Data Mining Final Report: Customer Churn and Associated Characteristics

Melissa L. Russell, Jonathan I. Voth, Nicholas J. Wawee

Bowling Green State University

Abstract

Credit card services would not survive without efforts to prevent customer churn. However, not all customers need the same level of effort regarding trying to get them to stay a customer. By identifying which customers will churn, credit card services can focus all efforts on those customers specifically. To identify which customers will churn, we need to build a predictive model. This current study aimed come up with a predictive model to classify which credit card customers of a bank will churn or not, with a secondary goal to evaluate the statistical significance of each feature. The dataset examined, *Credit Card Customers,* was found on the website Kaggle.com (Goyal, 2020), but unfortunately, the data collection process was not described. Our process included data cleaning, exploratory data analysis, implementing various cross-validated models (KNN, Decision Trees, Ensemble methods, etc.), evaluating each model by sensitivity, and finally inferring feature importance. Our main findings are that the Random Forest and ADABoost models (with downsampling) both were very successful models in predicting credit card customer churn with sensitivities of 0.94 and 0.95, respectively. We also found that the most important features in predicting customer churn is the number and amount of transactions a customer has. By uncovering this information, we have a better idea of which customers will churn and which will not, as well as the importance of each feature with respect to attrition. This can be used in the future to enable communication with customers predicted to churn in efforts to keep the customers.

Introduction

One of the most important metrics in the business world is customer churn. This is a measure of how many customers stopped buying a company's products during a certain timeframe. This is crucial to measure and understand, because by learning the characteristics of customers that have churned, companies can better predict who will churn in the future and how to prevent this, which leads to massive increases in profit. To do this, we first need to look at what past research on this topic has uncovered.

In 2009, Chih-Fong Tsai and Yu-Hsin Lu published an article that outlines their customer churn prediction model by using hybrid neural networks (Tsai & Lu, 2009). They note that neural networks are known for being able to predict customer churn well, but also that hybrid data mining techniques (combining two or more techniques) often perform better than single techniques alone. As a result, the authors used a mix of hybrid neural networks to predict customer churn. The neural networks techniques used were back-propagation artificial neural networks (ANN) and self-organizing maps (SOM). The hybrid aspects include using ANN combined with ANN and then using SOM combined with ANN. The results of these models showed that the hybrid models performed better than the single neural network baseline model in respect to errors (Types I and II) and prediction accuracy. The models were tested using the general testing set and two fuzzy testing sets that were based on the data that was filtered out by the ANN / SOM technique. Finally, the authors showed that the ANN + ANN hybrid model significantly outperformed both the SOM + ANN hybrid model and the ANN baseline model (Tsai & Lu, 2009). In regard to our project, this past research shows us the advantages of using hybrid data mining techniques over just a single technique as well as seeing how powerful neural networks can be in predicting customer churn.

In 2011, Chiun-Sin Lin, Gwo-Hshiung Tzeng, and Yang-Chieh Chin published an article that explained their attempts to predict customer churn in credit card accounts through a combined rough set theory and flow network graph (Lin, Tzeng, & Chin, 2011). Unlike the previous article discussed, this study uses rule-based decision-making techniques to extract rules related to customer churn, utilizes a path-dependent approach to then infer decision rules and variables, and then lastly draws conclusions about the relationships between rules and the different kinds of churn. According to this article, these methods can fully predict customer churn. The results of their flow network graph show that the top five characteristics of credit card customers that have voluntarily churned are "(1) does not attempt automatic debit transfers (1,033 supports); (2) number of annual pay-off times is under 1.84 times (1,033 supports); (3) marital status is single (900 supports); (4) annual average purchase amount is zero (900 supports); and (5) customers' age is between 30 and 39 (884 supports)," (Lin, Tzeng, & Chin, 2011). When testing this with validation data, the authors found the hit rates were 88.7% for the voluntary churn class of customers. Regarding our analyses, this shows us that rule-based decision-making techniques and flow network graphs can predict credit card customer churn with relative success. It also sheds light on what features may be important in predicting churn, such as our marital status and age variables.

