Final Report

Bank Credit Card Churning Customer Prediction

Dae Hyun Kim

Disclaimer

The data and idea came from following Kaggle website

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

Introduction

Just like in any other businesses, identifying customers who are likely to leave and accommodating for those customers will help maintaining the success of the business. From a bank's perspective, it is beneficial to identify clients discontinuing their credit card services, or churning customers. I obtained a data consisting 10,127 rows of customers and 19 features which include their age, salary, marital status, credit card limit, credit card category, card utilization etc. Within this data, 16.07% of customers have churned. The main objective of this project is to build a model to predict customers who are likely to churn and determine some features that are related to the client leaving the service.

Data Wrangling

There were 23 columns and 10,127 rows of data to begin with. Luckily the data was relatively clean, so I only had to work on minor details to cleaning the data. During the data wrangling process, I dropped last two columns labeled as Naive Bayes classifiers. These columns may have been from other analysis by the Kaggle publisher.

Other columns were all filled without repeating rows. Attrition_Flag, which is the target variable, and Gender columns were categorical but also binary, so I replaced to 0 and 1.

Exploratory Data Analysis

I used Tableau to help with the data visualization. For interactive data analysis, visit the following site
https://public.tableau.com/app/profile/dae.hyun.kim/viz/CapstoneThreeEDA/Summary

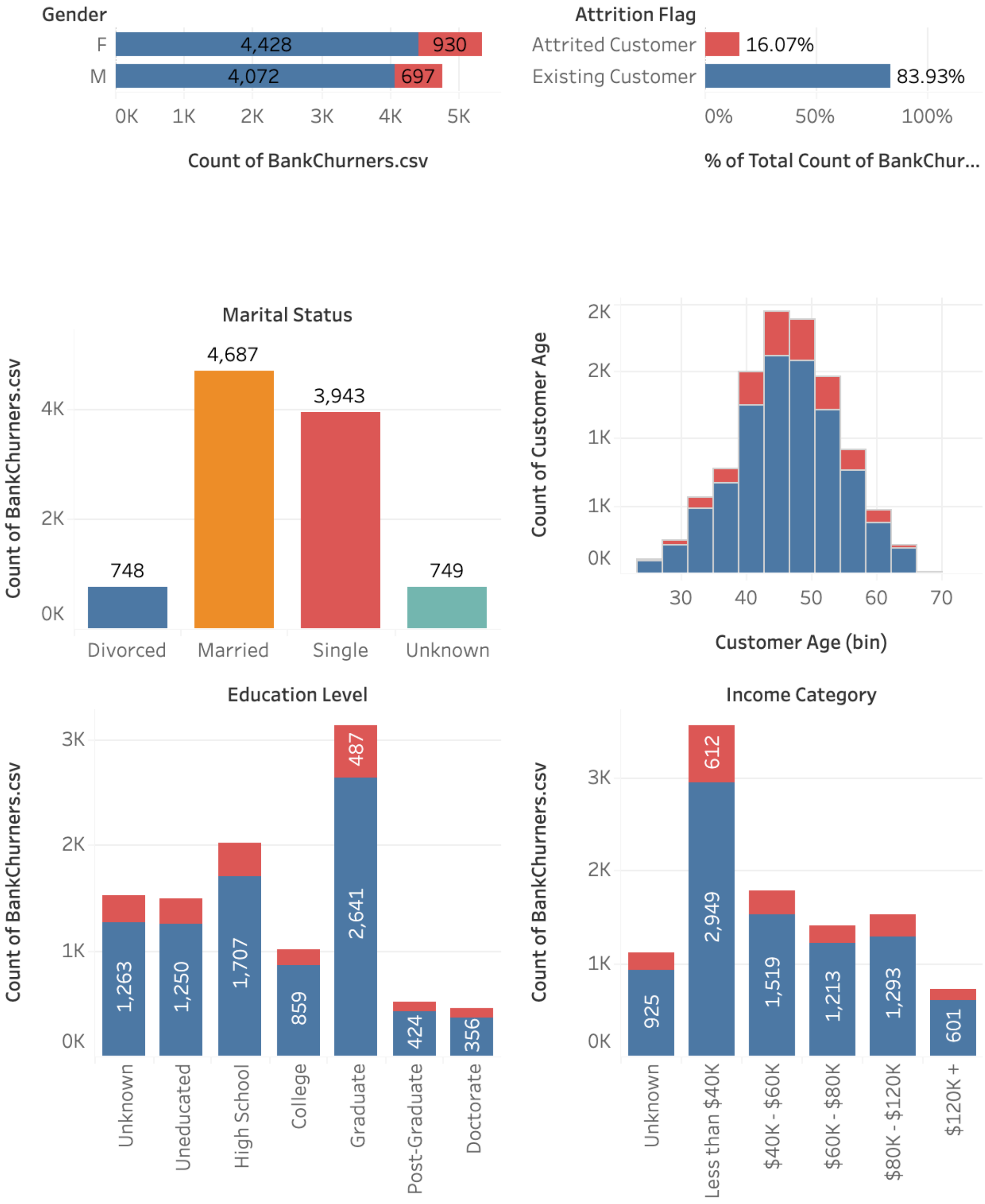
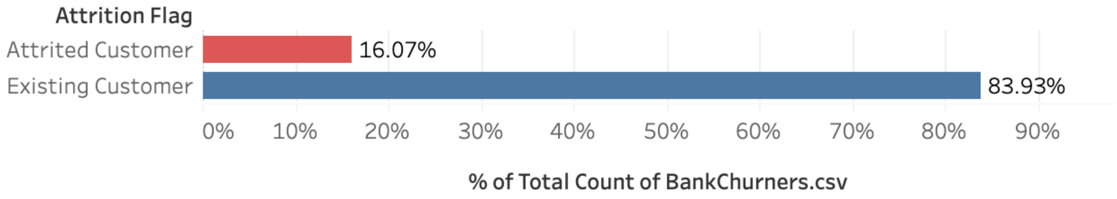


