香 港 大 學

THE UNIVERSITY OF HONG KONG

# COMP7103C Data Mining

Sem 2 2022/23

# Thera Bank Personal Loan Modeling

| Name | Student ID |
|---|---|
| KONG Xiangzhe | 3036033508 |
| YU Yang | 3036032281 |
| LAN Ting | 3036033405 |
| WU Qihan | 3036034590 |

***** For The detailed implementation please check our other submission files, or click this link: [https://drive.google.com/file/d/1M2-AprqRAglWpZerIFSk-R7vRPDrqQ50/view?usp=sharing](https://drive.google.com/file/d/1M2-AprqRAglWpZerIFSk-R7vRPDrqQ50/view?usp=sharing)**

# Table of Content

# 1. Project Description

### *1.1 Business Background*

A bank is nevertheless considered to have a growing customer base even if it only has a small number of borrowers for its lending products. As a consequence of this, the financial institution desires to expand its lending business, transform its depositors into loan borrowers, and widen the gap between its deposit and loan yields. In order to accomplish this goal, the retail credit section of the bank initiated a marketing campaign to stimulate personal loan business from a certain category of customers.[1] The department's other goal was to locate new customers who were in the industry of handling personal loan transactions through data analysis.

### *1.2 Analyze Problems*

(1) How many of the target market's consumers have engaged in the bank's pertinent business, and how are the clients of different businesses?
(2) What traits do clients who have dealt with the deposit and personal loan company often share?
(3) What factors deposit consumers have that help them become successful personal loan borrowers, as well as the conversion rate of deposit customers to personal loan borrowers?
(4) Based on the customer profile for personal loans, determine how the consumer manages the business.

### *1.3 Project Instruction*

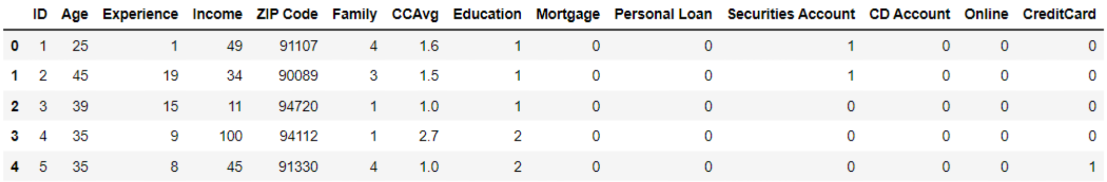
(1) Data Source: The kaggle site is where the data for this project was obtained. This data set displays 5,000 customer records for a bank's loan marketing campaign over the course of a certain year.
(2) Data Description:

| Feature | Description |
|---|---|
| ID | Customer ID |
| Age | Customer's age in completed years |
| Experience | Years of professional experience |
| Income | The annual income of the customer (USD'000) |
| ZIPCode | Home Address ZIP code |
| Family | Family size of the customer |
| CCAvg | Average spending on credit cards per month (USD'000) |
| Education | Education Level |

| | |
|---|---|
| Mortgage | Value of house mortgage if any (USD'000) |
| Personal Loan | If the customer accepts the personal loan. |
| Securities Account | If the customer has a securities account with the bank. |
| CD Account | If the customer has a certificate of deposit account with the bank. |
| Online | If the customer uses Internet banking facilities. |
| CreditCard | If the customer uses a credit card issued by the bank. |

Table 1.3.1 Data Description

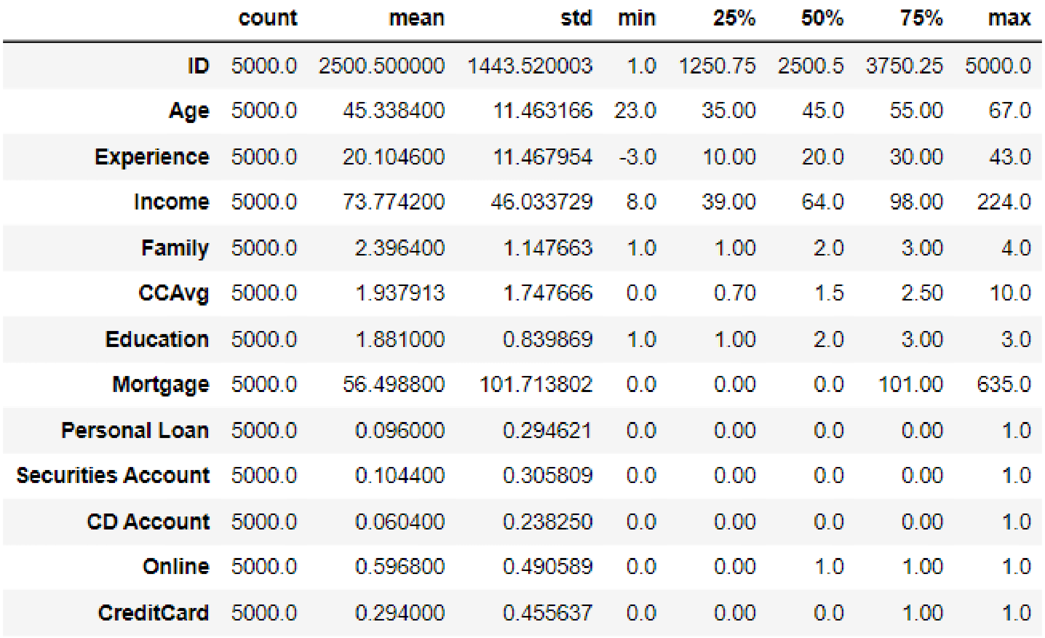
## 2. Data Cleaning and Preprocessing



| | ID | Age | Experience | Income | ZIP Code | Family | CCAvg | Education | Mortgage | Personal Loan | Securities Account | CD Account | Online | CreditCard |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 25 | 1 | 49 | 91107 | 4 | 1.6 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| **1** | 2 | 45 | 19 | 34 | 90089 | 3 | 1.5 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| **2** | 3 | 39 | 15 | 11 | 94720 | 1 | 1.0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| **3** | 4 | 35 | 9 | 100 | 94112 | 1 | 2.7 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| **4** | 5 | 35 | 8 | 45 | 91330 | 4 | 1.0 | 2 | 0 | 0 | 0 | 0 | 0 | 1 |

Table 2.1 Preview Data

We delete duplicate records and remove the postcode field-ZIP Code.



| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **ID** | 5000.0 | 2500.500000 | 1443.520003 | 1.0 | 1250.75 | 2500.5 | 3750.25 | 5000.0 |
| **Age** | 5000.0 | 45.338400 | 11.463166 | 23.0 | 35.00 | 45.0 | 55.00 | 67.0 |
| **Experience** | 5000.0 | 20.104600 | 11.467954 | -3.0 | 10.00 | 20.0 | 30.00 | 43.0 |
| **Income** | 5000.0 | 73.774200 | 46.033729 | 8.0 | 39.00 | 64.0 | 98.00 | 224.0 |
| **Family** | 5000.0 | 2.396400 | 1.147663 | 1.0 | 1.00 | 2.0 | 3.00 | 4.0 |
| **CCAvg** | 5000.0 | 1.937913 | 1.747666 | 0.0 | 0.70 | 1.5 | 2.50 | 10.0 |
| **Education** | 5000.0 | 1.881000 | 0.839869 | 1.0 | 1.00 | 2.0 | 3.00 | 3.0 |
| **Mortgage** | 5000.0 | 56.498800 | 101.713802 | 0.0 | 0.00 | 0.0 | 101.00 | 635.0 |
| **Personal Loan** | 5000.0 | 0.096000 | 0.294621 | 0.0 | 0.00 | 0.0 | 0.00 | 1.0 |
| **Securities Account** | 5000.0 | 0.104400 | 0.305809 | 0.0 | 0.00 | 0.0 | 0.00 | 1.0 |
| **CD Account** | 5000.0 | 0.060400 | 0.238250 | 0.0 | 0.00 | 0.0 | 0.00 | 1.0 |
| **Online** | 5000.0 | 0.596800 | 0.490589 | 0.0 | 0.00 | 1.0 | 1.00 | 1.0 |
| **CreditCard** | 5000.0 | 0.294000 | 0.455637 | 0.0 | 0.00 | 0.0 | 1.00 | 1.0 |

Table 2.2 Perform Descriptive Statistics

Analysis:

(1) Continuous Features: Age and work experience distributions are rather wide, as indicated by the fact that their means and medians are reasonably near but that their standard deviations are not small; other factors to consider include annual income, the average monthly use of credit cards, and collateral value. The standard deviations of annual income and collateral value are very large, and the degree of right skew is very

high. In particular, the median value of the collateral is 0; this shows that more than 50% of customers do not have mortgages on their homes. The mean values of all are higher than the median, indicating that they are all right-skewed distributions with extreme values.

