PREDICTION MODEL FOR CREDIT CARD CUSTOMER CHURN

INTRODUCTION

Problem Statement

What opportunities exist for the bank to improve its credit card user's retention rate through predicting churning customers, thus changing their decision by proactively providing them with better services.

Context

Credit card customer churn is a common problem since many card issuers offer attractive sign-up bonuses and rewards to get consumers to apply for new cards and spend a large amount right after account opening. Nowadays, customers can easily apply for new credit cards online in just a few minutes, which creating more challenges for the bank to retain their current users. With increasing numbers of new credit cards with attractive sign-up benefits, it can drive customers to switch to get a new credit card more often.

The bank experienced an increasing number of customers leaving their credit card services. Management wants to be able to predict who is going to get churned so that they can proactively go to the customers to provide them better services and turn customers' decisions in the opposite direction. In addition, they realized that the traditional approach of retaining a credit card customer needs to change. They have decided to study all the other features of a user to help them identify a churn credit card user more effectively.

Objective

In this project, we will perform an in-depth Exploratory Data Analysis that can help visualize what are the different attributes between existing and churning customers. Most importantly, we will build a predictive model, which can identify credit card customers who are getting churned. In addition, we will also need to determine some of the most influential factors that can lead to a customer's decision of leaving their credit card issuer. Our goal for this project is creating a model that can successfully predict 80% of churning customers.

DATASET

The “BankChurners” dataset was obtained from Kaggle. The dataset consists of 10,000 customers mentioning their status, which are either Existing Customer or Attrited Customer. The dataset also provided 19 independent features of each customers that can be utilized in predicting their status.

These are the features that are relevant for the project:

- Attrition_Flag: If the account is closed then 1 (Attrited Customer) else 0 (Existing Customer)
- Customer_Age: Customer's age in years
- Gender: M=Male, F=Female
- Dependent_count: Number of dependents
- Education_Level: Educational qualification of the account holder
- Marital_Status
- Income_Category: Annual income category of the account holder
- Card_Category: Type of card (Blue, Silver, Gold, Platinum)
- Months_on_book: Period of relationship with bank
- Total_Relationship_Count: Total number of products held by the customer
- Months_Inactive_12_mon: Number of months inactive in the last 12 months
- Contacts_Count_12_mon: Number of contacts in the last 12 months
- Credit_Limit: Credit limit on the credit card
- Total_Revolving_Bal: Total revolving balance on the credit card
- Avg_Open_To_Buy: Open to buy credit line (average of last 12 months)
- Total_Amt_Chng_Q4_Q1: Change in transaction amount (Q4 over Q1)
- Total_Trans_Amt: Total transaction amount (Last 12 months)
- Total_Trans_Ct: Total transaction count (Last 12 months)
- Total_Ct_Chng_Q4_Q1: Change in transaction count (Q4 over Q1)
- Avg_Utilization_Ratio: Average card utilization ratio

DATA WRANGLING

Data shape

The dataset contains 10,127 rows and 23 columns. The “CLIENTNUM” column was deleted because they do not contain any useful information to help predict the churn customers. In addition, the author of the dataset suggested to remove two columns named “Naive_Bayes_Classifier”. The final shape of the dataset is 10,127 rows and 20 columns, which included a dependent variable and 19 independent variables.

Data cleaning

The dataset was cleaned and only required a little cleaning effort. There was inconsistent format in the “Income_Category” column so we removed the “$” symbol to make the value in the column consistent with each other.