Figure 1. Customers' socioeconomic standings



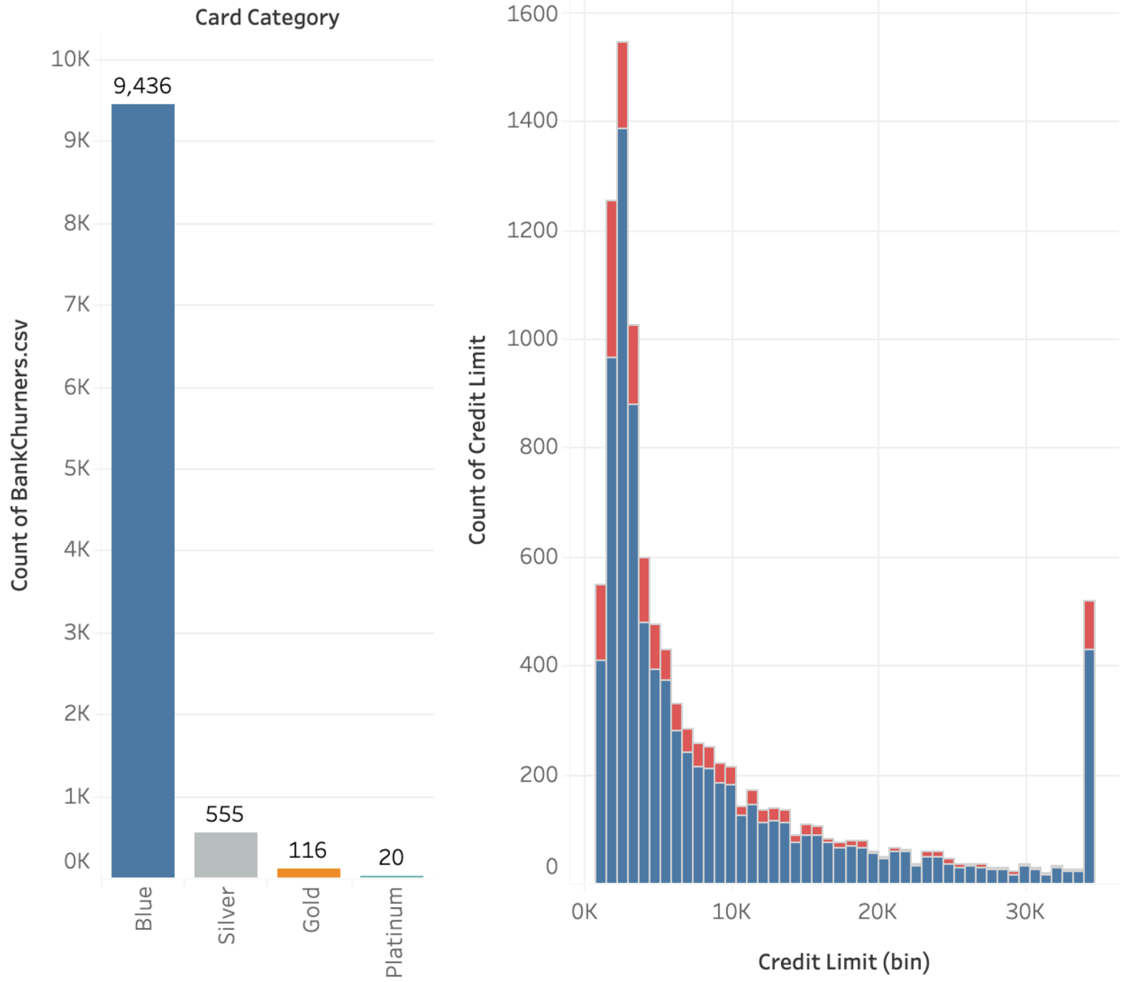
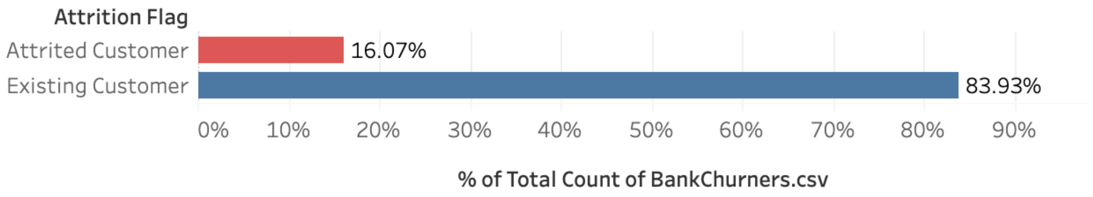


Figure 2. Customers' card category and credit limit

Customers' income level, education level, gender, marital status, and age didn't seem to affect the client leaving. Similarly, the card category and credit limit seemed irrelevant to the churning prediction.



