GR5291 Final Project

Machine Learning Case Study for

Clients' Credit Card Default Payments in Taiwan

May 2022

Professor: David Rios

Ruolin Chen, rc3411

Enliang Dong, ed2810

Yutong Yang, yy3167

Yuning Zhou, yz3922

Xinyi Zhu, xz3057

ADVANCED DATA ANALYSIS

COLUMBIA UNIVERSITY

TABLE OF CONTENTS

1 Introduction
1.1 Research Overview
1.2 Data Description

2 EDA before Feature Engineering
2.1 Categorical Variables
2.2 Numerical Variables

3 Feature Engineering
3.1 Categorical Variables
3.2 Numerical Variables
3.2.1 ***PAY_AMT1 - PAY_AMT6***
3.2.2 ***BILL_AMT1 - BILL_AMT6***

4 EDA after Feature Engineering
4.1 Numerical Variables
4.2 Categorical Variables
4.2.1 Sumpay
4.2.2 AGE

5 Model Construction
5.1 KNN Classification
5.2 Logistic Classification
5.3 Decision Tree Classification
5.4 Random Forest Classification
5.5 Gradient Boosting Classifier with Random Forest

6 Model Comparison
6.1 Precision, Recall, F1 score
6.2 Validation Curve
6.3 Sort the Importance Features
6.4 ROC Curve for Model which Has the Best Performance

7 Summary

References

1 Introduction

1.1 Research Overview

This research is aimed at studying the case of customers' default payments in Taiwan through different machine learning models, as well as comparing the predictive accuracy of probability of default among them. The expansion of new business in credit cards has driven banks to lower the requirements for credit card approvals, and young people have become their target customers although they tend not to have enough income. In the beginning of 2006, the Taiwanese credit card debt reached $268 billion USD (Wang, 2022). Therefore, revealing the predictors for individuals' financial behavior associated with credit card default is of importance to achieve reliable credit risk management.

This research collects 20000 instances related from clients' default payment which includes 24 variables, ranging from basic information of customers and information about their bill and payment conditions. Machine learning techniques for classification, including KNN, Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting methodologies, are applied to predict the correct class of individual's default behavior. In order to train different models and improve the accuracy of prediction, this project randomly split the dataset to train dataset (20000 * 0.7) and test dataset (20000 * 0.3). And thus, model evaluation metrics are developed by comparing the performance of the set of machine learning algorithms.

The structure of this report is organized following the process of the exploration in deriving insights of this data source. Interpretation of variables is conducted prior to further analysis. Data quality of missing data and data inconsistency is checked, and this process is followed by the exploratory data analysis (EDA). Feature engineering helps to reduce the dimension of data, which provides convenience in conducting machine learning study. Machine learning algorithms ranging from supervised learning to unsupervised learning algorithms are applied in gaining deeper insights on predicting the clients' default behavior through the dataset. The results of the algorithms mentioned above are compared and discussed in terms of a variety of methodologies, such as accuracy and ROCs, as well as proper visualization presentation. Finally, a conclusion is given at the end of this report. Among the 5 machine learning techniques, Gradient Boosting Classification has the best performance in can accurately predicting the default behavior based on bank account records.

1.2 Data Description

The dataset containing information on clients' default payment is taken from an important (a cash and credit card issuer) bank in Taiwan. It records the behavior of credit card holders of this bank from October 2005 to the following 6 months, whereas the third quarter of 2006 was seen as the peak of the Taiwan Credit Card Crisis (Yeh and Lien, 2009). Data mining for revealing the real default rate was conducted by Yeh and Lien in 2009, and their study in predictive accuracy has introduced this dataset to academy.

