Classification: Predict Churning Customers

Zixin Huang

July 13, 2021

1 Summary

In this project our goal is to help a bank predict churning customers, so the bank can reach out to those customers and provide better service and, therefore, reducing the number of customers who are leaving. Our models will be focus on prediction since the bank cares more about correctly identify the churning customers.
The data[1] contains customer information related to credit cards of a bank. It 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. There are 6 categorical features and 15 numerical features. There is no missing value in the data.

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

[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.

2 Data Cleaning and EDA

We first remove the client number from the data since it does not contain useful information for our prediction. The statistical summary of all numerical features and summary of categorical or binary columns is shown in Figure 1. It shows that the magnitudes of numerical features are very different, and some features, such as credit limit, have large standard deviation. Some of categorical columns, such as gender, are balanced, whereas others, such as card category, are not.

Figure 1: Summary of Features

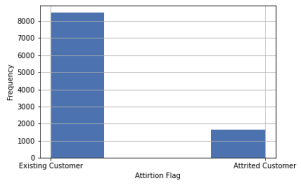
	count	mean	std	min	25%	50%	75%	max
Customer_Age	10127.0	46.33	8.02	26.0	41.00	46.00	52.00	73.00
Dependent_count	10127.0	2.35	1.30	0.0	1.00	2.00	3.00	5.00
Months_on_book	10127.0	35.93	7.99	13.0	31.00	36.00	40.00	56.00
Total_Relationship_Count	10127.0	3.81	1.55	1.0	3.00	4.00	5.00	6.00
Months_Inactive_12_mon	10127.0	2.34	1.01	0.0	2.00	2.00	3.00	6.00
Contacts_Count_12_mon	10127.0	2.46	1.11	0.0	2.00	2.00	3.00	6.00
Credit_Limit	10127.0	8631.95	9088.78	1438.3	2555.00	4549.00	11067.50	34516.00
Total_Revolving_Bal	10127.0	1162.81	814.99	0.0	359.00	1276.00	1784.00	2517.00
Avg_Open_To_Buy	10127.0	7469.14	9090.69	3.0	1324.50	3474.00	9859.00	34516.00
Total_Amt_Chng_Q4_Q1	10127.0	0.76	0.22	0.0	0.63	0.74	0.86	3.40
Total_Trans_Amt	10127.0	4404.09	3397.13	510.0	2155.50	3899.00	4741.00	18484.00
Total_Trans_Ct	10127.0	64.86	23.47	10.0	45.00	67.00	81.00	139.00
Total_Ct_Chng_Q4_Q1	10127.0	0.71	0.24	0.0	0.58	0.70	0.82	3.71
Avg_Utilization_Ratio	10127.0	0.27	0.28	0.0	0.02	0.18	0.50	1.00

	count	unique	top	freq
Attrition_Flag	10127	2	Existing Customer	8500
Gender	10127	2	F	5358
Education_Level	10127	7	Graduate	3128
Marital_Status	10127	4	Married	4687
Income_Category	10127	6	Less than $40K	3561
Card_Category	10127	4	Blue	9436

2.1 Attrition Flag

The distribution of the target variable is shown in Figure 2. It shows that we are dealing with an imbalanced data since only about 16% of customers are labeled as "Attrited Customer".

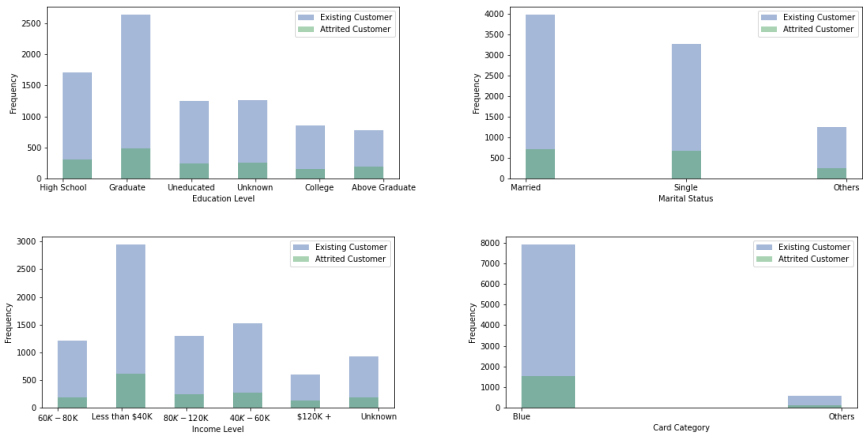
Figure 2: Distribution of Target Variable