Missing values:

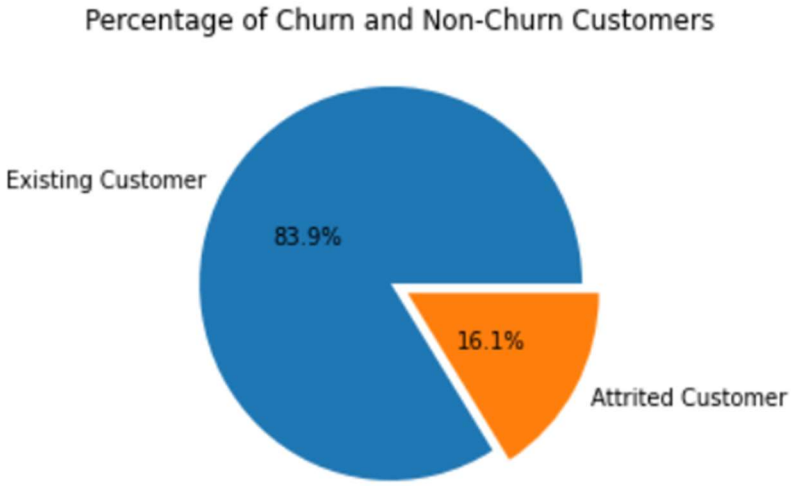
The dataset has no missing values in any columns. However, by examining the number of unique values in these columns, three columns including “Education_Level”, “Marital Status” and “Income_Category” contain many observations with the string value “Unknow”. These values have no useful information at all. Therefore, these values were converted to null values in this stage. It appears that these values were missing randomly in the dataset. However, we will address the null values after completing Exploratory Data Analysis stage, where we will explore the dataset to get more understanding about the null values.

Data distribution:

We examined the distribution of features to get more understanding on whether the values look sensible and whether there are any obvious outliers to investigate. In this stage, we only focused on whether the distributions look plausible or wrong. By drawing histogram for all the numerical features, it appears that the distributions of these features were reasonable. Although there are clearly some skewed distributions and outliers, these values are reasonable in the credit card industry.

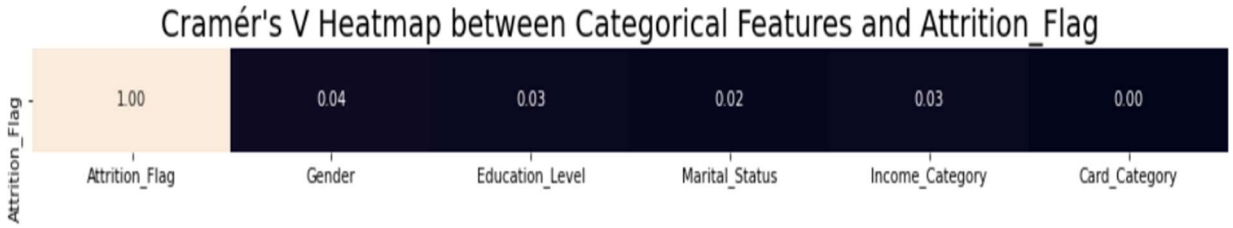
EXPLORATORY DATA ANALYSIS

Target Feature



As per the pie chart, only 16% of observations in the dataset are churn customers. This is an imbalanced dataset, which can pose classification problems where the classes are not represented equally. The imbalanced problem was expected since it is common in churn datasets, where the majority of customers stay with the business and a minority of users cancel their services. This problem may make a classification model to always predict a customer at one class and get an 84% accuracy score, whereas the model does not have any predictive power at all. We will address this issue in the preprocessing stage using Random Resampling techniques.

