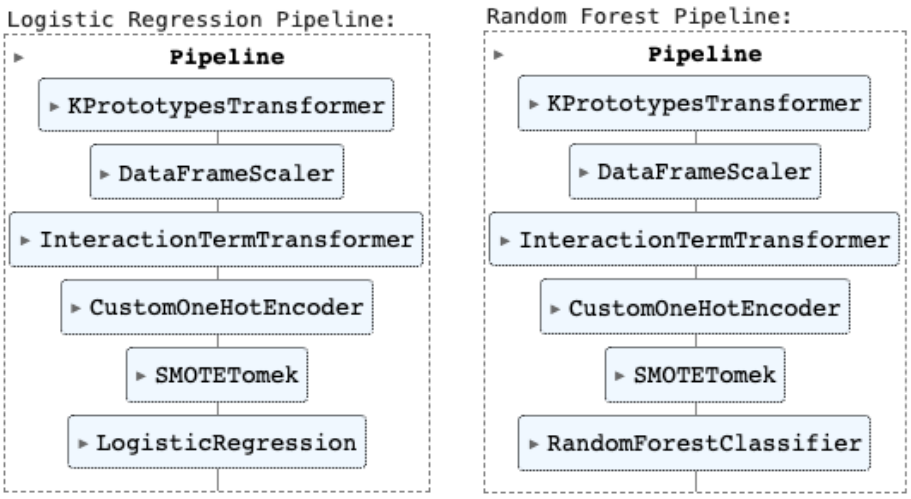
2.4 Predictive Analysis

The final portion of our research focused on obtaining a high-performing model in terms of its ability to predict customer churn. The data was split into three partitions: training set (50%), validation set (30%), and test set (20%). The partitions were stratified according to the label class proportions in the original dataset. Before applying the classification method, the data was transformed via a 5-step preprocessing pipeline. In the first step, the K-prototypes algorithm was applied to the original set of 12 features with k = 3. The resulting cluster labels for each observation were then added to the dataset as a new feature, “cluster_label”. In the second pipeline step, only the numeric features were standardized by subtracting the mean and dividing by the standard deviation. An interaction feature, “complaint:active_status” was added to the dataset in the third step. In the fourth step, the “cluster_label” feature, as a potentially non-ordinal categorical feature with 3 categories, was one-hot encoded with the reference category dropped. Lastly, the

SMOTETomek algorithm was applied to address the label's class imbalance. In SMOTETomek, the minority class (the positive class for "churn") is oversampled (i.e. new samples are created from the minority class using distance measures), balancing the class distribution and providing more instances for the classifier to train on. In addition, the SMOTETomek algorithm removes "noisy" instances of the majority class that are closer to instances of the minority class than they are to other instances of their own class. This makes the decision boundary between classes clearer, potentially leading to improved model performance. After preprocessing, one of seven classifiers was applied to the transformed dataset. An example of the complete pipeline, including the classifier step, can be found below (see Figure 1).

Figure 1

Pipeline Workflows with Preprocessing Steps and Classifier Step



The classification methods that were utilized include: Logistic Regression, Random Forest, Support Vector Machine, XGBoost, Histogram-based Gradient Boosting, CatBoost, and Symbolic Classifier (genetic programming). In logistic regression, the linear predictors are related to the log

odds of the dependent variable via a logit link function. The log odds are converted to probabilities of class membership via the sigmoid function and then class membership is determined based on a certain cutoff point. Random forests build multiple decision trees with different bootstrapped samples and varying subsets of the features. The results are averaged to provide an improved, ensemble-based prediction. Support vector machines iteratively construct the best hyperplane in higher-dimensional space that separates the different classes (Sarker, Kayes, & Watters, 2019). XGBoost involves iteratively enhancing the performance of weak-learner decision trees by sequentially fitting new trees to the residuals of the old tree ensemble and attempting to optimize scores on a specified loss function. It can incorporate bootstrapped samples and randomized feature subsetting (Aziz et al., 2020). Histogram-based gradient boosting involves the discretization of continuous feature values into bins, thereby making the model faster to train and more memory efficient. As with other boosting methods, there is an iterative process where decision trees are fitted to the residuals of the previous iteration (Nhat-Duc & Van-Duc, 2023). CatBoost is a gradient boosting method that excels at handling categorical data (Aziz et al., 2020). Symbolic classification (genetic programming algorithm) evolves a population of random symbolic expressions using operations such as mutation, crossover, and selection over generations. This ultimately improves the model's fitness for the classification task (Korns, 2018).

Hyperparameters were tuned for each classification method using a randomized search with stratified 5-fold cross validation. The tuning algorithm randomly selects subsets of the hyperparameter space for a specified number of iterations (we used 250 iterations). This can be faster and less computationally intensive than a full grid search for optimal hyperparameters. In addition, 5-fold cross validation involves the training set being divided into five parts and the model being trained and tested five times with different partition permutations for each

hyperparameter set. The cross validation is stratified in the sense that the label class in each of the five folds maintains the same class representation as the original training set. For each classifier method, the model with the best mean recall score was chosen as the best estimator. Recall is a measure of how well a model can correctly identify the positive class. We decided to optimize for recall ability because our goal is to reduce churn. Therefore, we need to be able to correctly identify a high number of potential churners, so that we may direct our retention efforts toward them. A model that does a poor job at identifying those at-risk of churning is not conducive to our goals. Ultimately, failure to identify potential churners can lead to an ineffective retention campaign and lost revenue. Of course, other metrics will be taken into consideration when assessing model performance. However, recall was selected as the metric we wish to prioritize when determining which estimator was the most performant. The best estimators for each classification method from the stratified 5-fold tuning process were then fit to the full training set. After training the models, they were evaluated on the holdout validation set based on five key metrics (accuracy, recall, precision, F1 score, and ROC-AUC score). A radar chart was produced to visually compare the results.

Next, the two best models were selected from among the seven. We combined the training and validation sets together to increase the data available for retraining the selected models. After retraining the top two models, we evaluated their performance on the holdout test partition using the same key metrics. The models were also visually compared using a combined ROC-AUC plot, diverging bar plot, confusion matrices, and permutation importance plots. Permutation-based feature importance differs from impurity-based feature importance. In the latter, mean decrease in impurity is used and a feature is considered highly important if it tends to decrease impurity the most across all trees. This method can often inflate the importance of continuous variables or high-

cardinality categorical variables due to its sensitivity to scaling, dimensionality, and outliers. On the other hand, the permutation importance algorithm randomly shuffles the values of a feature in the dataset and if the model's performance decreases significantly, the feature in question is determined to be important. The results of our analytical work will be discussed in the subsequent section.

3 Results

3.1 Statistical Analysis

Our analysis of the relationship between binary categorical variables and the binary dependent variable ("churn") involved creating contingency tables and performing chi-square tests of association. The chi-square statistic for the association between "complaint" and "churn" was 886.21 with a p-value very close to zero. For the association between "tariff_plan" and "churn", the chi-square statistic was 34.22 with a p-value very close to zero as well. The association between "active_status" and "churn" produced a chi-square statistic of 781.11 with a p-value very close to zero. All three associations are statistically significant at the 0.05 level. This suggests that we can reject the null hypothesis of no association for each of the three relationship pairs that we examined. The results are shown below (see Figure 2 and Figure 3). The strength of these associations will be discussed when we obtain the adjusted odds ratios from our final logistic regression model.

Figure 2

Contingency Tables and Chi-Square Test Results

Contingency Table for complaint by churn:

complaint	0	1
churn		
0	2614	41
1	295	200

Chi2 statistic: 886.21, p-value: 0.0000

Contingency Table for tariff_plan by churn:

tariff_plan	0	1
churn		
0	2416	239
1	489	6

Chi2 statistic: 34.22, p-value: 0.0000

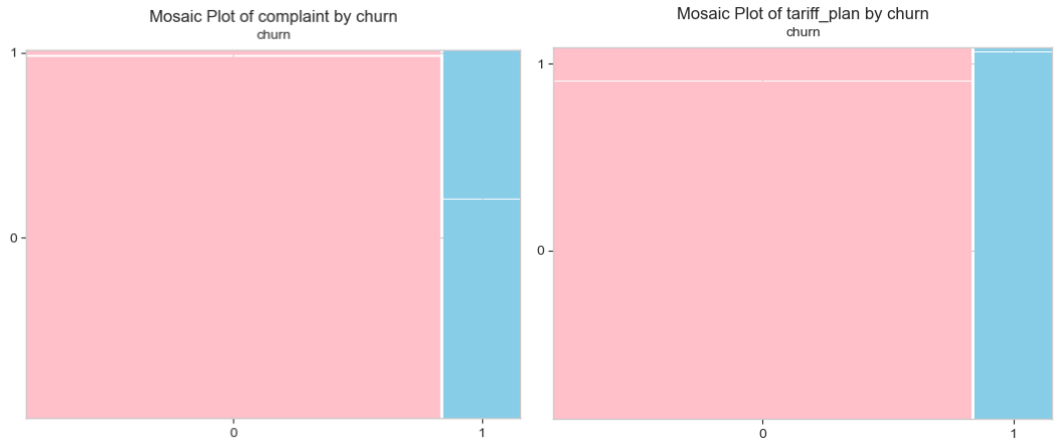
Contingency Table for active_status by churn:

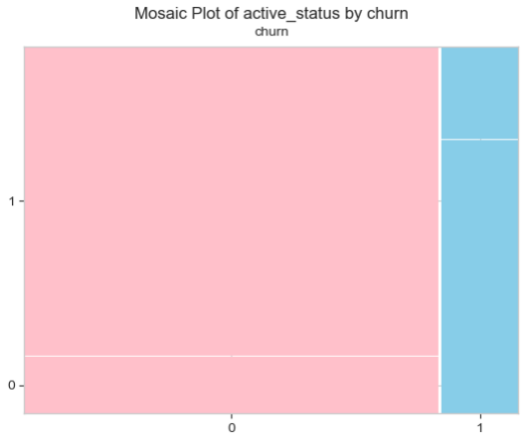
active_status	0	1
churn		
0	412	2243
1	370	125

Chi2 statistic: 781.11, p-value: 0.0000

Figure 3

Mosaic Plots of the Relationships





Associations between the ordinal variables and the binary dependent variable were tested using the Cochran-Armitage trend test (see Figure 4). The relationship between "charge_amt" and "churn" resulted in a test statistic of 5,511 with a near-zero p-value, which is statistically significant at the 0.05 level. For the relationship between "age_group" and "churn", the test statistic was 7,518 with a p-value of 0.39, which is not statistically significant at the 0.05 level. These results indicate that we can reject the null hypothesis of no trend for the relationship between "charge_amt" and "churn", but we cannot reject the null hypothesis for the relationship between "age_group" and "churn". This suggests that the "age_group" variable may not be an informative predictor variable in the modeling process.

Figure 4

Contingency Tables and CA Trend Test Results

Results for Variable: charge_amt
Contingency Table:

charge_amt	0	1	2	3	4	5	6	7	8	9	10
churn											
0	1347	574	372	192	75	30	11	14	19	14	7
1	421	43	23	7	1	0	0	0	0	0	0

Statistic: 5511.0000
Null Mean: 5174.8501
Null SD: 31.4262
Z-score: 10.6965
P-value: 0.0000

Results for Variable: age_group
Contingency Table:

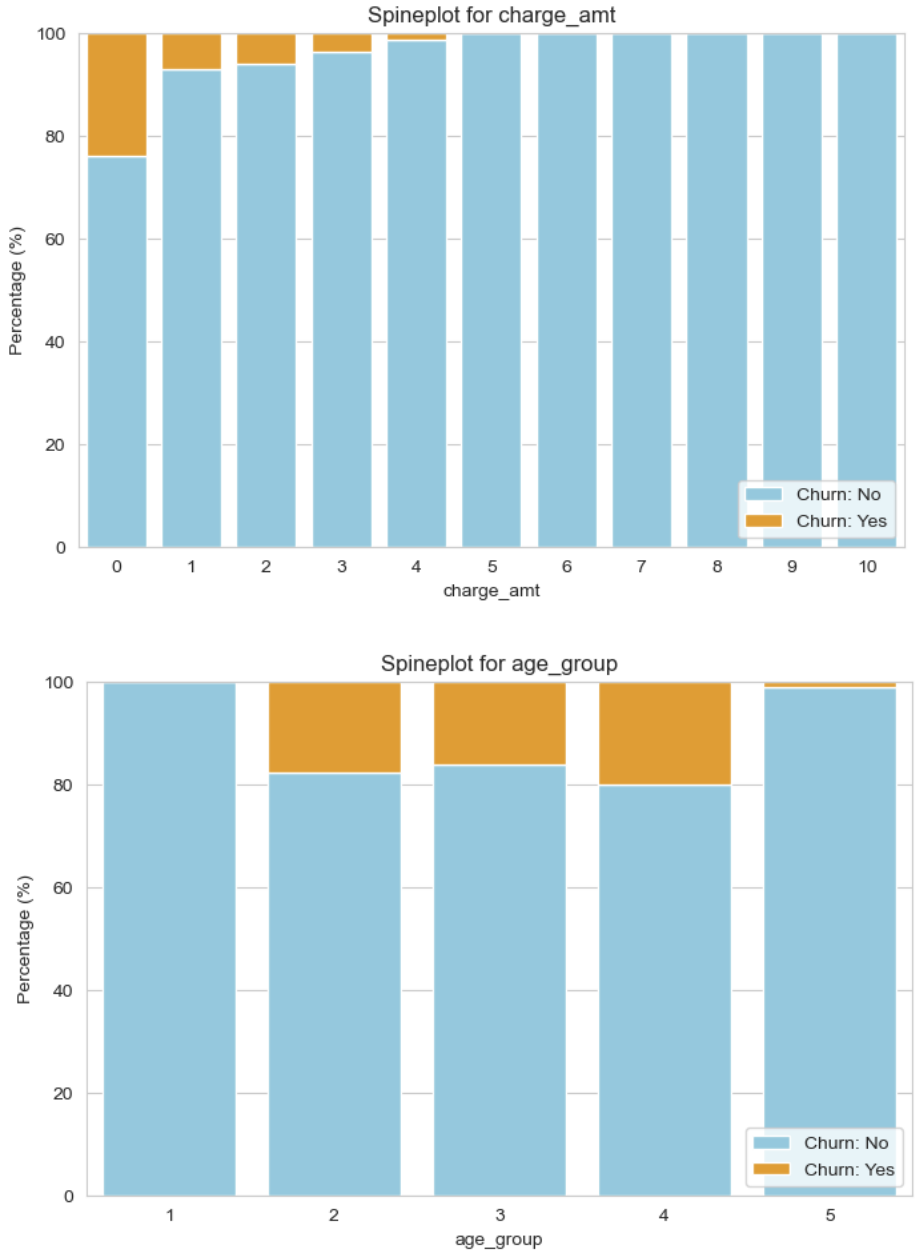
age_group	1	2	3	4	5
churn					
0	123	853	1195	316	168
1	0	184	230	79	2

Statistic: 7518.0000
Null Mean: 7502.3449
Null SD: 18.2435
Z-score: 0.8581
P-value: 0.3908

As we can see in the spineplots below (see Figure 5), with respect to "charge_amt", the proportion of churned customers appears to decrease as we move up the ordinal levels. However, for "age_group" the proportions appear to be relatively similar in the middle three levels, showing that no trend is present.

Figure 5

Spineplots of the Ordinal Variables vs. Dependent Variable

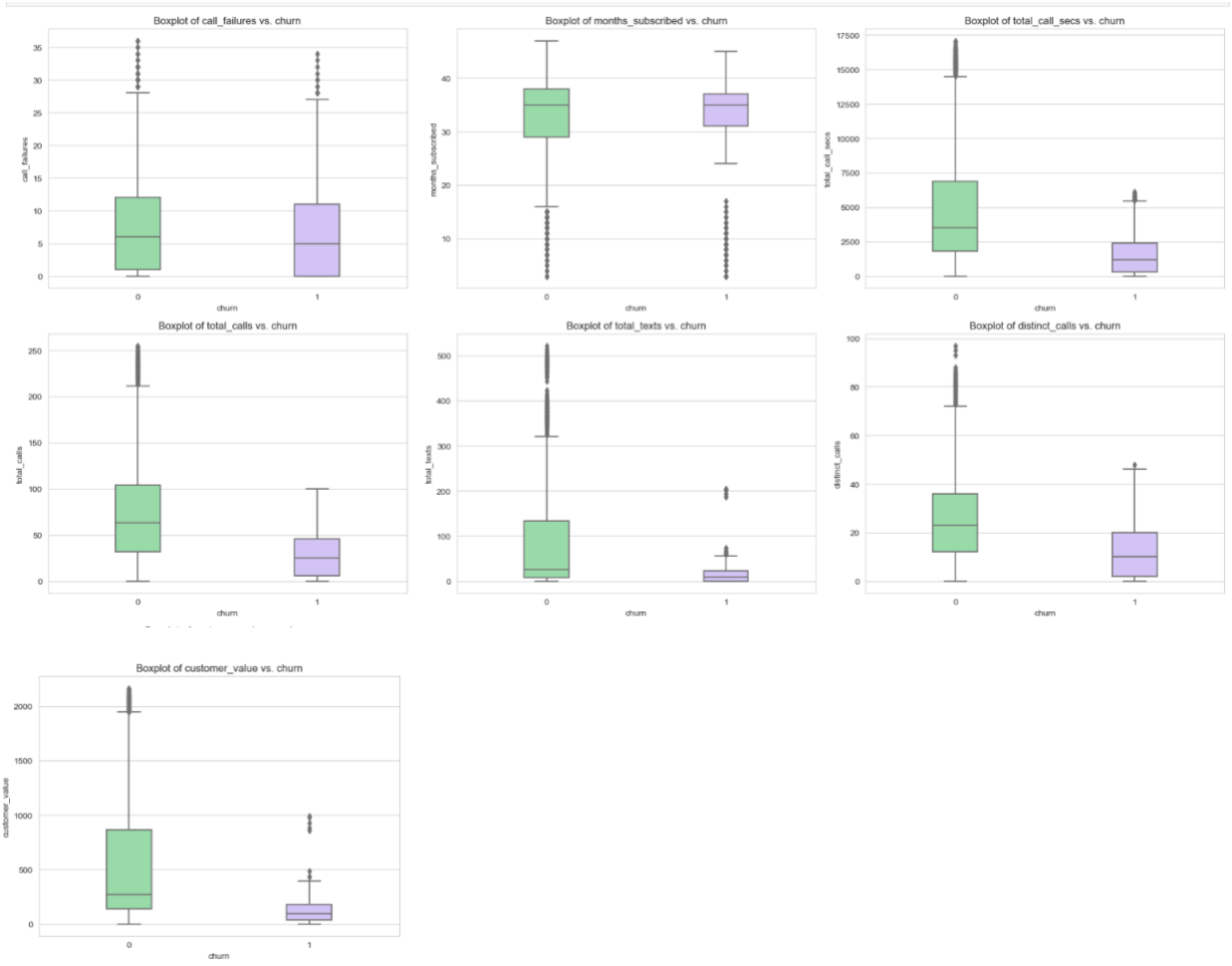


Continuous variables were examined with respect to the binary dependent variable using boxplots (see Figure 6). Outliers were present in virtually all the resulting plots, particularly for the non-churn groups. It appears that non-churners and churners experience similar amounts of