(2) Categorical Features: According to the average value of the five two-category fields, the proportion of sampled customers who have handled credit card business is approximately 29.40%; the proportion of sampled customers who have handled securities business is approximately 10.44%; the proportion of sampled customers who have handled personal loan business is approximately 9.60%; the proportion of sampled customers who have handled deposit business is approximately 6.04%; and the proportion of sampled customers who have used online banking is approximately 59.68%.

## 3. Overall Analysis of Features

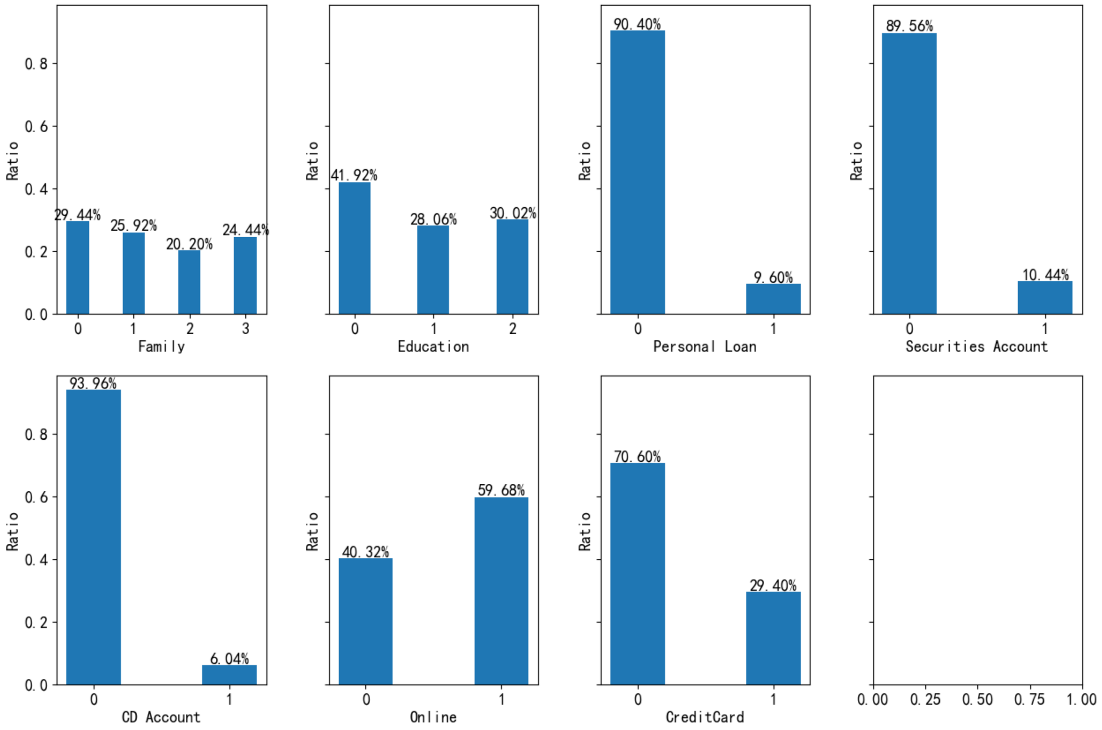


Figure 3.1 Overall Analysis of Features

Analysis:

(1) Family Size: The family size is one person, meaning that single customers are most prevalent; the next highest family size in terms of customers is two people, followed by four people, and then three people, in that order. Nevertheless, the customer counts for the four family sizes vary from one another. Between 20% and 30% of the overall number of consumers are small, and they are dispersed rather equally.

(2) Educational Level: The majority of clients have a bachelor's degree or less, although customers with a bachelor's degree or above made up more than half of the total. It is clear that the majority of the bank's clients are people with college degrees.

(3) Response Behavior: Customers who have not handled personal loans, stocks,

deposits, or credit card transactions outnumber those who have by a large margin. It is clear that just a tiny portion of the target population for this marketing effort has interacted with the pertinent bank business.

# 4. Deposit and Loan Business Customer Portrait

## *4.1 Deposit Customer Portrait*

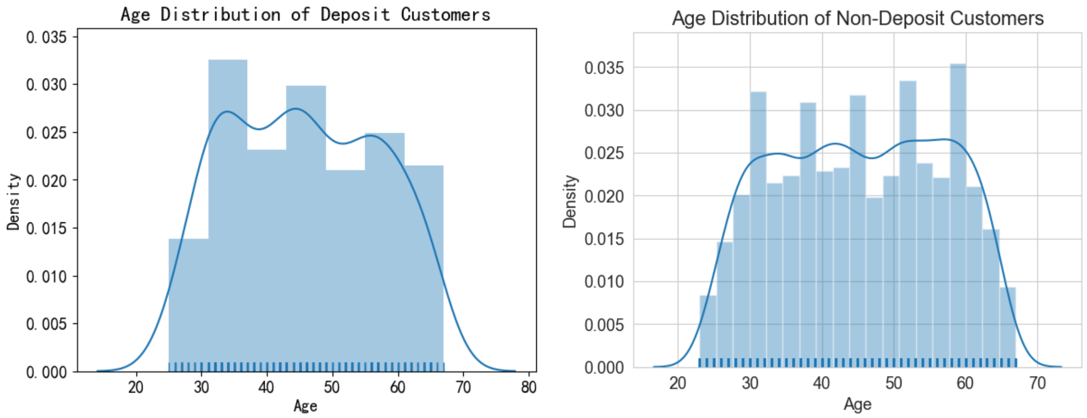


Figure 4.1.1 Age Distribution in Deposit

In terms of age distribution: Deposit clients are mostly between the ages of 33 and 55. Customers between the ages of 33 and 45 make up the majority of their clientele. Deposit clients tend to be youthful and middle-aged, as can be shown.

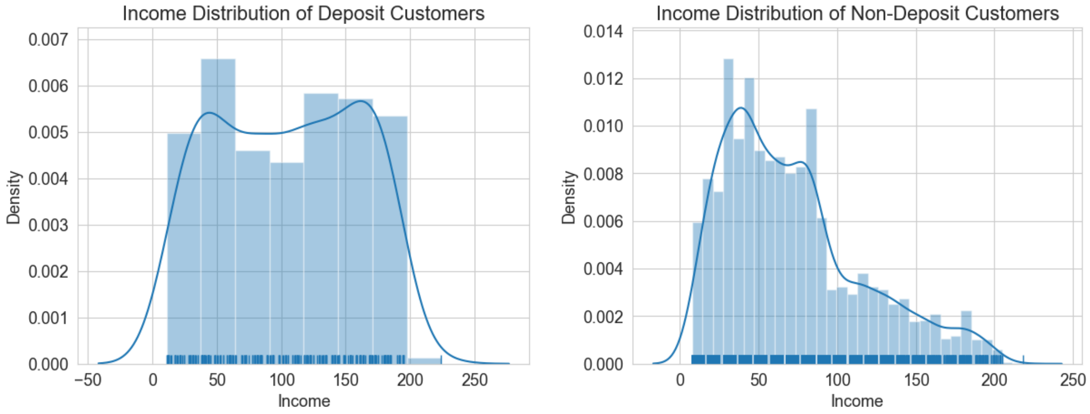


Figure 4.1.2 Income Distribution in Deposit

In terms of annual income distribution: The distribution of annual income of non-deposit customers is to the right, but the overall concentration is around 40,000 US dollars, indicating that non-deposit customers are primarily low- to middle-income; in contrast, the distribution of annual income of deposit customers is fairly broad and even. Focus on amounts between 40,000 and 160,000 US dollars.