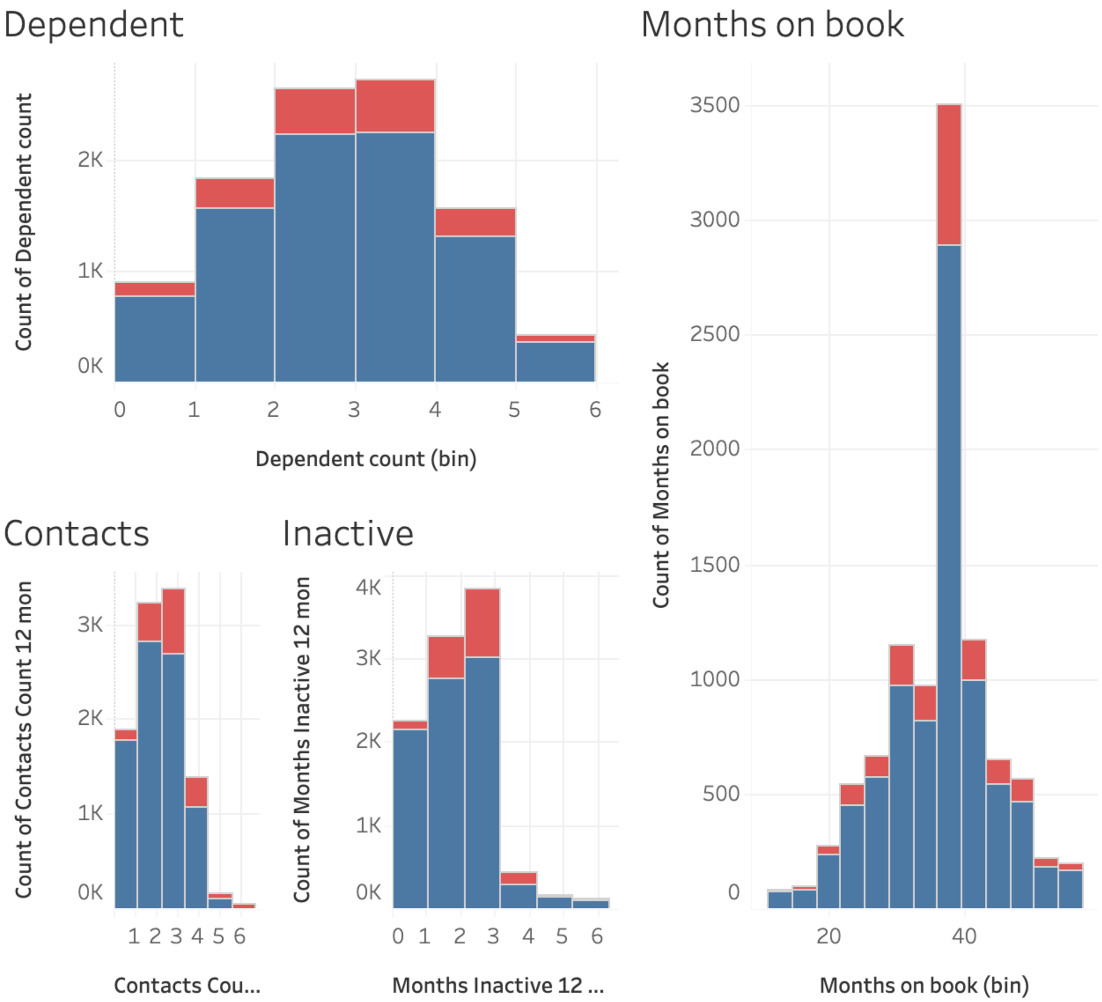
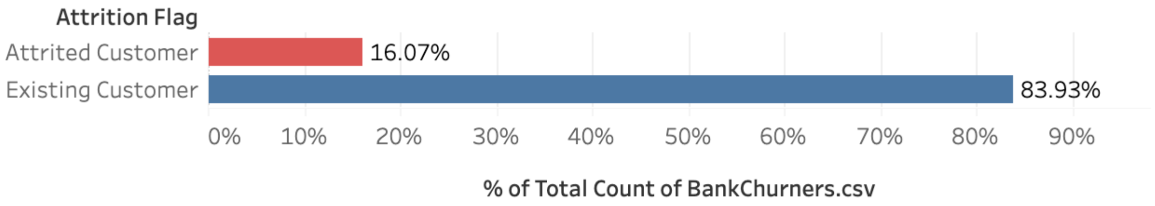


Figure 3. Customers' time active

From these visuals, one thing that stood out was the months inactive. Usually customers inactive 2-3 months tend to leave the service. Also, interestingly, there are small number of customers staying inactive for 5-6 months but still keeping the credit card. This may due to the fact that they forgot about the credit card altogether, or individuals who may not use credit card in daily basis.