call failures on average, and they are subscribed for a similar number of months on average. Individuals who churn appear to text less than their non-churning counterparts as evidenced by the smaller box, covering the middle 50% of the data. Churners have a lower number of total calls and a lower number of overall customer value as well.

Figure 6

Boxplots of Continuous Variables vs. Dependent Variable



The initial logistic regression model utilized all 12 independent variables, 4 of which did not have statistically significant p-values for their respective coefficients ("total_call_secs",

"distinct_calls", "age_group", and "tariff_plan"). This suggests that these variables may not have a meaningful association with the odds of churning. The model produced an AIC statistic of 1,411.13, a BIC statistic of 1,489.84, and a Deviance statistic of 1,385.13 (see Figure 7). The AIC and BIC statistics are measures of goodness of fit relative to model parsimony. They should be interpreted in the context of comparing two different models and determining which one may better maximize this fit vs. parsimony trade-off. The initial model will therefore serve as a baseline for comparison. The Deviance statistic may not be the best measure to evaluate the goodness of fit for a logistic regression model with continuous variables. It is intended to compare the current model's fit to that of a perfectly-fitting, saturated model. Conceptually, a saturated model is less clear when we have a set of continuous predictors that can take on a wide range of values or infinite values. Moreover, Deviance is sensitive to outliers, which appear to be present within many of our continuous predictors. Given these limitations, we will use the Hosmer-Lemeshow goodness-of-fit test instead (see Figure 7). This test produced a test statistic of 22.90 and a p-value of 0.003. Since the p-value is statistically significant at the 0.05 level, we must reject the null hypothesis that the model has an adequate fit that describes the observed data well. We also checked for multicollinearity using the Variance Inflation Factor (see Figure 8). There were 6 variables that had a VIF score above 10, indicating they are being affected by multicollinearity.

Figure 7

Goodness of Fit Statistics for the Initial Model

	Statistic	Value
0	AIC	1411.126529
1	BIC	1489.843580
2	Deviance	1385.126529

	Statistic	Value
0	Hosmer-Lemeshow Test H	22.904712
1	Degrees of Freedom	8.000000
2	p-value	0.003489

Figure 8

Variance Inflation Factor Scores for the Initial Model

	variable	VIF
11	customer_value	83.195806
6	total_texts	50.934524
5	total_calls	46.342451
4	total_call_secs	45.470945
8	age_group	12.463875
2	months_subscribed	11.202536
7	distinct_calls	6.873198
10	active_status	6.219737
0	call_failures	5.987179
3	charge_amt	4.133866
9	tariff_plan	1.629279
1	complaint	1.235967

Given the poor fit and multicollinearity issues with our initial model, a backward elimination procure was conducted to obtain a reduced model with only statistically significant variables. Ultimately, the four variables that were found to be statistically insignificant in the initial model were eliminated by our variable selection procedure. Upon evaluating the same metrics as before (see Figure 9), the AIC and BIC were lower for the reduced model than for the initial model (1,407.87 and 1,462.37, respectively). This suggests that the reduced model has a better fitness-to-parsimony balance. However, the Hosmer-Lemeshow test resulted in a test statistic of 28.74 with an even lower p-value of 0.00035. These results indicate that we must again reject the null hypothesis that the model fits the data well enough. An examination of VIF scores again (see Figure 10) shows that multicollinearity may be affecting our model, as three variables had VIF scores above 10.

Figure 9

Goodness of Fit Statistics for the Reduced Model

	Statistic	Value
0	AIC	1407.870918
1	BIC	1462.367338
2	Deviance	1389.870918

	Statistic	Value
0	Hosmer-Lemeshow Test H	28.742728
1	Degrees of Freedom	8.000000
2	p-value	0.000352

Figure 10

Variance Inflation Factor Scores for the Reduced Model

	variable	VIF
7	customer_value	39.794347
5	total_texts	26.089677
4	total_calls	11.285861
6	active_status	5.170407
0	call_failures	5.097420
2	months_subscribed	4.173426
3	charge_amt	2.434918
1	complaint	1.190960

In an effort to obtain a decently fitting model, we decided to experiment with adding an interaction term based on the interaction between “complaint” and “active_status”. All coefficients had statistically significant p-values associated with them. The interacted model resulted in lower AIC and BIC scores than both the initial and reduced models (1,384.35 for AIC and 1,444.90 for BIC) and the Hosmer-Lemeshow test produced a p-value greater than 0.05, indicating that we cannot reject the null hypothesis that the current model adequately fits the data (see Figure 11). VIF scores revealed that 3 variables still may be affected by multicollinearity (see Figure 12).

Figure 11

Goodness of Fit Statistics for the Reduced Model with Interaction Term

	Statistic	Value
0	AIC	1384.345561
1	BIC	1444.897138
2	Deviance	1364.345561

	Statistic	Value
0	Hosmer-Lemeshow Test H	12.490907
1	Degrees of Freedom	8.000000
2	p-value	0.130608

Figure 12

Variance Inflation Factor Scores for the Reduced Model with Interaction Term

	variable	VIF
7	customer_value	39.887448
5	total_texts	26.185168
4	total_calls	11.306887
6	active_status	5.528965
0	call_failures	5.138572
2	months_subscribed	4.395839
3	charge_amt	2.498476
1	complaint	1.972488
8	complaint:active_status	1.770423

The adjusted odds ratios are displayed in the figure below (see Figure 13). The adjusted odds ratio for “complaint” with respect to the dependent variable “churn” is 19.95 (see Figure 13). This indicates that individuals who have made a complaint are 19.95 times more likely to churn than individuals who have not made a complaint, holding all other factors constant. The 95% odds ratio confidence interval swings widely from 11.13 to 35.77, indicating a lack of precision in our point estimate. However, the interval does not include 1.00, therefore the odds ratio is statistically significant at the 0.05 level. The adjusted odds ratio for “active_status” is 0.23. This means that individuals who are active users are 0.23 times as likely to churn as those who are inactive, given all other variables are held constant. The confidence interval ranges from 0.16 to 0.35, indicating

better precision than with the "complaint" variable. Additionally, the interval does not include 1.00, so the odds ratio is statistically significant at the 0.05 level.

Figure 13

Adjusted Odds Ratios

	Coefficient	Odds Ratio	95% CI Lower	95% CI Upper
call_failures	0.128034	1.136592	1.100048	1.174351
complaint	2.993471	19.954826	11.130214	35.776050
months_subscribed	-0.025611	0.974714	0.955855	0.993946
charge_amt	-0.488091	0.613797	0.495538	0.760279
total_calls	-0.050362	0.950885	0.938894	0.963030
total_texts	-0.042214	0.958665	0.944028	0.973528
active_status	-1.449976	0.234576	0.157415	0.349560
customer_value	0.006947	1.006972	1.003375	1.010581
complaint:active_status	2.761162	15.818219	5.046324	49.583826

The plots of the Deviance residuals and the studentized Pearson residuals show patterns that indicate poor model specification (see Figure 14 and Figure 15). The large gap between the two lines and the sloping patterns are evidence of this. The model appears to overestimate and underestimate the data at certain points. This may be due to the fact that the model has very few predictors and it may require additional predictors, interaction terms, polynomial terms, spline functions, or other changes to adequately describe the data, capture non-linear relationships, and make better predictions. The presence of outliers and multicollinearity may also be affecting the model results.

