Exploratory Data Analysis Project

Zixin Huang

July 3, 2021

1 Data

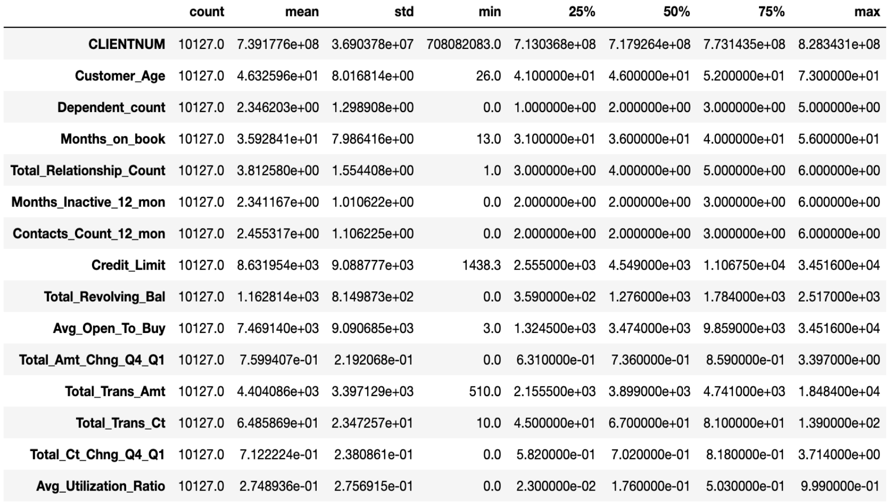
The data[1] contains customer information related to credit cards of a bank. The business problem is to predict who is going to get churned. The dataset contains 10127 observations and 21 features[2] such as customer age, gender, education, credit limit, total transaction amount. More detail of the columns are shown in Table 1 in Appendix. There are 6 categorical features and 15 numerical features. There are only 16% of the customers who are labeled "Attrited" and there is no missing value in the data.

2 Data Exploration and Data Cleaning

2.1 Numerical Features

The summary of all the numerical features in the data is shown in Figure 1 below. The table indicates that most of the features have a wide range and large standard deviation. Some features, such as total amount change and credit limit, have significantly different magnitudes.