In this project, the dataset we will be using consists of roughly 10,000 customers of a bank, and it was retrieved from Kaggle.com (Goyal, 2020). The dataset includes characteristics of each customer, such as age, gender, number of dependents, marital status, income, if they are an existing customer or have churned, etc. A description of the variables used in this project is shown below in **Table 1**. Unfortunately, Kaggle does not discuss how the dataset was collected, so that is unknown to us. The focus of the project involves challenges in customer churn regarding credit card services at the bank. The primary goal of this project is to classify which customers will churn and which will not. This predictive model will be used on future customers to enable communication with the ones who are predicted to churn in efforts to keep them. A secondary goal of this project is to evaluate the importance of each feature with respect to attrition. Based off existing literature, we predict that utilizing Random Forest, neural networks, and ADABoost will all yield successful predicting models for customer churn. We also predict that age, marital status, and income category will be the most significant features in our final model.

Table 1: Description of Variables

Name	Data Type	Type	Description
Attrition_Flag	Categorical (2, 16% Flagged)	Response	Did the customer leave?
Customer_Age	Integer	Feature	Age in years
Gender	Categorical (2)	Feature	Male or Female
Dependent_Count	Integer	Feature	How many dependents the customer has
Education_Level	Categorical (4)	Feature	How educated the customer is
Marital_Status	Categorical (3)	Feature	Marital Status
Income_Category	Categorical (4)	Feature	Income Bracket
Card_Category	Categorical (4)	Feature	What kind of card the customer posses
Months_on_book	Integer	Feature	How long the customer has had a relationship with the bank
Total_Relationship_Count	Integer	Feature	Total number of products a customer has
Months_Inactive_12_mon	Integer	Feature	Number of inactive months within the last 12
Contacts_Count_12_mon	Integer	Feature	How frequent the customer contacted the bank in the last 12 months
Credit_Limit	Integer	Feature	Credit limit in $
Total_Revolving_Bal	Integer	Feature	Total balance
Avg_Open_To_Buy	Integer	Feature	Limit - Balance
Total_Amt_Chng_Q4_Q1	Float	Feature	Change in transaction amount in Q4/Q1
Total_Trans_Amt	Integer	Feature	Total transaction amount
Total_Ct_Chng_Q4_Q1	Float	Feature	Change in transaction count in Q4/Q1
Avg_Utilization_Ratio	Float	Feature	Balance / Limit

Methods

Cleaning: The data was pre-processed before any analysis took place. There were several predictor variables that contained unknown values. For the purposes of this analysis, those were not included due to the large amount of data in the set. Additionally, the card category variable was removed. This variable was heavily skewed with over 93% of customers in the “Blue” card category. After cleaning, there were a total of 18 predictor variables, some of which are displayed in **Figure 1**. Customer age is relatively normally distributed. Credit limit and average utilization ratio are skewed right. Gender was split evenly, and most customers were married. The response variable, customer attrition, comprised of 15.7% customers that discontinued their services.