Figure 14

Plot of Deviance Residuals vs. Fitted Values

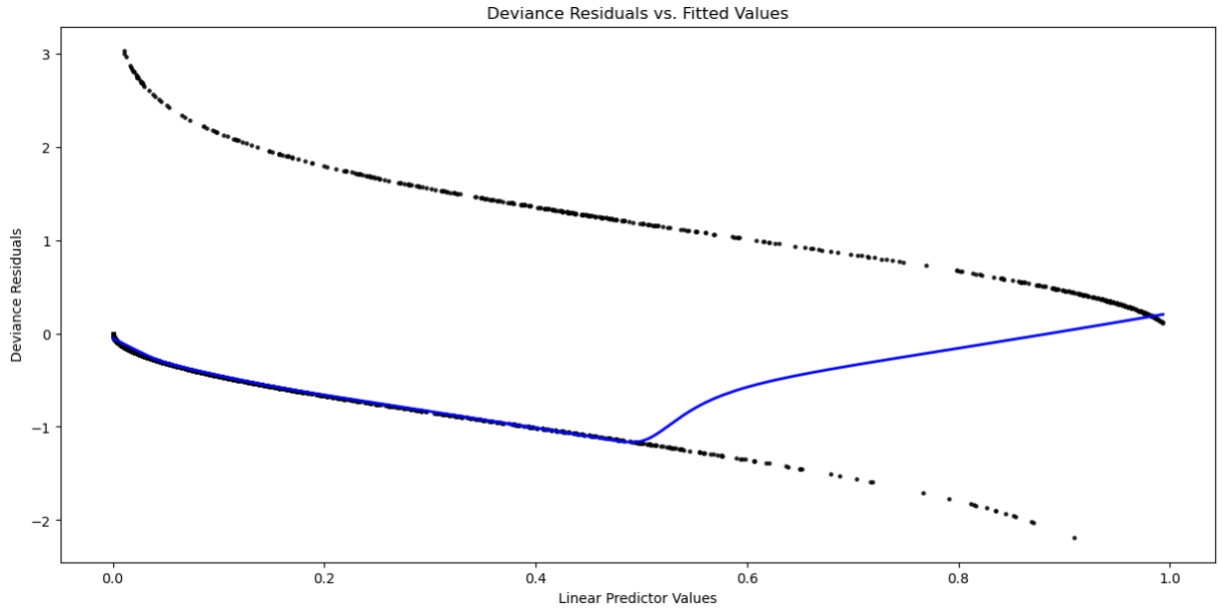
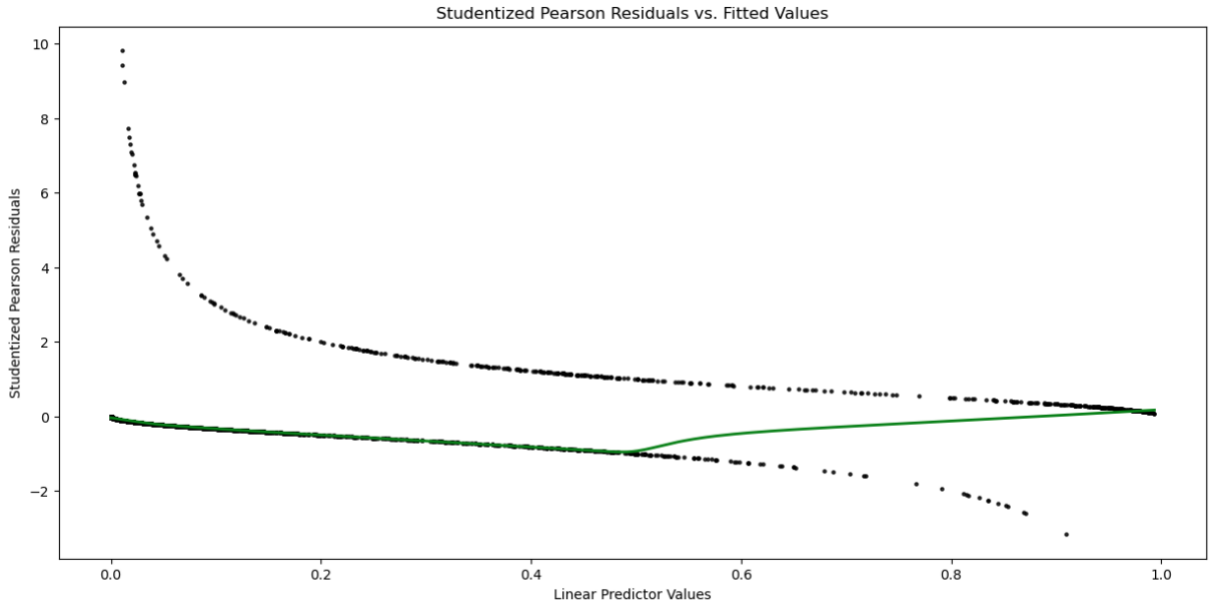


Figure 15

Plot of Studentized Pearson Residuals vs. Fitted Values



3.2 Cluster Analysis

K-Prototypes was the first clustering method we examined. This method requires the user to set the number of clusters (k). To find the optimal k value, we used elbow and silhouette score plots (see Figure 16 and Figure 17) that were modified to be compatible with mixed data. For the elbow plot, optimal k lies at the point where the line starts to look like an elbow rather than maintaining its smooth course across the plot. For the silhouette score plot, optimal k is indicated at the highest score between 0 and 1. The plots appeared to show that the optimal k value was most likely 2 or possibly 3.

Figure 16

Elbow Plot for K-Prototypes

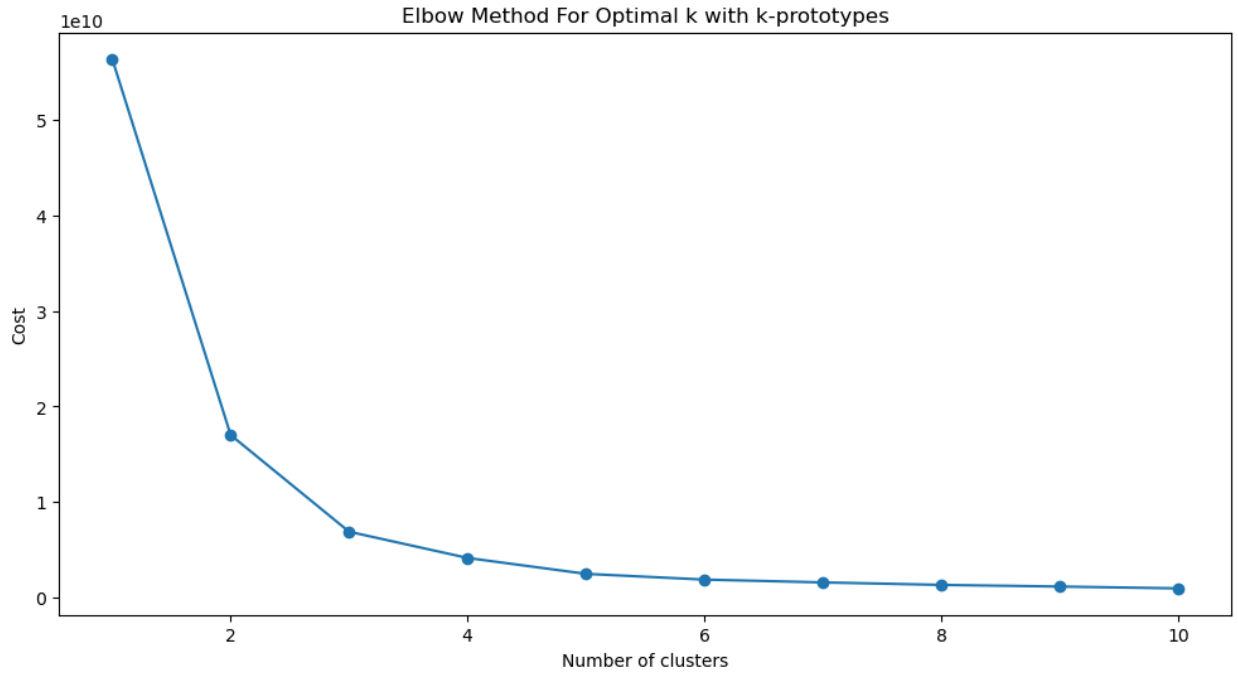
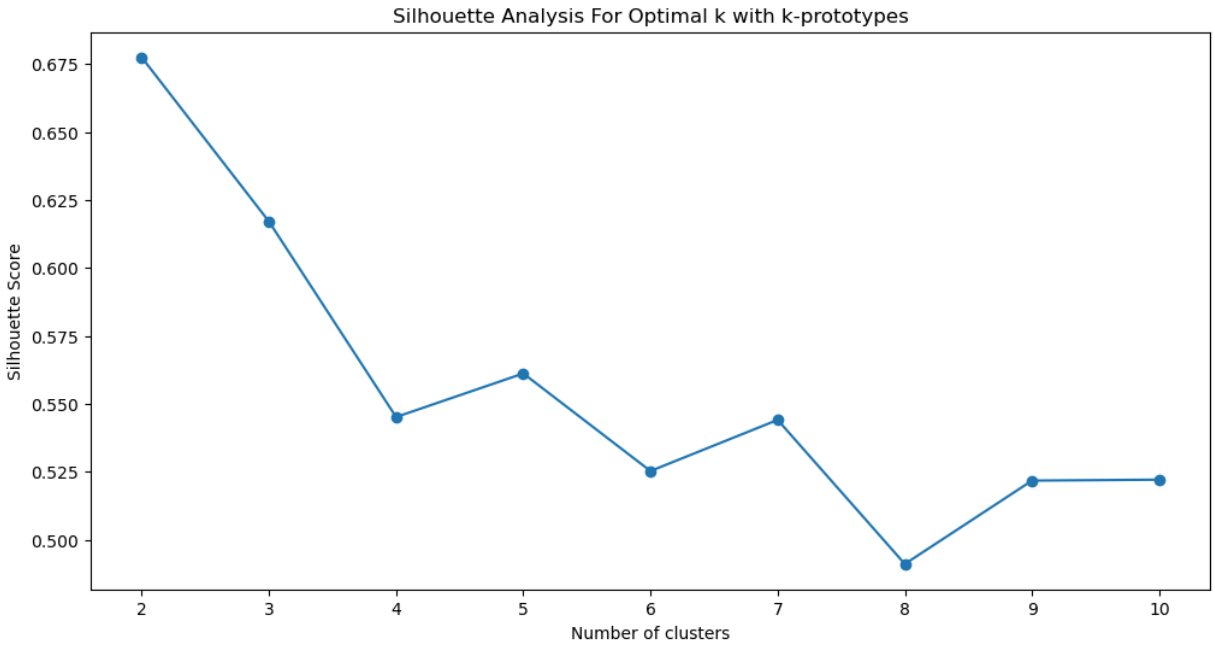