This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. The interpretation of the 23 explanatory variables is as below:

Table.1 Overview of the Explanatory Variables

Variables	Description
X1: Amount of the given credit (NT dollar)	It includes both the individual consumer credit and his/her family (supplementary) credit.
X2: Gender	1 = Male; 2 = Female
X3: Education	1 = Graduate school; 2 = University; 3 = High school; 4 = Others
X4: Marital Status	1 = Married; 2 = Single; 3 = Others
X5: Age	Year
X6 - X11: History of Past Payment (From April to September, 2005)	X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; X8 = the repayment status in July, 2005; X9 = the repayment status in June, 2005; X10 = the repayment status in May, 2005; X11 = the repayment status in April, 2005 -1 = pay early; 0 = pay duly; 1 = payment delay for one month;. . .; 9 = payment delay for nine months and above.
X12 - X17: Amount of Bill Statement (NT dollar)	X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; X14 = amount of bill statement in July, 2005; X15 =amount of bill statement in June, 2005; X16 = amount of bill statement in May, 2005; X17 = amount of bill statement in April, 2005;
X18 - X23: Amount of previous payment	X18 = amount paid in in September, 2005;

(NT dollar)	X19 = amount paid in August, 2005; X20 = amount paid in July, 2005; X21 = amount paid in June, 2005; X22 = amount paid in May, 2005; X23 = amount paid in April, 2005;

2 EDA before Feature Engineering

The dataset contains the information on default payment, demographic information, payment history, credit limit, and bill statement of Taiwan credit card clients from April 2005 to September 2005.
The dataset includes 20000 records and 24 variables and is overall clean and under a good quality in terms of limited amount of missing data and undocumented column values.

The purpose of this analysis is to identify how the variables may have an impact on the payment default, and to study the correlation between them. Graphical and statistical analysis is applied to check categorical variables, including *DEFAULT*, *SEX*, *AGE*, *EDUCATION*, and *MARRIAGE*.

2.1 Categorical Variables

Default:

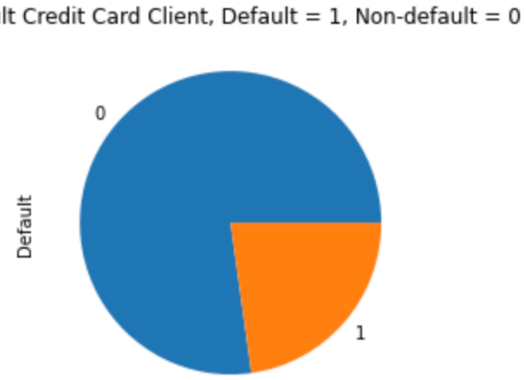

Fig.1 Pie Chart of Default and Non-default Clients

By marking default as 1 and non-default as 0, the number of default clients is 4558 and that of non-default clients is 15442, which means that 77% of clients are not expected to default payment whereas 23% of clients are expected to default the payment. This reveals the fact that most of the clients are paying bills on time.

Amount of Credit:

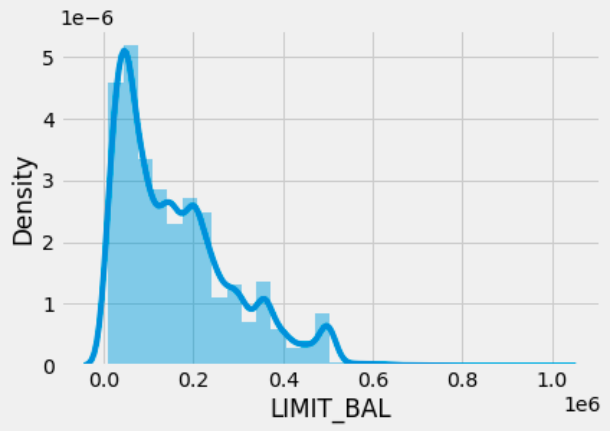

Fig.2 Histogram of LIMIT_BAL

The distribution of the amount of credit (NT dollar) is presented through the histogram in Fig.2. This shows that the distribution of credit amount is positively skewed: the median of credit is $130000, while there is only one person whose credit falls in the bracket of $1000000. The summary of quantiles of the amount of credit data is shown in Table.2 as below:

Table.2 Summary of amount of credit

MEAN	MIN	25%	50%	75%	MAX
163301.184	10000.000	50000.000	130000.000	230000.000	1000000.000

Genders:

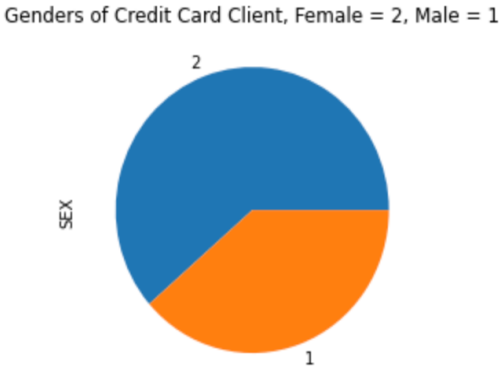

Fig.3 Pie Chart of Genders

By marking Female = 2 and Male = 1, we can see that females share a higher percentage in this sample than male. There are 12281 females and 7719 males in this sample. Whether the gender of a client will influence their behavior in credit card default requires further exploration in the following parts.

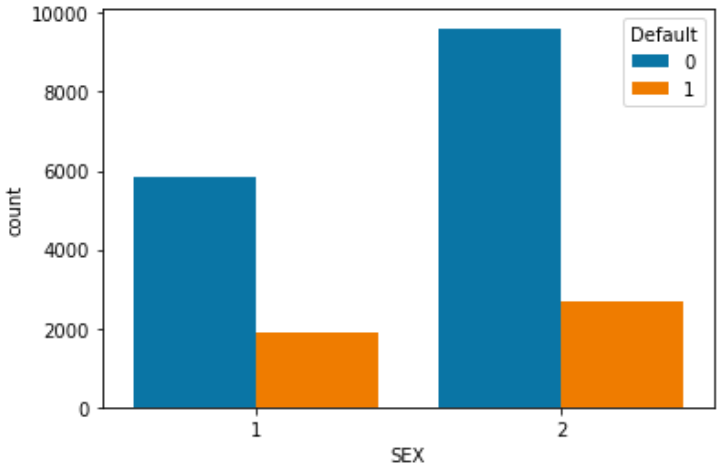

Fig.4 Histogram of Sex Associated with Default Behavior

While Female = 2 and Male = 1, and non-default = 0, default = 1, from the Fig.5 it is evident that females have overall less percentage of default payments with respect to males. Non-defaults have a higher proportion of Females (Sex = 2).

Education Levels:

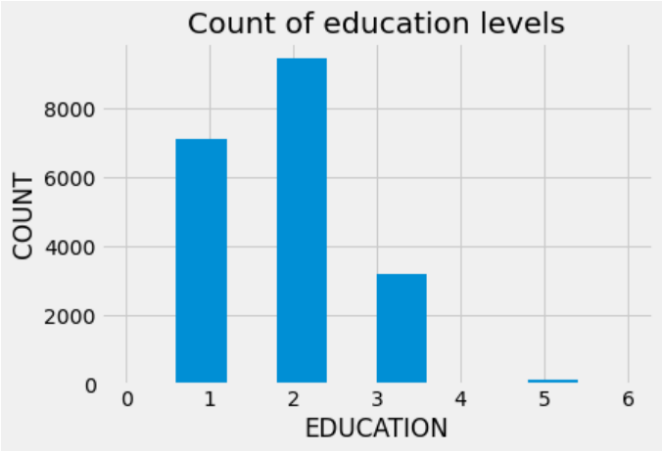

Fig.5 Histogram of Education Levels

The level of education is denoted by 1 = Graduate school, 2 = University, 3 = High school, 4 = Others, while 5 and 6 are unknown. By analyzing the information in Fig.6, most of the customers have an education background from graduate school or university.