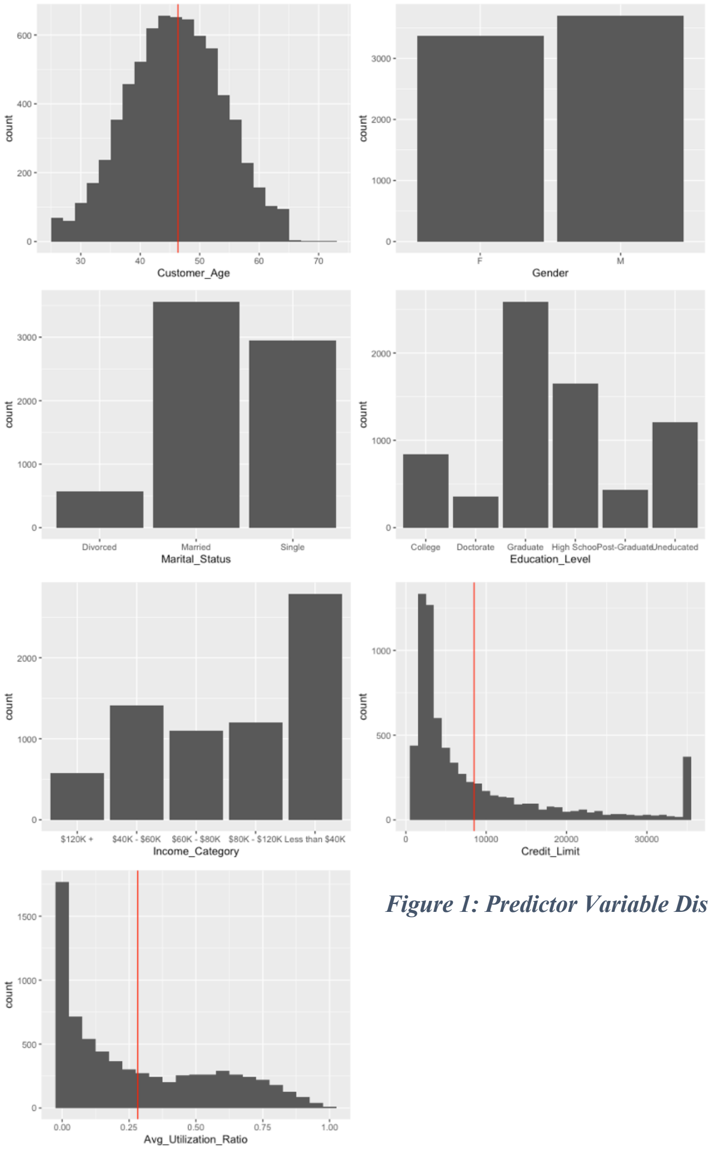

Figure 1: Predictor Variable Distributions

Model Fitting: Multiple classification models were fit to the customer data, predicting whether a customer is likely to churn or not. The data was partitioned into training and test samples. Of which, the training data was used to build the model. Because the response variable was not evenly distributed, traditional methods of model fitting did not bode well for this data. Various methods such as under sampling were used to fit a logistic model, classification tree, nearest neighbor, and neural network. Ensemble methods like Random Forest and ADABoost were also applied to both the whole data set and the under-sampled data. For all models, the cross-validation method with 10 folds was used on the training data to fit the model. Moreover, the tuning parameters for each model were specified by the user using “tuneGrid”. Using “tuneGrid”, the most optimal model is achieved because the parameters were not generated by the software. These powerful methods proved to be helpful with predicting the likelihood of customer churn.

Evaluation: Each model was evaluated using the test data partition. This comprised of 30% of the original data. Partitioning the data in this manner is an important step in the model building process because the model should be evaluated on data that it has not seen before – data that was not used to fit it. There are four important metrics to look at when evaluating a data model: accuracy, kappa, sensitivity, and specificity. The focus for this analysis was put on specificity because those are the customers that discontinued their services. The goal is to predict and identify those customers and prevent future attrition. The other metrics were also taken into consideration when choosing the best model.

Results

The accuracy, kappa, sensitivity, and specificity reflective of the cross-validated models being evaluated on the test data with and without down sampling is shown in **Figure 2**. ADABoost and Random Forest are top performers irrespective of the outcome imbalance. However, there is a decrease in the kappa value when training with the down-sampled data, as well as a slight decrease in accuracy in both models. Under sampling shows a dramatic increase in sensitivity in the K-nearest neighbors (KNN) model and a decrease in specificity. The down sampling hindered the neural networks performance by decreasing all the performance metrics.

