HOXHUNT SUMMER TRAINEE 2021

DATA ANALYSIS REPORT

Prepared by **Nhu Nguyen**

17th March 2021

Table of Contents

1. Introduction

Customer Attrition is one of the most important and challenging problems for businesses when there are many new established companies with competitive services. The reasoning of leaving can vary and would require domain knowledge in order to define properly, however some common ones are; lack of usage of the product, poor service and better price somewhere else. Regardless of the reasoning that can be specific for different industries, one thing applies for every domain is, it costs more to acquire new customers than it does to retain existing ones. This has a direct impact on operating costs and marketing budgets within the company.

In this data analysis, a bank manager is concerned that more and more customers are leaving their credit card services. The main purpose is to explore the dataset and find patterns/ sub-segments in order to give recommendations to the manager of how to detect customers who are at a higher risk of churning.

This data analysis project is done in Python and Jupyper Notebook, using the following libraries:

- Numpy, Pandas: loading the dataframe, conducting data cleaning and preparation
- Scikit-learn: providing a selection tools for machine learning
- Matplotlib, Seaborn and Plotly: data visualization tool

2. Exploratory Data Analysis

2.1. Data cleaning

As described on Kaggle, the two last columns are irrelevant so they are removed. The final dataset consists 21 columns and 10127 rows with non-null values and non-duplicated values. (isnull() and duplicated() are the methods for checking null values and duplicated values). The data information below summarizes all the information that we need to have an overview about the dataset: names of columns, non-null values count and types of the values.

