ABC Bank: Predicting the likelihood of credit card churn

By Phong Pham

1. Introduction

a) Problem statement

What challenges does ABC Bank face if they (a) increase the frequency of contacts to customers that are most likely to churn on their credit card and/or (b) have more promotional programs (e.g. cashback, rewards, etc.) to encourage customer to utilize their credit cards, in order to reduce the bank credit card churn rate to below 10% at the end of 2021?

b) Background

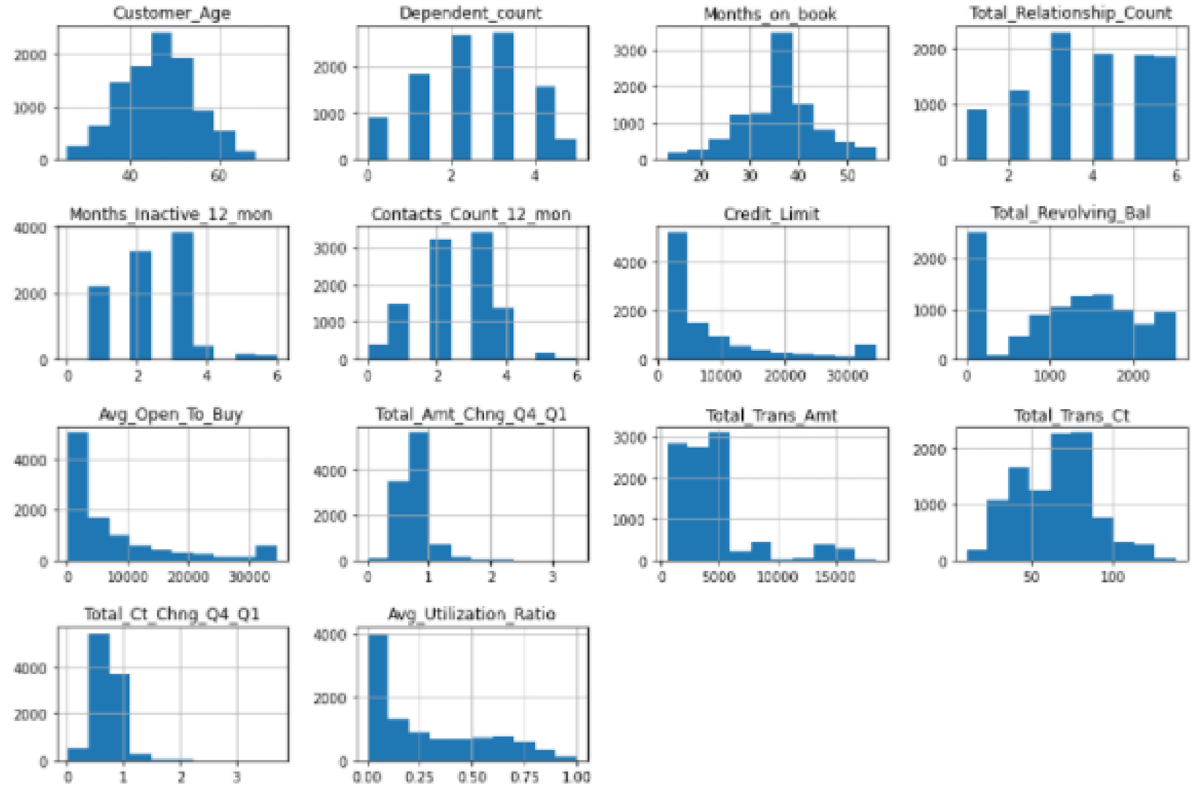
The management at ABC Bank is disturbed with more and more customers leaving their credit card services. They would like the data scientist team to come up with a model that could predict for them who is going to get churned so they can proactively go to the customer to provide them better services and turn customers' decisions in the opposite direction. Based on the data available from 10,000 credit card customers, the bank management wants to find data-driven solutions to encourage customers to continue using their cards.

c) Goal

This project aims to provide a prediction model for credit card churns using data-driven analysis. We use demographics, credit card utilization and other bank product details from the bank customer database to predict which customers are most likely to churn on their credit cards. Based on our prediction, the bank can come up with customized promotions for different groups of potentially churning credit card customers to make them change their minds.