Categorical features

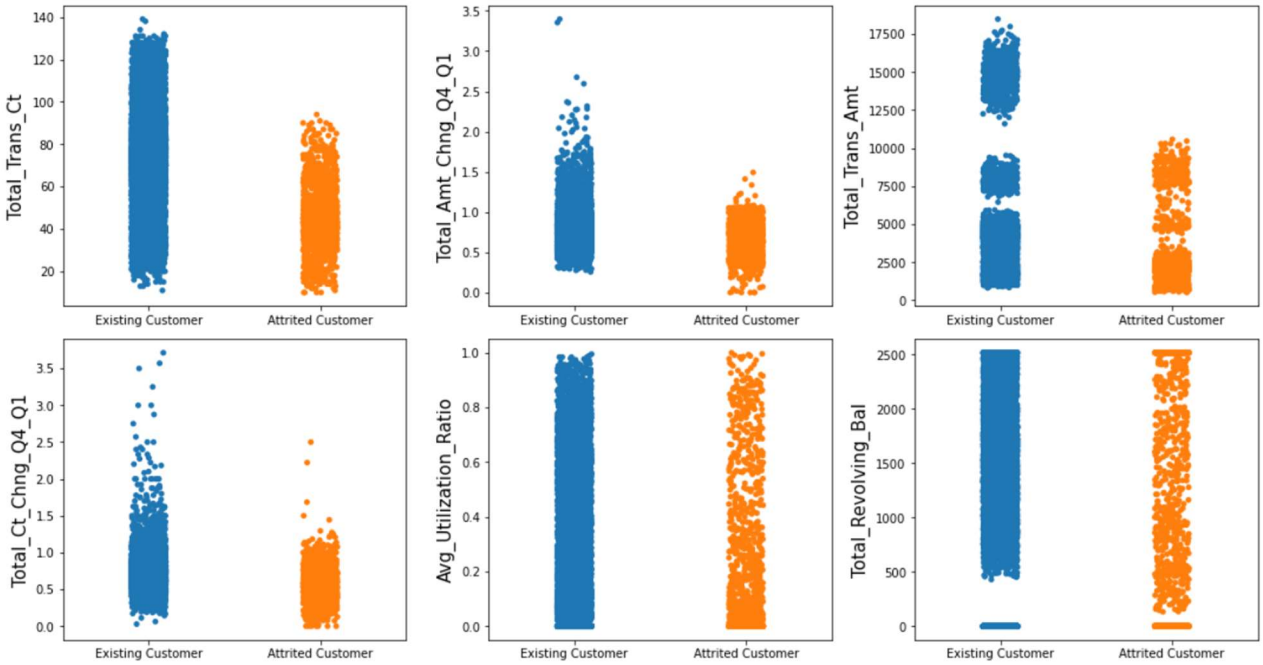


Cramér's V is a measure of association between two nominal variables, giving a value between 0 and 1. Similarly to correlation, the output is in the range of 0 and 1, where 0 means no association and 1 is full association.

The heatmap shows that churn or non churn customers have insignificant association with their Gender, Education_Level, Marital_Status and Income_Category. In addition, there was no relationship between Card_Category and Attrition_Flag. To prove that, we performed a Chi-square test on Card _Category and Attrition_Flag, the p-value was 0.53, so that we accept the hypothesis that these two features have no relationship to each other.

Numerical Features

Features with high correlation with Attrition_Flag



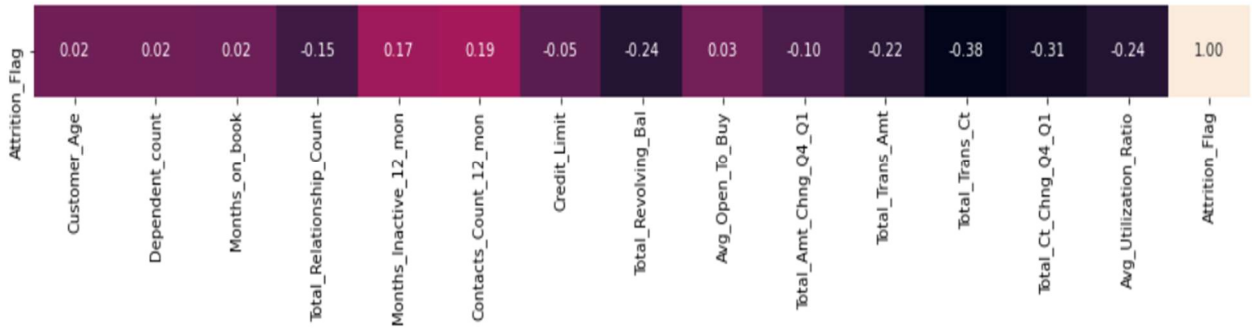
The strip plots show the different features' distribution between churn and non-churn customers. Apparently, churn customers have lower values in these features, which are Total_Trans_Ct, Total_Amt_Chng_Q4_Q1, Total_Trans_Amt, Total_Ct_Chng_Q4_Q1, Avg_Uitilization_Ratio and Total_Revolving_Bal.