Table 1 Data information

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10127 entries, 0 to 10126
Data columns (total 21 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   CLIENTNUM                 10127 non-null  int64
 1   Attrition_Flag            10127 non-null  object
 2   Customer_Age              10127 non-null  int64
 3   Gender                    10127 non-null  object
 4   Dependent_count           10127 non-null  int64
 5   Education_Level           10127 non-null  object
 6   Marital_Status            10127 non-null  object
 7   Income_Category           10127 non-null  object
 8   Card_Category             10127 non-null  object
 9   Months_on_book            10127 non-null  int64
 10  Total_Relationship_Count  10127 non-null  int64
 11  Months_Inactive_12_mon    10127 non-null  int64
 12  Contacts_Count_12_mon     10127 non-null  int64
 13  Credit_Limit              10127 non-null  float64
 14  Total_Revolving_Bal       10127 non-null  int64
 15  Avg_Open_To_Buy           10127 non-null  float64
 16  Total_Amt_Chng_Q4_Q1      10127 non-null  float64
 17  Total_Trans_Amt           10127 non-null  int64
 18  Total_Trans_Ct            10127 non-null  int64
 19  Total_Ct_Chng_Q4_Q1       10127 non-null  float64
 20  Avg_Utilization_Ratio     10127 non-null  float64
dtypes: float64(5), int64(10), object(6)
memory usage: 1.6+ MB
```

Table 2 Description of columns/features

Type of information	**Columns**	**Features**
Basic information	CLIENTNUM	Client numbers for the customer holding the account
Target	Attrition_Flag	The attrition customers and existing customers
Demographic variables	Customer_Age	Customer's age in years
	Gender	Customer's genders
	Dependent_Count	Number of dependents
	Education_Level	Education qualification of the account holder
	Marital_Status	The marital status of account holder
	Income_Category	Annual Income Category
Product variables	Card_Category	Type of card
	Months_on_book	The relationship period with the bank
	Total_Relationship_Count	Total number of products held by customer
