Capstone Project Report – Credit Card Customer Churn Prediction

Problem Statement

The credit card business in the bank possesses high risk and high profit. The existing customers leaving their credit card services are considered as churned customers. Customer Churn is one of the most important and challenging problems with banks. It is easier for banks to keep the existing customer than adding new ones. To support the bank, reduce churn rate, its crucial to predict customers that are high risk of churn.

The bank customer dataset was used to build machine learning model to find customers who are most likely to cancel the credit cards. This could help bank staffs to proactively follow up with customers to provide better services and turn their decision to keep them with the bank.

Data Preparation

The bank customer dataset was extracted from Kaggle dataset: Credit Card churn. This dataset contained 10127 records and 21 features about customers mentioning their age, salary, marital_status, credit card limit, credit card category, etc. This dataset didn't have any null values as this was obtained from the Kaggle and was already clean which made the preparation task easier. Though, this is not the case with the real-world data.

The target variable was 'Attrition_Flag' and almost 84% of the record belonged to 'Existing customer' category and 16% was 'Attired customer' category. Since, most of the record belong to one category the dataset had imbalanced class problem which was handled before training the model.

Exploratory Data Analysis

In this dataset, there were 5 categorical features. They were Gender, Education_level, Marital_status, Income_category and Card_cateogry. The proportion of gender count is almost equally distributed (52.9% male and 47.1%). Female customers churned more than male. Almost 9% of churned customers were female and 7% were Male. In the Education_level category, customer you who were 'Graduates' churned more than others. A high proportion of Education_level of attired customer was Graduate level (31%). Some of the customers didn't give the information of their education background. Similarly, there was an unknow category in the Marital_Status as well. The distribution for cancellation is similar for single and married customers. While looking at the income group, the customers in the income group 40K – 60K cancelled card service more than other income groups. Figure -1 shows the distribution of the numerical features in the dataset.

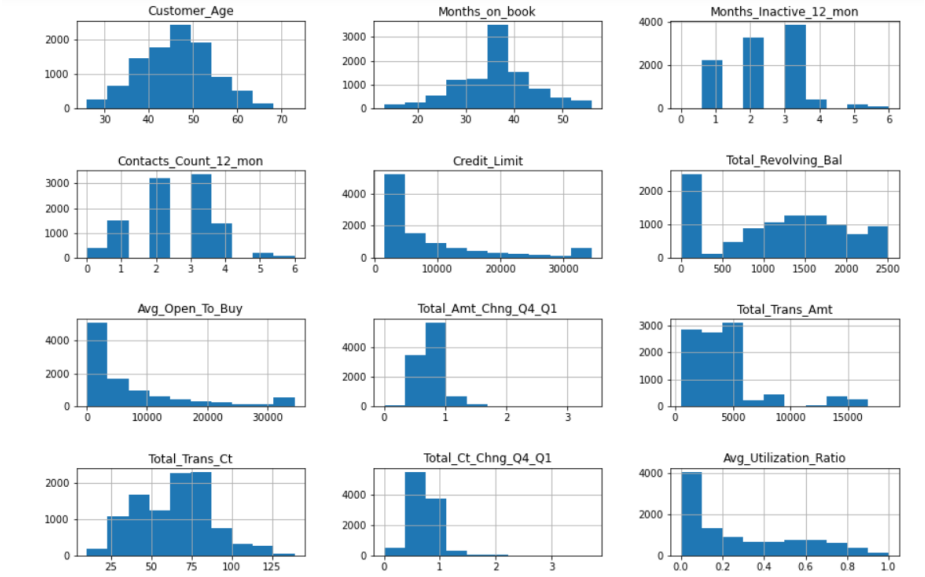

Figure 1

The Credit_Limit values were clustered toward lower end. The number of Months_inactive was between 1 to 6 months and the number of contacts counts was also in this range. The customer_age, total_trans_ct and Months_on_book were normally distributed.

