**Credit Card Customers**

**Data Mining Written Report (Group 15)**

Blake Murray, Julia Pysk, Jillian Kozlowski, Isaiah Moore

**1. Executive Summary**

This report contains the results of analysis performed on credit card customers, and the final model used to determine the optimal cutoff for which customers should be offered incentives to stay with Wells Fargo.

Wells Fargo management is concerned with the number of customers leaving their services, but they are unsure of which customers should be offered retention incentives. This is leading to more customers leaving the company, and ultimately an inefficient decision boundary deciding which customers are being offered incentives. With our available data, we aim to provide insight as to the specific traits that identify a customer likely to churn and provide a model which can calculate the likelihood a given customer will leave Wells Fargo.

Our data set contains over 10,000 customers and for our analysis we used attrition-flag as the target variable, as it indicates whether a customer is still with the company’s credit card services. Upon performing exploratory analysis, we noticed that a few features seemed to have an outsized impact on the classification of a customer. Lower utilization ratio, revolving balance, and change in spending from quarter 1 to 4 were all indicative of a customer who was more likely to churn. Though our final model included these factors, we feel they are still important to consider independently for influencing customer behavior to increase retention.

To determine which classification model was able to best predict the classification of each instance, we tuned the hyperparameter of several different models and calculated the resulting AUC. After cross-validation of classification models, the adaboost model resulted in the best performance. To properly weigh the cost of offering incentives with the value of retention, we applied a cost matrix to the adaboost model and plotted the resulting cost curve. The matrix we used in creating the curve associated a cost of $100 with the incentive, and a net value of $300 for retaining a customer who would have otherwise left.

After analyzing the cost curve, we determined that profits could be optimized by offering an incentive to any customer with a 9% chance of churning or greater. The biggest limitation of our analysis is the assumptions built into the cost matrix. One of the most valuable next steps would be to determine the true cost to retain a customer and the value of that customer, resulting in a decision boundary more specific to Wells Fargo.

**2. Problem Description**

As managers are concerned with the number of customer’s leaving the credit cards services, Wells Fargo needs to determine the likelihood a given customer will leave their services. We aim to identify the customers most likely to leave, allowing for better decision-making regarding incentive targeting.

2.1 **Background**

Wells Fargo generated over $4 billion in revenue from its credit card services in 2020. However, management is concerned with the company’s ability to remain competitive in an industry filled with so many large competitors. Though Wells Fargo has offered retention incentives in the past, they are concerned that their previous methods of determining who receives these incentives are not accurate, and a better method of determining who should receive incentives would lead to less churn and more profit.

2.2 **Business Goal and Data Mining Goal**

Our business goal is to identify which customers will leave Wells Fargo. Ultimately, we want to discover which customers should be offered incentives to increase customer loyalty and make more efficient use of incentives. Our data mining goal is to create a model that will accurately predict whether a customer will churn. This prediction target is attrition\_flag, which indicates whether a customer still belongs to Wells Fargo’s credit card services. Additionally, we will utilize clustering techniques to compare the characteristics of customers who churn with that of customers who stay with Wells Fargo. Though this will not directly impact our model, it may provide insight to managers for how to optimize the nature of incentives or provide more value to customers. To effectively target incentives, we will compare the cost of retaining a customer with that customer’s value to optimize the cutoff for who should be offered incentive. The incentive used in our analysis is a statement credit of $100 if the customer spends over $1000 in the past 3 months. Not only do we think this will make a customer stay, but we also believe it will incentivize them to use their Wells Fargo credit card over a competitor.