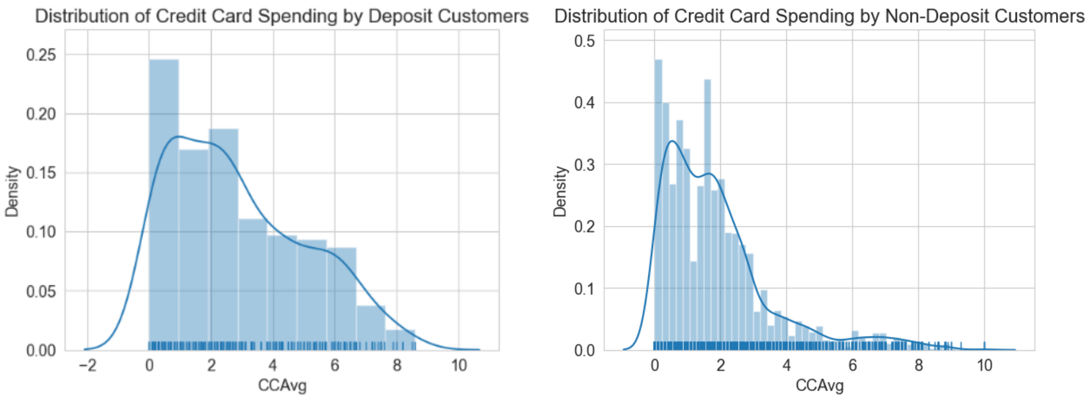


Figure 4.1.3 Distribution of Credit Card in Deposit

In terms of distribution of average monthly credit card spending: The average monthly credit card consumption of depositors and non-deposit customers is centered at around 1,000 US dollars, indicating that the total level of credit card consumption of the two groups of clients is not large. There are consumers who consume a lot, and the distribution is right-skewed.

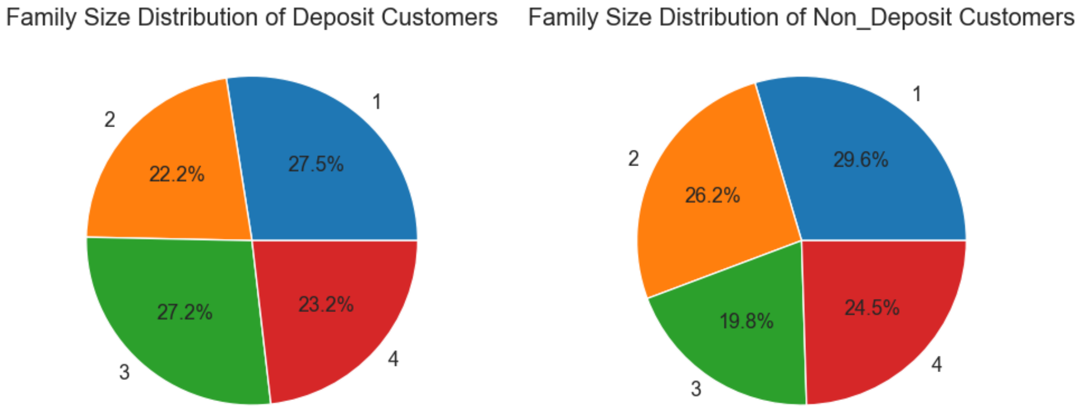


Figure 4.1.4 Family Size Distribution in Deposit

There is an average of 1 person in each family, which means that single and divorced customers make up the majority of depositors and non-depositors; however, the average family size among depositors is three or four people, i.e. married with children. Customers with a family size of two, or married without children, have a higher proportion of customers.

### *4.2 Personal Loan Customer Portrait*

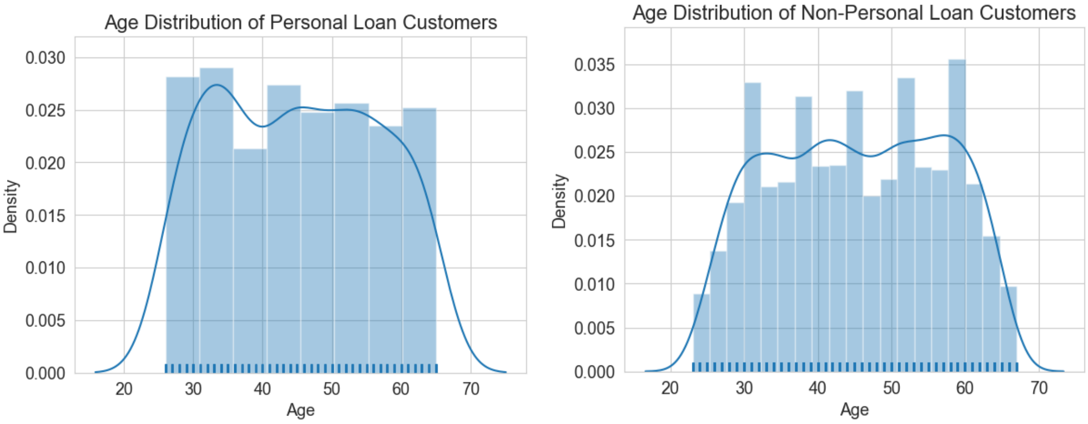


Figure 4.2.1 Age Distribution in Loan

In terms of age distribution: The majority of personal loan borrowers are between the ages of 32 and 50; those between the ages of 32 and 38 and those between the ages of 40 and 50 have higher loan balances than those in other age brackets. Nearby clients of similar ages are less prevalent.

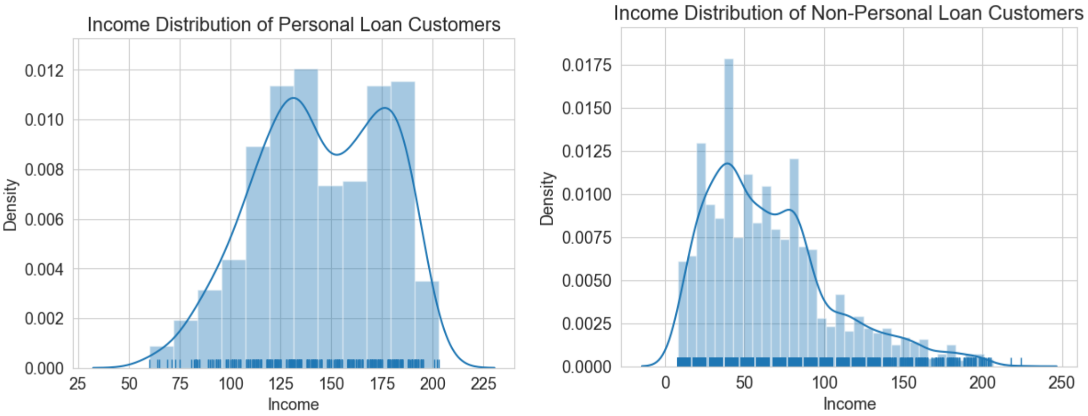


Figure 4.2.2 Income Distribution in Loan

In terms of annual income distribution: While the annual income of personal loan customers is relatively evenly distributed, concentrated between 130,000 US dollars and 175,000 US dollars, primarily middle-to-high income; at the same time, on the right side of 98,000 US dollars, the kernel density curve of personal loan customers is rising sharply, indicating that the majority of these customers have annual incomes between 40,000 and 80,000 US dollars. It is clear that the cutoff limit is an annual income of 98,000 USD, and consumers whose income exceeds that amount are more likely to obtain personal loans.