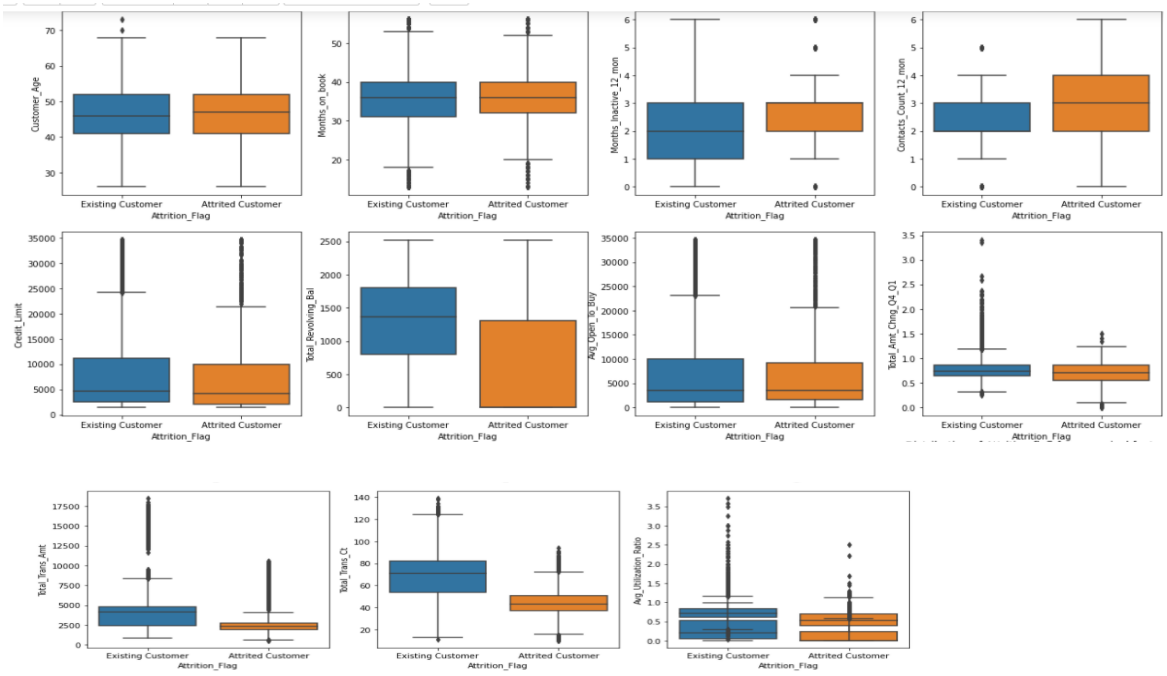

Figure 2

From the above box plot, credit_limit, Avg_open_to_buy, total_amt_change_Q4_Q1, total_trans_amt, utilization ratio all these columns had outliers in both categoies. The contact_count_12_mon had wide range of Attired customers, and it's seemed to be less for the existing customers. Months on book had similar distribution for both categories.

It is important to check if the variables in the dataset have any meaningful correlation between. Attrition_Flag had reasonable negative correlation with Total_trans_ct. The heatmap below (figure 3) didn't suggest any strong correlation between target and other features.