Figure 17

Silhouette Score Plot for K-Prototypes



We reasoned that only exploring 2 clusters for this analysis would not yield very interesting results. We also wondered whether the clusters would merely be the observations separated into the non-churn and churn groups or some other simplistic separation, which might not provide any worthwhile insights or insights that could be used for customer segmentation. Consequently, we performed the K-Prototypes clustering with 3 clusters and then observed the data grouped by cluster labels for any patterns. The first cluster appeared to show customers who were very active users (high total call seconds, total calls, and other measures of activity), the second cluster showed customers who were slightly less active, and the third cluster showed customers who were hardly active. The mean “months_subscribed” and the mean “age_group” were all relatively similar. In essence, the clustering algorithm appeared to have grouped the data by activity level where some individuals were very-high-frequency users, some were moderate-to-high-frequency users, and

some were low-frequency users. Clustering algorithms can achieve different results on different runs and with different values of k, so perhaps the results would be different if we had experimented with the algorithm for longer. It is unknown then, whether this clustering pattern would be reproduced.

Next, we investigated the use of Factor Analysis of Mixed Data (FAMD) with the K-Means clustering algorithm. To find the optimal number of clustering for the k parameter, we used an elbow plot with the FAMD-reduced data and created silhouette plots with the FAMD-reduced data for various values of k (see Figure 18 and Figure 19). Based on the plots, the optimal number of clusters appeared to be 2 or 3, which were the same results we obtained with the K-Prototypes method.

Figure 18

Elbow Plot for FAMD with K-Means

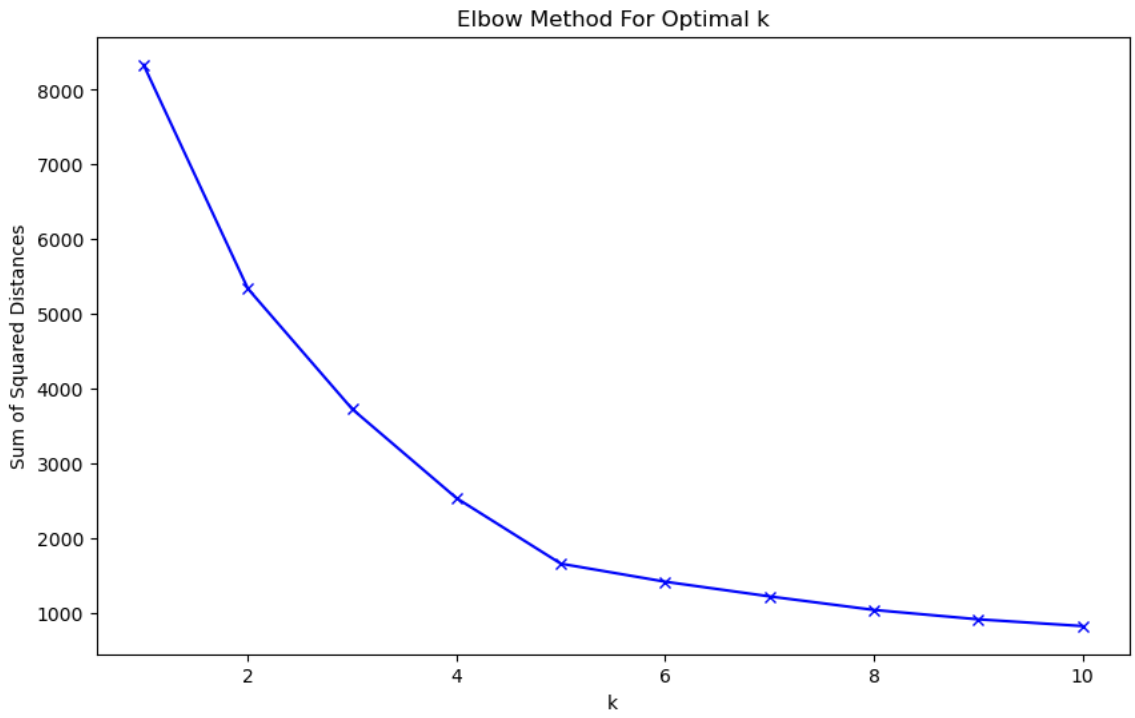
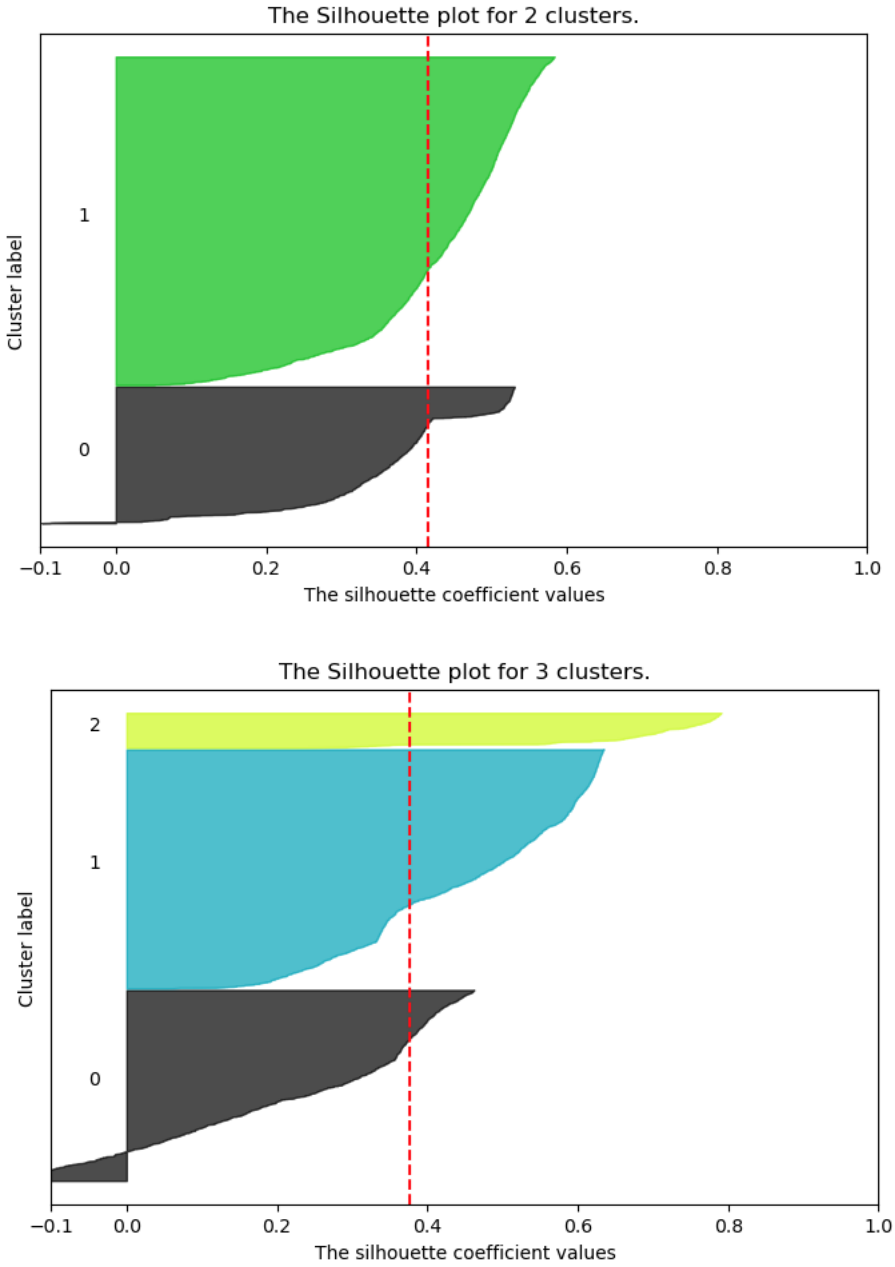