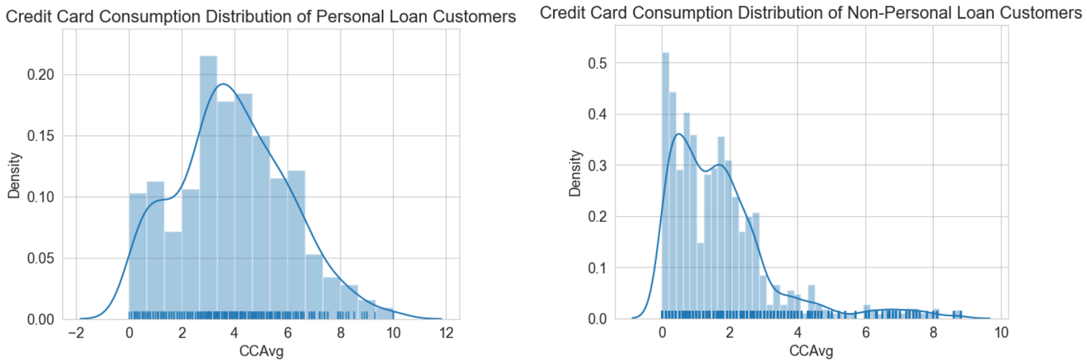


Figure 4.2.3 Credit Card Consumption Distribution in Loan

In terms of monthly average credit card consumption: The distribution of average monthly credit card consumption is relatively even, and it is concentrated at about 3.5 thousand US dollars, mainly in the middle and high consumption levels. The average monthly credit card consumption of non-personal loan customers is distributed to the right, but it is concentrated between 0.5 and 1.7 thousand US dollars, and the consumption level is generally low. The kernel density curve of customers who take out personal loans simultaneously climbs significantly and soon achieves its top on the right side of USD 2.8 thousand, whereas the kernel density curve of customers who take out non-personal loans drastically falls. As can be observed, a cut-off value of USD 2.8 thousand and the typical monthly credit card usage more likely to do personal loan business are customers with an amount above this threshold.

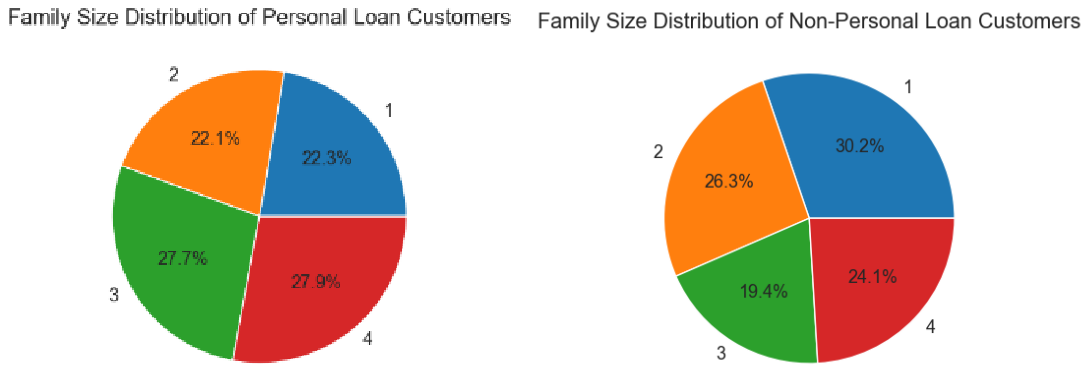


Figure 4.2.4 Family Size Distribution in Loan

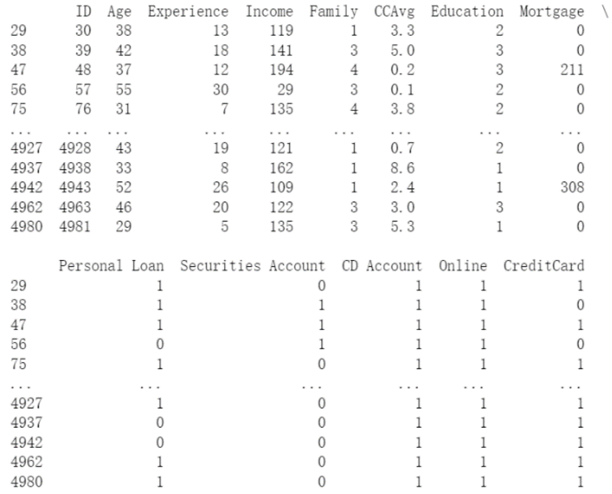
Non-personal loan customers are mainly single or married without children, while personal loan customers are mainly married with children.

# 5. Correlation Analysis

## *5.1 Different customers' percents in CDAccount customer*

The next part is about correlation analysis. We need to analyze some relations between data of several columns. There are several attributes: ID, Age, Experience, Income , Family , CCAvg, Education , Mortgage, Personal Loan , Securities Account , CD Account, Online, CreditCard.

First, our purpose is to gain more personal loan customers. Therefore, we need to investigate some relations between some features. Our purpose is to convert CDAccount customers to personal loan customers. We get the data of CDAccount customers at the first step. Then, we calculate the percentages of several kinds of customers in CDAccount customers. We separately calculate the percentages of Personal Loan customers, online customers and credit card customers in CD Account customers. And the results are 46.36%, 93.71%, 79.47% separately. Thus, the percentage of Personal loan customers in CDAccount customers is too low. Then, we need to do some further research.



```
        ID  Age  Experience  Income  Family  CCAvg  Education  Mortgage  \
29      30   38          13     119       1    3.3          2         0
38      39   42          18     141       3    5.0          3         0
47      48   37          12     194       4    0.2          3       211
56      57   55          30      29       3    0.1          2         0
75      76   31           7     135       4    3.8          2         0
...    ...  ...         ...     ...     ...    ...        ...       ...
4927  4928   43          19     121       1    0.7          2         0
4937  4938   33           8     162       1    8.6          1         0
4942  4943   52          26     109       1    2.4          1       308
4962  4963   46          20     122       3    3.0          3         0
4980  4981   29           5     135       3    5.3          1         0

      Personal Loan  Securities Account  CD Account  Online  CreditCard
29                1                   0           1       1           1
38                1                   1           1       1           0
47                1                   1           1       1           1
56                0                   1           1       1           0
75                1                   0           1       1           1
...             ...                 ...         ...     ...         ...
4927              1                   0           1       1           1
4937              0                   0           1       1           1
4942              0                   0           1       1           1
4962              1                   0           1       1           1
4980              1                   0           1       1           1
```

Figure 5.1.1 CDAccount customers data

### *5.2 Research about features which affect personal loan*

Second, we did some further research about the differences between these two kinds of customers: CD Account and Personal Loan customers, CD Account but not Personal loan customers. We want to compare these two kinds of customers to find out which features account for the differences. We selected several features including Age, Income, CCAvg, Mortgage, Family, Education. We wrote theses codes: “plt.rcParams['font.sans-serif'] =['SimHei'] plt.rcParams['axes.unicode_minus'] = False plt.rc('font',size=15)”to use plt library to set some parameters. Also, we used the sns library to compare the performance of these two kinds of customers on these features. Then, we use histograms to show the results. For age, the difference is not obvious. However, for other features, the differences are extremely different. High income customers are easy to become personal loan customers because more loan customers have high income. High CCAvg customers are easy to become personal loan customers because more loan customers have high CCAvg. High mortgage customers are easy to become personal loan customers because more loan customers have high mortgages. Big family size customers are easy to become personal loan customers because more loan customers have big family size. Customers who receive better education are easy to become personal loan customers because more loan customers receive better education.