2. Data Wrangling

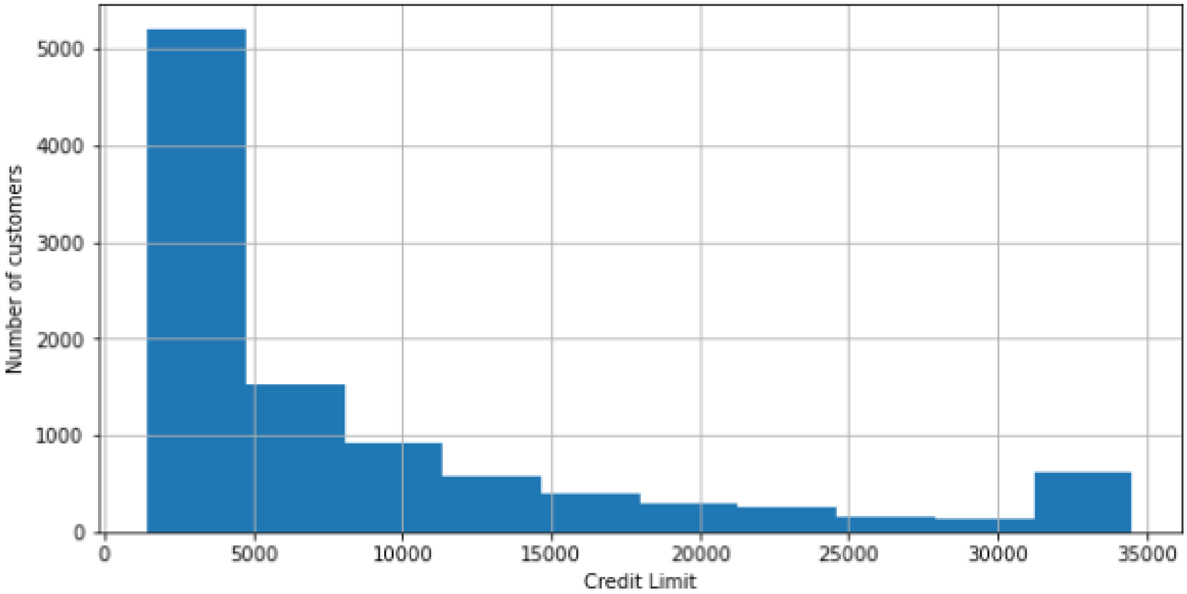
The original dataset from Kaggle contained 10,127 rows with 23 columns. It was relatively a clean dataset without any null values. However, some of the demographics details were unknown, and as these play a part in the prediction model, I dropped the records with "Unknown" values for all 3 demographics variables (Education level, Income, and Marital status). Only 7 records were dropped, leaving us with 10,120 rows.



There seems to be no clear outliers in any of the numerical features. Some columns have values closer to 0, such as Credit Limit, Total Revolving Balance, Average Open To Buy, and Average Utilization Ratio.

There are more than 5351 people earning $60K or less per year. Since a credit card limit usually correlates with an applicant's annual income, it is plausible to have more than 5,000 people with credit limits on the lower end (less than $5,000).

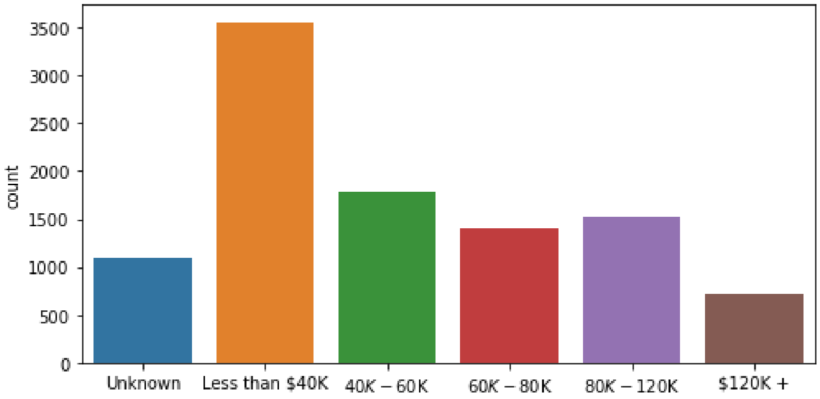
Also, it is not unusual that a lot of people might not use their credit cards, or they might pay off the amount owed on their cards before the end of the billing cycle, hence the low card revolving balance, open-to-buy, and utilization ratio.



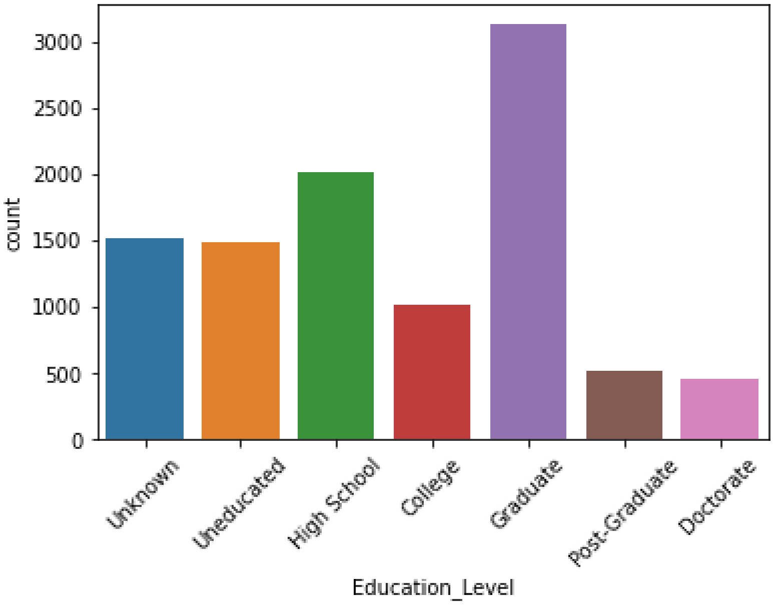
It is noticeable that the credit limit chart has a spike at the end, meaning the number of customers with credit limit over $30,000 is comparably higher than a few other ranges that come before them in the chart. Upon investigation, we can see that most of the customers with credit limit over $30,000 have annual incomes higher than $80,000.