2.2 Categorical Features

Other categorical features that have imbalanced distributions are education level, marital status, income level and card category. For education level, there are only around 9% of the customers have a degree higher than graduate level, therefore we combine those customers as one category "Above Graduate". For the same reason we combine less represented categories of marital status and card type as "Others". The distribution of these 4 features and their relationships with the target variable can be seen in Figure 3.

Figure 3: Relationships between Categorical Features and Attrition Flag

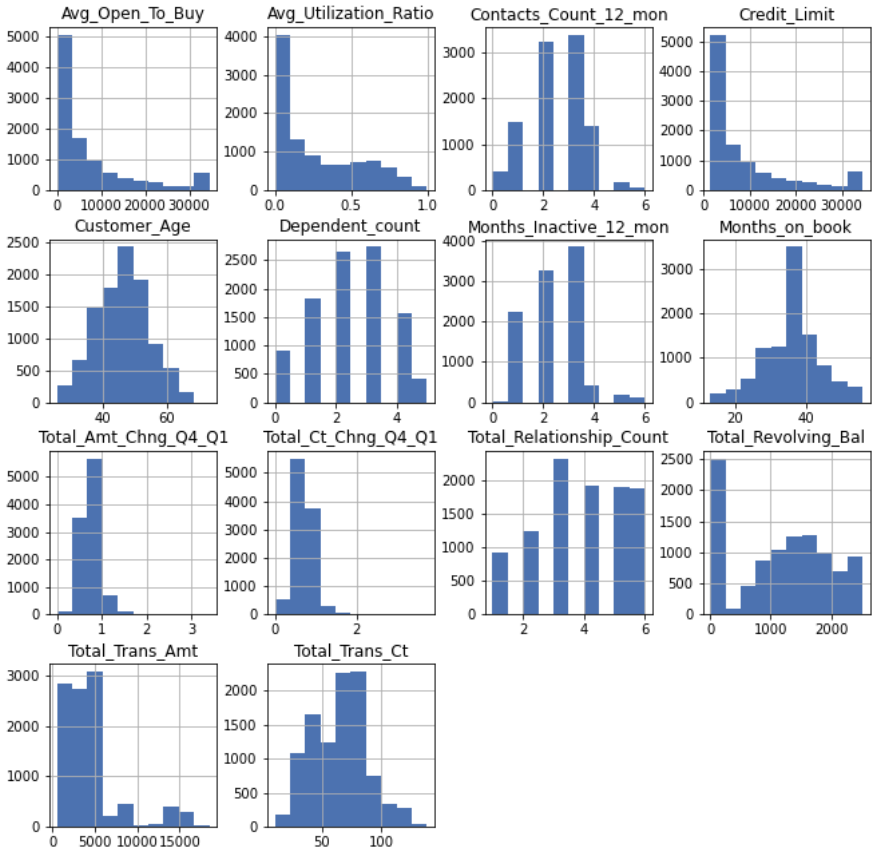


We can see that most of the customers are married, hold a graduate degree, have an income less than $40000 annually, and use a Blue Card. In particular, a customer who is single might be more likely to give up the credit card service than a customer who is married. For other categorical features there is no clear difference between the distribution of churning and existing customer.