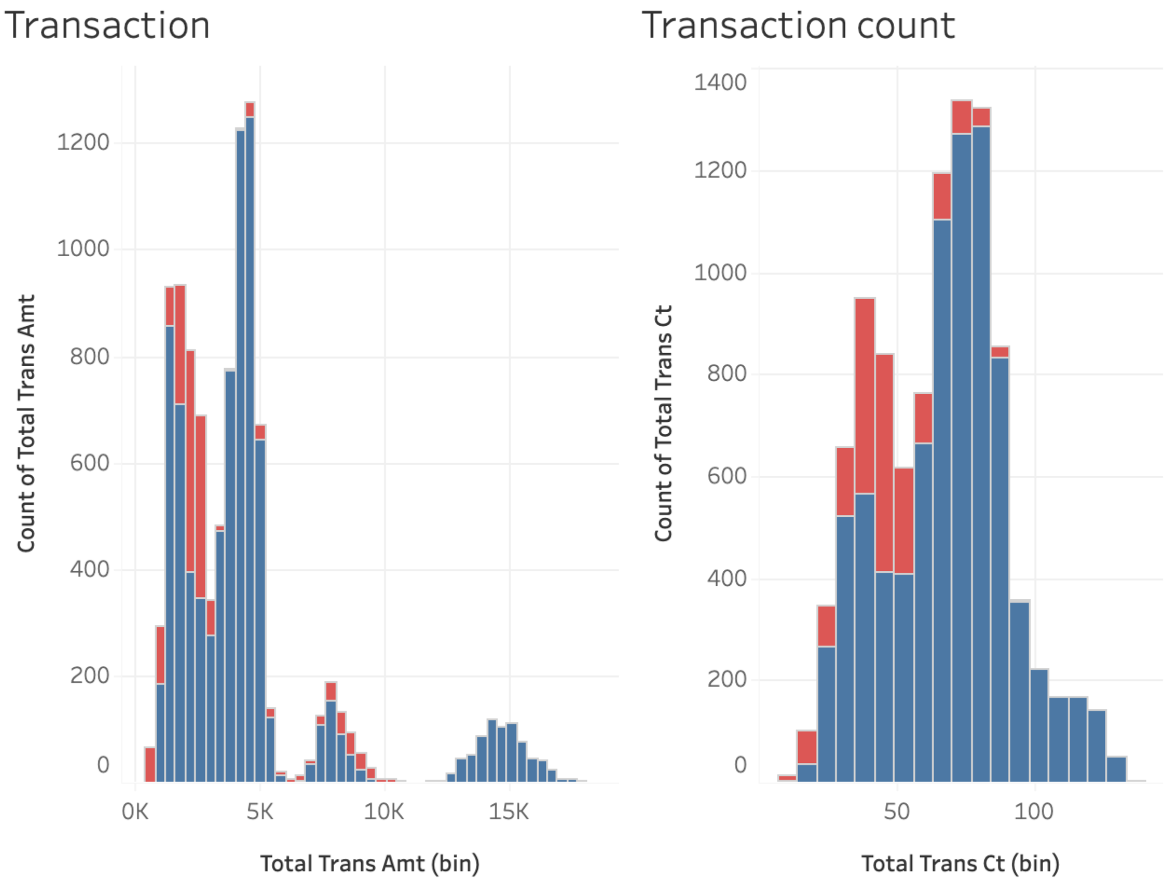


Figure 4. Customers' Transactions

In contrast, the majority of attrited customers were those who didn't utilize their credit card much. Roughly speaking $3k total transaction amount and 50 total transaction count may be a good indicator for churning customers.

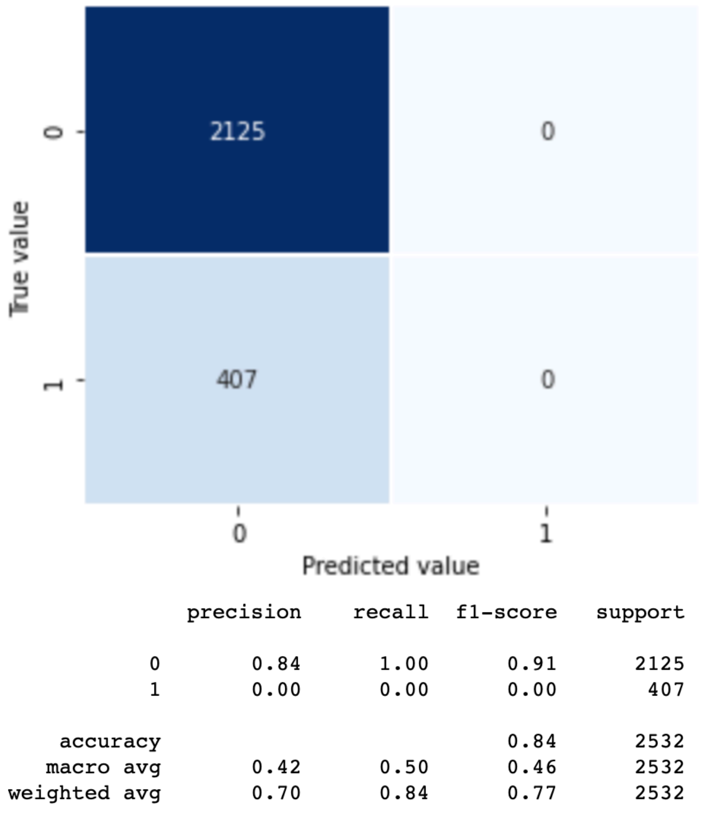
From the Exploratory Data Analysis, I could suggest bank to look out for customers who were inactive for 2-3months and customers who does not exceed $3k total transaction amount and 50 total transaction count. However, this is roughly speaking based on the visualizations.