Figure 1: Summary Statistics of Numerical Features



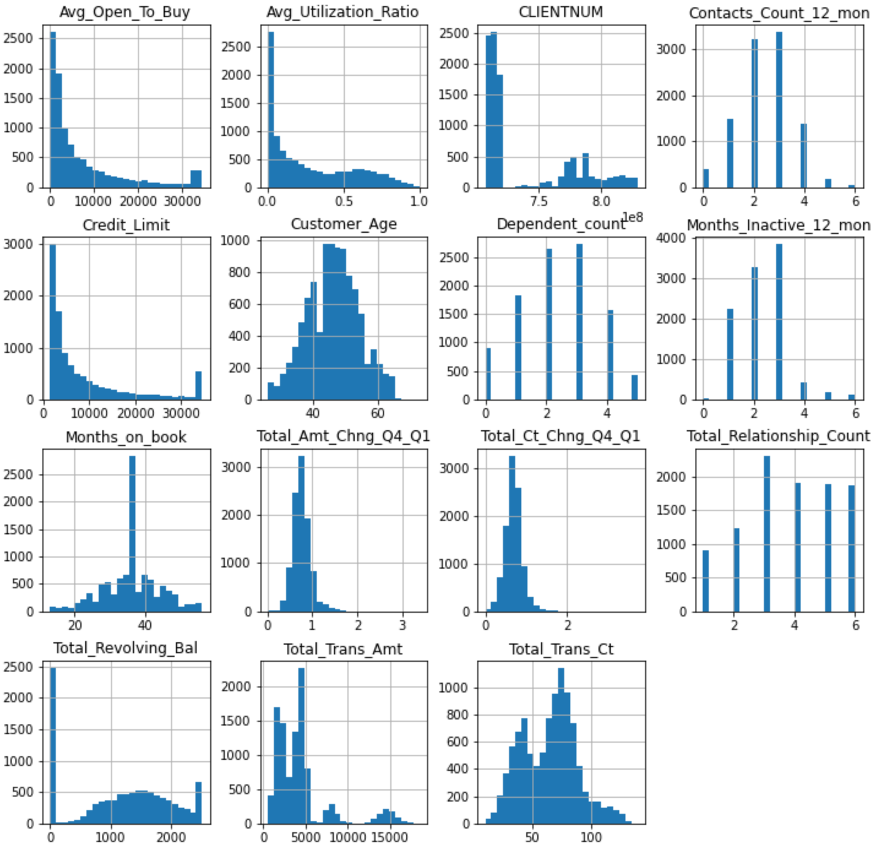
	count	mean	std	min	25%	50%	75%	max
CLIENTNUM	10127.0	7.391776e+08	3.690378e+07	708082083.0	7.130368e+08	7.179264e+08	7.731435e+08	8.283431e+08
Customer_Age	10127.0	4.632596e+01	8.016814e+00	26.0	4.100000e+01	4.600000e+01	5.200000e+01	7.300000e+01
Dependent_count	10127.0	2.346203e+00	1.298908e+00	0.0	1.000000e+00	2.000000e+00	3.000000e+00	5.000000e+00
Months_on_book	10127.0	3.592841e+01	7.986416e+00	13.0	3.100000e+01	3.600000e+01	4.000000e+01	5.600000e+01
Total_Relationship_Count	10127.0	3.812580e+00	1.554408e+00	1.0	3.000000e+00	4.000000e+00	5.000000e+00	6.000000e+00
Months_Inactive_12_mon	10127.0	2.341167e+00	1.010622e+00	0.0	2.000000e+00	2.000000e+00	3.000000e+00	6.000000e+00
Contacts_Count_12_mon	10127.0	2.455317e+00	1.106225e+00	0.0	2.000000e+00	2.000000e+00	3.000000e+00	6.000000e+00
Credit_Limit	10127.0	8.631954e+03	9.088777e+03	1438.3	2.555000e+03	4.549000e+03	1.106750e+04	3.451600e+04
Total_Revolving_Bal	10127.0	1.162814e+03	8.149873e+02	0.0	3.590000e+02	1.276000e+03	1.784000e+03	2.517000e+03
Avg_Open_To_Buy	10127.0	7.469140e+03	9.090685e+03	3.0	1.324500e+03	3.474000e+03	9.859000e+03	3.451600e+04
Total_Amt_Chng_Q4_Q1	10127.0	7.599407e-01	2.192068e-01	0.0	6.310000e-01	7.360000e-01	8.590000e-01	3.397000e+00
Total_Trans_Amt	10127.0	4.404086e+03	3.397129e+03	510.0	2.155500e+03	3.899000e+03	4.741000e+03	1.848400e+04
Total_Trans_Ct	10127.0	6.485869e+01	2.347257e+01	10.0	4.500000e+01	6.700000e+01	8.100000e+01	1.390000e+02
Total_Ct_Chng_Q4_Q1	10127.0	7.122224e-01	2.380861e-01	0.0	5.820000e-01	7.020000e-01	8.180000e-01	3.714000e+00
Avg_Utilization_Ratio	10127.0	2.748936e-01	2.756915e-01	0.0	2.300000e-02	1.760000e-01	5.030000e-01	9.990000e-01

[1] Data Source: https://www.kaggle.com/sakshigoyal7/credit-card-customers?select=BankChurners.csv

[2] The original data has 23 columns, but we removed the last two columns initially as instructed by the provider of the data.

The distributions of these features are shown in Figure 2. These two figures indicate that there are potential outliers in the total amount change ratio and the total transaction count change ratio. The boxplot of these two features are shown in Figure 3.

Figure 2: Distribution of Numerical Features



In this project, we use the 1.5 IQR (Inter-Quartile Range) rule[3] to determine the outliers in these two columns. The IQR is the range between the first quartile (Q1) and the third quartile (Q3), and the upper and lower boundaries to identify outliers can be calculated by

$$lower = Q1 - IQR * 1.5$$

$$upper = Q3 + IQR * 1.5$$

We replace the data outside of this range by the lower and upper bounds. Figure 4 shows the boxplot after the values of outliers being changed. We also removed the client number column since it is not useful in our analysis.