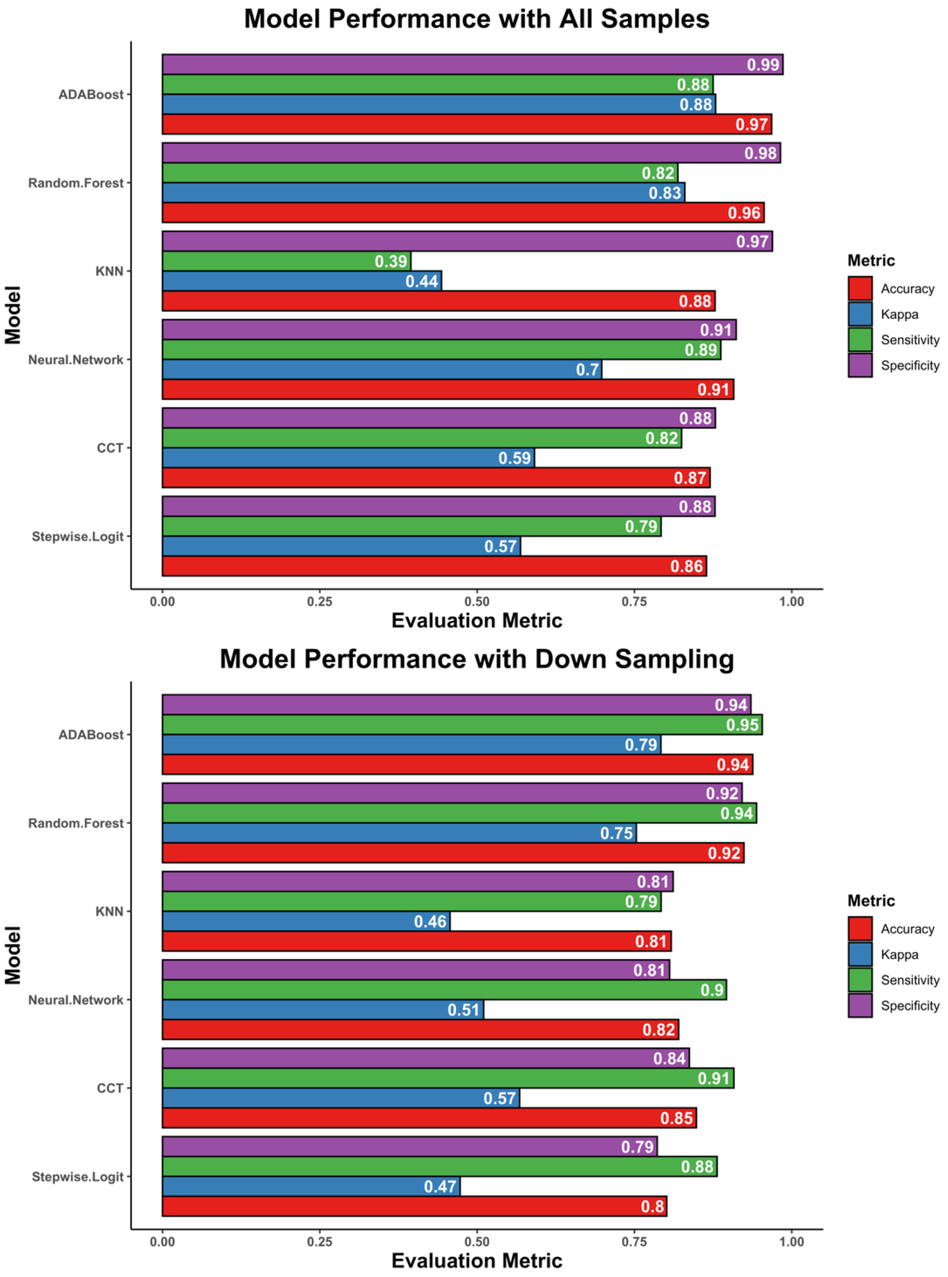

Figure 2: Summary of Model Evaluation Result

Optimization Metric: Kappa, ***Train-Test Split:*** *70%-30%*

The cost complexity tree (CCT) has a slight increase in sensitivity with down sampling and a slight decrease in specificity. The stepwise logistic regression suffered a decrease in accuracy, kappa, and specificity with down sampling but has an increase in sensitivity.