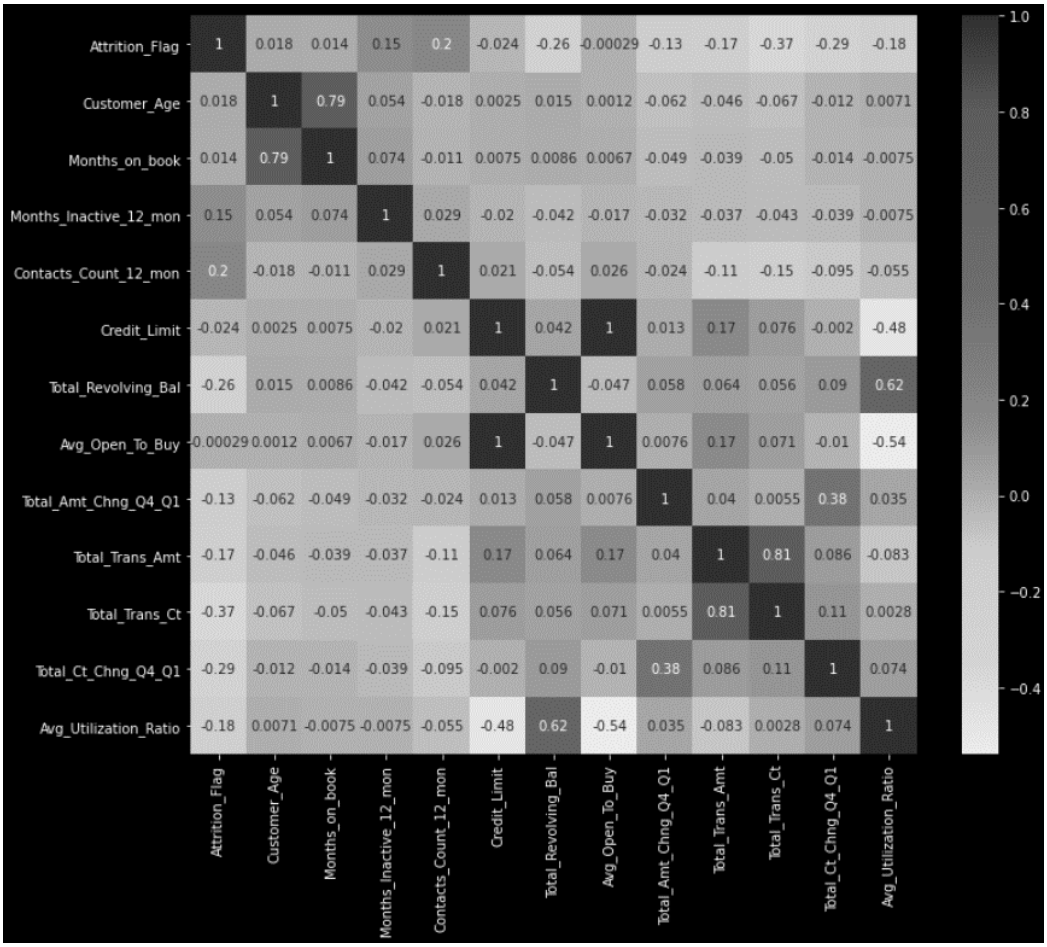

Figure 3

Data Preprocessing

The data set didn't have any missing values and didn't require any further transformation. Another important step was to drop feature that didn't add any value to modeling. In this case, 'CLIENTNUM' column was dropped. Dummy variable was created for the categorical variables Gender', 'Education_Level', 'Marital_Status', ' Income_Category', 'Card_Category'. Standardizing a dataset

involves rescaling the distribution of values so that the mean of observed values is 0 and the standard deviation is 1. All the variables were scaled using the StandardScalar method in scikit-learn. Then the data set was partitioned to 70% training and 30% testing set.

The next important step was to check the balance of target variable and this dataset had imbalanced class, i.e., almost 84% of the record belongs to Existing customer category and 16% belongs to the Attired customer category. Due this problem any machine learning model tend only to predict the majority class. In more technical words, if we have imbalanced data distribution in our dataset then our model becomes more prone to the case when minority class has negligible or very lesser recall. SMOTE (Synthetic Minority Oversampling Technique) – Oversampling is one of the common methods for oversampling the minority class. It generates virtual records of the minority class selecting one or more K-nearest neighbors for each example of the minority class.

SMOTE method was used to resample the training dataset so that we have equal records for both 'Existing' and 'Attired' customers. After performing this step, the next step was developing different machine learning classifiers on the training dataset.

Modeling

The Logistic Regression classifier was as a base line model and various machine learning models were developed and results were compared to choose the best model. In Logistic Regression, C is the regularization strength (equivalent to $1/\alpha$ from the Ridge case), and smaller values of C mean stronger regularization. The regularization tries to prevent features from having terribly high weights, thus implementing a form of feature selection. Various values for C (0.01, 0.1, 1, 10, 100) were passed to GridSearch to select the best Logistic Regression model. The optimal value of C was '0.1' with a model accuracy of 84%. The AUC – ROC curve is a performance metric for classification problems. Basically, it tells how much the model could distinguish between 0 and 1. The higher value of AUC, the better model that predicts 0s as 0 and 1s as 1. The ROC curve was plotted between True Positive Rate and False Positive Rate. The AUC score for Logistic Regression was 0.84. Then, hyper-parameter tunning was carried using GridSearch CV for various model such as RandomForest, Gradient Boosting and XGBoosting to find the best hyper-parameter for each model. All the other models performed better the Logistic Regression.