**3. Data Description**

3.1 **Data**

The dataset used in this report includes information about current credit card users as well as past customers who have since left the company. This dataset consists of 10,127 rows of data. There are 21 features used, 6 being categoric and 14 being numeric. The categoric features primarily describe the customer, while the numeric features provide more insight on how they use their card and their relationship with the credit card company. Features include: clientnum, attrition\_flag,customer\_age, gender, dependent\_count, education\_level, marital\_status, income\_category, months\_on\_book, total\_relationship\_count, months\_inactive\_12\_mon, contacts\_count\_12\_mon, credit\_limit, total\_revolving\_bal, avg\_open\_to\_buy, total\_amt\_chng\_q4\_q1, total\_trans\_amt, total\_trans\_ct, total\_ct\_change\_q4\_q1, avg\_utilization\_ratio. The target variable for our analysis is attrition\_flag, which showcases if a customer still has an account with the credit card company (Existing Customer) or if they are no longer with the company (Attrited Customer.)

Link: <https://www.kaggle.com/sakshigoyal7/credit-card-customers>

3.2 **Exploratory Analysis**

Our team gained valuable insights from our dataset that contributed to our data mining goal. We utilized multiple methods of analysis to ensure we obtained as much value from the data set as possible. We first used clustering to separate the data into segments to identify similarities and differences across the customer base. Clustering provided a unique perspective by being a form of unsupervised learning. Next, our team generated feature importance based on logistic regression equations where we could see the contribution to the target variable through the magnitude of the coefficients. Lastly, we created graphs of meaningful variables to analyze the distributions to see if they affect the target variable.

Clustering

Our clustering process started with prepping the data and ensuring that no target variable was selected to maintain the principles of unsupervised learning. The method we chose to use was k-means clustering. After determining the optimal number of clusters through the elbow method was 3, the group decided to instead use only 2 clusters since this provided an obvious distinction between which class a customer was likely to fall into. We have included a plot under Appendix Item 1 which illustrates how in some cases the data more naturally falls into 3 classes, but since 2 of these classes had similar amounts of churn, this distinction was less beneficial to our analysis. Information about each segment came from comparing the cluster means to the overall data means. If the values differed, this was an indicator that the feature is important in determining the class of a customer.

Segment 1 consisted of a larger portion of customers that were lost to attrition, meaning they are no longer using the company’s credit card services. Features that were different from the data means included average utilization ratio and credit limit. This segment had a lower utilization ratio indicated by a cluster mean of 0.18 compared to the data mean of 0.27. A lower utilization ratio implies that the segment doesn’t use much of their credit limit. This cluster mean of credit limit is 0.35 which is higher than the data mean of 0.21. Though these features are related, they are both important to consider for what make this segment different.

Segment 2 contained primarily existing customers, which means they are less likely to churn and will continue to use the services. Features that contributed to the identity of this segment include included the inverse from segment 1. Having substantially having higher utilization ratio and lower credit limits. Another interesting feature included gender which is showing that this segment may be composed of more females. This implies there may be more females that continue with the services while men were more likely to churn.

Logistic Regression Importance

To better understand the impact different features had on the likelihood a customer would stay with Wells Fargo, we performed a logistic regression. After prepping and transforming the data, we built a logistic regression model that provides an equation to identify the coefficients of each feature. The coefficient can be insightful for several reasons. First, whether the value is positive or negative shows how a change in that feature will affect the probability of the attrition flag to be the positive class, indicating an existing customer. Secondly, the higher the value, the more influence it will have on the classification of that customer. We have included the results of the most significant features from our logistic regression under Appendix Item 2.

The generated coefficients indicated both positive and negative features that need to be considered. Two features with large positive effect on the resulting probability are total transaction count and quarter 1 to quarter 4 change. The coefficient for total transaction count is 15.07. This indicates that the more transactions a customer has in a year, the more likely they are to stay with Wells Fargo. To take advantage of this, Wells Fargo could consider finding ways to encourage customers to reward customer for transaction volume over other factors. Quarterly change has a coefficient the next largest coefficient with a value 10.21. This feature identifies that the customer has increased use of the credit services as the year progresses. Going forward, Wells Fargo could consider more proactive incentive offers towards customers that have decreased their card usage in recent months.