Preprocessing

Before implementing models, I worked on little bit of preprocessing. Again, the original data was relatively clean. First of all, I dropped the CLIENTNUM column because it was the primary key. Next, for categorical variables: Education_Level, Marital_Status, Income_Category, and Card_Category I created dummy variables. Finally, I separated the target variable Attrition_Flag and scaled the data using the standard scaler.

Model Selection

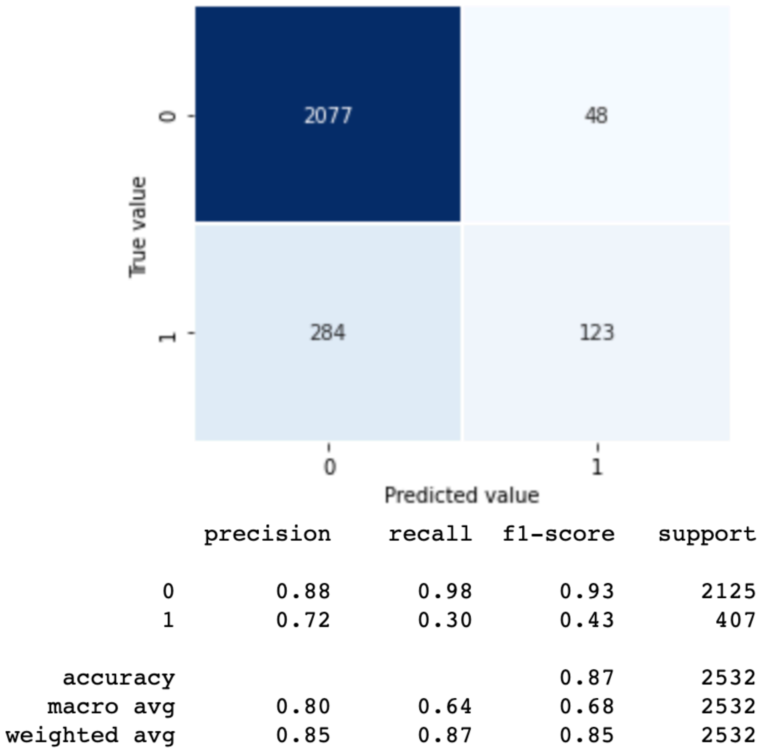
As my first model, I established the baseline model which predicts every customer will remain in the system. Since the data consisted 16.07% of attrited customers, this predictor will have about 84% of accuracy. The confusion matrix and classification report verify my initial guess.



```
              precision    recall  f1-score   support

           0       0.84      1.00      0.91      2125
           1       0.00      0.00      0.00       407

    accuracy                           0.84      2532
   macro avg       0.42      0.50      0.46      2532
weighted avg       0.70      0.84      0.77      2532
```

Figure 5. Baseline dummy model’s confusion matrix and classification report

Secondly, I used the K-Nearest Neighbor classification. With hyperparameter tuning, 5 neighbors resulted in 87% accuracy. Normally, 87% would be an impressive result, but this was only 3% improvement compared to the baseline model. The confusion matrix and classification report can be seen below.