	Months_Inactive_12_mon	Number of inactive months in the last 12 months
	Contacts_Count_12_mon	Number of contacts in the last 12 months
	Credit_Limit	The limitation of credit card
	Total_Revolving_Bal	Total revolving balance
	Avg_Open_To_Buy	Open to buy credit line
	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

2.2. Target variable

The Attrited_Flag column is set as the target for prediction. There are two categories: Attrited Customer (those who left) and Existing Customer (those who stayed).

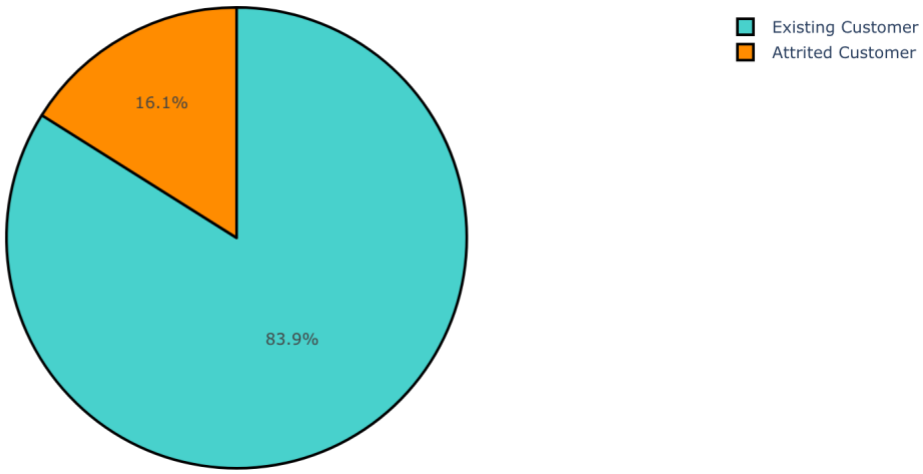

Figure 1 Attrited Customer vs Existing Customer

It's clear that the majority of our customers (83.9 %) stay. The churned customers are less than 20% of the total customers.

2.3. Demographic variables

The demographic variables include age, gender, marital status, education level, dependent count, and income category. Pie charts and box plots are chosen to illustrate these features, the reasons are:

Box plots:

- Describe the distribution of values (mean, standard deviation, etc.)
- Detect the outliers

Pie charts:

- The contribution of each slide to the whole
- Compare the groups of different sizes (see Figure 1)
- Not too many categories in one feature

The figures below illustrate the distribution of each category from demographic variables between attrited customer and existing customer.

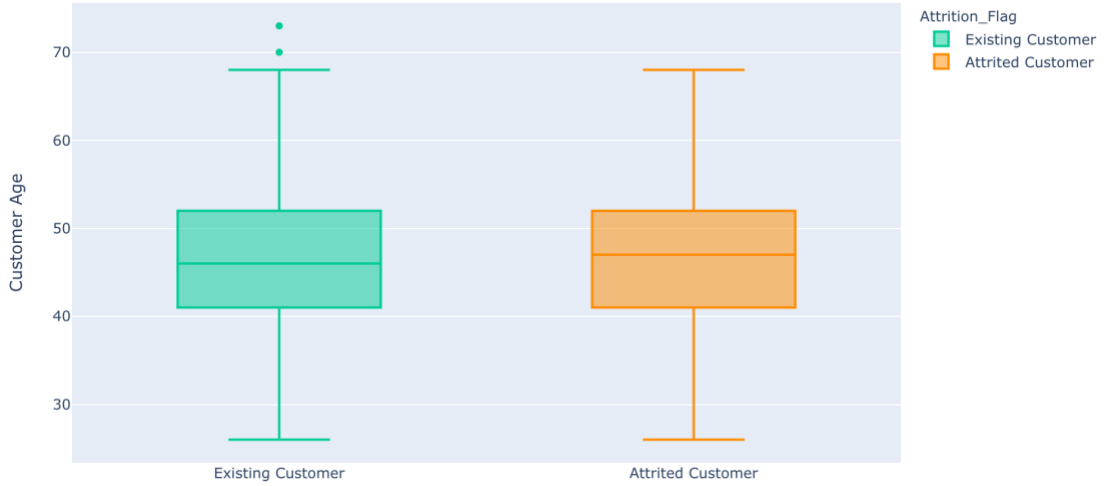

Figure 2 Customer Age Range.

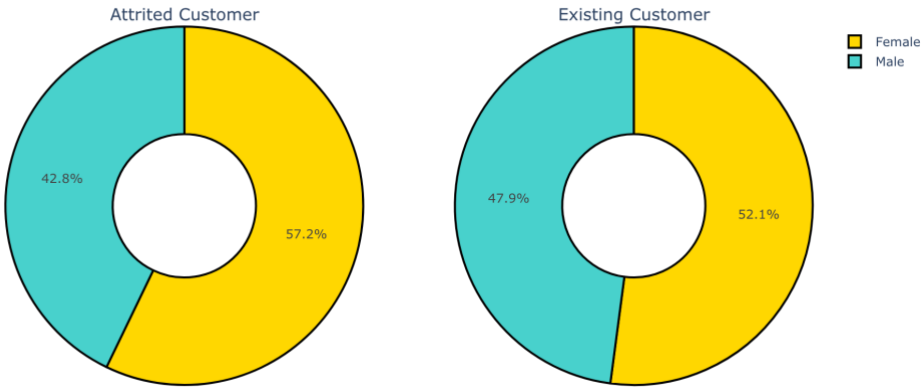

Figure 3 Gender of Attrited Customer vs Existing Customer

.

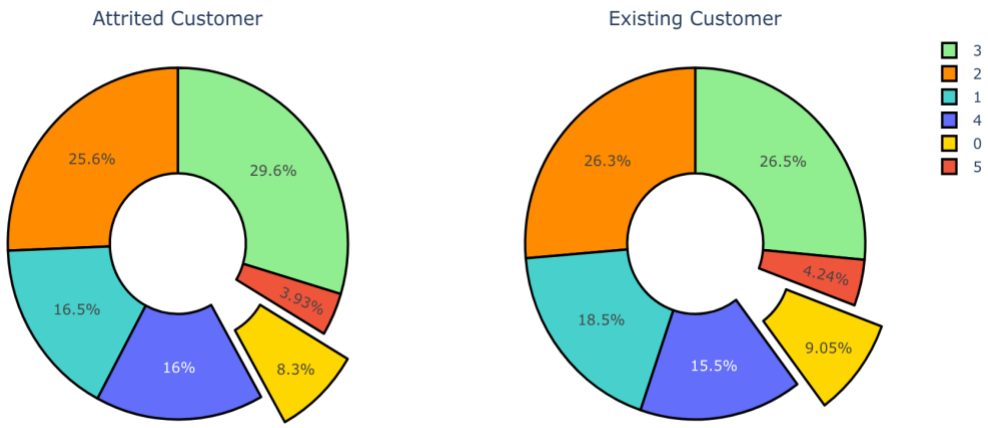

Figure 4 Dependent count of Attrited Customer and Existing customer

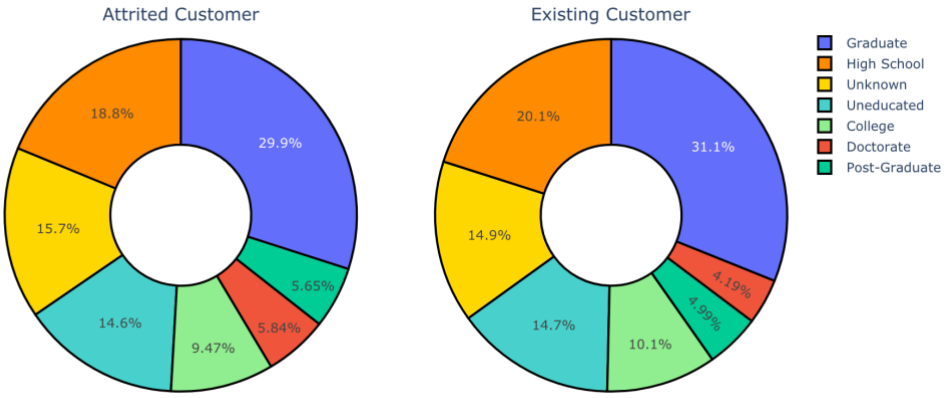

Figure 5 Education Level of Attrited Customer vs Existing Customer

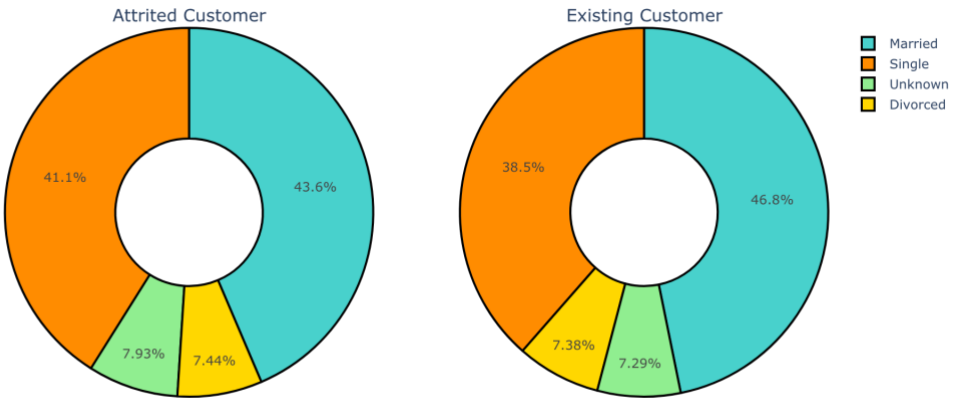

Figure 6 Marital Status of Attrited Customer vs Existing Customer

Figure 7 Income Category of Attrited Customer vs Existing Customer

The gender feature shows a slightly difference, however, **it is too small to be assumed that one gender is more eager to leave**. The same applies to other features. Thus, we can assume that the demographic variables do not give any signals about the high risk of churning.