2.3 Numerical Features

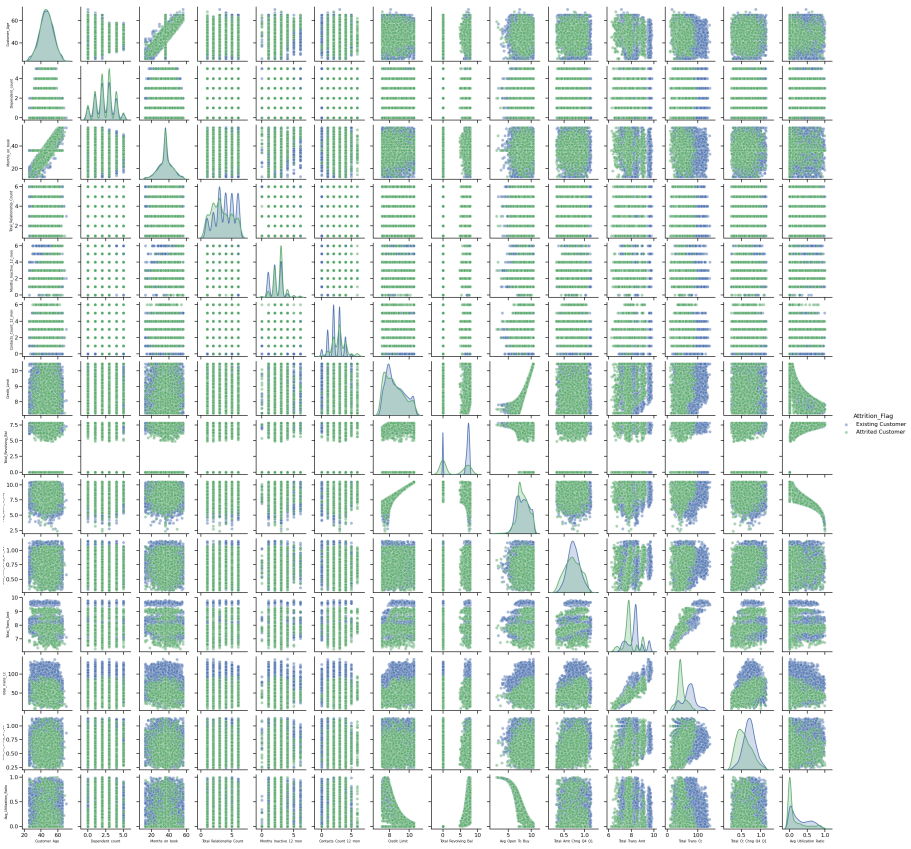
The distribution of numerical features is shown in Figure 4. We can see that monetary values such as credit limit and average open to buy, have larger scales than other features and are right skewed. Therefore, we preform log transformation to those monetary features.
We also notice that there might be outliers with regard to total change of transaction amount and transaction count. Thus we remove observations that exceed the upper and lower bounds of these two features according to the 1.5 IQR rule. After removing the outliers we have 9318 observations, and there are still about 16% of the customers are labeled as "Attrited Customer".

Figure 4: Distribution of Numerical Features



From Figure 5 we can further examine the relationship between numerical features and the target variable as well as between features themselves. It seems that features such as customer age, number of dependent, and number of periods since join the service do not have significant impact on whether the customer is churning. Features that seems to be related to a customer's decision are change of transaction amount, change of transaction count, average utilization ratio and total revolving balance. In addition, some features appear to be correlated. For example, average open to buy amount and average utilization ratio appear to be negatively correlated, and total transaction amount and total transaction count seem to be positively correlated.