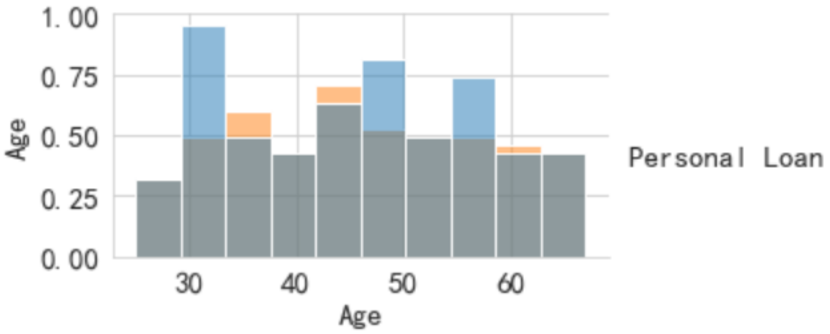


Figure 5.2.1 Age

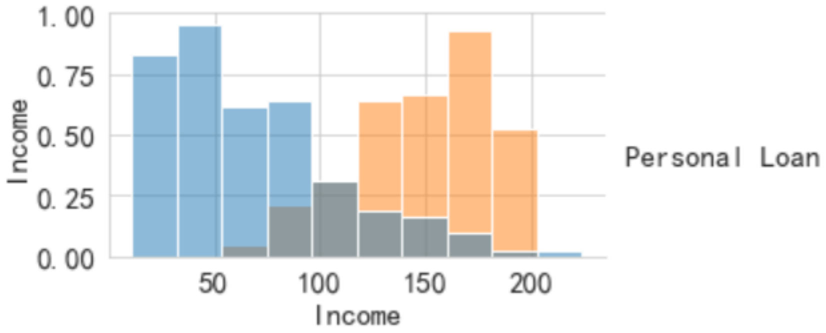


Figure 5.2.2 Income

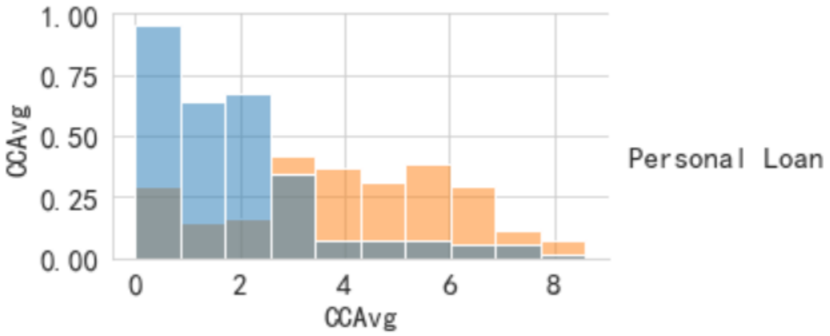


Figure 5.2.3 CCAvg

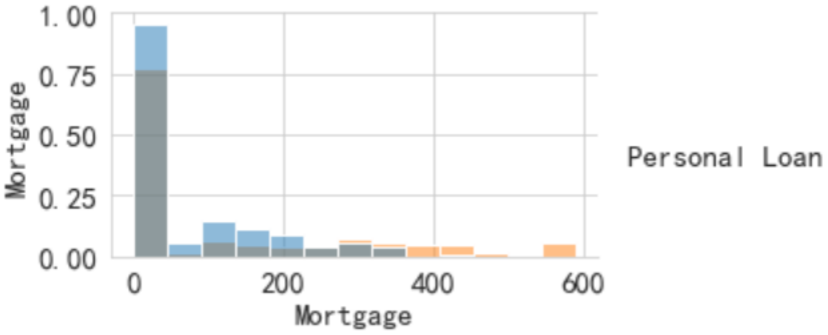


Figure 5.2.4 Mortgage

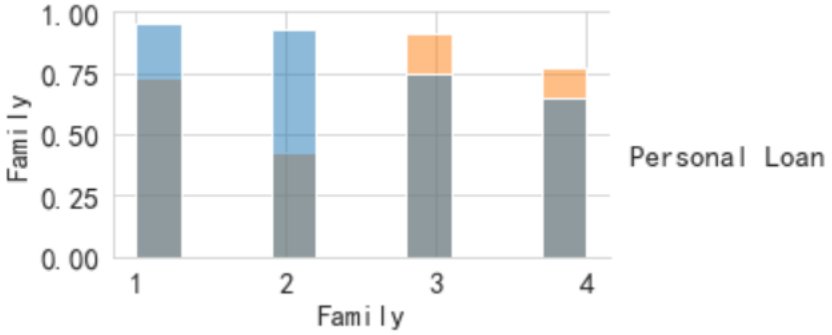


Figure 5.2.5 Family

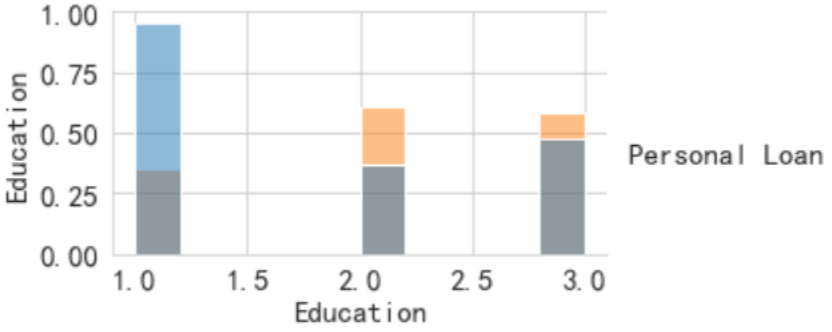


Figure 5.2.6 Education

### *5.3 Suggestions*

The vast majority of deposit customers handle credit card or online banking at the same time, but few handle personal loan business at the same time, and the conversion rate from deposit customers to personal loan customers is not high.

Suggestions:

(1) The deposit business is the basic business of the bank, which has an important driving and derivative effect on other businesses; but only 6.10% of the audience of the marketing campaign has opened a deposit account in the bank, so deposits should be further expanded Increase the market share of the business and expand the deposit customer base.

(2) There is a high degree of distinction between personal loan and non-personal loan customers among deposit customers, so we should make full use of and revitalize deposit customers, and encourage them to pay more attention to personal loan customers by increasing their personal loan marketing efforts and providing personal loan discounts. transform.

(3) In view of the vast majority of deposit customers' habit of using online banking, online banking can be fully utilized as a platform and medium to push personalized services to deposit customers and induce deposit customers to handle more related businesses.(4) Precise marketing of deposit and personal loan business should be

carried out according to customer portraits to reduce marketing costs and improve customer acquisition efficiency.

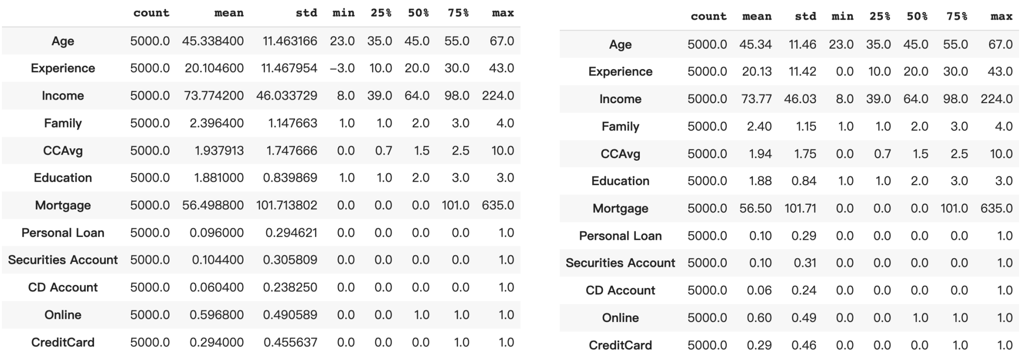
These suggestions are based on analysis. Thera Bank can use these suggestions to achieve their own goals and increase the conversion rate from depositors to personal loan users.

# 6. Modeling

In this part, we're going to build the classification models using different algorithms. Before feeding data to the model, a feature engineering is utilized to process the dataset in order to improve the model performance.
We use the scikit-learn library to implement six classification models: Decision Tree, Random Forest, AdaBoost, SVM , KNN and Logistic Regression. The classification result of each model is analyzed in detail in section 6.2 to 6.7.

## *6.1 Feature Engineering*

Firstly , outlier treatment is done to the dataframe.From the description chart of the dataset below, it can be observed that the minimum value of experience is -3.0. To treat this kind of outlier, we take the absolute value of the experience.After the treatment, we can observe from the figure 6.1.1(b) that the experience feature doesn't include negative value now.



| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Age | 5000.0 | 45.338400 | 11.463166 | 23.0 | 35.0 | 45.0 | 55.0 | 67.0 |
| Experience | 5000.0 | 20.104600 | 11.467954 | −3.0 | 10.0 | 20.0 | 30.0 | 43.0 |
| Income | 5000.0 | 73.774200 | 46.033729 | 8.0 | 39.0 | 64.0 | 98.0 | 224.0 |
| Family | 5000.0 | 2.396400 | 1.147663 | 1.0 | 1.0 | 2.0 | 3.0 | 4.0 |
| CCAvg | 5000.0 | 1.937913 | 1.747666 | 0.0 | 0.7 | 1.5 | 2.5 | 10.0 |
| Education | 5000.0 | 1.881000 | 0.839869 | 1.0 | 1.0 | 2.0 | 3.0 | 3.0 |
| Mortgage | 5000.0 | 56.498800 | 101.713802 | 0.0 | 0.0 | 0.0 | 101.0 | 635.0 |
| Personal Loan | 5000.0 | 0.096000 | 0.294621 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| Securities Account | 5000.0 | 0.104400 | 0.305809 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| CD Account | 5000.0 | 0.060400 | 0.238250 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| Online | 5000.0 | 0.596800 | 0.490589 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 |
| CreditCard | 5000.0 | 0.294000 | 0.455637 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 |