2.4. Product variables

Product variables will describe the product activities of customer. In this section, I use pie chart, box plot, bar chart and scatter plot to see the distribution of each feature in two target groups.

First of all, we can have a look at types of credit cards and the relationship period each customer group have.

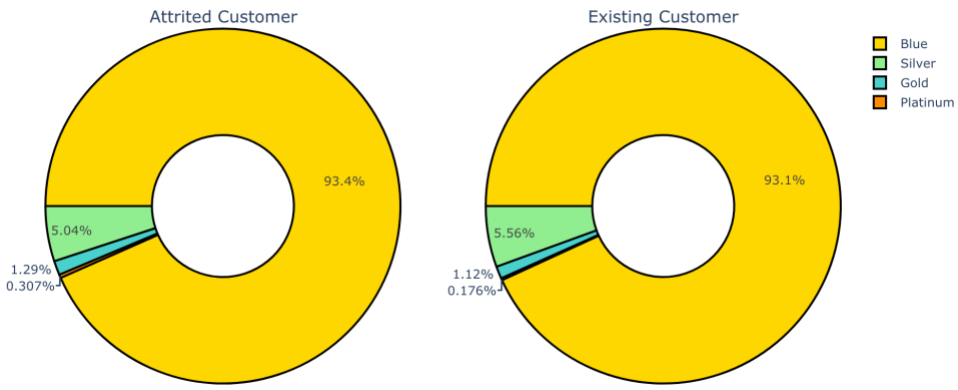

Figure 8 Types of credit cards

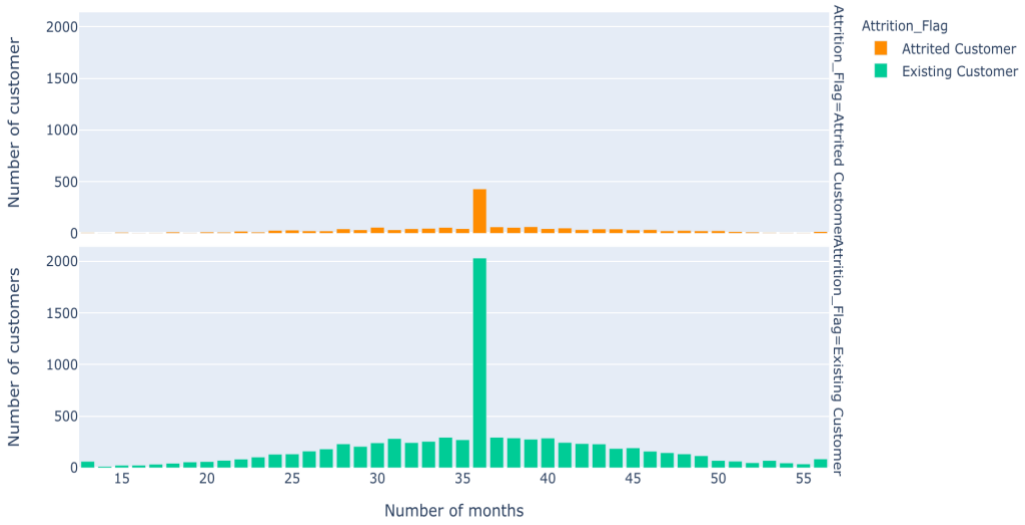

Figure 9 The relationship period with the bank

We can see that two groups share the same distribution. Blue card is the most popular one among others and the relationship period with the bank often falls in 36 months.

However, the difference can be seen in number of product held by customers.

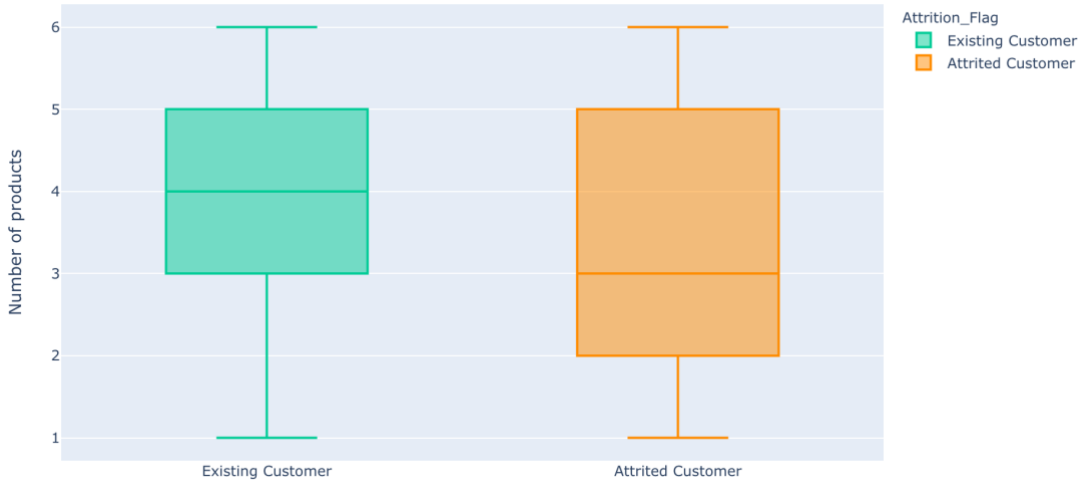

Figure 10 Number of product held by customer

Existing customers own more products (4 products on average) than attrited customers (3 products on average).

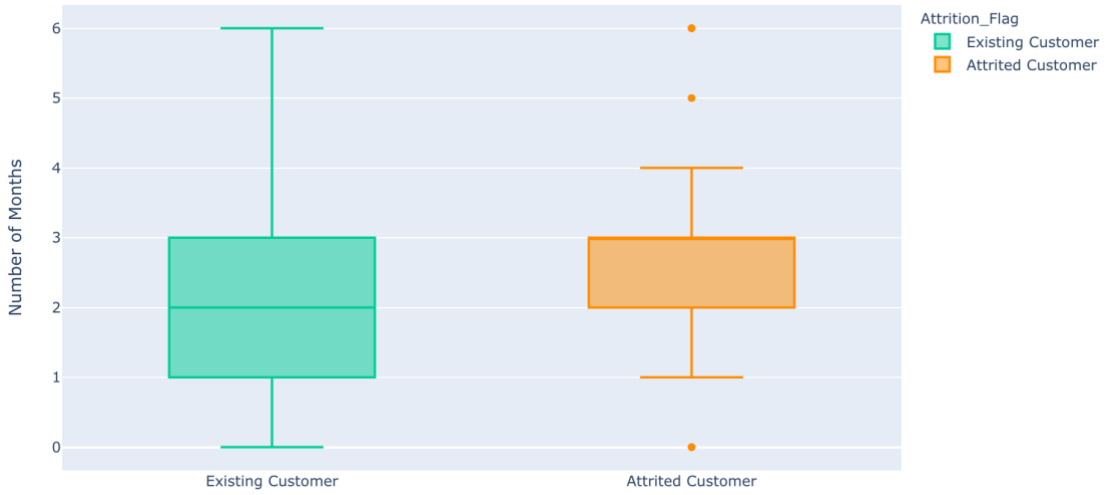

Figure 11 Number of inactive months

While the average inactive months of existing customers is 2, the average of attrited customers is 3. This could suggest that **customers who have longer inactive time will have more intention to leave**.

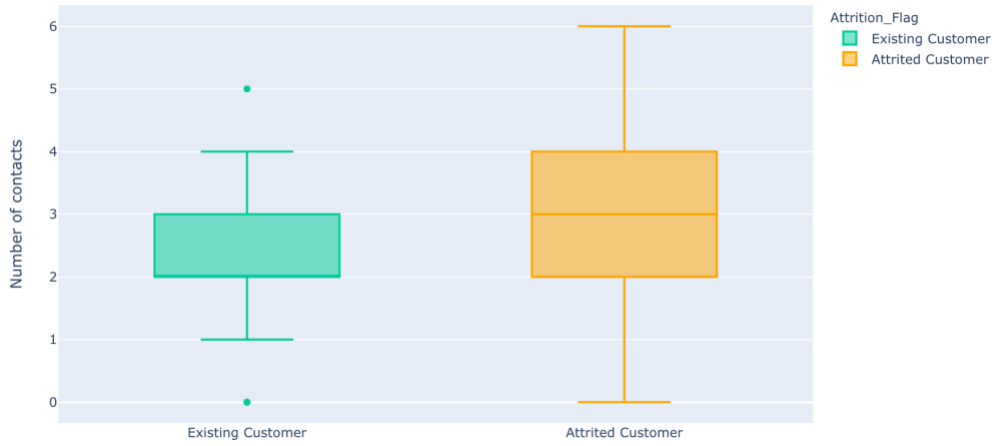

Figure 12 Number of contact in the last 12 months

The attrited customers are most likely to have more contacts than the existing customer when they have a higher number of contacts.

Figure 13 Credit limitation

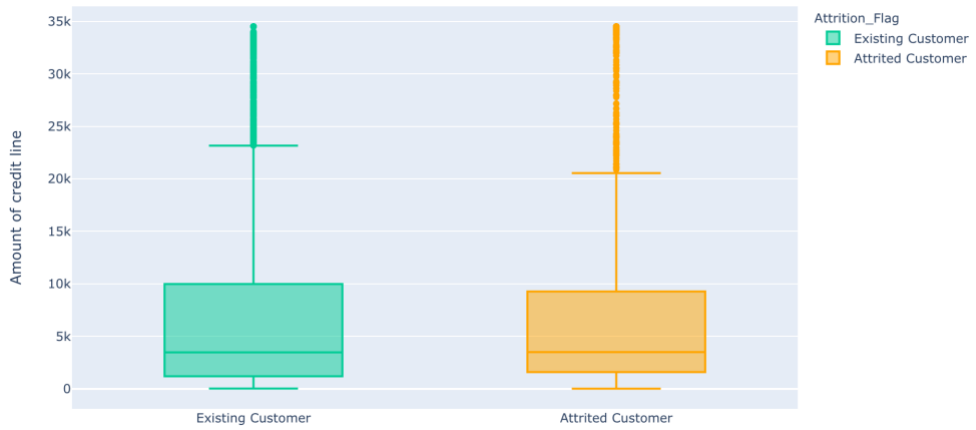

Figure 14 Open line of credit

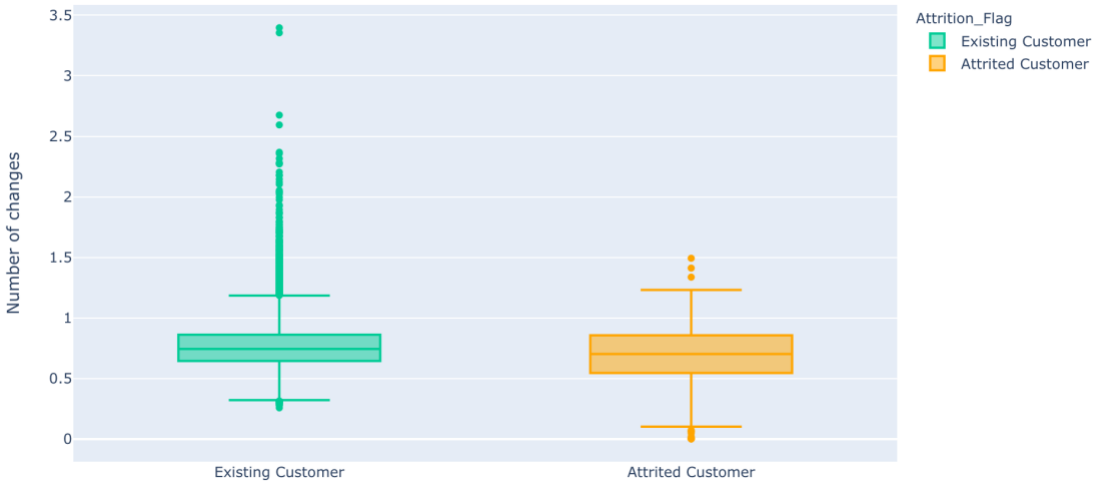

Figure 15 Change in transaction amount (Q4 over Q1)

The three features above (Figure 13, 14,15) share a similar pattern between two groups. Only the change in transaction amount (Q4 over Q1) has a slightly difference in the minimum value and also in the outliers.

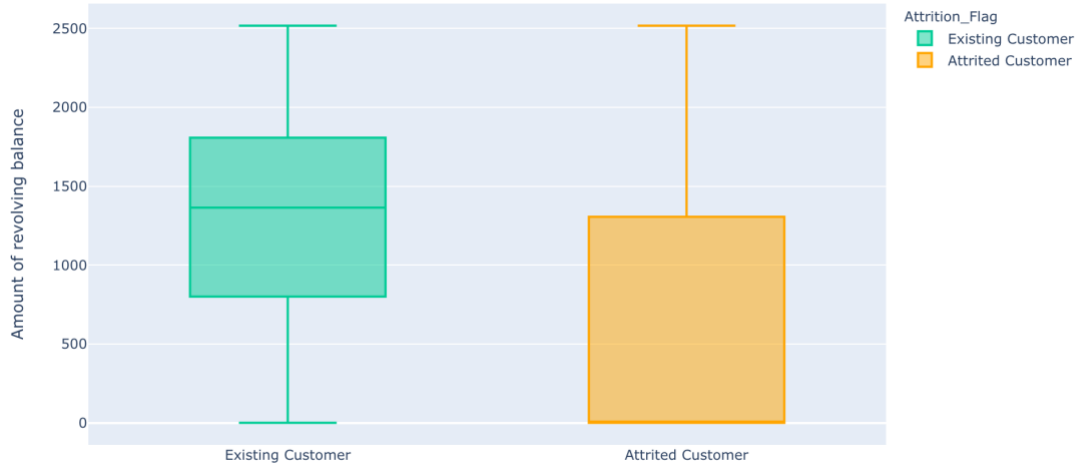

Figure 16 Revolving balance