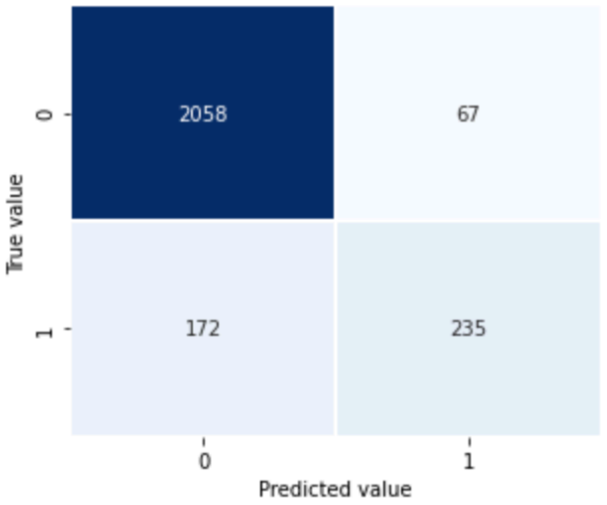


```
              precision    recall  f1-score   support

           0       0.88      0.98      0.93      2125
           1       0.72      0.30      0.43       407

    accuracy                           0.87      2532
   macro avg       0.80      0.64      0.68      2532
weighted avg       0.85      0.87      0.85      2532
```

Figure 6. K-Nearest Neighbor (k=5) model's confusion matrix and classification report

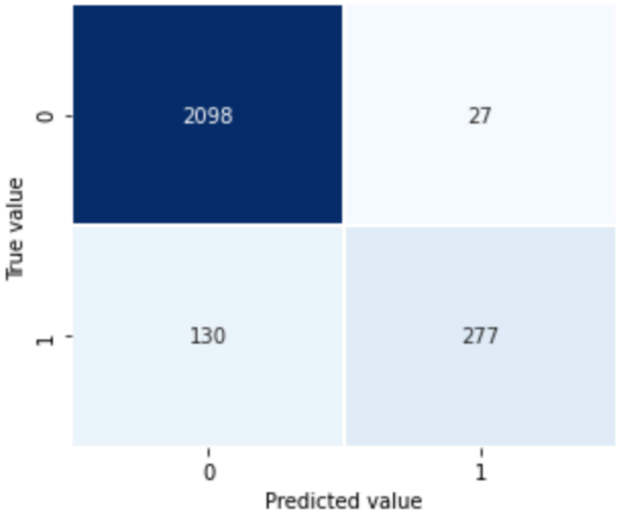
Next, I used the logistic regression. With this method I was able to achieve 91% accuracy. The result can be seen in Figure 7.

Finally, for my last model, I used the RandomForest Classification. The final model was superior compared to all the other methods with the 94% accuracy and the greatest number of true positives and true negatives.



```
              precision    recall  f1-score   support

           0       0.92      0.97      0.95      2125
           1       0.78      0.58      0.66       407

    accuracy                           0.91      2532
   macro avg       0.85      0.77      0.80      2532
weighted avg       0.90      0.91      0.90      2532
```

Figure 7. Logistic regression model’s confusion matrix and classification report



```
              precision    recall  f1-score   support

           0       0.94      0.99      0.96      2125
           1       0.91      0.68      0.78       407

    accuracy                           0.94      2532
   macro avg       0.93      0.83      0.87      2532
weighted avg       0.94      0.94      0.93      2532
```

Figure 8. RandomForest Classification model’s confusion matrix and classification report

Takeaways

After finding the final model, I decided to check the feature importance as shown in Figure 9. Similar to the conclusion I got from the exploratory data analysis, the transaction amount and the transaction count were top two features in prediction. Followed by the recent transactions in Q4.

With other top features being self-explanatory and expected, another column that stood out was the Total_Relationship_Count which is the number of bank products held by the customer. It seems like this can be a good indication to see how loyal the customer is to the credit card company.

Finally, it was interesting to note the minimal influence of the socioeconomic status of the customer. The education level, gender, age, income category, and marital status didn't have much impact on churning prediction.

