Term Project Report – Executive Summary

Ruby Link, Ellie Strande, Jonathan V

Problem Description and Motivation

Credit card attrition, also known as customer churn, refers to the phenomenon of customers discontinuing their relationship with a business or organization. In the context of the credit industry, credit attrition occurs when customers close their credit accounts, stop using credit cards, or switch to competitors' services. The ability to accurately predict credit attrition is crucial for businesses in the financial sector to proactively identify customers who are at high risk of attrition and take appropriate measures to retain them. When customers attrite, it leads to revenue loss and negatively impacts profitability. Moreover, acquiring new customers is typically more expensive than retaining existing ones. By accurately identifying customers who are at high risk of attrition, businesses can proactively implement targeted strategies to reduce churn and retain valuable customers. These strategies may include personal offers, loyalty programs, improved customer service, or other retention initiatives.

A predictive model for credit attrition can provide several benefits. Firstly, it allows businesses to focus their resources and efforts on customers who are most likely to close their accounts, enabling them to allocate their retention budget more efficiently. Secondly, it helps businesses gain insights into the factors that contribute to attrition, such as changes in spending patterns, payment behavior, or external factors like economic conditions. This knowledge can be leveraged to optimize customer experience, develop better products, and enhance customer satisfaction. Overall, the development of a reliable and accurate predictive model for credit attrition is vital for businesses to mitigate revenue loss, enhance customer retention, and optimize resource allocation. By targeting customers at risk of attrition, businesses can effectively implement retention strategies and maintain a competitive edge in the credit industry.

Data Insights

The customer attrition dataset contains a variety of features that provide valuable insights into customer behavior and characteristics. The dataset includes information such as customer age, representing the age of the customer, and months on book, indicating the length of the customer's relationship with the bank. Additionally, it includes the period of the relationship with the bank, giving a measure of how long the customer has been associated with the bank. The dataset also captures information about customer inactivity, including the number of months inactive within the past 12 months and the total revolving balance. Furthermore, the dataset provides details on customer transaction behavior, such as the total transaction count and the total transaction count in the last 12 months. These features give an understanding of the customer's engagement with the bank's services. Other demographic features like gender and income category are also included, providing additional context about the customer profile.

The target variable of the dataset is attrition flag, a pivotal feature that plays a vital role in training predictive models to effectively identify customers at risk of attrition. The attrition flag provides critical information regarding the customer's status, enabling a clear distinction between those who are existing customers and those who have attrited or left the bank. However, it is important to recognize the inherent nature of the attribution flag's distribution within the dataset. The distribution of the attrition flag variable exhibits noticeable skewness, characterized by a significantly larger number of observations associated with non-attrited customers compared to attrited customers with the data containing 7575 cases and only 1440 positives. This class imbalance presents challenges in developing accurate predictive models, as it can lead to biased predictions and a disproportionate focus on the majority class. Consequently, the predictive modes may be more inclined to predict non-attrited instances while neglecting the identification of attrited customers, which is crucial for effective retention strategies. Addressing the class imbalance is paramount in the modeling process to ensure the model's effectiveness in both existing and attrited customers with equal attention and precision.

Our analysis began with a comprehensive data cleaning process, where we diligently removed 14 unnecessary variables from the data frame, ensuring a streamlined dataset for further analysis. Additionally, we transformed gender, education level, marital status, and income category into factor variables, allowing for effective modeling. To facilitate our predictive models, we converted the target variable, attrition flag, into a binary format consisting of 0s and 1s. Finally, we conducted feature engineering on the credit limit variable, which initially exhibited a highly skewed distribution. By applying log transformation to the credit limit, we successfully mitigated the impact of extreme values or outliers, resulting in a significantly improved distribution. This data cleaning process concluded with the achievement of an optimized dataset ready for subsequent analysis in the three predictive models we chose to implement.

Logistic Regression Results and Analysis

For the logistic regression model's results, we used the summary function to see the magnitude of the coefficients for each predictor and to see which were statistically significant at an alpha of 0.1. Gender, months inactive, total revolving balance, and total transaction count had the highest level of statistical significance, joined by credit limit and income but with smaller significance. It is important to point out that while not every range of income had a p-value above the 0.1 threshold, the majority of them did, so we ultimately decided to keep income in the model; enough of the variable was statistically significant and overall, it had a notable amount of predictive power. Banks should look at these predictors to assess whether someone is likely to churn, and from there, banks can take proactive steps to retain customers. The most influential variable with the largest coefficient was total transaction count. To interpret the coefficient, we exponentiated the results and found that for every additional transaction, a customer is 5.85% less likely to churn. Again, this information is useful to banks because they can flag customers

with low transaction counts and take preventative measures to deter them from withdrawing by providing additional incentives or fixing current problems.