(a) before treatment

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Age | 5000.0 | 45.34 | 11.46 | 23.0 | 35.0 | 45.0 | 55.0 | 67.0 |
| Experience | 5000.0 | 20.13 | 11.42 | 0.0 | 10.0 | 20.0 | 30.0 | 43.0 |
| Income | 5000.0 | 73.77 | 46.03 | 8.0 | 39.0 | 64.0 | 98.0 | 224.0 |
| Family | 5000.0 | 2.40 | 1.15 | 1.0 | 1.0 | 2.0 | 3.0 | 4.0 |
| CCAvg | 5000.0 | 1.94 | 1.75 | 0.0 | 0.7 | 1.5 | 2.5 | 10.0 |
| Education | 5000.0 | 1.88 | 0.84 | 1.0 | 1.0 | 2.0 | 3.0 | 3.0 |
| Mortgage | 5000.0 | 56.50 | 101.71 | 0.0 | 0.0 | 0.0 | 101.0 | 635.0 |
| Personal Loan | 5000.0 | 0.10 | 0.29 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| Securities Account | 5000.0 | 0.10 | 0.31 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| CD Account | 5000.0 | 0.06 | 0.24 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| Online | 5000.0 | 0.60 | 0.49 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 |
| CreditCard | 5000.0 | 0.29 | 0.46 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 |

(b) after treatment

Figure 6.1.1 dataset description

After the data has been standardized using the *StandardScaler( )* Function, boxplots are used to check for outliers. From the boxplot of all features, we can see that the three features *CCAvg_std,Income_std* and *Mortage_std* contain outliers. as shown in the figure 6.1.2 and figure 6.1.3, the outliers are removed after z-score treatment.

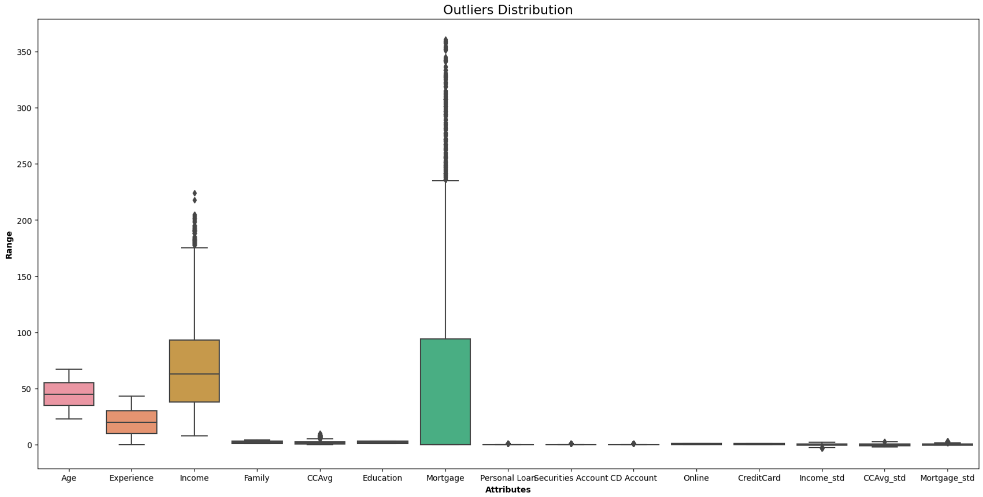


Figure 6.1.2 boxplot of dataset

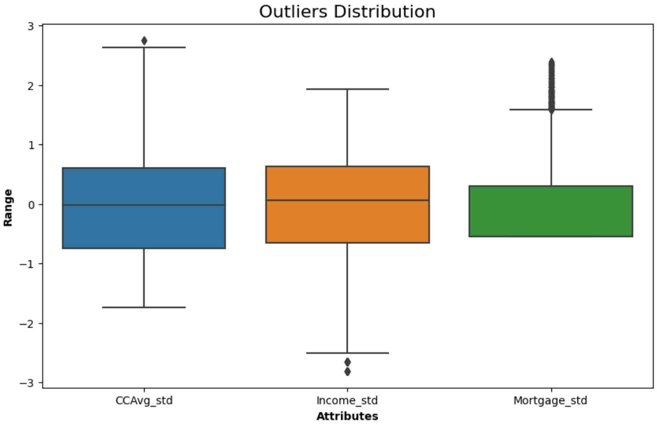


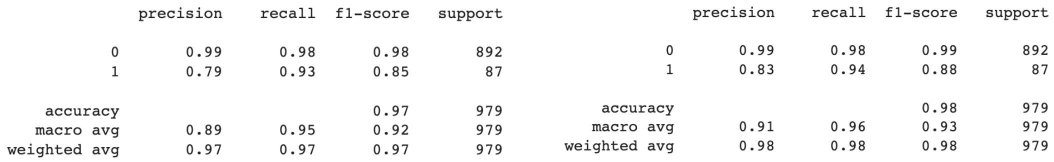
Figure 6.1.3 boxplot after z-score treatment

Next,we split the dataset into a training dataset and a test data set. The test dataset has a proportion of 0.2. We counted the number of samples. The total number of samples was 4895, of which 346 were positive training labels and 3570 were negative labels.Since the number of negative samples far exceeded the number of positive samples, the bias of the imbalance dataset can influence the effectiveness of the machine learning algorithms. Usually, data that are unbalanced in class are categorized in favor of the majority class(Rok Blagus & Lara Lusa,2013). Thus, oversampling is needed.

The technique we use for treating imbalance data is Synthetic Minority Oversampling TEchnique (SMOTE) .After this treatment, the number of positive and negative training samples is the equal, both are 3570. Now the dataset is balanced and we can feed this dataset to our classifiers.

### *6.2 Decision Tree*

Firstly, a decision tree with default configuration is created by *DecisionTreeClassifier algorithm.* The results are shown in the figure 6.2.1.According to the results, this classifier can achieve reasonable results,with accuracy of 0.97. Also, for the positive labels, the precision is 0.79 and recall achieves 0.93. Then we pruned the decision tree classifier to achieve optimal results.The depth of decision tree is limited to 4. After pruning, the accuracy improved from 0.97 to 0.98 and the precision of the positive samples increased from 0.79 to 0.83.The visualization of the two decision tree classifiers are shown in the figure 6.2.2.



```
              precision    recall  f1-score   support

           0       0.99      0.98      0.98       892
           1       0.79      0.93      0.85        87

    accuracy                           0.97       979
   macro avg       0.89      0.95      0.92       979
weighted avg       0.97      0.97      0.97       979
```

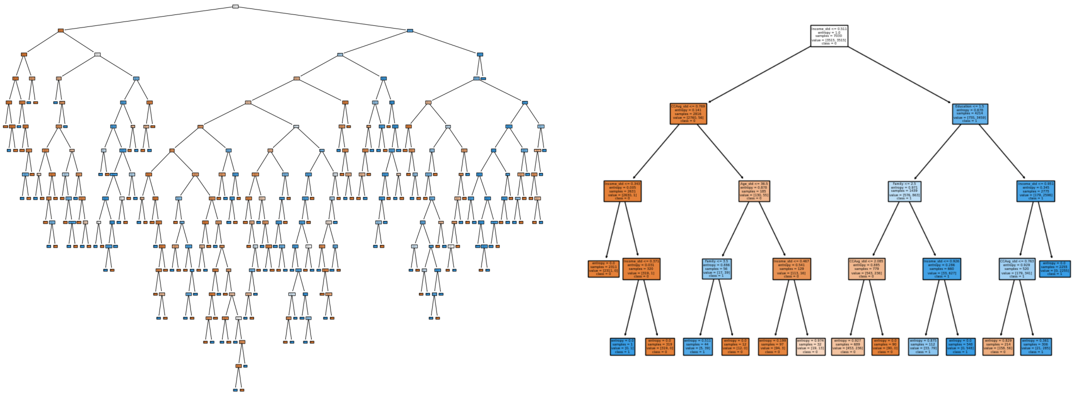
```
              precision    recall  f1-score   support

           0       0.99      0.98      0.99       892
           1       0.83      0.94      0.88        87

    accuracy                           0.98       979
   macro avg       0.91      0.96      0.93       979
weighted avg       0.98      0.98      0.98       979
```

(a) default classifier result (b) pruned classifier result

Figure 6.2.1 results of Decision Tree



(a)Default version (b)Pruned version

Figure 6.2.2 Decision tree visualization

The importance of each feature is shown as the figure. As shown in the figure 6.2.3, the top five important features are Education, Income and Family, Credit Card Average and Age.