The difference in the revolving balance is significant. While the existing customers intend to have a high amount of revolving balance, the attrited customers have a pretty low one.

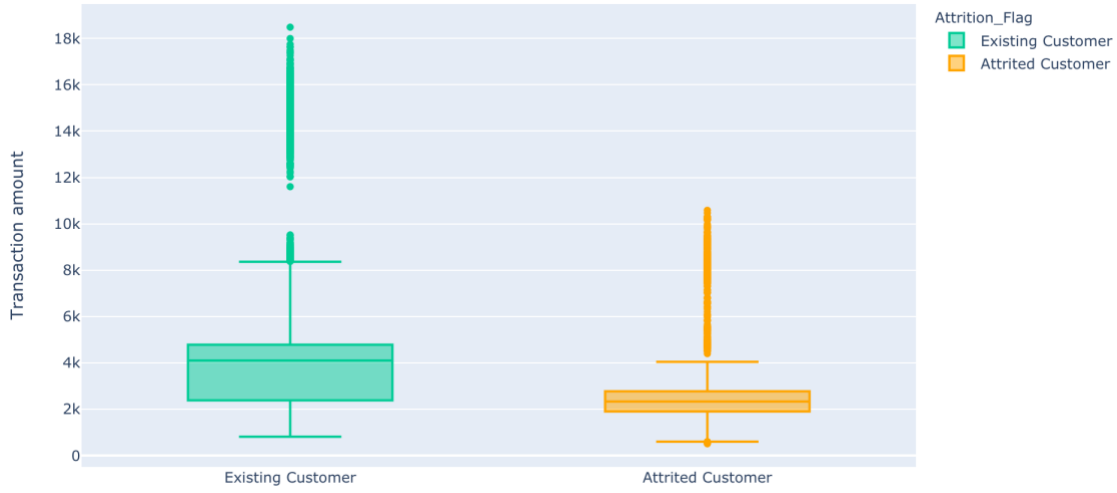

Figure 17 Total transaction amount

Similar to revolving balance, total transaction amount in attrited customers group is higher than existing customers (4k and more than 2k respectively)

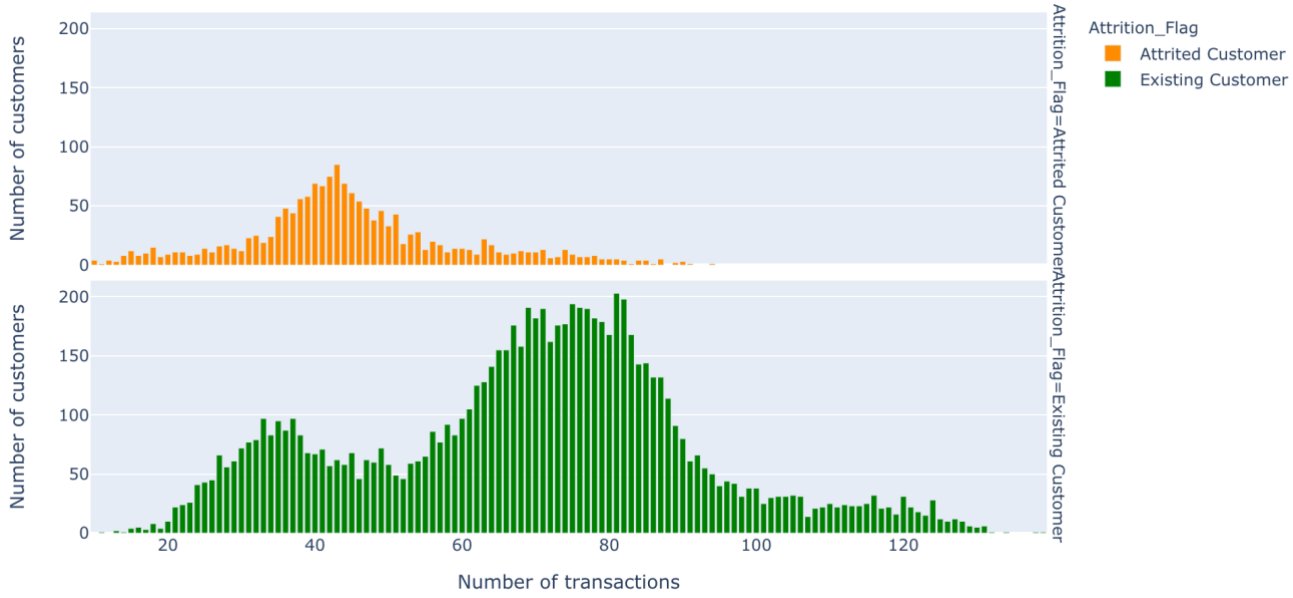

Figure 18 Total transaction count

The difference in total transaction count is noticeable. Number of transaction of attrited customer often falls between 30 to 50 while its of existing customer is from 70 to 90.

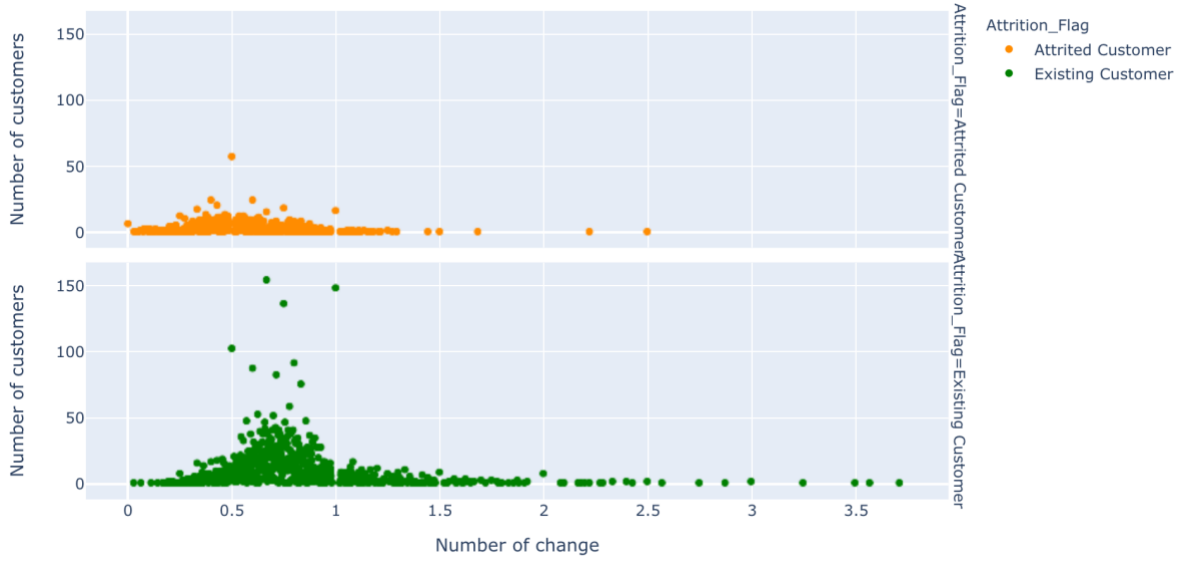

Figure 19 Transaction count

Similar to the total transaction counts, the change in transaction count chart also shows that the number of change focused mostly from 0.5 to 1 while the other is from 0.3 to 0.6.

Figure 20 The average utilization ratio

The last feature is the average utilization ratio. We also notice the different between two groups based on their average value and values range. The attrited customers have a low utilization ratio.

The table below shows average values of each feature in the product variables that show big differences between Existing Customer and Attrited Customer. This summary will help to select which features can be used to detect the customer intention to leave.

	Existing Customer	**Attrited Customer**
Number of products	4	3
Inactive months	2	3
Number of contacts	2	3
Revolving balance	1364	0
Total transaction amount	4100	2229
Total transaction count	71	43
Total count change	0.72	0.53
Utilization ratio	0.21	0

However, as checking the feature one-by-one is time consuming and (sometimes) not effective. Thus, building a prediction model is a must.

3. Prediction Model

3.1. Data preprocessing

There are six columns with object type that need to be label encoded (transform the text to number): Attrition_Flag, Gender, Education_Level, Marital_Status, Income_Category and Card_Category.