Figure 19

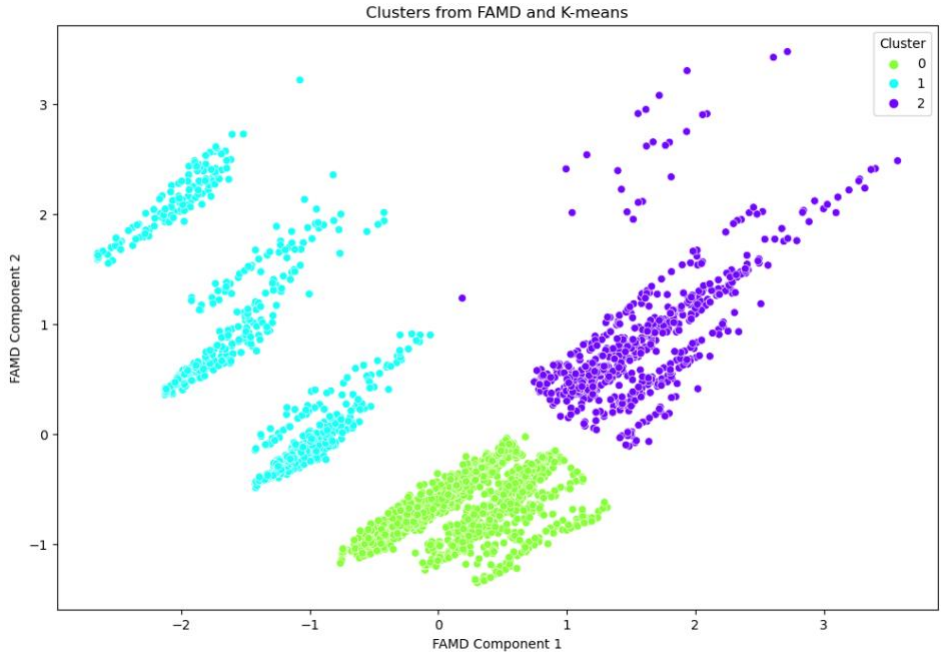
Silhouette Plots for FAMD with K-Means



Again, we settled on 3 clusters in the hopes of gleaning more interesting insights than would be achieved with merely 2 clusters. The same patterns emerged from the data grouped by cluster labels as were observed with the K-Prototypes algorithm. One group had very high numbers for usage and activity measures like total call seconds and total calls. Another group had lower activity and usage levels than the first group, but still had fairly high levels overall. Lastly, the third group was very low in terms of their usage and activity levels. The consistency across clustering methods lends credence to the idea that clustering the data with 3 clusters will result in an ordinal grouping of the customers by activity level. The plot of the FAMD principal components shows a fair level of separability between the 3 clusters (see Figure 20), indicating that our choice for the number of clusters may have been appropriate. However, there appears to be additional separation within the second cluster, suggesting that trying other cluster numbers may be warranted in future experimentation.

Figure 20

FAMD with K-Means Components Plot



3.3 Predictive Analysis

The final component of our work involves building, selecting, and evaluating a model to predict customer churn. The emphasis for this model should be on maximizing recall, so that stakeholders can correctly identify all the individuals who may be at-risk of churning and direct their retention efforts effectively. After completing the preprocessing pipeline, optimizing hyperparameters, training the best estimators, and applying the models to the holdout validation set, we evaluated the five key metrics described earlier for each classification method. The Symbolic Classifier and Support Vector Machine Classifier performed the best in terms of recall, but their other metric scores were extremely poor. The XGBoost classifier and Histogram-based Gradient Boosting classifier both performed satisfactorily in terms of recall while still maintaining

adequate or above-adequate scores on the other four metrics. The metric scores and radar chart are displayed below for comparison (see Figure 21 and Figure 22).

Figure 21

Model Metric Scores

Metrics for Logistic Regression:

	accuracy	recall	precision	f1_score	roc_auc
0	0.684656	0.858108	0.314356	0.460145	0.823722

Metrics for Random Forest:

	accuracy	recall	precision	f1_score	roc_auc
0	0.839153	0.905405	0.492647	0.638095	0.935675

Metrics for SVM:

	accuracy	recall	precision	f1_score	roc_auc
0	0.437037	0.993243	0.216814	0.355932	0.803982

Metrics for XGBoost:

	accuracy	recall	precision	f1_score	roc_auc
0	0.912169	0.824324	0.681564	0.746177	0.953241

Metrics for HistGB:

	accuracy	recall	precision	f1_score	roc_auc
0	0.93545	0.831081	0.773585	0.801303	0.962028

Metrics for CatBoost:

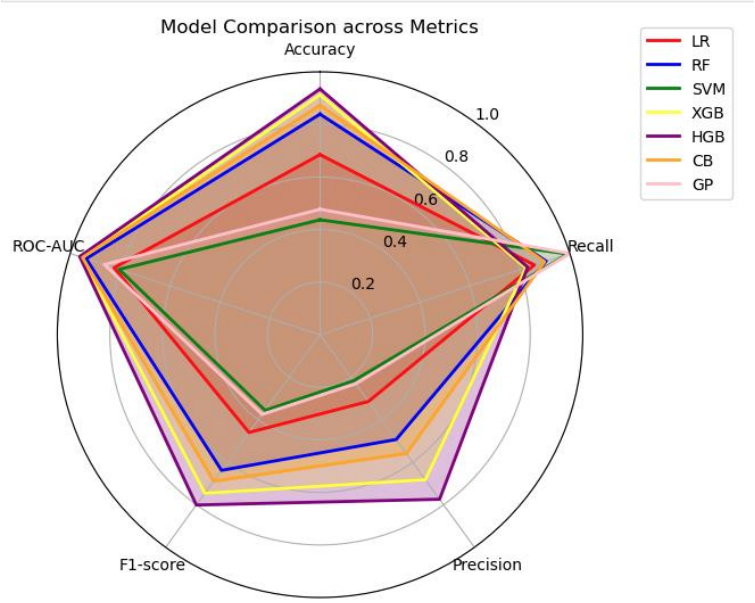
	accuracy	recall	precision	f1_score	roc_auc
0	0.873016	0.898649	0.558824	0.689119	0.95144

Metrics for Genetic Programming:

	accuracy	recall	precision	f1_score	roc_auc
0	0.477249	1.0	0.23053	0.374684	0.863428

Figure 22

Radar Chart Comparing Model Metrics



Next, we retrained the XGBoost and Histogram-based Gradient Boosting models on the combined training and validation data and applied them to the holdout test set. Upon reviewing the results (see Figures 23-26), we saw that the Histogram-based Gradient Boosting model had slightly higher recall. However, the metric scores on all five metrics were extremely close. Both models performed quite well in terms of recall, accuracy, and ROC-AUC scores, while maintaining a fair degree of precision. Either model may allow us to correctly identify a high proportion of customers who may be at-risk of churning, and our predictions as to who may be at-risk are likely to be correct almost 75% of the time.

Figure 23

Metric Scores for Top Two Models

```
Test Results for HistGB:
accuracy: 0.9380952380952381
recall: 0.9292929292929293
precision: 0.7419354838709677
f1: 0.8251121076233184
roc_auc: 0.9763453746504593


Test Results for XGBoost:
accuracy: 0.9349206349206349
recall: 0.9090909090909091
precision: 0.7377049180327869
f1: 0.8144796380090498
roc_auc: 0.9764880442846545
```