3. Exploratory Data Analysis

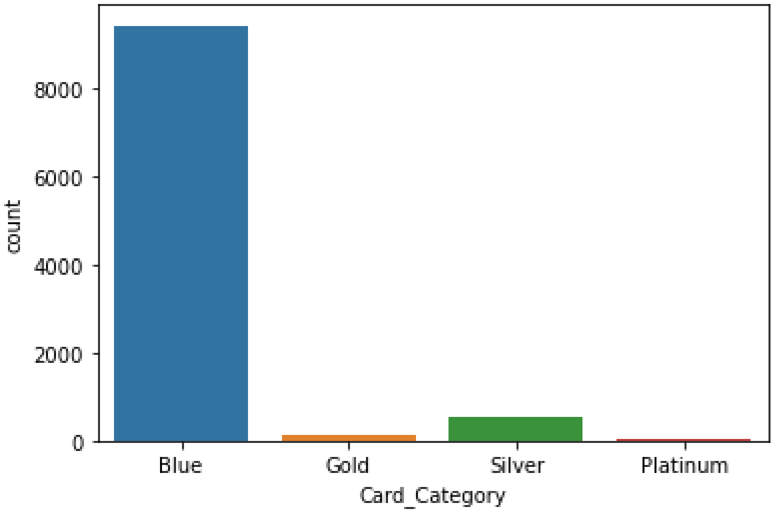
In this section, we explore the demographics and card ownership of ABC Bank customers.



About 35% of customers have an income of less than $40,000. There are about 1000 customers who did not report their income, which might make it harder for the bank during data analysis.

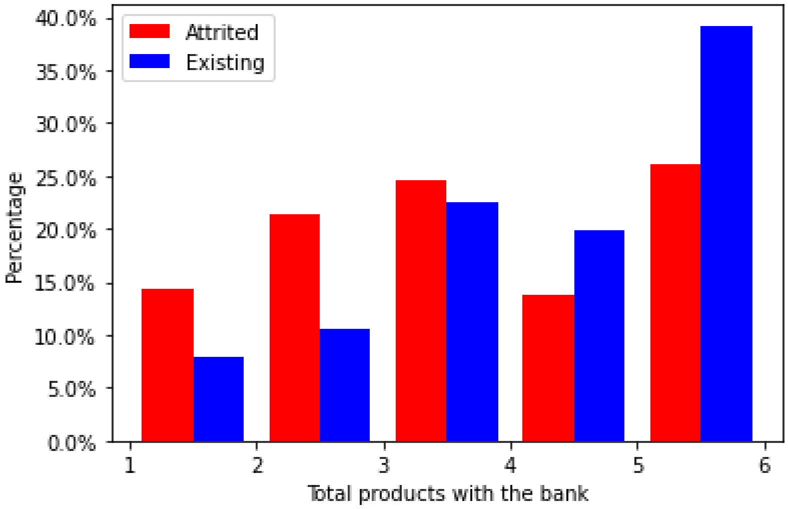


Looking at education level, people with a graduate degree account for the highest number of customers, followed by High School diploma holders. Again, the number of people with an unknown education level is quite high (around 1500), so this missing data might be something the bank should look into updating in the future.

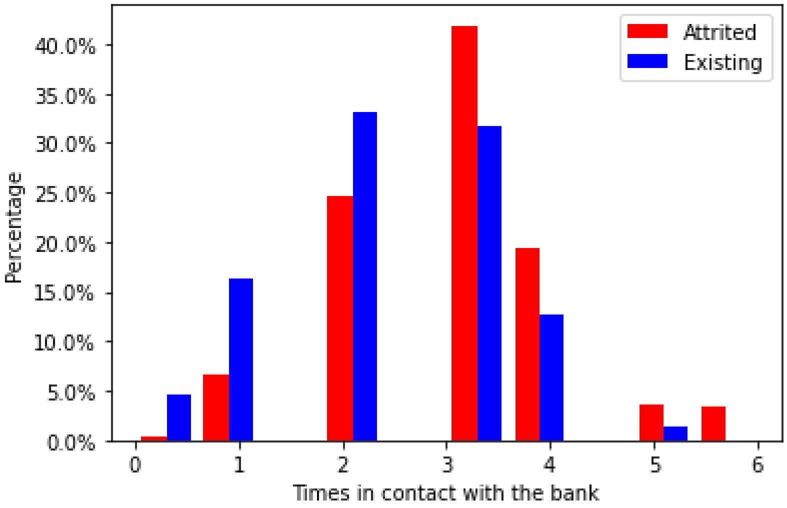


The majority of customers hold a Blue credit card (more than 9000), followed by Silver card holders.

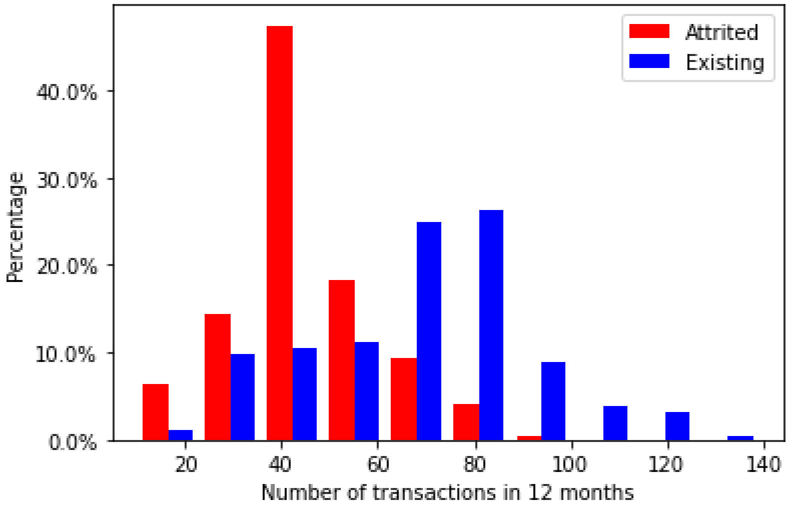
The dataset only has about 16% of customers that have churned on credit cards. Therefore, in order to scale the data for comparisons of each numerical variable, we will get the percentage of each value when compared to the total number of existing/attritted customers, then plot them on a histogram.