	CLIENTNUM	Attrition_Flag	Customer_Age	Gender	Dependent_count	Education_Level	Marital_Status	Income_Category	Card_Category
0	768805383	0	45	1	3	2	3	3	0
1	818770008	0	49	0	5	4	1	1	0
2	713982108	0	51	1	3	4	3	4	0
3	769911858	0	40	0	4	2	0	1	0
4	709106358	0	40	1	3	1	3	3	0
...	...	...	...	...	...	...	...	...	...
10122	772366833	0	50	1	2	4	1	2	0
10123	710638233	1	41	1	2	0	4	2	0
10124	716506083	1	44	0	1	2	3	1	0
10125	717406983	1	30	1	2	4	0	2	0
10126	714337233	1	43	0	2	4	3	1	2

Table 3 Label encoded features

The target is to build the binary classification model, then, it is not necessary to check the multi-collinearity. However, I still conduct this step as the collinear features maybe less informative of the outcome than the other (non-collinear) and as such they should be considered for elimination from the features set anyway.

	CLIENTNUM	Attrition_Flag	Customer_Age	Gender	Dependent_count	Education_Level	Marital_Status	Income_Category	Card_Categ
CLIENTNUM	1.000000	-0.046430	0.007613	0.020188	0.006772	-0.006946	0.003556	0.026295	0.006
Attrition_Flag	-0.046430	1.000000	0.018203	-0.037272	0.018991	0.008796	-0.021125	-0.013577	-0.003
Customer_Age	0.007613	0.018203	1.000000	-0.017312	-0.122254	-0.002369	0.017330	0.023508	-0.019
Gender	0.020188	-0.037272	-0.017312	1.000000	0.004563	-0.005087	0.003725	0.786608	0.080
Dependent_count	0.006772	0.018991	-0.122254	0.004563	1.000000	0.000472	0.007235	0.066278	0.023
Education_Level	-0.006946	0.008796	-0.002369	-0.005087	0.000472	1.000000	0.016192	-0.011677	0.014
Marital_Status	0.003556	-0.021125	0.017330	0.003725	0.007235	0.016192	1.000000	0.012028	-0.043
Income_Category	0.026295	-0.013577	0.023508	0.786608	0.066278	-0.011677	0.012028	1.000000	0.074
Card_Category	0.006235	-0.003687	-0.019172	0.080198	0.023310	0.014565	-0.043431	0.074855	1.000
Months_on_book	0.134588	0.013687	0.788912	-0.006728	-0.103062	0.006613	0.015199	0.022122	-0.014
Total_Relationship_Count	0.006907	-0.150005	-0.010931	0.003157	-0.039076	0.000766	0.022253	-0.003202	-0.081
Months_Inactive_12_mon	0.005729	0.152449	0.054361	-0.011163	-0.010768	0.005761	-0.003597	-0.016310	-0.016