Marital Status:

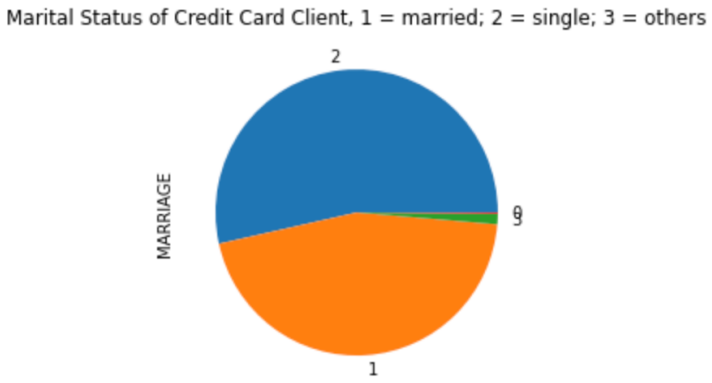

Fig.6 Histogram of Marriage Status

By denoting 1 = married, 2 = single, 3 = others, and 0 for unknown, most clients are married or single in this sample. There are 10702 (53.5%) persons married, 9033 (45.2%) persons single, while the remaining are classified as others or unknown.

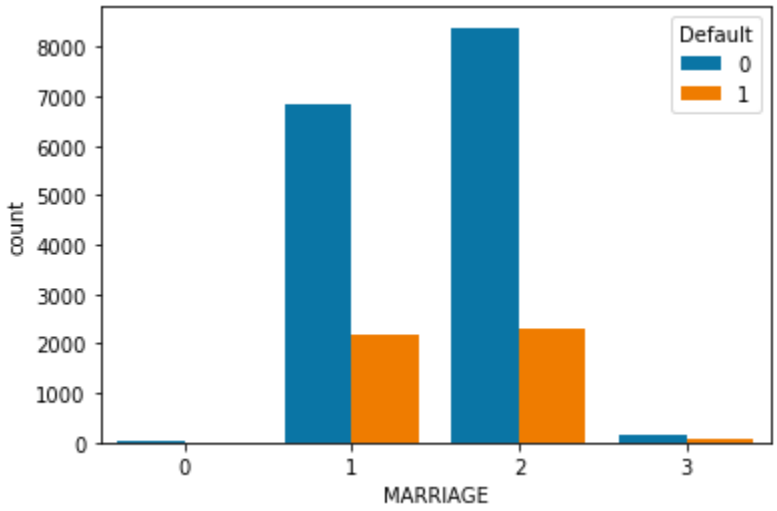

Fig.7 Histogram of Marriage Associated with Default Behavior

With a similar number of defaults in marital status “married” and “single”, Fig.7 shows that single clients are less likely to default when compared to married couples.

2.2 Numerical Variables

Age:

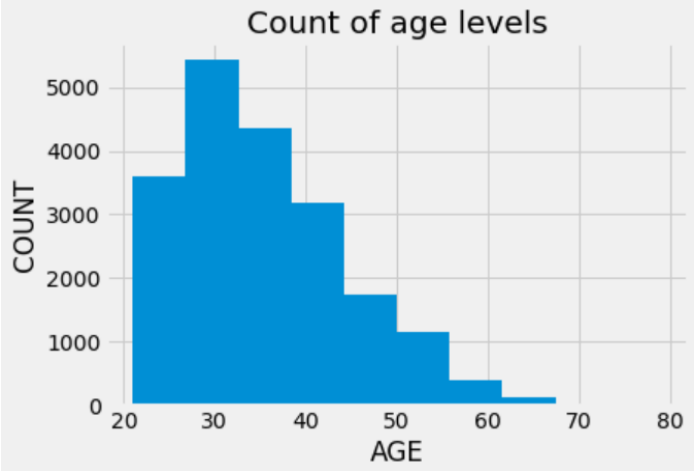

Fig.8 Histogram of Age Levels

According to the histogram Fig.8, the distribution of age is right skewed; most customers' ages fall in the range between 20 and 50. The median of age is 34 and the age level with the highest proportion is 29. More detailed analysis about the relationship between age and credit card default is conducted as below.