[3] https://online.stat.psu.edu/stat200/lesson/3/3.2

Figure 3: Boxplots of Total Amount Change and Total Transaction Count Change

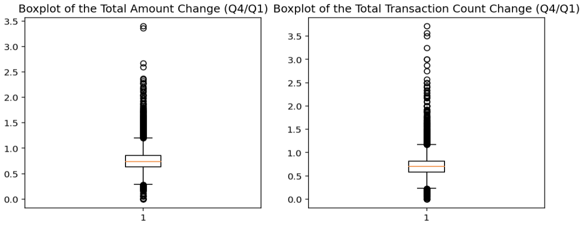
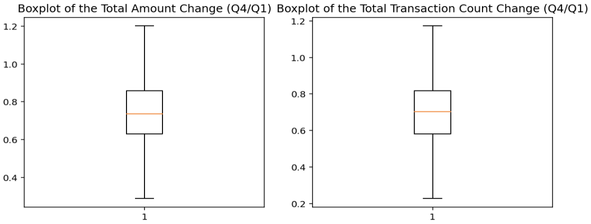
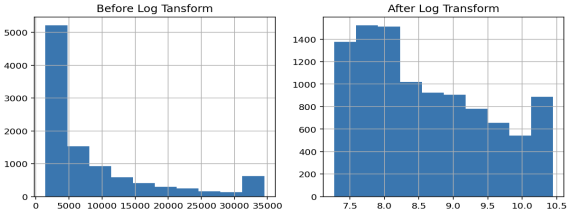


Figure 4: Boxplots of Total Amount Change and Total Transaction Count Change After Change Outliers



Since some of the distributions in Figure 2 appear to be right-skewed, we perform log transform on features that have skewness greater than 0.75. Those features are credit limit and average open to buy value. An example of before and after log transformation can be seen in Figure 5.

Figure 5: Distribution of Credit Limit Before and After Log Transformation



To visualize the relationship between features and the label as well as between different features, we create a pair plot. First we replace the label, which is the attrition flag with 0 and 1, where 0 stands for existing customer and 1 stands for attrited customer. The pair plots is shown in Figure 9 in the Appendix.

2.2 Categorical Features

Distribution of categorical features are shown in Figure 6 & 7. The distribution of the label indicates that we have an unbalanced dataset, whereas the gender distribution is almost balance. Most of the customers seem to have the Blue card, either married or single, hold a graduate degree, and have annual income less than $40000.

Figure 6: Distribution of Attrition Flag and Customer Gender

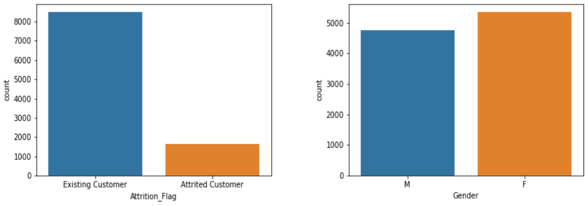
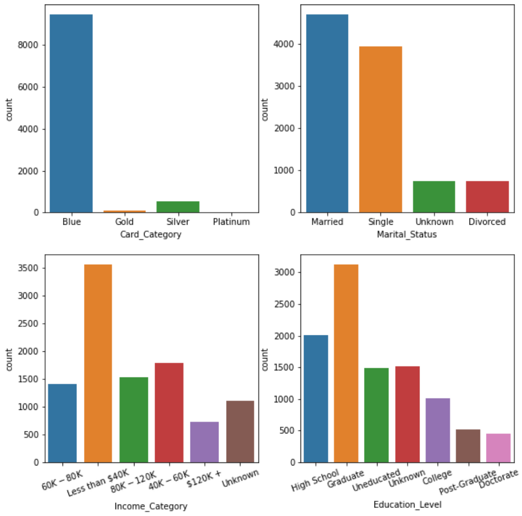


Figure 7: Distribution of Other Categorical Features



3 Feature Engineering

3.1 One-hot Encoding

First we map the less represented categories in card category, which are ”Gold”, ”Silver”, and ”Platinum” into ”Others”. For the same reason we map ”Divorced” and ”Unknown” in marital status into ”Others”. Then we convert customer gender into 0 and 1 with 0 stands for female and 1 stands for male. Finally we one-hot encode all categorical features excluding the attrition label.

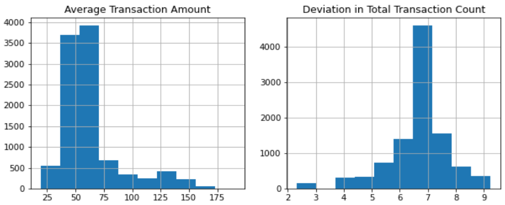
3.2 Feature Interaction and Distribution

We create one more feature capturing the average transaction amount . This feature can be calculated as

$$Avg\_Tans\_Amt = \frac{Total\_Trans\_Amt}{Total\_Trans\_Ct}$$

Since the pair plots indicate that the total transaction counts of existing customers are more diverse, whereas that of attrited customers are concentrated around the small numbers. Therefore, we add another feature capturing the number of the total transaction as well as the deviation of this count

Figure 8: Distribution of Newly Created Features



from the mean, which can be computed by

$$Dev\_Trans\_Ct = Total\_Trans\_Ct * |Total\_Trans\_Ct - \overline{Total\_Trans\_Ct}|$$

We then perform log transformation on this feature. The distribution of the newly created features are shown in Figure 8.