Contacts_Count_12_mon	0.005694	0.204491	-0.018452	0.039987	-0.040505	-0.006280	-0.002073	0.023113	-0.000
Credit_Limit	0.005708	-0.023873	0.002476	0.420806	0.068065	-0.002354	-0.037908	0.475972	0.497
Total_Revolving_Bal	0.000825	-0.263053	0.014780	0.029658	-0.002688	-0.006800	0.031368	0.034718	0.018
Avg_Open_To_Buy	0.005633	-0.000285	0.001151	0.418059	0.068291	-0.001743	-0.040712	0.472760	0.495
Total_Amt_Chng_Q4_Q1	0.017369	-0.131063	-0.062042	0.026712	-0.035439	-0.010040	0.042835	0.011352	0.004
Total_Trans_Amt	-0.019692	-0.168598	-0.046446	0.024890	0.025046	-0.007460	-0.051350	0.019651	0.185
Total_Trans_Ct	-0.002961	-0.371403	-0.067097	-0.067454	0.049912	-0.004307	-0.092786	-0.054569	0.123
Total_Ct_Chng_Q4_Q1	0.007696	-0.290054	-0.012143	-0.005800	0.011087	-0.016692	0.003004	-0.012657	-0.005
Avg_Utilization_Ratio	0.000266	-0.178410	0.007114	-0.257851	-0.037135	-0.001849	0.033536	-0.246476	-0.208

Table 4 Example of correlation table

As can be seen in the correlation chart above, there is no multi-collinearity with the target column. Thus, I keep all the features for further analysis.

3.2. Handling imbalanced data

Data imbalance usually reflects an unequal distribution of classes within a dataset. As presented in the previous part, the data is imbalance because the existing customer takes 83.9% while churned customer is only 16.1%. If we train a binary classification model without fixing this problem, the model will be completely biased. I use the oversampling method, with the SMOTE technique to solve the problem.

3.2. Building model and evaluation

In the next step, I fit a classification model to predict if a customer will leave the bank and the main signals of their leaving. After testing with several models (Logistic Regression, SVC, Random Forest and AdaBoost), I decided to choose Random Forest to build the classification model. GridsearchCV is also used for the hyper parameter tuning in order to improve the performance of model. The best parameters and best cross-validation scores are given as below:

```
Best params: {'max_depth': 15, 'n_estimators': 200}
Best cross-validation score: 0.98
```

Printing the confusion matrix and classification report of the model.

```
[[1657   28]
 [  50 1665]]
Accuracy on test set: 0.98
---------------------------------------------------
              precision    recall  f1-score   support

           0       0.98      0.97      0.98      1707
           1       0.97      0.98      0.98      1693

    accuracy                           0.98      3400
   macro avg       0.98      0.98      0.98      3400
weighted avg       0.98      0.98      0.98      3400
```

Evaluation metrics can be applied such as:

- **Confusion Matrix:** a table showing correct predictions and types of incorrect predictions.
- **Precision:** the number of true positives divided by all positive predictions. Precision is also called Positive Predictive Value. It is a measure of a classifier's exactness. Low precision indicates a high number of false positives.
- **Recall:** the number of true positives divided by the number of positive values in the test data. Recall is also called Sensitivity or the True Positive Rate. It is a measure of a classifier's completeness. Low recall indicates a high number of false negatives.
- **F1-Score:** the weighted average of precision and recall.

The confusion matrix and classification report of Random Forest model shows that the accuracy is 98% with the F1-score for both two classes are the same 98%. **The result is extremely good for our prediction.**

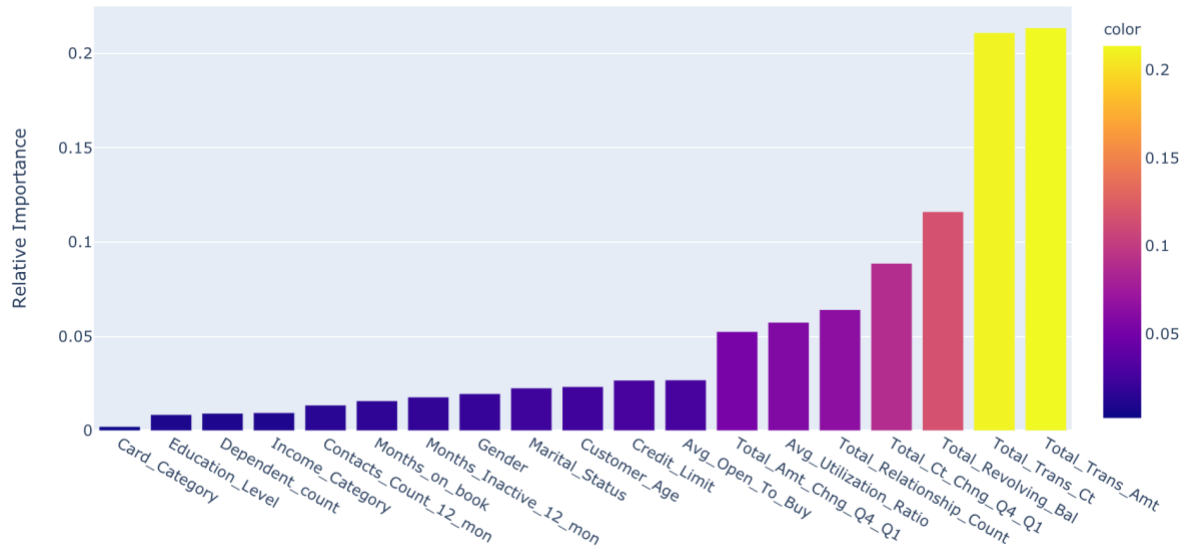

Figure 21 Random Forest feature importance

I use feature importance method of Random Forest to figure out what predictor variables the random forest considers most important. Feature importance can give us insight into a problem by telling us what variables are the most discerning between classes. According to the Random Forest feature importance figure (see Figure 21), the top three importance figures to the leaving of customer are total transaction amount, total transaction count and total revolving balance.

4. Conclusions and Recommendations

It clearly shows in both EDA and Random Forest feature importance figure (see Figure 21) that the product variables have made significant signals to the churning of customer.

Recommendations:

- It is necessary to pay attention in changing in product variables. The customer who will have high risk of churning:
    - Low number of products (2-5 average products)
    - High inactive months (average 2-3 months)
    - High number of contacts (2-4 contacts on average)
    - Low revolving balance amount (under 600)
    - Low in total transaction amount (under 2300)
    - The total transaction range from 30-50 times
    - Change in transaction range from 0.25-0.75 times.

    - Low utilization ratio (almost 0)
- Use the prediction model to predict the customer at highest risk of churning.

Improvements:

- Building more models to make a comparison.
- Including other feature selection methods (chi-square, variance threshold, etc).
- Handling the outliers if using other models.