COMP5310 Principles of Data Science

Assignment 1 Report - Credit Card customers (Predict Churning customers)

Student name: Xinyi Cui	Student name: [redacted]
Student ID: [redacted]	Student ID: [redacted]
Unikey: [redacted]	Unikey: [redacted]

Problem

Customer churn means that customers terminate various business with the bank through credit cards. There is no doubt that a certain amount of customer loss will bring huge losses to the bank. Given that the cost of avoiding a customer loss is likely to be much lower than the cost of finding a new one, the analysis and prediction of customer loss is critical.

The analysis model of bank customer loss can provide effective early warning for bank business. Marking customers who are prone to churn can make banks more targeted in customer management through analysing the causes for leaving, improve customer retention rate, and reduce the loss caused by customer churn.

In stage 2 of the project, we will establish the classification model of credit card customers and adopted decision tree and logistic regression algorithm for the user loss warning model, so that we can analyse the causes for losing customers. The following answers are that we want to solve:

- First, we also evaluate the customer value through analysing features in pie charts or histograms, which customers are worth retaining or not.
- Secondly, we analyse the warning signs of churn, which customers are about to leave for several factors.
- At the same time, our business objective is to analyse the reasons why customers give up, predict the probability of customer churn and the value of customer retention.

Approach

Only 16% of data represented churn customers, so we used SMOTE to sample the churn to match the sample size of our regular customers. This will solve the problem of data imbalance. It changes the data distribution of an unbalanced dataset by adding the generated few class samples. After that, we will use PCA (principal component analysis) to reduce the dimension of the variables, thereby reducing the variance. We compared the F1 score by the random forest, Adaboost and SVM models.

For clustering, we can extract those customers who deserve retaining. For example, the type of high value and low card cancellation probability: a customer with $80K-$120K Income, post graduate education degree, holding a Blue card. There is a high possibility for him to upgrade to Gold card, which is a high value customer.

Data

Method of obtaining data

The selected data is titled "Credit Card customers", which was downloaded from the Kaggle website with the URL as, https://www.kaggle.com/sakshigoyal7/credit-card-customers.

Once unzipping from the zip file, the folder contains a CSV file with a total size of 2.4MB. The actual data set consists of 23 columns and 10128 rows of information. The dataset records 10,127 customers, of which only 16.07% of customers who have churned, and the remaining customers are still existing customers. Each customer's information description mentions their age, salary, marital status, credit card limit, credit card category, etc. There are nearly 19 functions.

Description of data from a general perspective

According to the supporting information (Credit Card customers | Kaggle, 2020), it is recommended to ignore the last two columns of Naive Bayes Classifier before processing the dataset. The information provided in the table is mainly composed of three types of variables, which are personal information characteristics of the account holder (6 columns) and type of credit card (1 column) and credit card usage situations (12 columns). For some variables that will be analysed graphically, the data was cleaned to remove nulls. The full list and explanation of the tables is provided in Appendix 1.

Exploratory analysis we have done

Based on their personal information and credit card usage situations, we want to infer the potential reason why they want to leave the credit card service . Of particular interest was data pertaining to total transfer amount and total transfer count on the credit card relating to customer activities shows as a scatter graph in Appendix 2.

In the exploratory analysis stage, we plot distribution of data for different variables relating to customer activities, one examples of histograms is shown in Appendix 3. A histogram of credit card holder income category is generated, the number of people in the interval of less than 40k is the largest and does not show a normal distribution. It is also necessary to analyse other variables such as usage variables to figure out what the real potential influencing factors are.

Description of some data preparation

To facilitate subsequent processing, rearrange the fields, we will place the targeted variable in the first column, and remove the useless fields. The other fields can be divided into classification variables and numerical variables, and the transformation and other operations can be carried out in the data processing.

New feature column is created when its necessary, e.g., the percentage of income a customer deposits with the bank are also an important character to consider, so a new feature is added, which is 'the deposit-income ratio', the higher the deposit-income ratio, the easier it is to lose.

Description of the tools used to clean and explore the data set

Data preparation is done by transferring the CSV files. In the cleaning stage, we handled missing values and outliers, for example, if data does not need to deal with missing values, omitted. In the exploratory analysis stage, we plotted some figures/tables by using library tool such as 'plotly' and 'scikit-plot' in Python, which enables us to have a better understanding in distribution of data for different variables relating to customer activities.