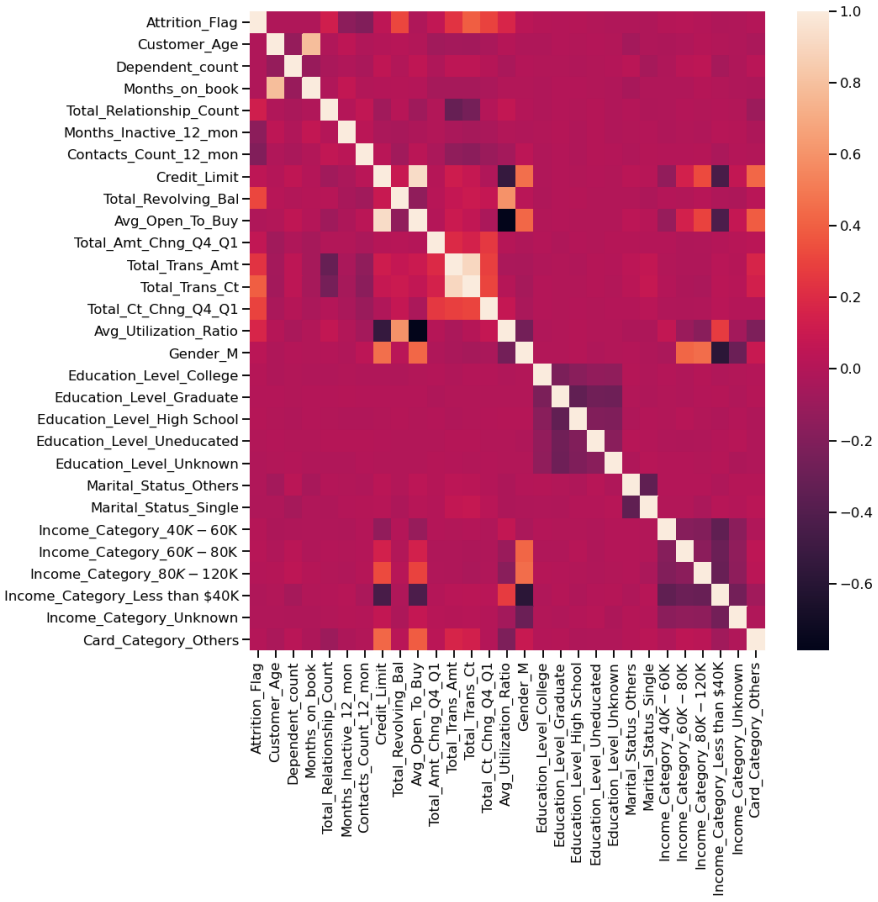
Figure 5: Pair Plot between Numerical Features and the Target Variable



3 Feature Engineering

First we label customers who are marked as "Existing Customer" with 1 and label customers who are marked as "Attrited Customer" with 0. Then we one-hot encode the categorical features, with the first category of each feature dropped to avoid multicollinearity. Now we have 29 columns with a label and 28 feature columns. The heatmap which describe the correlations between these columns is shown in Figure 6. We can see that whether the customer is male and whether the customer has annual income less than $40000 is strongly correlated. Thus we remove the feature that indicate customer gender. We also remove average open to buy amount since it is strongly correlated with average utilization ratio as well as annual income.
Finally, we have a dataset that contains 9318 rows and 27 columns, including the label.

Figure 6: Heatmap of All Columns



4 Training Models

4.1 Data Preparation

Since we have an imbalanced dataset, we use stratified split to split training and test set to ensure the ratio between two labels are similar in both sets. Then we scale the all features to the range between 0 and 1 to avoid the affect of magnitudes on decision boundaries. For the training set, we over-sample the minority class, which is "Attrited Customer", using SMOTE so that there are equal number of observations for two classes. Next we randomly split the training data into 3 folds to perform cross validation. We do not over-sample the test set.

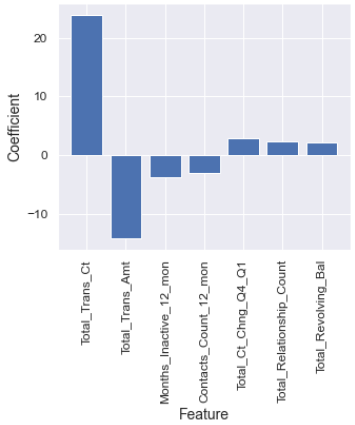
4.2 Logistic Regression

We first use logistic regression with $L2$ penalty as a baseline model. The regression can be written as

$$p = \frac{e^{\beta_0 + \boldsymbol{\beta} X}}{1 + e^{\beta_0 + \boldsymbol{\beta} X}}$$

where p is the probability of $y = 1$. The model will predict $y = 1$ if $p > 0.5$. After fitting the training data, the 7 most important features and their coefficients can be seen in Figure 7.

Figure 7: Coefficients of the 7 Most Important Features



4.3 Linear SVM

The second model we use is linear SVM (Support Vector Machines). We use cross validation to tune the regularization parameter C. The F Score for C equals to 0.1, 1, and 10 is shown in Figure 8. We can see a drastic improvement of the model performance when the value of C increases from 0.1 to 1, and the performance barely improved when C increase from 1 to 10. Therefore, we choose $C = 1$ for out linear SVM model, and fit the model to the training data. The 7 most important features when using the linear SVM model are the same as those when using the logistic regression (see Figure 9).