4 Key Findings and Insights

The pair plots show that for some features, whether the customer is an existing customer does not affect their distribution. For example, as shown in Figure 9, the distribution of the age of customers are almost the same for existing and attrited customers. It suggests that most of the customers are around age 45, but it does not provide much information for predicting customer churn.

Other features have different distributions among existing customers and attrited customers. For example, if a customer has a high total revolving balance or a large amount of total transaction, this customer are more likely to be an existing customer rather than an attrited customer. These features are useful in our analysis to predict the customer churn[4].

Some features have relationship between themselves. For example, the period relationship of a customer with the bank appears to be linearly correlated with the age of a customer. Although a relationship of 37 months with the bank seems to be very common and exists in all age groups. Another example is the polynomial relationship between the average open to buy credit line and the average utilization ratio. An higher average open to buy credit line is associated with a lower average utilization ratio. However, these features do not appear to be very useful in our analysis.

5 Hypothesis Testing

5.1 Hypothesis I

To check whether there is a relationship between customer attrition and the total revolving balance on the credit card. First we run regression

$$Y = \alpha + \beta X + \varepsilon$$

where Y is the customer attrition flag and X is the total revolving balance. Then we define null hypothesis and alternative hypothesis

$$H_0 : \beta = 0$$

$$H_1 : \beta \neq 0$$

[4] See Hypothesis I and Hypothesis II

Next we compute the t-statistics

$$t\_stat = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

where $\hat{\beta}$ is the estimated parameter and $SE(\hat{\beta})$ is the standard error of this estimated parameter. Since the absolute value of t_stat in this case is larger than 1.96, it is outside of the 95% confidence interval, we can reject the null and assume the total revolving balance is useful for predicting the customer attrition with 95% confidence.

5.2 Hypothesis II

To examine whether there is a relationship between customer attrition and total amount of transaction, we conduct hypothesis testing similar to that in Hypothesis I. We change the X in the regression in Hypothesis I to the total amount of transaction, then compute t_stat using the same method. Since the t_stat is greater than 1.96, we can assume, with 95% confidence that the total amount of transaction is useful in predicting the customer attrition.

5.3 Hypothesis III

Another hypothesis we propose is that the distribution of customer age is normal. We test whether its skew and kurtosis are different from those of normal distribution[5]. The p value we have is smaller than 0.05, indicating that with 95% confidence we can reject the null that distribution of customer age is normal.

6 Next Steps

To better prepare the data for analysis, the next step will be scale the features to avoid unbalanced impacts of the magnitudes on results[6]. Then we can split the data into training set and test set. We may need to split the data to multiple folds to perform cross-validation to avoid overfitting. Meanwhile, we need to consider the models we will use to learn the pattern and make prediction.

7 Quality Summary

In general, the quality of this dataset is fair. The information appears to be accurate and, mostly, relevant. There is no missing value in the dataset. However, we may need other information such as the customer register time or the location of customers. In addition, we are not sure about the timeliness of the data.

[5] https://het.as.utexas.edu/HET/Software/Scipy/generated/scipy.stats.normaltest.html