The table below shows the Accuracy and AUC score for different models.

	classifiers	Accuracy	AUC
0	LogisticRegression	84.63	0.840552
1	RandomForest	95.39	0.916896
2	GradientBoosting	96.31	0.920945
3	XGBoost	96.58	0.939196

The AUC score of 3 out of 4 model were above .91 as they have similar performance in predicting 0 and 1. The XGBoost classifier had the best performance with AUC score of almost .94. The comparison of the ROC curve for all the models were shown in figure-4 below.

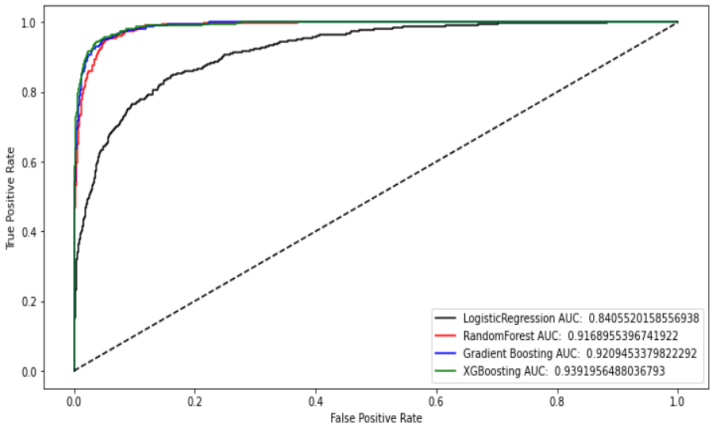

Figure 4

Feature Importance from XGBoost

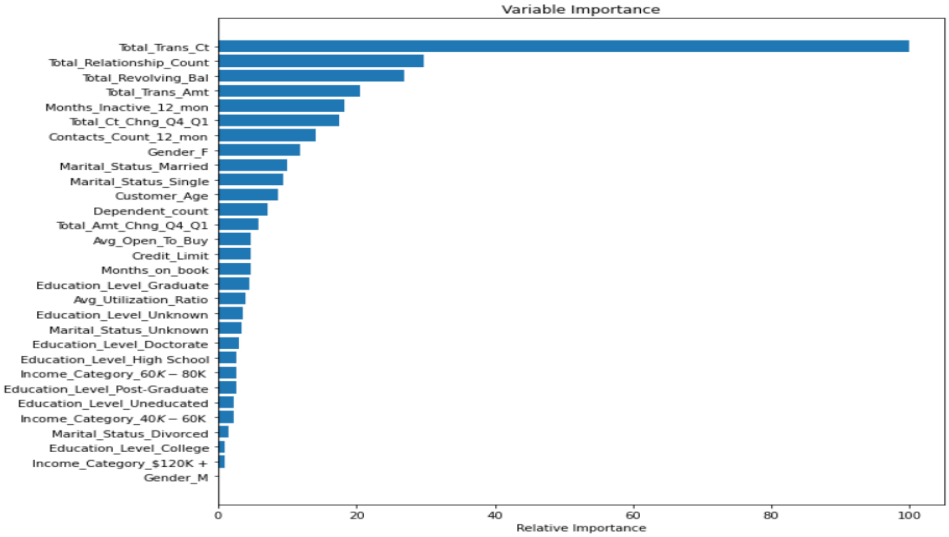

Figure 5

XGBoost had a feature that allows to select the important feature in the dataset and the above figure-5 visualizes the contribution of each variable. The total_Trans_ct almost contibuted 100% and Total_relationship_count was close to 40%. These were the variables that contributed the most in classifying the dataset. The other variables like income_category, Marital_status, Eduction_level didn't contribute to the model.

The Total_tran_ct was one of important feature from all the models. If a customer had not transacted in while with less total_trans_ct are more likely to chrun. Then, Total_relationship_count was another important variable. It would be to better to proactively reach out to the customer to improve a personalize customer experience.

Conclusion

In this project, the main foucs was to give the bank with knowledge regarding their customers. This predictive model was not considered as the final solution. Instead, the bank staff could use the information to make infomed decision about their customers. The bank could collect more information about the customers as how long they are with the service and satisfaction survey about the service. These features could help improvise the model and could gain futher insights about the customers.