Proposal

Our goal for stage 2 of the project is to analyse, visualize and quantify the possible relationship. Variables involved in the relationship analysis are roughly divided into three types: demographic variable and product variable and usage variable, as follows:

- Customers activity versus demographic variable (personal information characteristics of the account holder) which including age, gender, educational level, etc.
- Customers activity versus product variable (type of credit card)
- Customers activity versus usage variables in the last 12 months such as period of relationship with bank, total number of products held by the customer, credit limit, etc.

Depending on the outcomes above, visualizing main distribution of provided information. Through the corresponding information left by the customer, it would be reasonable to find out reason why customer wants to leave the credit card service and leverage the same to predict customers who are likely to drop off.

To be able to extract data set efficiently, it is supposed to use SQL querying statements, statistic functions, then implement Python for in-depth data analysis and visualization. If strong relationships are quantifiable between customers activity and other variables, it is proposed to develop a data model to predict churning customers and then test with the data set.

Conclusion

After the model is established, we can understand which customers are more likely to lose in the future and the value of their potential customers. What we need to do is to identify the worthiest customers to retain and find the key factors for losing customers.

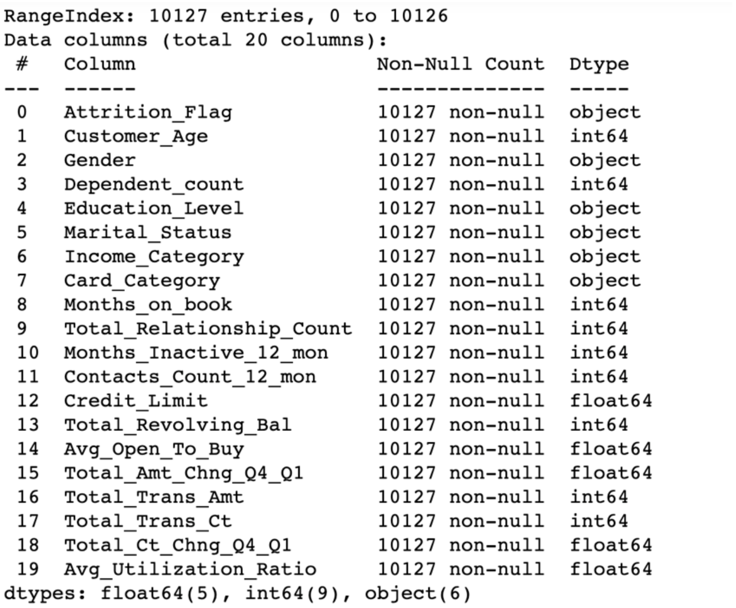
References

1. Kaggle 2020, Credit Card customers, Kaggle, viewed 10 April 2021, https://www.kaggle.com/sakshigoyal7/credit-card-customers
2. Catboost Classifier w Feature Import+Vizualization, Kaggle, viewed 14 April 2021, https://www.kaggle.com/dmitriyveselov/catboost-classifier-w-feature-import-vizualization
3. Bank credit card customer churn forecast (Kaggle), https://zhuanlan.zhihu.com/p/33640276
4. Predicting credit card customer churn in banks using data mining, Dudyala Anil Kumar, Vadlamani Ravi profile imageV. Ravi, 2008, https://dl.acm.org/doi/10.1504/IJDATS.2008.020020
5. Developing a prediction model for customer churn from electronic banking services using data mining, Abbas Keramati, Hajar Ghaneei & Seyed Mohammad Mirmohammadi, 2016, https://jfin-swufe.springeropen.com/articles/10.1186/s40854-016-0029-6
6. Behavioral attributes and financial churn prediction,Erdem Kaya, Xiaowen Dong, Yoshihiko Suhara, Selim Balcisoy, Burcin Bozkaya & Alex “Sandy” Pentland, 2018, https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-018-01655
7. Developing a prediction model for customer churn from electronic banking services using data mining,Abbas Keramati, 2016, https://www.researchgate.net/publication/306388481_Developing_a_prediction_model_for_customer_churn_from_electronic_banking_services_using_data_mining

Appendices

1. Data table summary
2. Aggregate data visualisation (customer activities, total transfer amount and total transfer count on the credit card)