Figure 24

ROC Plot for Top Two Models

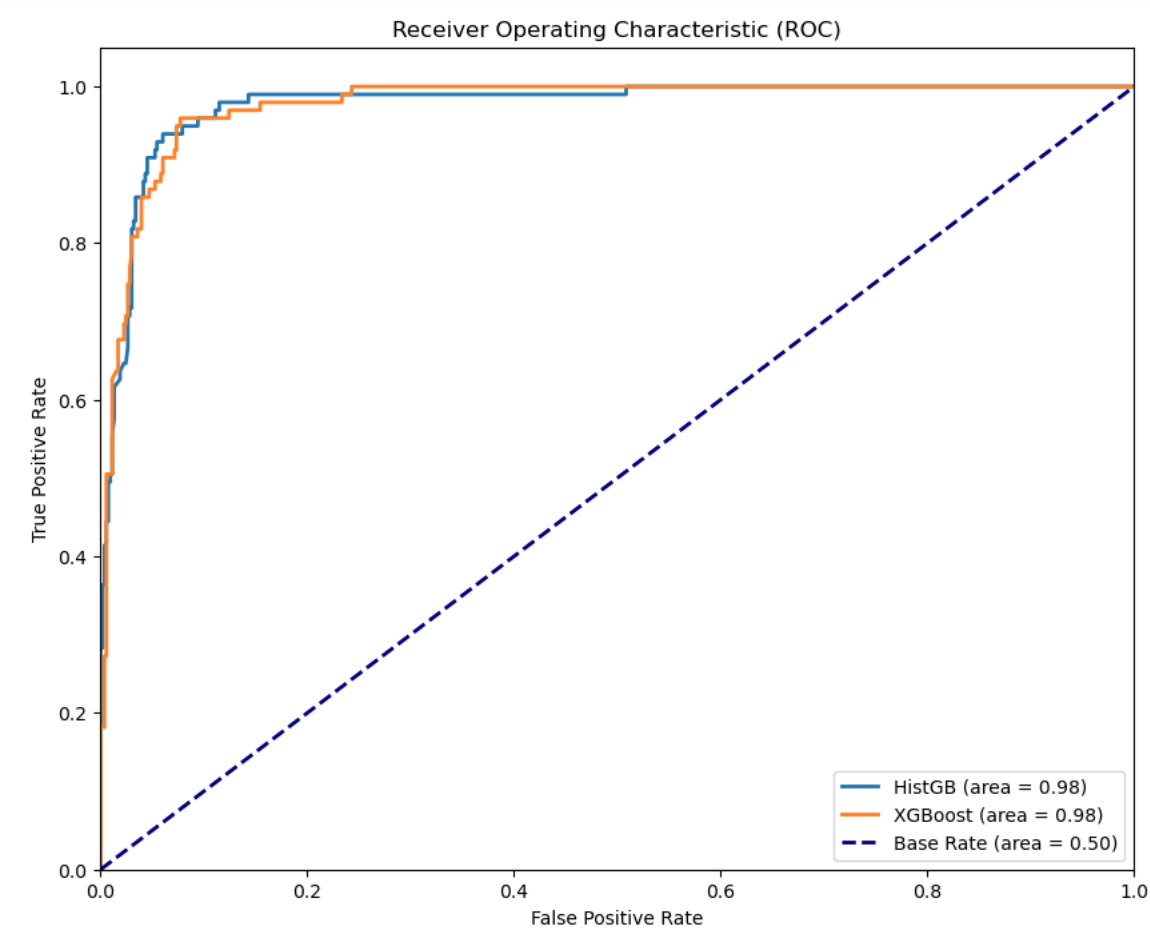


Figure 25

Diverging Bar Plot for Top Two Models

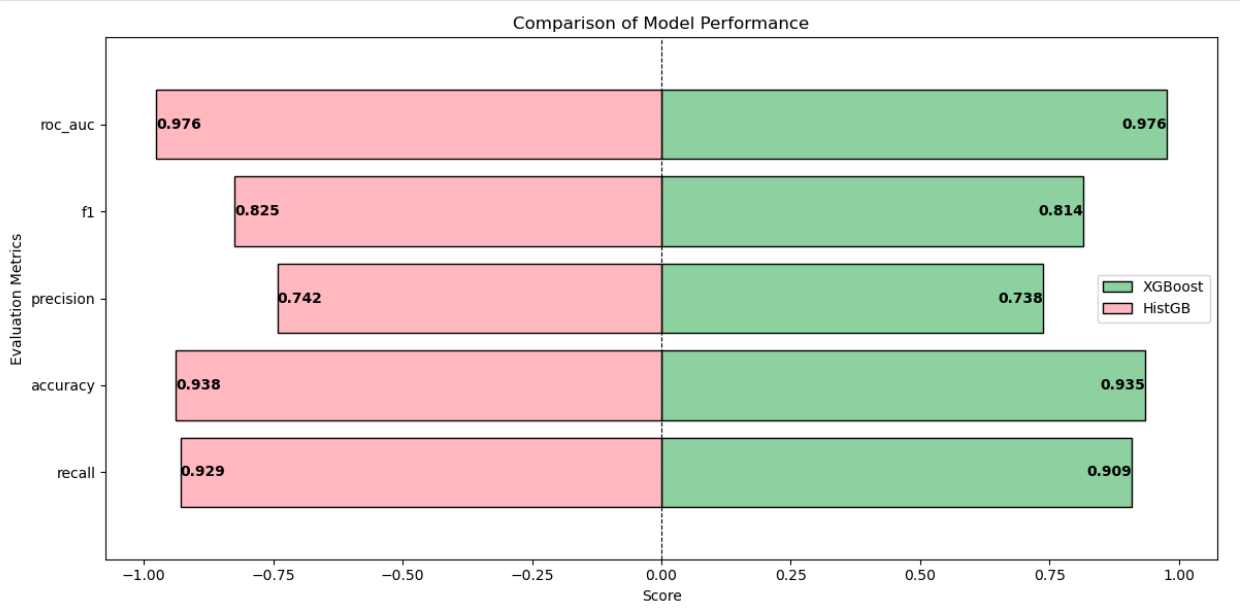
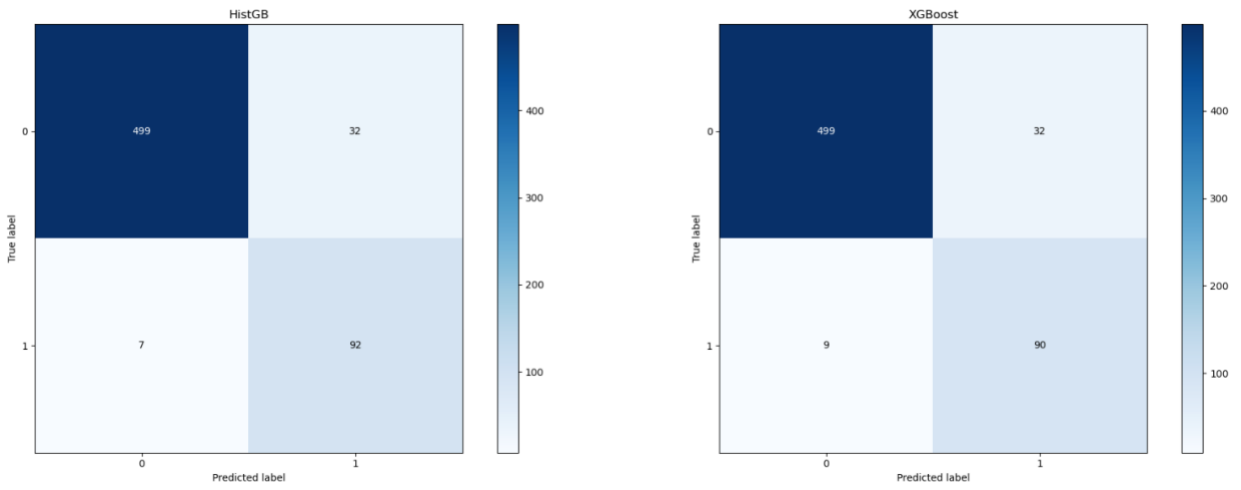


Figure 26

Confusion Matrices for Top Two Models

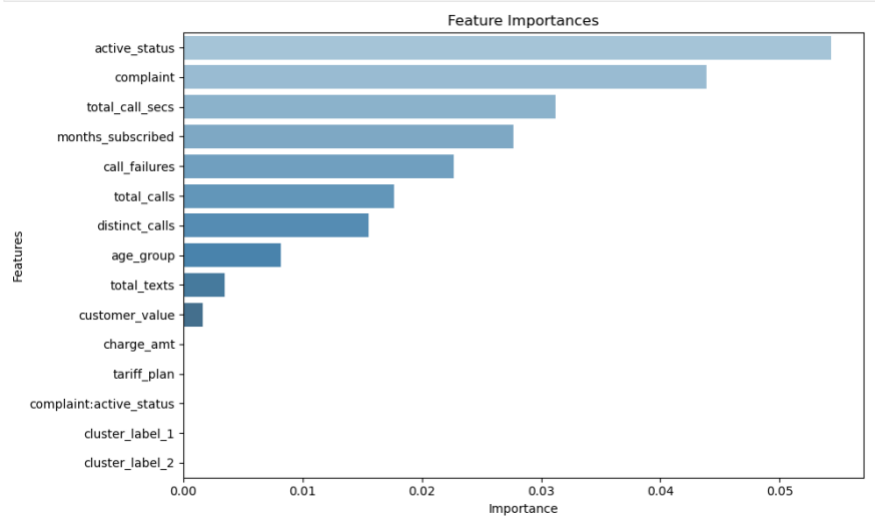


We also sought to examine the permutation feature importances for both models (see Figure 27). The XGBoost and Histogram-based Gradient Boosting classifiers identified many of the same features as important ones. The "complaint" and "active_status" features were in the top 3 important features for the two models. "Months_subscribed" was fourth in importance for both. In addition, the cluster-label-based features and the interaction term feature were determined to be the least important features. The permutation feature importances for the Histogram-based Gradient Boosting classifier show that the "customer_value" feature has negative importance. This means that randomly shuffling this feature resulted in an improvement in model performance, thereby suggesting that this feature may be noisy or redundant rather than being informative to the model.