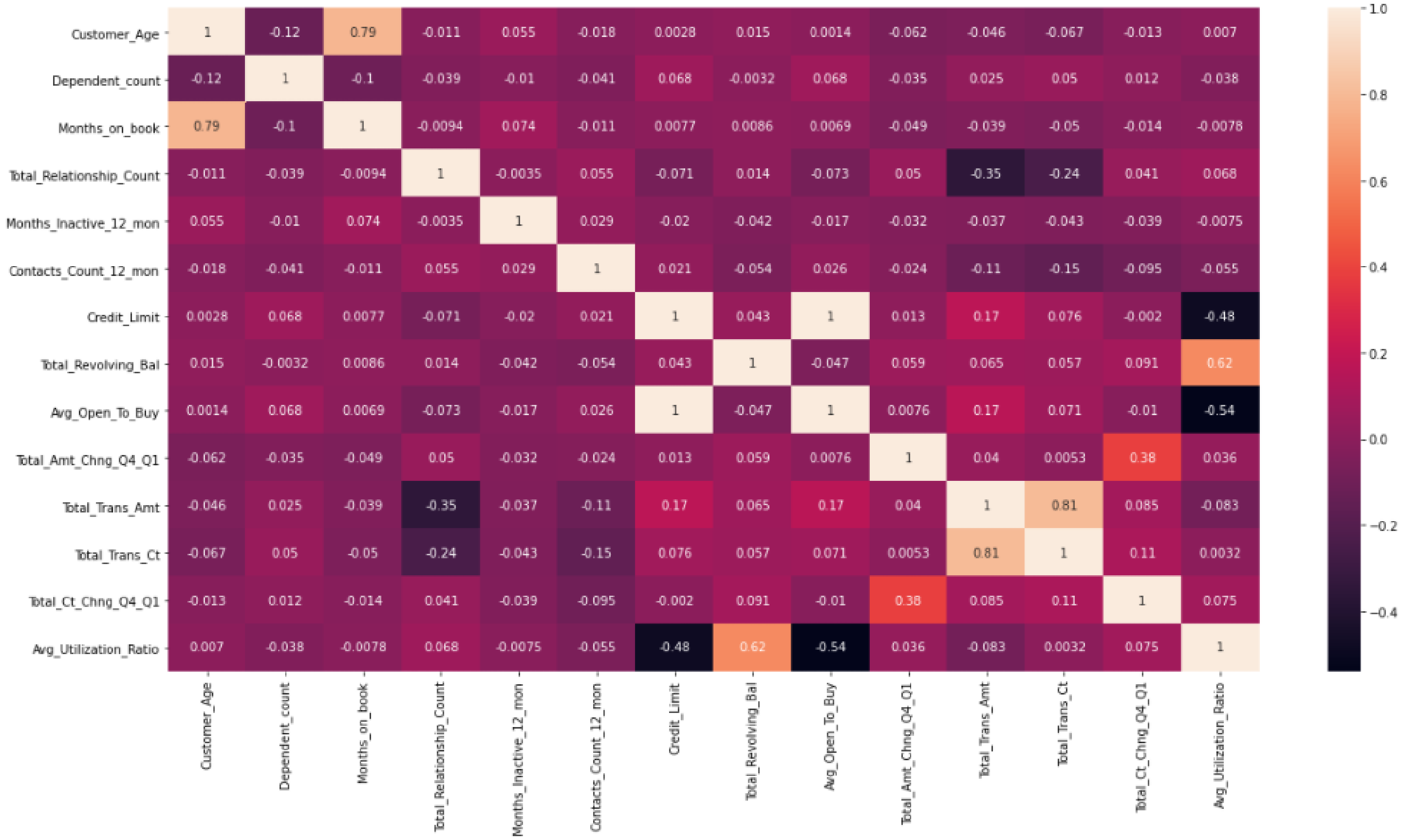
When comparing the percentage of attritted and existing customers, people with fewer products with the bank (4 or fewer) look more likely to churn on their card.



We can see that customers that have contacted the bank 3 or more times will be more likely to churn on their cards. This suggests that the reasons for contacting the bank might be of something they are not happy about regarding the products offered by the bank.



We can clearly see that attrited customers are more likely to be the one who use their cards less often (60 times or fewer in 12 months).



Upon exploring the correlations between columns, we can see fairly strong correlations between a customer age and their time being a customer of the bank (0.79), total transfer amount and total transaction count (0.81), and a moderately strong correlation between total revolving balance and average utilization ratio (0.62). To avoid collinearity affecting the accuracy of our churn prediction models, we will remove one column from each highly correlated pair. The removed columns being removed were Customer_Age, Total_Trans_Amt and Avg_Utilization_Ratio.

On the other hand, there are moderately strong negative correlations between credit limit and average utilization ratio (-0.48), and average open to buy credit line (-0.54).