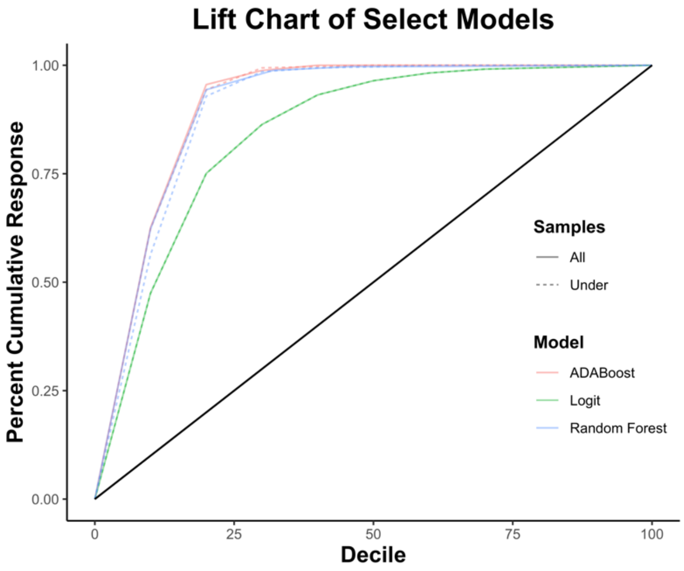

Figure 3: Lift Chart for Select Models

The ADABoost, Random Forest, and stepwise logit models were investigated further because of their performances as well as their ability to explain the predictor relationship to the outcome. **Figure 3** displays each models performance via a lift chart. The under sampling had little impact on the performance regarding the capture of cumulative response at lower deciles, with the slight exception of the Random Forest model. Both the tree models outperformed the stepwise logit model by capturing almost all the attritted customers in the top 25% of customers, in comparison to the logistic regression model where it would take at least the top 60% of customers to capture all the ones that left the bank. The stepwise logistic regression model is included in subsequent analyses because of its ability to provide insight regarding customer attributes and attrition.

The under sampled models are investigated further because of their increased performance in sensitivity and specificity. **Figure 4** depicts the specifics of the under-sampled model predictions via confusion matrices. All models have fairly low numbers of false negatives (the number of predicted customers who stay but actually leave) in comparison to the number of