There are no Existing Customers with Total_Revolving_Bal less than $429 (except for $0). It is interesting that if customers have Total_Revolving_Bal less than $429 (except for $0), they are extremely likely to churn.

Scatter plot of Attrition_Flag, Total_Trans_Ct and Total_Trans_Amt

As per the scatter plot, Existing Customers tend to spend more time and higher amount compared to Attrited Customers. Based on the plot, we can see that customers will stay with the services if their annual transaction count is above 100 and their annual transaction amount is above $12,500.

Heatmap between numerical features and Attrition_Flag



<u>Strong negative correlation:</u>

Total_Relationship_Count: Customers are less likely to churn if they have more products with the card issuers.

Total_Revolving_Bal: Customers, who keep their unpaid balance low, are more likely to churn.

Total_Trans_Amt, Total_Trans_Ct, Total_Ct_Chng_Q4_Q1: This is rational since customers with low usage have higher chance of churning.

Avg_Utilization_Ratio: This variable is also indicating customers with low usage rate are more likely to churn.

Strong positive correlation:

Months_inactive_12_mon: This makes sense intuitively that the more inactive months within a year, the more likely customers cancel their services.

Contacts_Count_12_mon: This is rational since customers usually call due to problems with their cards and services. The more contact counts, the more problems customers have with their credit cards.

PRE-PROCESSING

The missing values in categorical columns like Education_Level, Marital_Status and Income Category were filled with a new class - "Other". As per finding in the Exploratory Data Analysis, all categorical features have insignificant association with churn or non-churn customers. Therefore, imputing missing values with any information would not have any noticeable effect to the outcome of prediction model.

In addition, Label Encoder was used to transform all categorical features into numerical values so that the dataset can be fit into a predictive model. Each categorical features' values were transformed to values between 0 – n-1 (n is the number unique value in each feature).

Next, we split the dataset into 2 sets: train set and test set so that we can fit a predictive model to the train set and still have the test set to evaluate model performance. By doing that, we can avoid overfitting the model to the whole dataset.

As mentioned in the Exploratory Data Analysis, the dataset is imbalanced. Therefore, two Random Sampling techniques, including Random Oversampling and Random Undersampling were applied to the dataset to address the imbalanced problem. In this dataset, SMOTE performs better when combined with undersampling of the majority class, such as random undersampling. We oversampled the examples in the minority class, then undersampled the examples in the majority class in the training dataset. This can balance the class distribution but does not provide any additional information to the model. As result, we generated a train set which includes 1,784 observations for churn customers and 1,784 observations for non-churn customers.

As the features measured in many different units, with numbers that vary by orders of magnitude, we applied StandardScaler to scale each feature to zero mean and unit variance so that all features were on a consistent scale.

MODELING

The following classification models were applied to the dataset:

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

We will use the cross-validation technique to evaluate the performance of each model. In addition, each model will be assessed by several evaluation metrics for performance review. The evaluation metrics includes accuracy, precision, recall, confusion matrix.