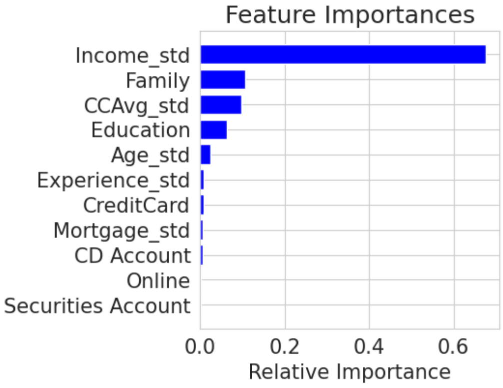
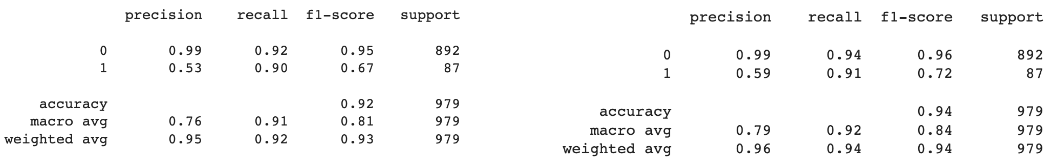


Figure 6.2.3 Relative importance of features

### *6.3 AdaBoost*

To tune the parameters, TPE (Tree-structured Parzen Estimator) algorithm is used to find the optimal number of estimators and learning rate. After 100 trials, the optimal number of estimators is 483 and the optimal learning rate is 0.467.Then we feed the best parameters to the model and get better results. The results of the AdaBoost classifier are shown in the figure 6.3.1.



```
              precision    recall  f1-score   support

           0       0.99      0.92      0.95       892
           1       0.53      0.90      0.67        87

    accuracy                           0.92       979
   macro avg       0.76      0.91      0.81       979
weighted avg       0.95      0.92      0.93       979
```

```
              precision    recall  f1-score   support

           0       0.99      0.94      0.96       892
           1       0.59      0.91      0.72        87

    accuracy                           0.94       979
   macro avg       0.79      0.92      0.84       979
weighted avg       0.96      0.94      0.94       979
```

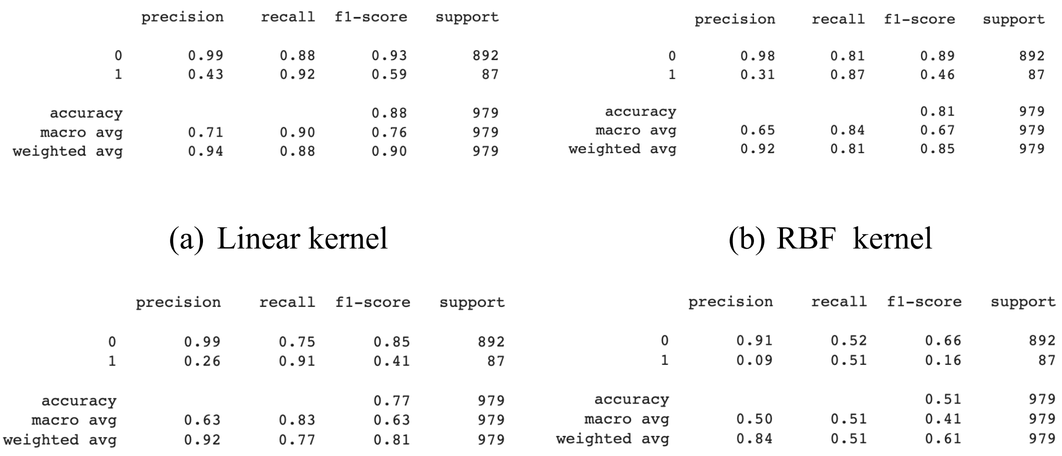
(a) results with default parameters

(b) results with optimal parameters

Figure 6.3.1 classification results of AdaBoost

### *6.4 Support Vector Machine*

There’re three Hyperparameters in SVM( Support Vector Machine): kernel, regularization and gamma. In this experiment, four types of kernels (including linear, polynomial, radial basis function (RBF), and sigmoid) are tried to optimize the classification result. The comparison of results is shown in the figure 6.4.1 below.



```
              precision    recall  f1-score   support

           0       0.99      0.88      0.93       892
           1       0.43      0.92      0.59        87

    accuracy                           0.88       979
   macro avg       0.71      0.90      0.76       979
weighted avg       0.94      0.88      0.90       979
```

(a) Linear kernel

```
              precision    recall  f1-score   support

           0       0.98      0.81      0.89       892
           1       0.31      0.87      0.46        87

    accuracy                           0.81       979
   macro avg       0.65      0.84      0.67       979
weighted avg       0.92      0.81      0.85       979
```

(b) RBF  kernel

```
              precision    recall  f1-score   support

           0       0.99      0.75      0.85       892
           1       0.26      0.91      0.41        87

    accuracy                           0.77       979
   macro avg       0.63      0.83      0.63       979
weighted avg       0.92      0.77      0.81       979
```

(c) Polynomial kernel

```
              precision    recall  f1-score   support

           0       0.91      0.52      0.66       892
           1       0.09      0.51      0.16        87

    accuracy                           0.51       979
   macro avg       0.50      0.51      0.41       979
weighted avg       0.84      0.51      0.61       979
```

(d ) Sigmoid kernel

Figure 6.4.1 results of SVM classifier

Among the four kernels, the best performing one is the linear kernel, which achieves an accuracy of 0.88. However, none of them can achieve good precision of positive training samples, especially the one with sigmoid kernel. As the kernel serves as a function to transform the training set of data into required form(Gupta, 2021), the difference in classification performance between different kernels may lie in the ability of the kernel to better fit the input data. In this case, the linear kernel  defined by the dot product of vectors can transform the data better so it’s easier to distinguish between categories in the high dimensional space.

### *6.5 KNN*

In this algorithm, we need to choose the number of neighbors first. As shown in the figure 6.5.1, the model performs best with the number of neighbors equal to one.

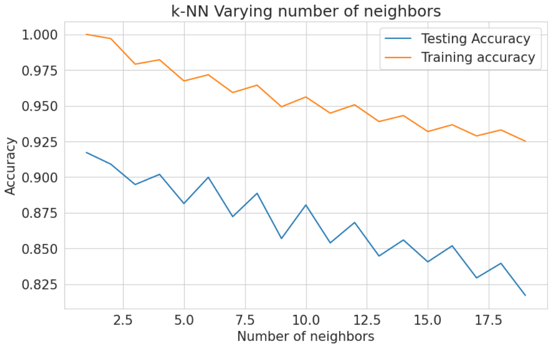
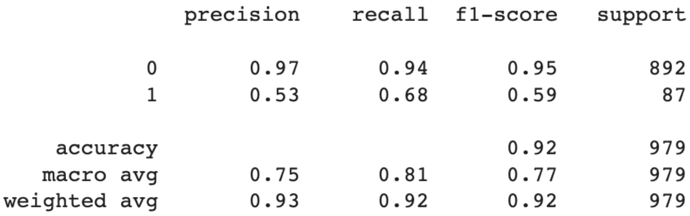


Figure 6.5.1 K-NN neighbors

The results and confusion matrix of the KNN model are shown in figure 6.5.2 and figure 6.5.3. In the KNN model, the recall is low. This may be due to the fact that we have a heavily imbalanced dataset. Although we used SMOTE on the training dataset, the test data set is still imbalanced. So the performance of the model is affected by the bias. According to the confusion matrix, 28 positive cases out of 87 total positive labels are classified to the wrong category, which is not a satisfactory result.



```
              precision    recall  f1-score   support

           0       0.97      0.94      0.95       892
           1       0.53      0.68      0.59        87

    accuracy                           0.92       979
   macro avg       0.75      0.81      0.77       979
weighted avg       0.93      0.92      0.92       979
```

Figure 6.5.2 results of KNN

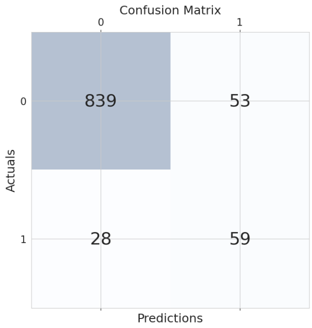


Figure 6.5.3 confusion matrix

### *6.6 Random Forest*

Empirically we choose k=5, by 5-fold validation, the result shows:

Best Hyper Parameters: {'max_samples': 4000}

Best Accuracy: 0.9250713175238925

It means that the bootstrap size is 4000, because the accuracy reaches 0.92507131752389250, which is the highest value seen in Figure 6.6.1.