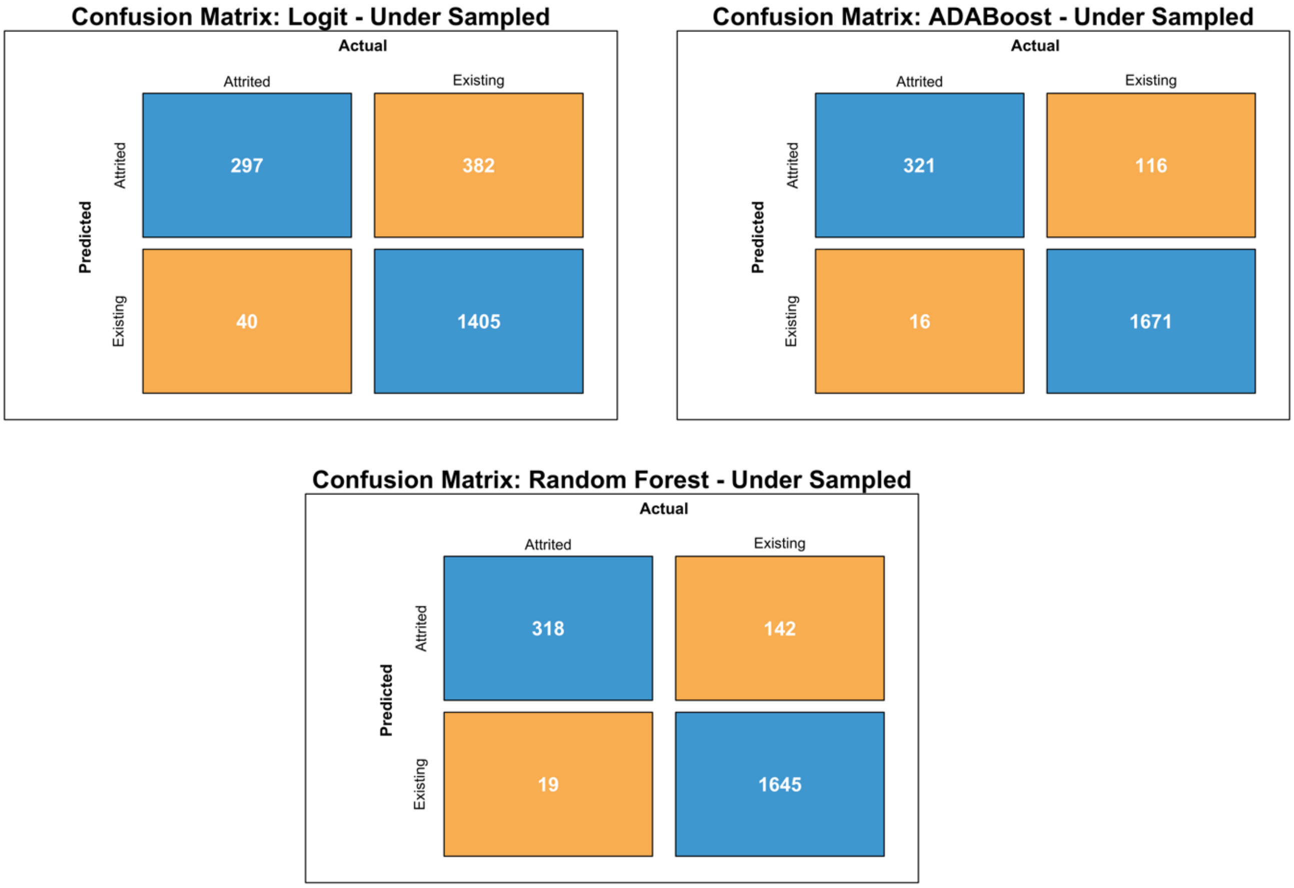

Figure 4: Confusion Matrices for Select Models

ADABoost Optimal Parameters: Iterations = 150, Maximum Tree Depth = 6, Learning Rate = 0.15
Random Forest Optimal Parameters: Number of Predictors = 8, Number of Trees = 50

true positives, which attributes to their high sensitivity. Similarly, the number of false positives is large in comparison to true negatives which attributes to the high specificity in the models. The logit model has twice the number of false negatives and positives as the selected tree models in **Figure 4**. The increased performance demonstrated by the ensemble models because of their ability to combine several models into a single prediction of the test data. However, the logistic regression model provides valuable insight as to how the customer's attributes impact the attrition prediction.

Figures 5 and **6** show the predictor importance and regression coefficients for the selected tree models and stepwise logit model, respectively. The predictor importance score for the tree models is derived from which predictor decision rules result in the maximum impurity reduction. The logistic regression coefficient estimates were predicted using an ordinary least squares approach. The total transaction number and amount appeared to be the most important in both the ADABoost and Random Forest models, followed by the total revolving balance. As seen in **Figure 5**, there are several variables that are more important in the ADABoost model in comparison to the Random Forest model. Interestingly, the customer's age is somewhat important in the Random Forest model, but not at all in the ADABoost model. The months inactive and contact count within the previous 12 months both appear be important in all models, but more so in the logit and ADABoost models.