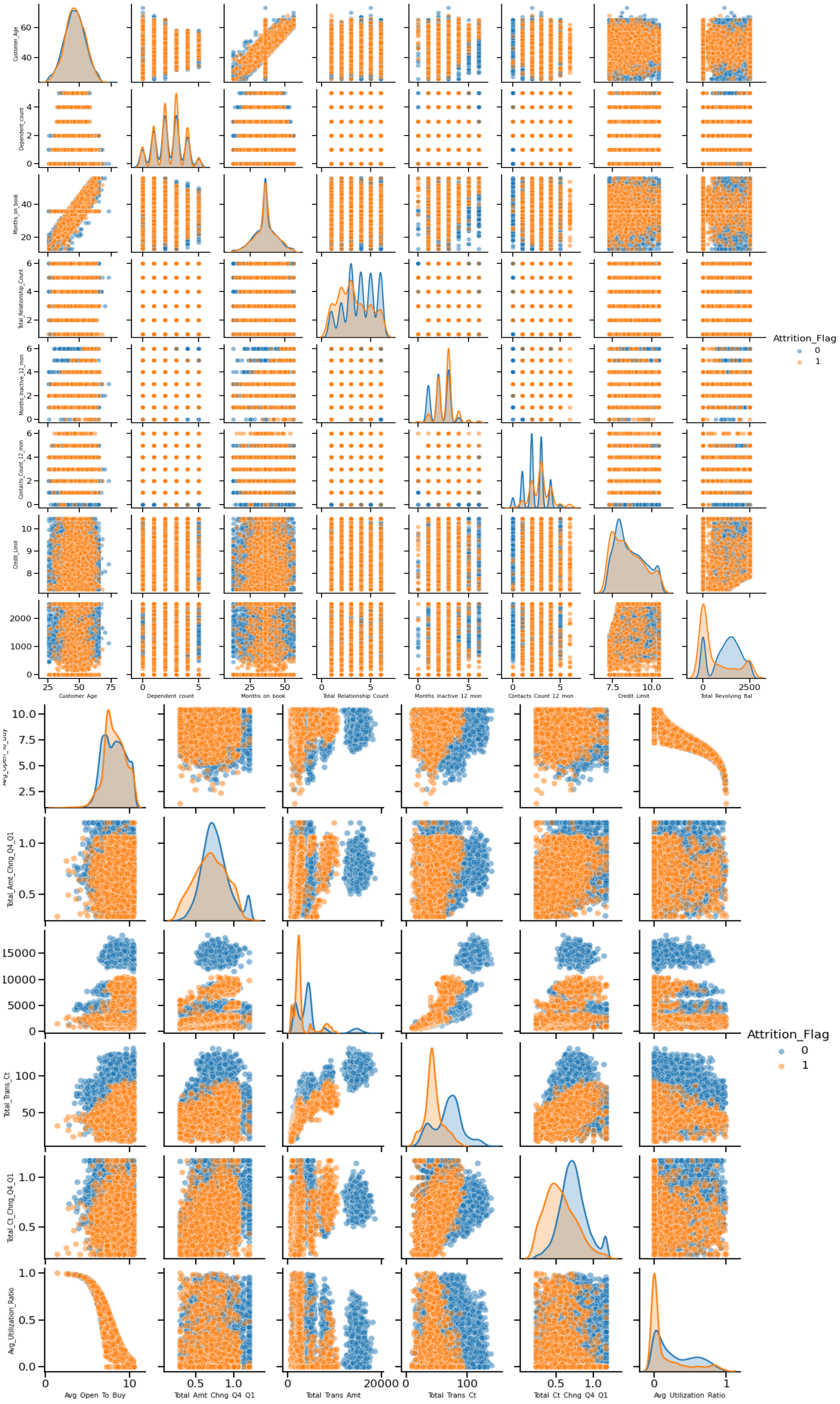
[6] https://scikit-learn.org/stable/auto_examples/preprocessing/plot_scaling_importance.html

8 Appendix

Table 1: Columns and Explanation of the Data

Column	Detail
CLIENTNUM	Client identification number
Attrition_Flag	Weather the customer account has been closed
Customer Age	Age of a customer
Gender	Customer gender
Dependent_Count	Number of dependents
Education_Level	Education Qualification of account holder
Marital_Status	Married, single, divorced, or unknown
Income_Category	Annual income category of a customer
Card_Category	Type of card
Month_on_book	Period of relationship with the bank
Total_Relationship_Count	Total number of products hold by the customer
Month_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 card
Avg_Open_To_Buy	Last 12 months average of open to buy credit line
Total_Amt_Chng_Q4_Q1	Change in transaction amount (Q4 over Q1)
Total_Trans_Amt	Total transaction amount in last 12 months
Total_Trans_Ct	Total transaction count in last 12 months
Total_Ct_Chng_Q4_Q1	Change in transaction count (Q4 over Q1)
Avg_Utilization_Ratio	Average card utilization ratio

Figure 9: Pair Plots for All Numerical Features and Attrition Label


Customer_Age
Dependent_count
Months_on_book
Total_Relationship_Count
Months_Inactive_12_mon
Contacts_Count_12_mon
Credit_Limit
Customer Age
Dependent count
Months on book
Total Relationship Count
Months Inactive 12 mon
Contacts Count 12 mon
Credit Limit
Total Revolving Bal
Attrition_Flag
Avg_Open_To_Buy
Total_Amt_Chng_Q4_Q1
Total_Trans_Amt
Total_Trans_Ct
Total_Ct_Chng_Q4_Q1
Avg_Utilization_Ratio
Avg Open To Buy
Total Amt Chng Q4 Q1
Total Trans Amt
Total Trans Ct
Total Ct Chng Q4 Q1
Avg Utilization Ratio
Attrition_Flag