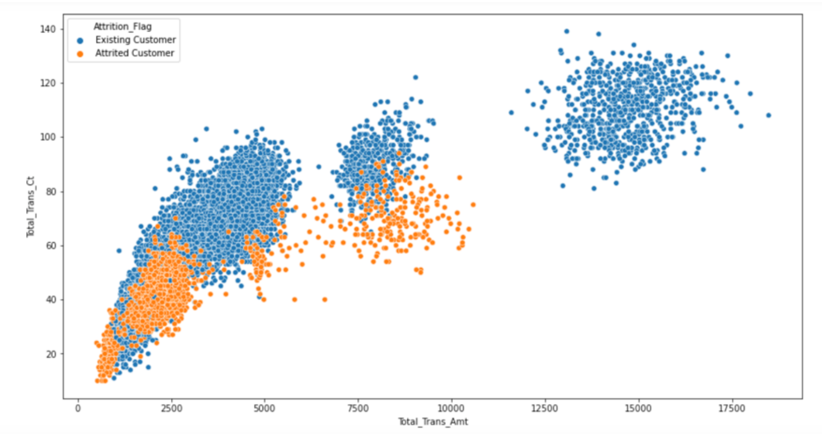
Appendix 1 - Data table summary



```
RangeIndex: 10127 entries, 0 to 10126
Data columns (total 20 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Attrition_Flag            10127 non-null  object
 1   Customer_Age              10127 non-null  int64
 2   Gender                    10127 non-null  object
 3   Dependent_count           10127 non-null  int64
 4   Education_Level           10127 non-null  object
 5   Marital_Status            10127 non-null  object
 6   Income_Category           10127 non-null  object
 7   Card_Category             10127 non-null  object
 8   Months_on_book            10127 non-null  int64
 9   Total_Relationship_Count  10127 non-null  int64
 10  Months_Inactive_12_mon    10127 non-null  int64
 11  Contacts_Count_12_mon     10127 non-null  int64
 12  Credit_Limit              10127 non-null  float64
 13  Total_Revolving_Bal       10127 non-null  int64
 14  Avg_Open_To_Buy           10127 non-null  float64
 15  Total_Amt_Chng_Q4_Q1      10127 non-null  float64
 16  Total_Trans_Amt           10127 non-null  int64
 17  Total_Trans_Ct            10127 non-null  int64
 18  Total_Ct_Chng_Q4_Q1       10127 non-null  float64
 19  Avg_Utilization_Ratio     10127 non-null  float64
dtypes: float64(5), int64(9), object(6)
```

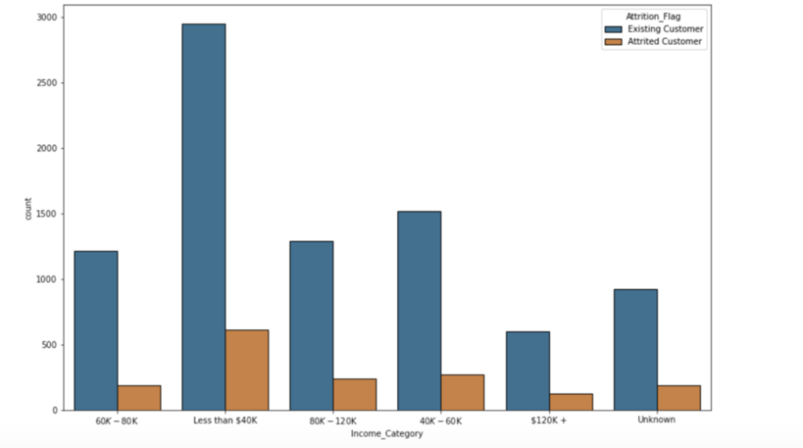
Appendix 2 - Aggregate data visualisation

The scatter graph in the figures below show distribution of data for total transfer amount and total transfer count on the credit card relating to customer activities.



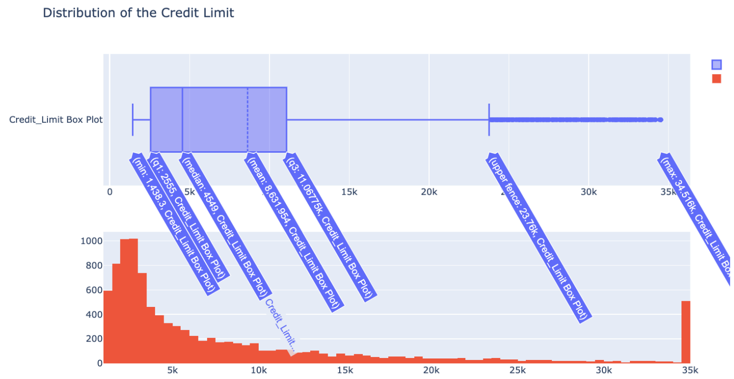
Appendix 3 - data visualisation using histogram

The histogram below shows distribution of data for Income category relating to customer activities.