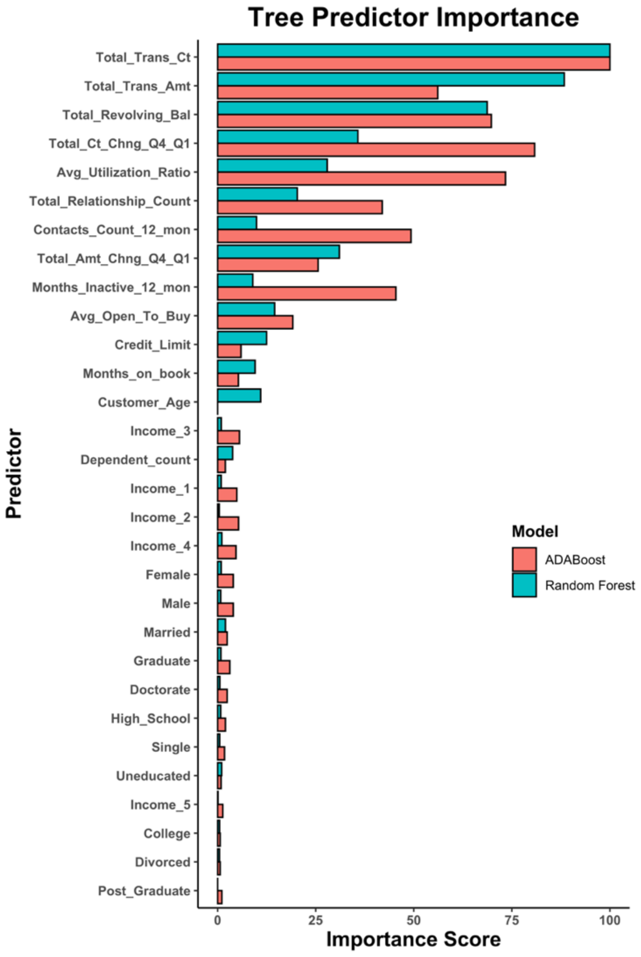

Figure 5: Random Forest and ADABoost Predictor Performance

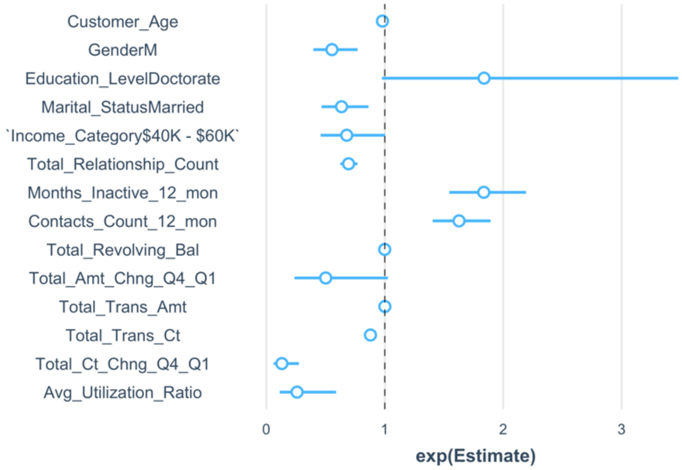

Figure 6: Logistic Regression Coefficients

Attrited Customer = 1, Existing Customer = 0

Table 2: Stepwise Logistic Regression Coefficient Statistics

Term	Estimate	Standard Error	Z	P-Value
(Intercept)	7.9562	0.7742	10.2765	0.0000
Education_LevelDoctorate	0.6088	0.324	1.8788	0.0603
Months_Inactive_12_mon	0.6076	0.0894	6.7966	0.0000
Contacts_Count_12_mon	0.4868	0.0761	6.3971	0.0000
Total_Trans_Amt	6e-04	0.0000	11.5472	0.0000
Total_Revolving_Bal	-7e-04	1e-04	-5.5821	0.0000
Customer_Age	-0.0201	0.0098	-2.0499	0.0404
Total_Trans_Ct	-0.1289	0.0080	-16.0779	0.0000
Total_Relationship_Count	-0.3652	0.0530	-6.8848	0.0000
`Income_Category$40K - $60K`	-0.3879	0.1999	-1.9406	0.0523
Marital_StatusMarried	-0.4545	0.1575	-2.8864	0.0039
GenderM	-0.5911	0.1695	-3.4885	5e-04
Total_Amt_Chng_Q4_Q1	-0.6913	0.3736	-1.8503	0.0643
Avg_Utilization_Ratio	-1.3507	0.4206	-3.2115	0.0013
Total_Ct_Chng_Q4_Q1	-2.0349	0.3832	-5.3097	0.0000

Figure 6 depicts the exponentiated logit coefficients with their associated 95% confidence intervals to best illustrate their multiplicative impact to the odds ratio. **Table 2** lists further details regarding the logistic regression coefficients. Customer attributes that decrease the odds ratio include the male gender, if the customer is married, the total number of products the customer has with the bank, transaction count ratio of quarter four to quarter one, the total number of transactions, and the balance to limit ratio. Characteristics that increase the chances of turnover include inactivity of customers in the previous 12 months, and the total transaction amount.

Discussion / Conclusion

Our data mining techniques and analyses gave us many interesting and important results. While we had expected some of the results, others were a surprise to us. The models that we fit to the data included stepwise logit, K-nearest neighbor, cost-complexity tree, neural network, Random Forest, and ADABoost, and they all were applied to both the whole data set and the under-sampled data. Confusion matrices revealed that the Random Forest and ADABoost models performed best. Since the goal of the project was to predict customers that churned, we focused our model evaluations on sensitivity. The ADABoost model had a sensitivity of 0.88 with all samples and 0.95 with down sampling, while the Random Forest model had a sensitivity of 0.82 with all samples and 0.94 with down sampling. Clearly, both of these models performed quite well and are able to successfully predict which customers will churn. Based off the work of Tsai & Lu (2009), we had predicted that a neural network model would successfully predict customer churn. The neural network model yielded a sensitivity of 0.89 for all samples and 0.90 for down sampling. So, while this can be considered a successful predicting model in regard to its sensitivity, it does not perform as well as both the Random Forest and ADABoost models for other measures (accuracy, Kappa, and specificity), which is shown in **Figure 2**. Our prediction that Random Forest and ADABoost models would perform well was correct.