Figure 8: F Score for Different C Values

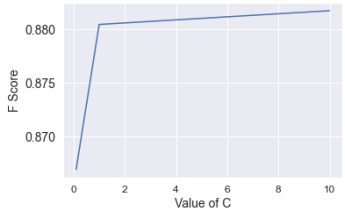


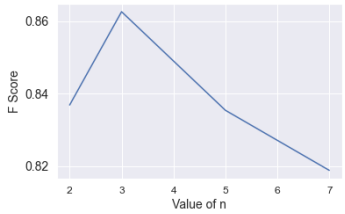
Figure 9: Most 7 Important Features using Linear SVM

	Features	Coefficient	Abs_Coef
10	Total_Trans_Ct	15.645347	15.645347
9	Total_Trans_Amt	-9.139311	9.139311
4	Months_Inactive_12_mon	-2.492016	2.492016
11	Total_Ct_Chng_Q4_Q1	2.048462	2.048462
5	Contacts_Count_12_mon	-1.871008	1.871008
3	Total_Relationship_Count	1.404781	1.404781
7	Total_Revolving_Bal	1.285014	1.285014

4.4 KNN

The last model we use is KNN (K Nearest Neighbors). To tune the parameter n, which specifies the number of observations in the neighbourhood, we use the same cross validation folds. The cross validation F Score using different values of n can be seen in Figure 10.

Figure 10: F Score for Different n Values



Since the model performs best when $n = 3$, we use $n = 3$ to train our model on the whole training data.

5 Prediction

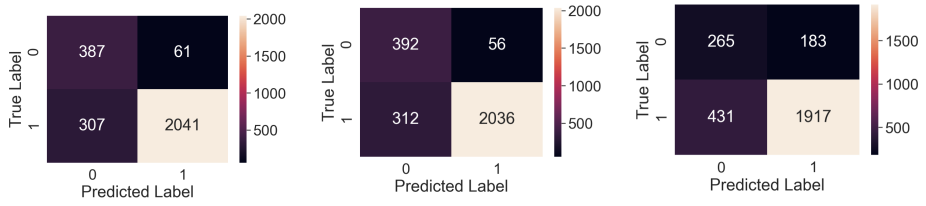
In this step we predict the label using the test data, and compare the results of different models. The precision, recall and F Score of the three models is shown in Table 2.

Table 2: Model Performance on Test Set

Model	Logistic Regression	Linear SVM	KNN
Precision	0.905	0.907	0.828
Recall	0.868	0.868	0.780
F Score	0.879	0.879	0.798

The confusion matrix for these three models is shown in Figure 11. We can see that the linear SVM performs slightly better than logistic regression in terms of precision. For recall and F Score, the performance of logistic regression and SVM are very similar, and KNN is slightly worse. KNN also has the lowest score on precision.

Figure 11: Confusion Matrix of Logistic Regression (Left), SVM, and KNN (Right)



From the confusion matrices we can see that KNN performs worst. Linear SVM has the minimum Type I error, whereas logistic regression has the minimum Type II error. We also noted that using the over-sampled training set will lead to more of Type II error but less of Type I error compared to using the original training data. We decide to use the over-sampled training data since we

want to identify the churning customers as precise as possible, so the bank can reach out to them and ultimately reduce the churning rate. Missing churning customers would do more harm than misclassify some existing customers.
For the same reason, we recommend using linear SVM as final model for this purpose.

6 Findings and Insights

From logistic regression and linear SVM we find that the most important features that related to customer churning are total number of transactions, total amount of transactions, number of months inactive, number of contacts, change in transaction count, total number of product hold, and total revolving balance. More specifically, churning customers tend to have a higher total transaction amount, and their actions with the bank tend to be more concentrated. Whereas the existing customers tend to have higher number of transaction, hold more products of the bank, have higher total revolving balance and greater change in transaction counts.

7 Suggestions for Next Steps

After finding the most important features, we may do more feature selection, then using only selected features on models such as SVM or boosting. We may also consider using decision tree model to achieve a better explanation. In addition, we can try different method for over-sampling the training data to see which method works best.