To assess the model's overall performance, we calculated the accuracy, sensitivity, and specificity. Initially, we picked a prediction cutoff of 0.5 – if the model predicted a score above 0.5, it was classified as a positive case (attrition), or otherwise a negative case (no attrition). For the training data, the model scored 87.2% for accuracy, 37.2% for sensitivity, and 96.2% percent for specificity. The test scored 87.8% accurate, 39.2% for sensitivity, and 96.6% for specificity. Overall the accuracy and the false positive rate are very strong, but as discussed in the "Data Insights" section above, there is a large class imbalance in the dataset which is causing the model to be unsuccessful in predicting positive causes, also leading to some overfitting. To remedy this, we printed out a ROC curve to find a better prediction threshold. We found that the best threshold was 0.2 and with the new cutoff, we had much stronger results. The accuracy for the training and test predictions were 79.5% and 79.%, respectively, 72.3% and 75.6% for sensitivity, and finally, 80.9% and 79.7% for specificity. We can see that now there is almost no overfitting in the model, and while the accuracy and true negative rate decreased, the true positive rate increased and has nearly the same success rate as the other two metrics.

Decision Tree Results and Analysis

For the decision tree we used the same predictors as the linear regression model. The tree that was outputted was very large, with originally 14 terminal nodes. The path that led a customer to be flagged for attrition was if their total transaction count is less than 54.5, their total revolving balance is less than $656.5, and they were inactive for more than 1.5 months over a 12-month period. This information is very useful in helping banks predict if someone will churn. It gives banks a detailed path of what they should look for when identifying customers who may be at risk of leaving the bank with specified thresholds and turnover points. However, because the tree is so large and nearly all of the paths, pruning the tree would provide a more straightforward path for banks to look at while still outputting confident results.

To determine the ideal size of the tree, we printed out a graph comparing error to the tree size and determined that a tree of size 4 would have the same success rate as the previous model with 14 leaves. We pruned our tree to have 4 terminal nodes and used the model to predict results for our test data. We found that for both the training and test data, accuracy was about 90%, sensitivity was around 50%, and specificity sat at 97%. Compared to the previous logistic regression model, the decision tree is less effective at predicting positive cases due to the significant class imbalance in the dataset, explaining why we see such a low sensitivity score.

While pruning the tree did not affect the model's performance, banks should use the smaller tree to analyze potential customer churn if used in a real-world scenario. It has the same effectiveness as the larger tree but is much more straightforward to interpret for someone who may not be well-versed in data science or statistical analysis.

Random Forest Results and Analysis

Unsatisfied with the result of the previous models, we utilized a Random Forest model to improve the results. A random forest model is a model that takes in multiple decision trees and captures the individual trees' unique differences. Taking in slightly different versions of the training data for each tree enables the model to reduce overfitting. The model isn't exposed to the same data for each iteration, and it can better generalize the data trends. With the random forest model, the train and test data results were both around 90% accuracy, sensitivity increased to approximately 75%, and specificity was near 92%. The random forest model drastically improved the sensitivity compared to the decision tree and logistic regression results, which were as low as 37%. For the other two models, accuracy and specificity were continually high. Still, sensitivity was low, and with the logistic regression model, we had to sacrifice some of the false positive rate to bring up the true positive rate, leaving every score at around 75%. However, the random forest model retains over 90% accuracy and specificity, with a 75% sensitivity score. This indicates that our random forest model can now effectively predict if an individual will close their credit card accounts 75% of the time.

To analyze the model further, when utilizing the distribution of minimal depth graph from the Random Forest model, we can observe that predictors Total Transaction Counts, Total Revolving Balance, Months Inactive, and Credit Limit are the most important in predicting whether or not a customer will churn. These variables were consistently put at the trees' root, which implies that they make the largest overall improvement compared to the predictors placed below the root. These results are consistent with our decision tree model and logistic regression in determining which predictors have the most considerable impact when assessing attrition.

Conclusion

The Random Forest model is the most successful and practical of the three and in a business setting, should be the model that is utilized for bank churn rates. The model increased the accuracy of true positive cases – individuals who were predicted to churn and actually did. The sensitivity increased from approximately 50% in the decision tree model to around 75%, as observed in our random forest model. Additionally, when comparing the random forest model to the logistic regression, the overall accuracy increased from 80% to 90%, which is a 10% jump, and immensely improved both the sensitivity and specificity rates.

Our data analysis has effectively demonstrated what predictors are important when assessing churn and how banks should use the information from each of the three models to mitigate high attrition rates. As seen in all models, the total transaction count is the highest indicator of credit card attrition and should be closely monitored by banks. In the decision tree model, we were able to observe that total transaction counts under 55 put a customer most at risk for credit card attrition. Other important variables include total revolving balance, months inactive in the last 12 months, and credit limit. Banks can use the discussed findings to target

uneasy customers and provide incentives like credit card rewards or special deals to deter churn and increase or continue a bank's success.