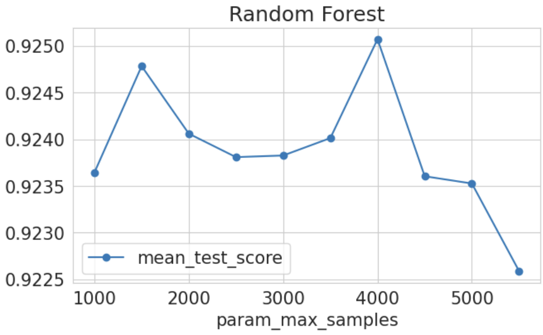


Figure 6.6.1 param_max_samples

In Figure 6.6.2 we can see there's no drop off in Mean validation accuracy seen as the training samples increase, so the final model will be fitted with all the training samples and the max_samples parameter is set to 4000.( Harris, C,2018)

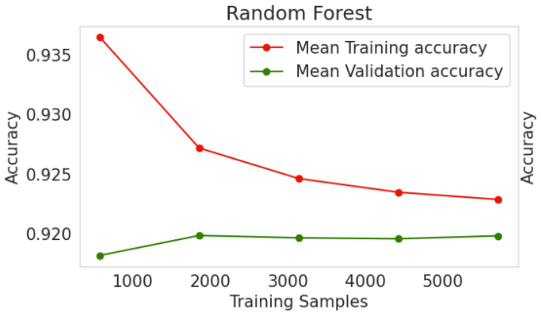
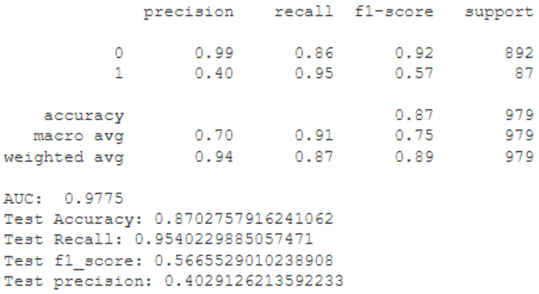


Figure 6.6.2 Training Samples

The result of the Random Forest Algorithm shows in Figure 6.6.3. The precision value is 0.40, which may be due to the imbalanced value as there's only 87 positive labels which accounts for a small percentage. Although we used SMOTE technique to over sample, this may still affect the result.



```
              precision    recall  f1-score   support

           0       0.99      0.86      0.92       892
           1       0.40      0.95      0.57        87

    accuracy                           0.87       979
   macro avg       0.70      0.91      0.75       979
weighted avg       0.94      0.87      0.89       979

AUC:  0.9775
Test Accuracy: 0.8702757916241062
Test Recall: 0.9540229885057471
Test f1_score: 0.5665529010238908
Test precision: 0.4029126213592233
```

Figure 6.6.3 Result of Random Forest

### *6.7 Logistic Regression*

The last algorithm we used is Logistic Regression. Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables.(What is logistic regression? (no date) IBM. ) First we use the 5-fold validation to find the best hyper parameters. We can see the result from Figure 6.7.1.

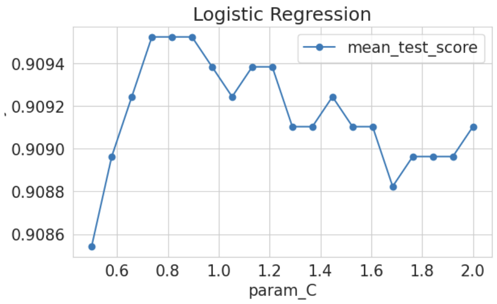


Figure 6.7.1 Param_C

Best Hyper Parameters: {'C': 0.7368421052631579}
Best Accuracy: 0.9095238095238095

In Figure 6.7.2 it shows that variance decreases as training samples increase, so to best generalize unseen data the final model will be fitted with all the training samples and the inverse regulation parameter C is set to 1.0526315789473684.

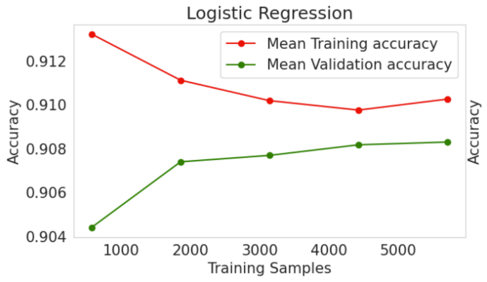
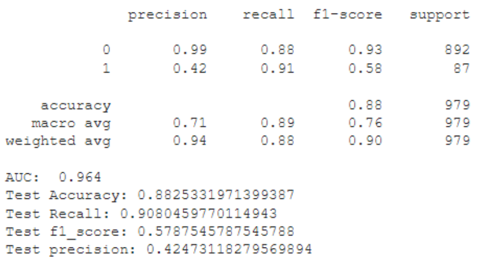


Figure 6.7.2 Training Samples

The result of the Random Forest Algorithm shows in Figure 6.7.3. The precision is 0.42, which is still low.



```
              precision    recall  f1-score   support

           0       0.99      0.88      0.93       892
           1       0.42      0.91      0.58        87

    accuracy                           0.88       979
   macro avg       0.71      0.89      0.76       979
weighted avg       0.94      0.88      0.90       979

AUC:  0.964
Test Accuracy: 0.8825331971399387
Test Recall: 0.9080459770114943
Test f1_score: 0.5787545787545788
Test precision: 0.42473118279569894
```

Figure 6.7.3 Result of Logistic Regression

### *6.8 Comparison and Conclusion*

| | Accuracy | Precision | Recall | f1 score |
|---|---|---|---|---|
| Decision Tree | 0.98 | 0.83 | 0.94 | 0.88 |
| AdaBoost | 0.94 | 0.59 | 0.91 | 0.72 |
| Support Vector Machine | 0.88 | 0.43 | 0.92 | 0.59 |
| K-Nearest Neighbors | 0.92 | 0.53 | 0.68 | 0.59 |
| Random Forest | 0.87 | 0.40 | 0.95 | 0.57 |
| Logistic Regression | 0.88 | 0.42 | 0.91 | 0.58 |

Table 6.8.1 Comparison of the result

In conclusion, Decision Tree and AdaBoost outperform other algorithms in terms of all four metrics including accuracy, precision, recall and f1 score.

## *Reference*

Blagus, R., Lusa, L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics 14, 106 (2013). https://doi.org/10.1186/1471-2105-14-106

Gupta, A. (2021, June 1). *Kernel tricks in support Vector Machines*. Medium. Retrieved April 20, 2023, from https://medium.com/geekculture/kernel-methods-in-support-vector-machines-bb9409342c49

Harris, C., & Peer, W. (2018). A Comparative Analysis of Random Forest and Logistic Regression for Weed Risk Assessment. University of Maryland, College Park.

Mulani, S. (2022, August 3). *Using StandardScaler() function to standardize python data*. DigitalOcean. Retrieved April 19, 2023, from https://www.digitalocean.com/community/tutorials/standardscaler-function-in-python

What is logistic regression? (no date) IBM. Available at: https://www.ibm.com/topics/logistic-regression#:~:text=Resources-,What%20is%20logistic%20regression%3F,given%20dataset%20of%20independent%20variables. (Accessed: April 19, 2023).

## *Contribution Table*

| | |
|---|---|
| KONG Xiangzhe | Data preprocessing, Customer Portrait |
| WU Qihan | Correlation Analysis |
| YU Yang | Decision Tree,  Adaboost, SVM, KNN |
| LAN Ting | Random Forest, Logistic Regression |