4. Preprocessing and modeling

There are a total of 6 categorical variables in the dataset:

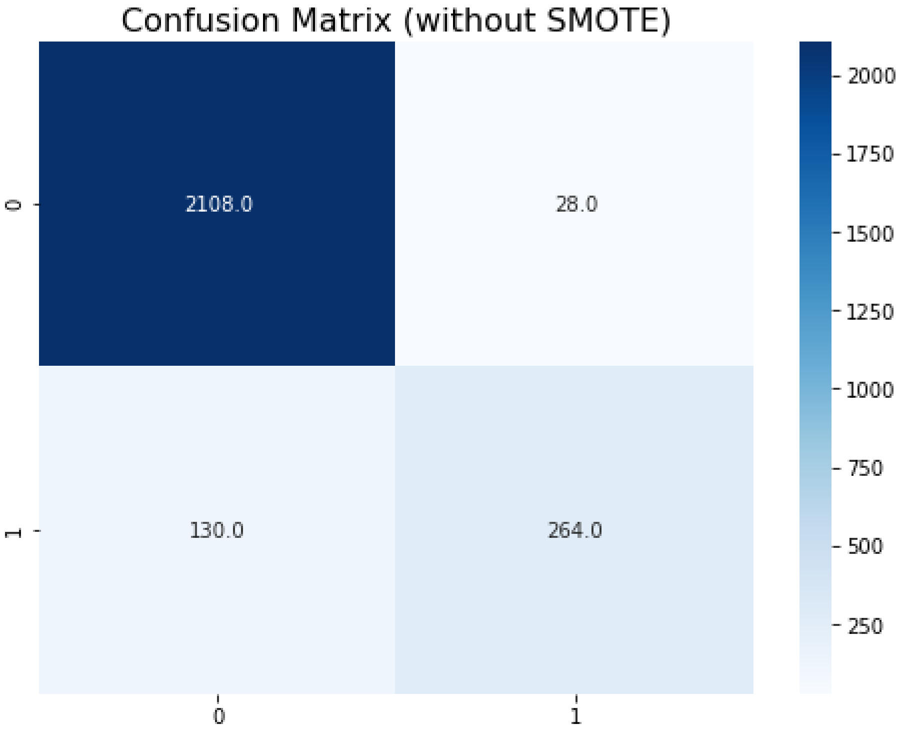
- Attrition_Flag
- Gender
- Education_Level
- Marital_Status
- Income_Category
- Card_Category

In order to use these columns in our prediction models, we converted them into dummy/indicator variables using Pandas one-hot encoding method get_dummies. After this step, the original 18 columns were transformed into 31 columns.

We then standardized numerical columns using a range of [0,1] as they were on different scales.

As there are only about 16% of customer who churned, there is a class imbalance in this dataset. We needed to handle this imbalance using SMOTE to oversample the minority class.

For comparison, we created a Random Forest model on the dataset and ignored the class imbalance. This resulted in an accuracy score of 94% and a recall score of 67%.



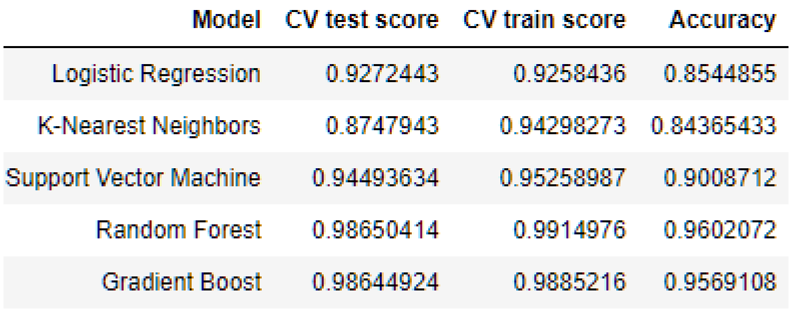
As we can see, it can correctly classify almost all customers who don't churn. But it also classified 33% of churned customer as existing customer.

Using SMOTE, we oversampled the number of churned customers. The original dataset with 10,120 rows became a balanced one with 16,988 rows. When applying a Random Forest model to this new dataset, the accuracy slightly improved to 96% and the recall score dramatically shot up to 95%. With SMOTE, only 5% of churned customers in the test set were misclassified as existing, so this is much better for modeling purposes.

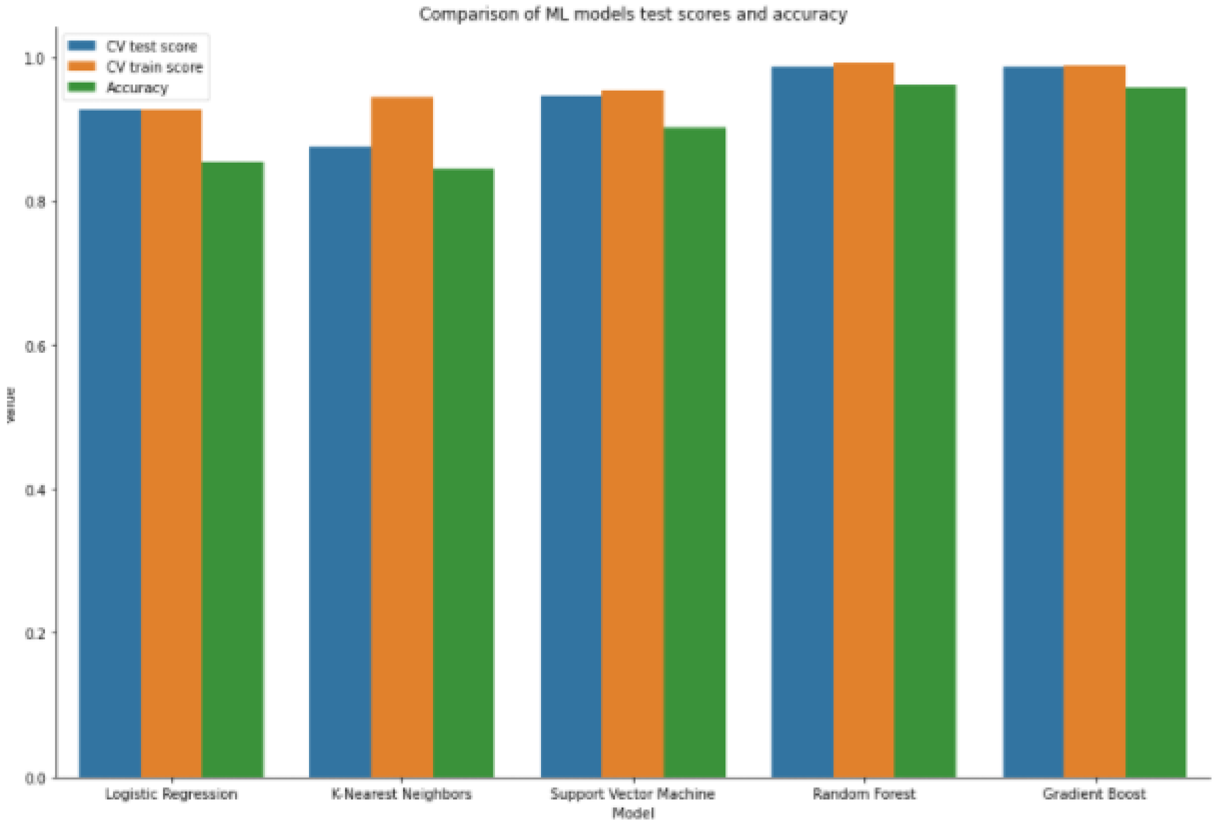
As this is a classification problem aiming to predict which customers churn their credit cards, we will use the following machine learning models:

- Logistic Regression
- K-Nearest Neighbor (KNN)
- Support vector machine (SVM)
- Random Forest
- Gradient Boost

We split the dataset into a training and a testing set with a 75:25 ratio, then applied different ML models and evaluated them using ROC-AUC scores. Here are the cross validation scores and accuracy of all models:



Model	CV test score	CV train score	Accuracy
Logistic Regression	0.9272443	0.9258436	0.8544855
K-Nearest Neighbors	0.8747943	0.94298273	0.84365433
Support Vector Machine	0.94493634	0.95258987	0.9008712
Random Forest	0.98650414	0.9914976	0.9602072
Gradient Boost	0.98644924	0.9885216	0.9569108

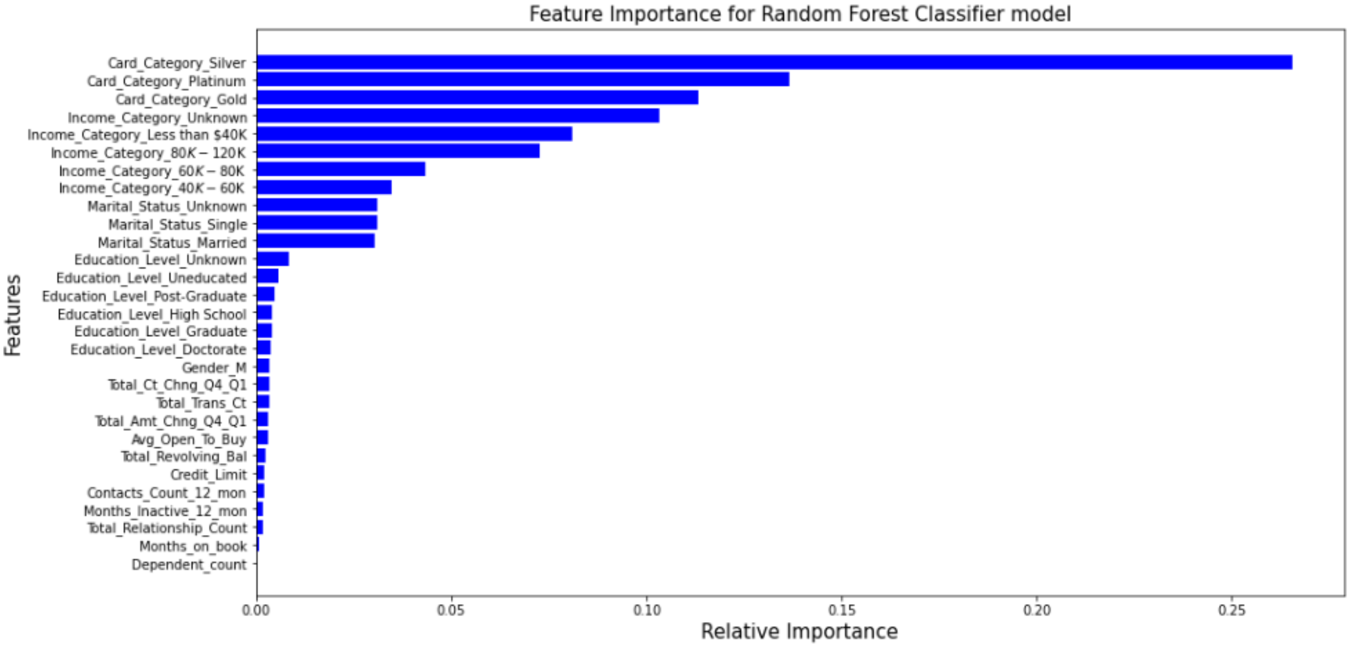


From the barplot, we can see that Random Forest and Gradient Boost are the best models to use for our credit card churn predictions. Therefore, we used these 2 models with optimal hyperparameters to predict the attritted flag value.