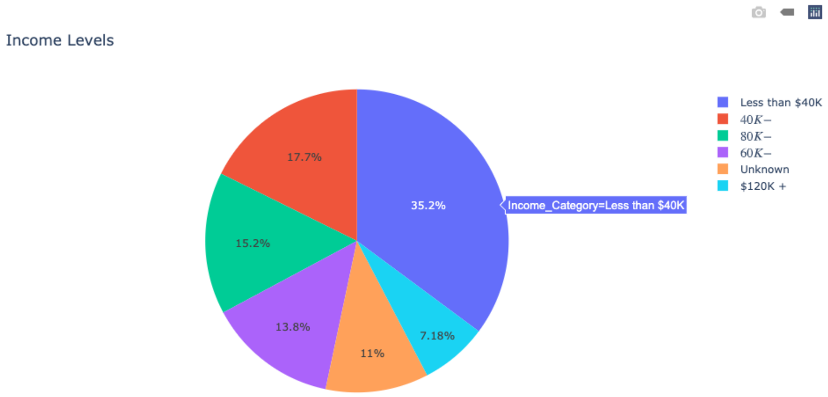
Appendix 4 Distribution of total transaction volume of customers

The distribution of total transaction volume shows the distribution of "multiple groups". If we cluster customers into different groups according to this separation, the similarity between them can be obtained



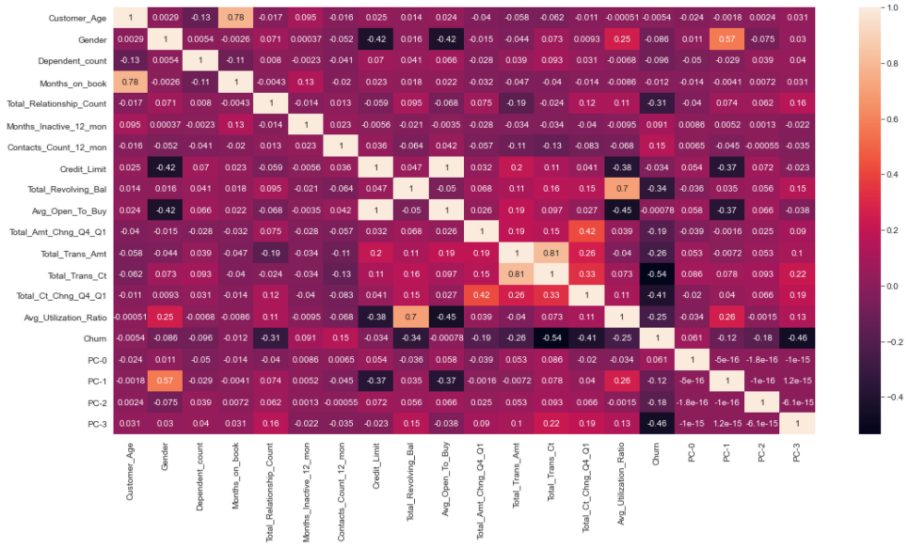
Appendix 5 The educational level of the client

More than 70% of customers have formal education, which is about 35% for master's degree or above, and 45% for a bachelor's degree or above.

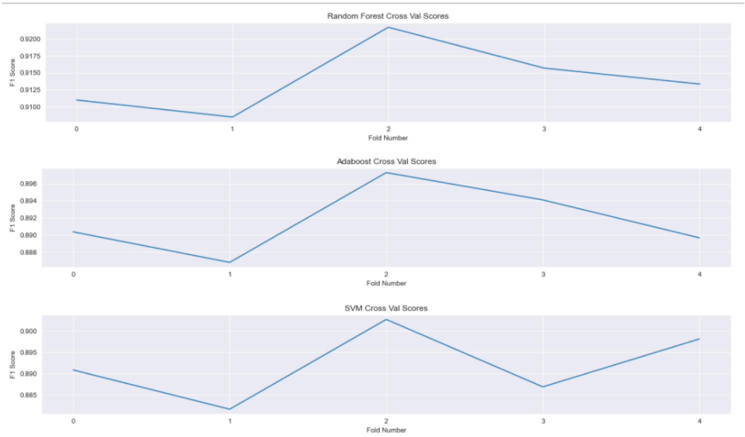


Appendix 6 heat map

Heat map shows correlation between features.



Appendix 7 Random forest, Adaboost and SVM models



The F1 score of random forest was the highest, reaching 0.92