The secondary goal of this project was to examine the importance of each feature with respect to customer attrition, and we had predicted that age, marital status, and income category would all be significant features in our predicting model. Lin, Tzeng, and Chin (2011) had found that being single and between the ages of 30-39 were two of the five most important characteristics of credit card customers that had voluntarily churned, which is why we predicted that age and marital status would be significant features. Our results yielded that the total transaction number and amount appeared to be the most important in both the ADABoost and Random Forest models, which we did not predict to be the most important. Surprisingly, the variable for customer's age was in the top 10 most significant features in the Random Forest model but was not significant at all in the ADABoost model. It's quite surprising to see this stark difference in the two models since they both perform so well. Also, the married and single variables have much less significance than we had anticipated, both for the Random Forest model and the ADABoost model. Income also was only moderately significant for both models. Lin, Tzeng, and Chin (2011) had also found that two major characteristics of customers that voluntarily churned were being single and having an average purchase amount of zero. While we did not have any predictions for this, our data does support this result. Our logit results shown in **Figure 6** show that being married decreases the change of churning, while having a small total transaction amount increases the chance of churning. This aligns with Lin, Tzeng, and Chin's results – being single and not spending much both are major characteristics of customers that have churned. So, our predictions for which models would be good predictors were mostly correct, but our predictions for which features would be important were mostly incorrect.

The results of the current study lend more support for using Random Forest and ADABoost models for predicting credit card customer churn in the future. For our dataset in particular, our results show the significant characteristics of customers that have churned, such as the total transaction number and amount. This is important, because the bank management can use this information to prevent customer churn in the future. Since they know the number and amount of transactions is important, they can identify which customers have small numbers/amounts of transactions and focus on getting communication/advertising to them more to prevent attrition. This is also true for the logit results – bank management can focus more on customers that are single and have not spent much on their credit cards. Preventing customer churn is immensely important to companies, because this can have a massive impact on profit. In conclusion, our data mining analyses yielded very successful predicting models for credit card customer churn and the associated features.

Limitations & Future Work

One major limitation of this study is the lack of knowledge of how the data was collected and what bank the data came from. The study would have benefited from the knowing where the data came from to make sense of the context of the problem and where to make recommendations. Another limitation is the attributes of each customer, some studies had other descriptive predictors such as automatic debit transfer enrollment. The addition of features may provide additional information regarding the customers who left the bank.

Future directions of the customer churn analysis include the utilization of advanced sampling methods to balance the attrition outcome. Methods such as Edited Nearest Neighbor (ENN) to down sample would benefit the study by selecting customers that are more representative of the population. Synthetic Minority Oversampling Technique (SMOTE) could also be implemented to generate artificial data to improve the response balance. Incorporation of dimensionality reduction such as principal component analysis (PCA) prior to prediction would benefit the study by potentially improving evaluation metrics of some models. Use of unsupervised clustering of the customer base would elucidate different groups within the customer base. Identifying specific characteristics of customers who left the bank would allow for them to be targeted to enable solutions to be made to extend banking relationships.

References

Goyal, Sakshi (2020). *Credit Card Customers*, Version 1. Retrieved from https://www.kaggle.com/sakshigoyal7/credit-card-customers.

Lin, C., Tzeng, G., & Chin, Y. (2011). Combined rough set theory and flow network graph to predict customer churn in credit card accounts. *Expert Systems with Applications*, *38*(1), 8-15. doi:https://doi.org/10.1016/j.eswa.2010.05.039

Tsai, C., & Lu, Y. (2009). Customer Churn Prediction By Hybrid Neural Networks. *Expert Systems with Applications*, *36*(10), 12547-12553. doi:https://doi.org/10.1016/j.eswa.2009.05.032

Acknowledgements

We would like to thank Dr. Shuchismita Sarkar for her help with the data mining process in general by teaching the MSA 6440 course at Bowling Green State University.

Code Availability

The code, data, and plots are available via GitLab at https://gitlab.com/nickwawee/customer_churn.