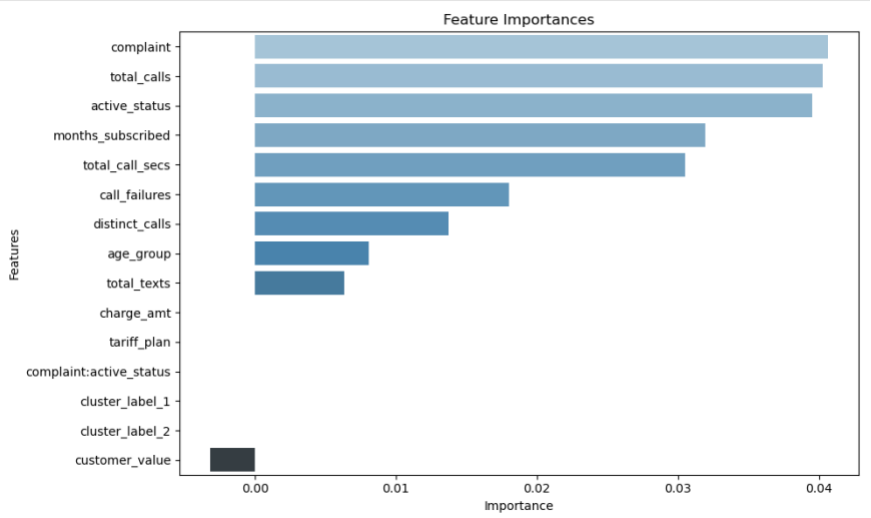
Figure 27

Plots of Permutation Feature Importance for Top Two Models

XGBoost



Histogram-based GB



4 Discussion

4.1 Statistical Analysis

The results of our statistical analysis initially suggested that all three binary predictors ("complaint", "tariff_plan", and "active_status") had significant associations with "churn". Then, our baseline logistic regression model showed that "tariff_plan" was not statistically significant in terms of predicting the log odds of churn and it was subsequently removed in our backward elimination procedure. Statistical significance does not necessarily equate to practical significance, however. A customer's tariff plan type may still be meaningful to analyze, but more data and research is required to make a robust conclusion.

The results of the Cochran-Armitage test and the baseline logistic regression model suggest that age group may not be a relevant factor for consideration. It was found not have no trend with respect to the dependent variable and a statistically insignificant relationship with the log odds of churn. Age, more generally, may not play a meaningful role in determining churn behavior among

customers. Future research should attempt to corroborate this conclusion and experiment with using age as a continuous variable rather than an ordinal categorical variable as was used here.

The models that were created in the Statistical Analysis section potentially suffered from two major issues. Each may have been significantly affected by the presence of outliers and multicollinearity. Outliers are a problem in logistic regression because they can influence the estimated probabilities produced by the model, skewing the results, and possibly leading to misinterpretation. They can also affect the convergence of the logistic regression algorithm, leading to distorted coefficients, inaccurate predictions, or failure to converge. Multicollinearity is an issue because it makes it difficult for the model to discern the effect of each predictor on the outcome probability. This can result in inaccurate coefficient and standard error estimates and potential for model misinterpretation. Our goal was to merely understand the relationships in the data as is, so outliers and multicollinearity were not specifically addressed in our work. However, the persistent and potentially impactful issues presented by outliers and multicollinearity suggest that addressing these two factors is warranted in future modeling attempts. Methods to handle these issues in the future may include robust regression, outlier removal or replacement (Winsorizing, imputation, etc.), data transformation, and regularization techniques such as Ridge or Lasso regression. In addition, the models displayed poor overall fit and may benefit from additional interaction terms, spline functions, or other specifications to describe the data more adequately.

The adjusted odds ratio for “complaint” with respect to “churn” was very high (point estimate = 19.95). Whether or not a customer makes a complaint appears to be a significant factor in determining whether they will churn. Service providers should identify the most common complaints that customers make, formulate action plans to address these issues wherever possible, and proactively implement their service-improvement strategies. By preemptively remedying

potential complaints, the organization may avoid unnecessary complaints and the satisfied customer may be less likely to churn. The adjusted odds ratio for “active_status” with respect to “churn” was low (point estimate = 0.23). Since, active customers are less likely to churn than inactive customers, service providers should identify ways to reduce customer inactivity and encourage customer activity. Doing so may result in less churn and greater customer retention.

4.2 Cluster Analysis

The results of the cluster analysis showed that the optimal number of clusters for the data was 2 or possibly 3. We decided to implement the selected clustering algorithms with 3 clusters to obtain a greater number of possible customer segments, more than a mere dichotomous separation. The resulting clusters for both algorithms we implemented showed that the data may have been grouped into clusters based on activity and usage levels. There appeared to be an ordinal-like pattern based on measures related to usage frequency. This makes sense since many of our features were continuous variables that indicated activity or usage level (total calls, total texts, etc.). Including other demographic information about customers (income level, education level, location, personality metrics, etc.) may have resulted in more useful information for cluster-based customer segmentation efforts. With the current data and clustering results, all we can conclude is that customers differ primarily based on activity and usage frequency. Perhaps retention strategies can be created to account for the user’s activity level. Ultimately, more research should be conducted with a richer set of customer-related variables. This will facilitate the discovery of more meaningful customer segmentation results.

4.3 Predictive Analysis

The results of the predictive analysis showed that the gradient boosting models performed the best on the classification task at hand. Specifically, the XGBoost classifier and Histogram-based Gradient Boosting classifier were the two top performing models. These classifiers obtained a high level of recall while maintaining a satisfactory degree of precision. As such, they may allow stakeholders to correctly identify a high percentage of customers who are at-risk of churning. Moreover, their predictions regarding who is likely to churn may be relatively reliable. The Histogram-based Gradient Boosting model performed the best out of all the models examined, and it narrowly outperformed the XGBoost model in the final test evaluation. It should be noted that the Histogram-based Gradient Boosting model is also optimized for faster processing due to its discretization methodology, making it the best-performing classification method in our study and a highly efficient one that can help to mitigate computational overhead. As such, stakeholders may wish to implement this model for their churn prediction applications. Future research should examine the performance of other classification methods for churn prediction, including neural networks and ensembling approaches (for example, a stacking classifier based on our top two models with a final meta-model applied to the predictions).

The permutation feature importance results showed that the cluster label and interaction term features were not useful features for the top two models. This may be because the features are redundant. The cluster analysis results yielded clusters that were based on measures of user activity and usage frequency. Plenty of other features in the model already provide this information. Therefore, a cluster label feature would likely not provide any new information that could be useful to the model. In addition, the gradient boosting tree-based models already compute interactions between features. Consequently, including an explicit two-way interaction term may

not offer any additional information that the algorithms cannot access already. The exclusion of these features in future modeling efforts appears to be warranted due to their potential for redundancy and low permutation-based importance. The “complaint” and “active_status” features were found to be important by each of the top two models. This supports our earlier conclusions from our statistical analysis, which found that the likelihood of churning was meaningfully and strongly related to these two variables. As stated previously, service providers may wish to address the most common customer complaints and galvanize customer activity to reduce churn, improve retention, and ultimately increase profitability.

5 Conclusion

In sum, we have used statistical methods, clustering techniques, and predictive modeling to complete a comprehensive churn analysis of the selected dataset. Our findings indicate that service providers should implement strategies that address or preempt customer complaints and encourage customer activity to effectively reduce churn and increase revenue. Our customer segmentation efforts were hindered by the lack of customer-relevant information, including demographics and other personal characteristics. Data of this nature should be collected and included in future segmentation analysis endeavors, especially for those that utilize clustering methods. The classification method that performed the best (highest recall while maintaining strong performance on other metrics and adequate performance on precision) was the Histogram-based Gradient Boosting model. Although it only narrowly outperformed the XGBoost model, the Histogram-based Gradient Boosting model is renowned for its computational efficiency. Hence, stakeholders may wish to implement this model in their applications due to its slight performance edge and its efficient use of computational resources.

6 References

Ahn, J.-H., Han, S.-P., & Lee, Y.-S. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. Telecommunications Policy, 30(10-11), 552-568. https://doi.org/10.1016/j.telpol.2006.09.006

Almana, A. M., Aksoy, M. S., & Alzahran, R. (2014). A Survey On Data Mining Techniques In Customer Churn Analysis For Telecom Industry. International Journal of Engineering Research and Applications, 4(5), 165-171.

Aziz, N., Akhir, E. A. P., Aziz, I. A., Jaafar, J., Hasan, M. H., & Abas, A. N. C. (2020). A Study on Gradient Boosting Algorithms for Development of AI Monitoring and Prediction Systems. In 2020 International Conference on Computational Intelligence (ICCI). IEEE. https://doi.org/10.1109/ICCI51257.2020.9247843

Celik, O., & Osmanoglu, U. O. (2019). Comparing to Techniques Used in Customer Churn Analysis. Journal of Multidisciplinary Developments, 4(1), 30-38.

Keramati, A., & Ardabili, S. M. S. (2011). Churn analysis for an Iranian mobile operator. Telecommunications Policy, 35(4), 344-356. https://doi.org/10.1016/j.telpol.2011.02.009

Korns, M. F. (2018). Genetic Programming Symbolic Classification: A Study. In Genetic Programming Theory and Practice XV (pp. 39-54). Springer. https://doi.org/10.1007/978-3-319-90512-9_3

Nhat-Duc, H., & Van-Duc, T. (2023). Comparison of histogram-based gradient boosting classification machine, random Forest, and deep convolutional neural network for pavement raveling severity classification. Automation in Construction, 148, 104767. https://doi.org/10.1016/j.autcon.2023.104767

Sarker, I. H., Kayes, A. S. M., & Watters, P. (2019). Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. Journal of Big Data, 6, Article 57.