For the Random Forest model, we used the following hyperparameters:

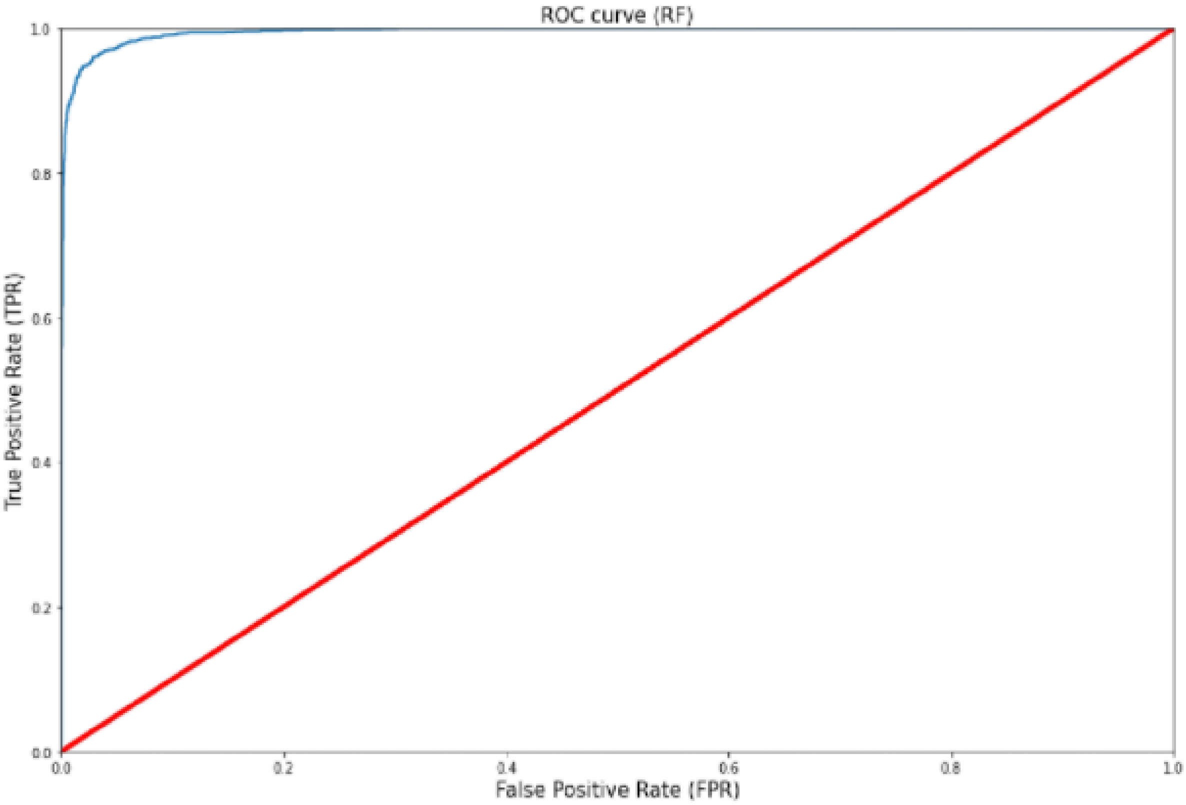
RandomForestClassifier(bootstrap=False, max_depth=32, max_features='sqrt', n_estimators=1000, random_state=42)

These optimal parameters resulted in an accuracy improvement of 0.4% when compared to the Random Forest model with default hyperparameters (96.42% vs. 96.02%). The recall score also improved from 95.46% to 96.12%.



We can see that a customer holding a Silver credit card (or not) is an overly important indicator of whether they churn on their card.

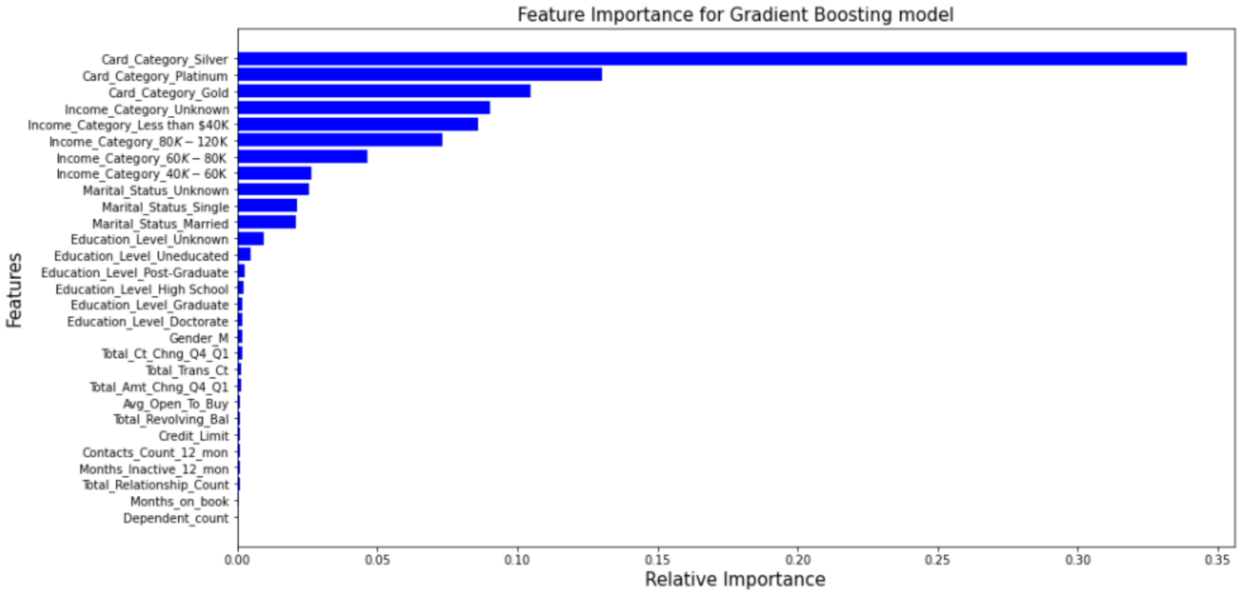
This model also has very high ROC-AUC and F1 scores of 0.995 and 0.964, respectively. Below is the ROC curve of the model.