The most significant negative features in our analysis include transaction amounts and contacts to the bank. The coefficient for transaction amount is -8.65 and for contacts to bank, it is -2.88. Customers who have a large dollar value of transactions are more likely to exit the company’s services, and we believe that performing future analysis on the reasoning behind this could prove very insightful to Wells Fargo. Additionally, providing more automated tech support could lead to customer’s having to contact the bank less frequently, leaving them more satisfied with the bank’s services.

*Total Revolving Balance Distribution*

Chart, histogram

Description automatically generatedDistribution Comparisons

A feature that provided good insight regarding our

target variable is Total Revolving Balance. The revolving balance is a variable that shows how much credit the customer is carrying over from one month to the next. From the histogram we can see that existing customers have a higher density at higher revolving balance values. This is a feature that can be used to help separate the classes. The graphic to the right shows these distributions.

3.3 **Data Pre-processing**

Though our dataset from Kaggle was mostly prepared for analysis, we made sure to carefully analyze the features to ensure they were all relevant to our data mining goal. We also made sure that Rattle correctly identified the data type of all our features, as incorrectly identifying a data type could drastically reduce the effectiveness of our models. We transformed all categorical variables into indicators, as some models were unable to function properly if we left them as categoric. The final, and arguably most important, step performed in our analysis was ensuring that all numeric variables were rescaled from 0 to 1. This ensures that the models to not unjustly identify variables with larger ranges as more important than other variables.

**4. Data Mining Solution**

To determine which model was most effective for our specific problem, we tuned a number

a variety of models, and then tested the AUC of each model to decide which was the best predictor

for our data set. To use this model in decision making, we applied a cost matrix to the

model with the best AUC.

4.1 **Models**

Our task of predicting customer churn is a classification problem. Our team ultimately wants to find a model with the best performance to predict the class of the target variable for each customer. When searching for the best models we narrowed down our choices to the ones that deal with classification problems. These include random forests, decision trees, artificial neural networks, support vector machines, logistical regression, and adaboost. Each model was made by tuning all the relevant hyperparameters. Once the model had the optimal parameters based on the validation set, we then moved on the comparing the final model for each technique.

4.2 **Performance Evaluation**

The area under the curve is a metric used to measure the ability of a classifier to distinguish between classes. A higher AUC is indicative of better performance. We calculated the AUC for each model based on the results of the model on the test set, using the parameters we determined based on the validation set. Overall, the adaboost model had the best performance based on an AUC of 0.9897. The table below shows the cross-validation matrix of the classification models.

|  |  |
| --- | --- |
| **Models** | **Cross Validation – Test Set AUC** |
| Random Forest | 0.9890 |
| Decision Tree | 0.9455 |
| ANN | 0.9666 |
| SVM Radial | 0.9576 |
| SVM Polynomial | 0.9295 |
| Logistic Regression | 0.9362 |
| Adaboost | 0.9897 |

4.3 **Cost Matrix**

After finding the best model, we felt that the optimal way to use it in decision making was to apply a cost matrix which weighs the cost associated with the incentive with the additional revenue gained from retaining a customer. The cost of the incentive is a $100 statement credit when that customer spends $1,000 in the next 3 months. We calculated the yearly revenue for each customer to be $400. Half of this revenue came from taking the average revolving balance of the existing customers ($1,250) and multiplying it by an interest rate of 16%. The 2020 Wells Fargo financial statements say that in addition to interest revenue, they generate roughly $1 in services fees for each dollar of interest revenue, so the total revenue for each customer is roughly $400. Based on these assumptions, the value of keeping a customer who would have otherwise left is $300 after the incentive expense, and the cost of offering an incentive to a customer who would not have left is $100.