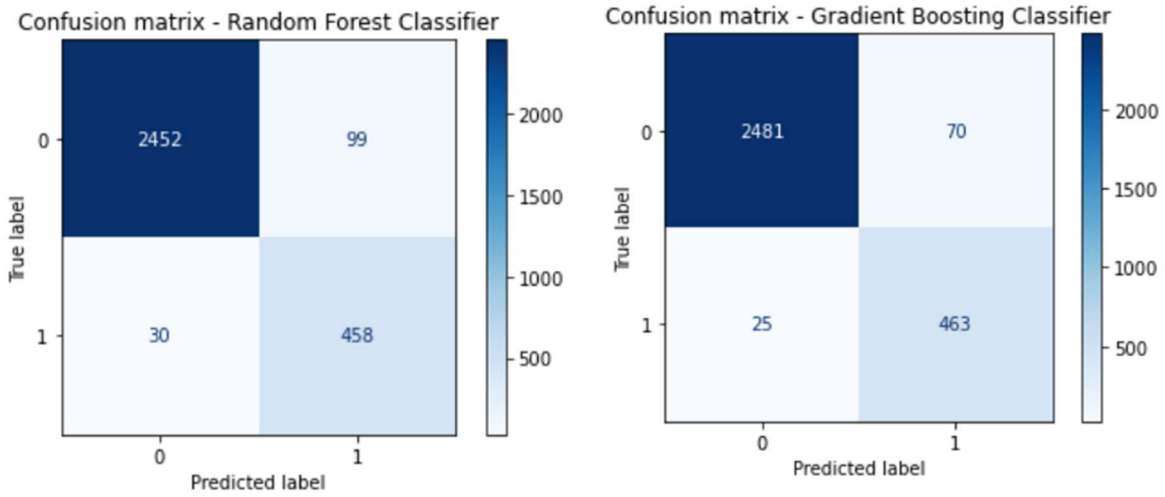
Our top priority in this business problem is to identify customers who are getting churned. Even if we predict non-churning customers as churned, it will not harm the business. But predicting churning customers as Non-churning will do. Therefore, we aimed to create and use a model with the highest recall score.

Model Comparison

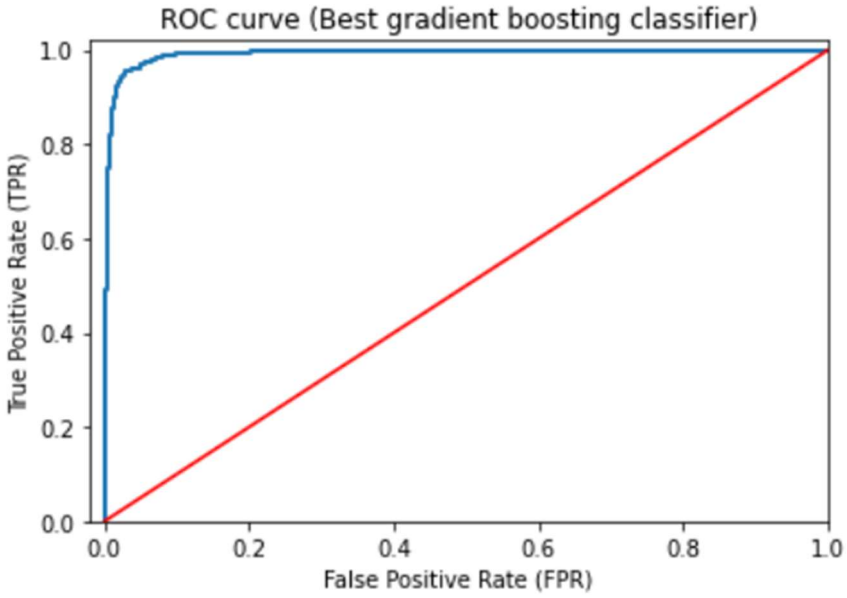
	Model	Accuracy	Precision	Recall	ROC_AUC
0	KNN	0.84	0.49	0.84	0.91
1	LogisticRegression	0.85	0.52	0.84	0.92
2	SVM	0.89	0.62	0.87	0.96
3	RandomForestClassifier	0.96	0.82	0.94	0.99
4	GradientBoostingClassifier	0.96	0.82	0.95	0.99

Random Forest Classifier and Gradient Boosting Classifier have the best performance. Both models have accuracy scores of 96%. In addition, the models have good recall scores of 94%. Although precision scores

are only 82%, the most important metric for this project is recall, which helps the business to identify all potential churn customers.