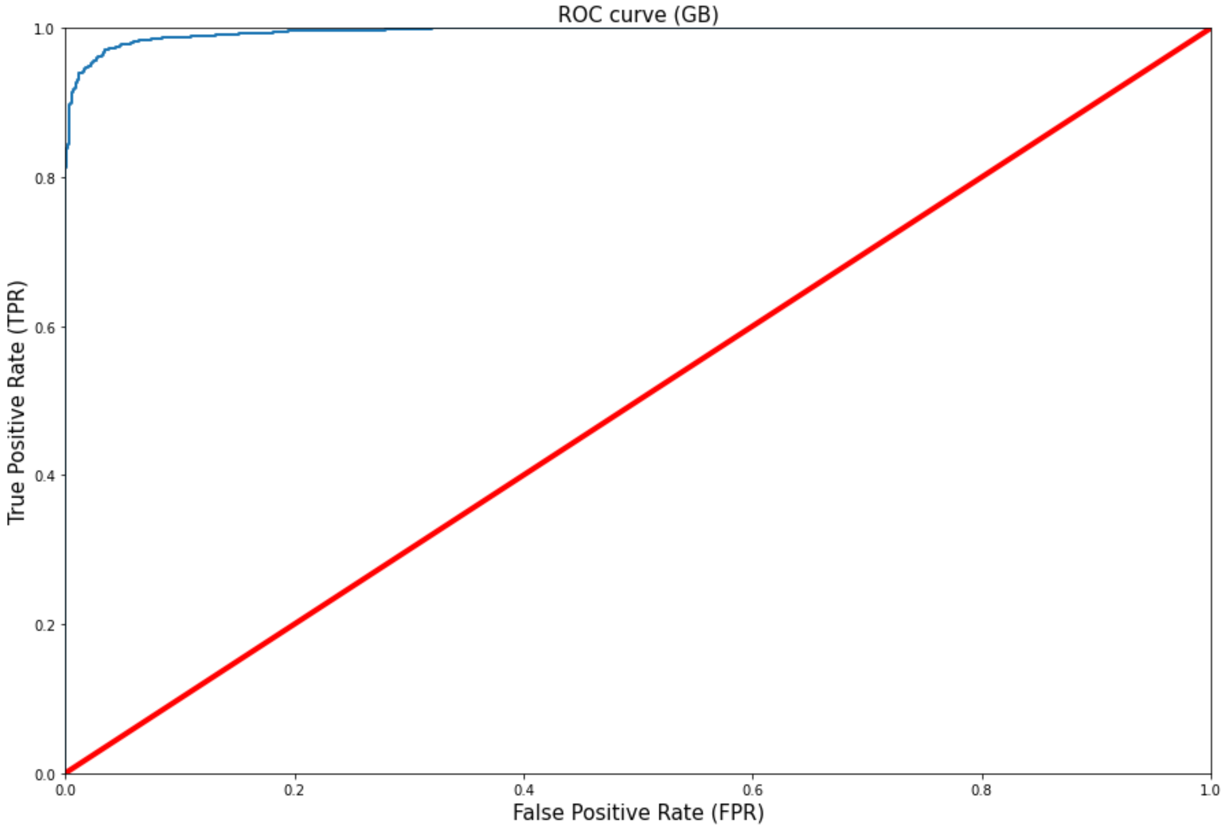


For the Gradient Boosting model, we used the following hyperparameters:

```
GradientBoostingClassifier(n_estimators=200, min_samples_split=2,  min_samples_leaf=1,
                           max_features='sqrt', max_depth=8, learning_rate=0.1)
```

These optimal parameters resulted in an accuracy improvement of 0.9% when compared to the Random Forest model with default hyperparameters (96.6% vs. 95.7%). The recall score also improved from 94.84% to 95.98%.





Again, Card_Category_Silver leads in feature importance when trying to predict credit card churns. The ROC-AUC and F1 scores are also high: 0.9952 and 0.9657, respectively.

5. Conclusion

In order to predict the credit card churning rate, here we have considered a bunch of 29 features engineered from the original dataset. From the feature importance graphs, we can see that a bank customer possessing a silver credit card is an overly important feature to predict whether he/she churns on the credit card or not.

This is a classification problem. Here we have used the following classification models:

- Logistic Regression
- K-Nearest Neighbor (KNN)
- Support vector machine (SVM)
- Random Forest
- Gradient Boost

We have evaluated each models in terms of model accuracy score, and 'ROC-AUC' score for both the training and test data, and plotted them. The two best performing models are Random Forest and Gradient Boosting. Both are ensemble models, based on decision trees.

Using optimal parameters for our Random Forest and Gradient Boosting models, we can see that Gradient Boosting is a little bit more accurate than Random Forest (GB 96.6% vs RF 96.4%), but the recall score of Random Forest is higher than Gradient Boosting (recall GB 95.7% vs. RF 96.1%). Since we need to focus on keeping the number of false negatives (i.e. the number of people predicted as "not churn" but in fact churn on their credit cards) as low as possible, we will select the Random Forest model in this case for future predictions.