Chart, line chart

Description automatically generated

Likelihood a Customer Leaves Wells Fargo

$300

-$100

$0

$0

After applying the cost matrix to the adaboost model, we obtained the above cost curve. Using this curve, we determined that even if all customers are offered the incentive, net revenues will still increase for Wells Fargo. This highlights how valuable customers are, and why management is justified in their churn with the amount of customer turnover. However, the return from the incentives can be optimized by offering incentives only to customers with a 9% chance of leaving the credit card services or greater. If there are limitations on the amount of funds allocated to this project offering incentives to any cutoff point above this 9% will still offer a positive return.

**5. Conclusion**

5.1 **Recommendations**

One of the major takeaways from our initial exploration was a few features seemed to have an outsized impact on the likelihood a customer would leave the credit card services. These features were low revolving balances, utilization ratio, and transaction volume. Finding ways to encourage customers to spend more frequently and use a larger portion of their credit balance could lead to less churn going forward. By using the adaboost model in combination with the cost matrix, we calculated that net revenue can by 10% after offering a $100 incentive to any customer having a greater than 9% change of leaving. Based on Wells Fargo’s 2020 earnings this could lead to an additional $450 million.

5.2 **Limitations and Future Work**

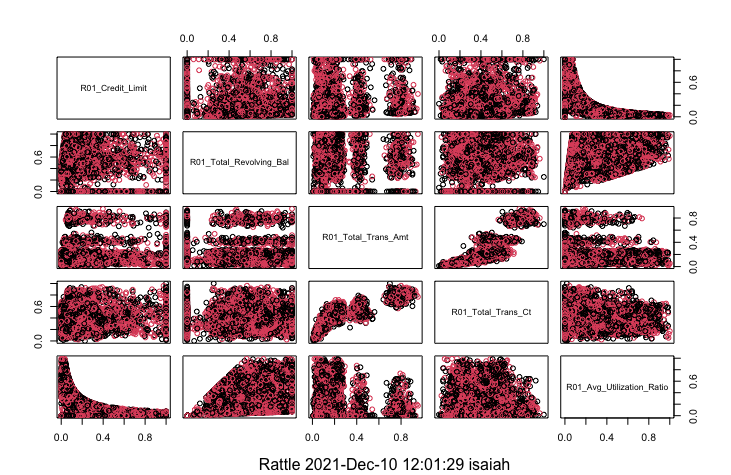
One of the largest limitations of our model is that we had to build a lot of assumptions into our cost matrix based on past financial statements. By consulting with managers in the credit card department to obtain more accurate data on the value of a customer and the cost of retaining them, we could build a model which provides a more accurate decision boundary based around Wells Fargo’s specific business model.

Additionally, we feel that adding more features could allow for a more accurate model. One feature we believe would be helpful is a count of how many positive and negative contacts a customer had with the bank. When performing exploratory analysis, we noticed that, in some cases, many contacts to the bank made a customer more likely to leave. However, there seemed to be some cases where large number of contacts did not indicate a customer was likely to leave. If the contacts were instead to be broken up into the reason behind the contact, it might be more informative for the model. Another feature we thought would be valuable is how many other credit cards a customer has. We hypothesize that if a customer has more cards, they may be more likely to leave Wells Fargo. Though we would need the data to decide if this is true, we think it could potentially make the model even more informative. Other features we believe could be informative include number of total accounts with Wells Fargo, number of authorized users on the account, and whether any recurring charges are currently set up on the card. We believe that all these features will allow the model to distinguish more efficiently between who should be offered an incentive, leading to a more efficient used of capital.

In the future, we recommend that data for the model is collected and updated quarterly, so that the data remains relevant. When Wells Fargo is looking to send out promotions or retention incentives, the marketing team should insert customer data into our model and offer incentives to any customer that has a 9% chance of leaving. If at any point the company has a limitation on the funds allocated to incentives, using any cutoff above 9% will still result in a positive return for the company.

**Appendix**

**Item 1:**

****

**Item 2:**

**A screenshot of a computer

Description automatically generated with low confidence**