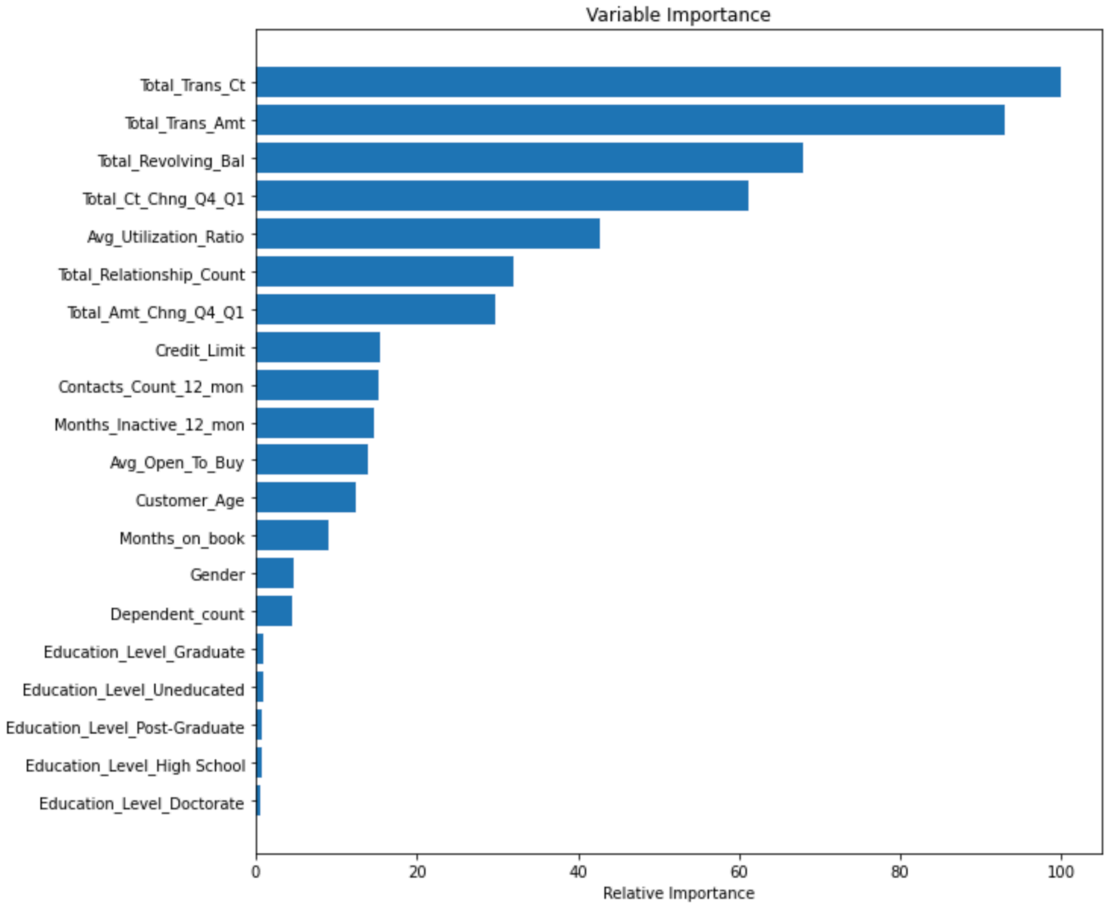


Figure 9. Relative feature importance from RandomForest Classification model

For bank managers who are concerned about losing credit card customers, promoting higher percentage of cashback to encourage credit card usage from customers can be an immediate solution. Furthermore, as the customer use the credit card, also introducing other products to increase the customer loyalty may be appropriate.

Future Direction

Fortunately, the data I got for this problem was clean, and the prediction model reached stunning 94% accuracy. However, one thing I would like to improve on is to come up with more concrete warning signs or indicators to tell which customers are likely to dropout. So far, I could only suppose an estimated guess from my exploratory data analysis graphs. I wonder if there are more systematic tools to come up with limits for each feature. This may require more extensive data visualization with more breakdowns before and after implementing machine learning models.