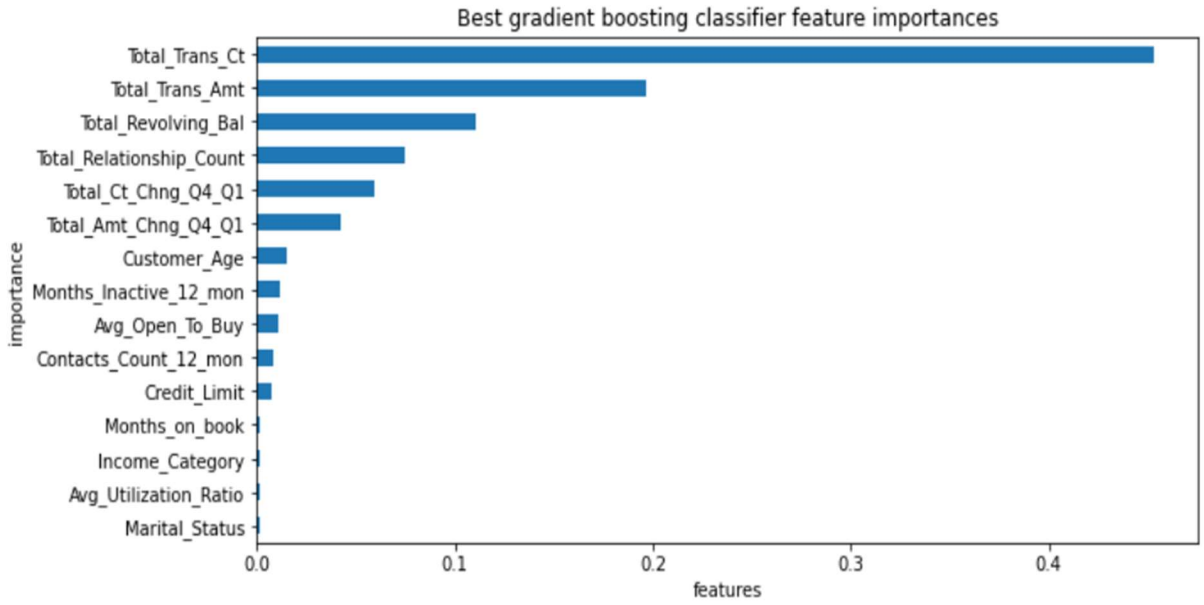


As per the confusion matrix, both models performed very well in predicting churn customers. The Gradient Boosting model has a slightly better performance, which successfully identify 463 churn customers, only 25 churn customers have the wrong label.



Gradient boosting model also has a 99% score in Area under curve of ROC. The score indicates Gradient Boosting Classifier is an excellent predictive model for this dataset.

Therefore, we will choose Gradient Boosting model as it is the model with best performance. The predictive models can be implemented to help identify churn customers in the early stage, so that the business can proactively approach customers with better services turn customers' decisions in the opposite direction.



The above plot ranks the importance of each feature in predicting a churn customer. As expected in the Exploratory Data Analysis, the most important features to predict churn customers are: Total_Trans_Ct, Total_Trans_Amt, Total_Revolving_Bal, Total_Ct_Chng_Q4_Q1, Total_Relationship_Count, Total_Amt_Chng_Q4_Q1.

The categorical features do not have much impact in predicting churn customers. Therefore, the business cannot use any categorical features like Gender, Education_Level, Marital_Status, Income_Category or Card_Category to predict churn customers since the categorical features have almost no values in identifying customers with high possibility of churning. This also implies that the business can only rely on actual annual transaction information to effectively predict churn customers.

FUTURE IMPROVEMENT

We can improve the precision score of the model so that it can have a higher chance of accurately predicting churn customers. We will need to collect more churn customer samples so that the model can be trained with a more balanced dataset.

This dataset contains only customers stayed with the business for 1 – 5 years. Therefore, this predictive model may only be applicable to predict churn customers int the same time range. We can collect more customer observations out of this range so that we can build a model that can perform well in different range.

Blue Credit Card Category is accounted for 93% of the total card holders in this dataset. We can use the transaction information to identify customers with high spending pattern and recommend upgrading their credit cards.