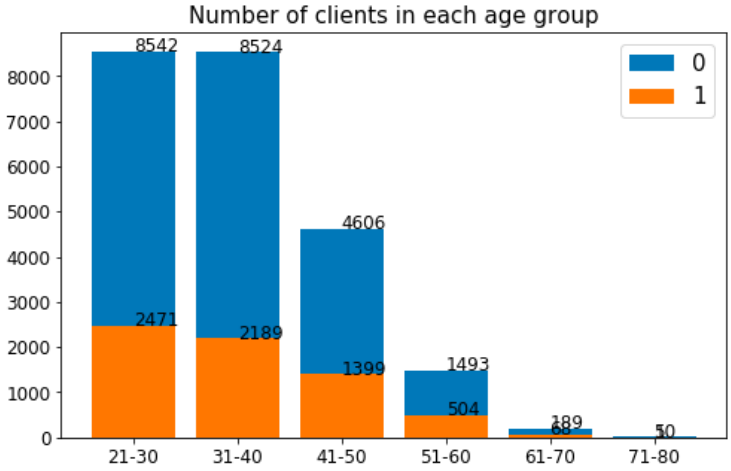

Fig.9 Histogram of Age Levels Associated with Default Behavior

Fig.9 shows a study in the relationship between age level and the corresponding number of defaults. This can provide information in determining whether *AGE* should be a feature of importance in default analysis. By observation, with the growth in age, the number of clients that will default the payment next month decreases. Therefore, we can see that *AGE* could be an important feature in the prediction of default payment next month.

Correlation:

Then the correlation is checked within the continuous variables from 2 groups, historical bill amount and historical payment amount data. The correlation coefficient of each variable is presented through a heat map, which shows a pattern of high correlation between variables in the same groups of variables (*PAY_X* and *BILL_AMT_X*).

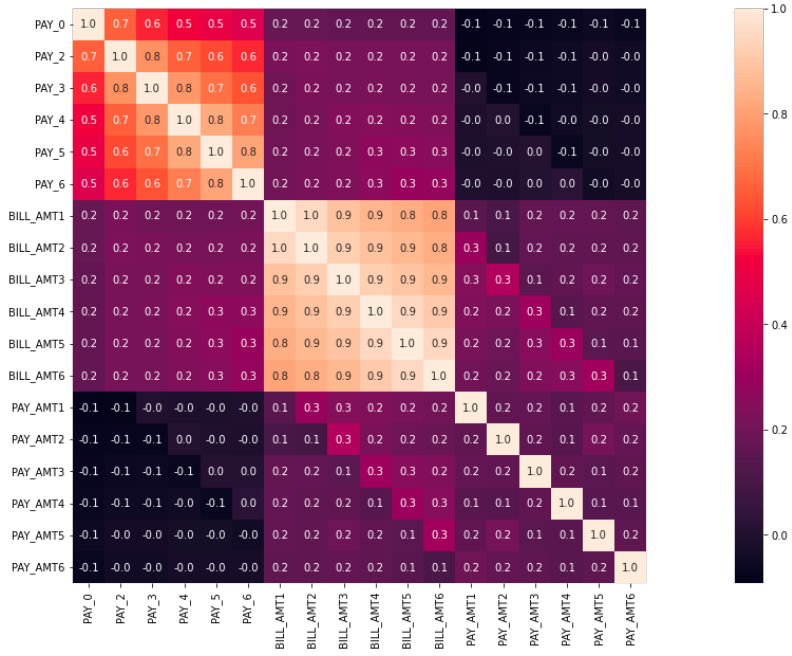

Fig.10 Heat Map between Continuous Variables

To discover the relationships inside the two groups (*PAY* and *BILL_AMT*) of variables, we can further make scatter plots among variables in group *PAY_X* and in group *BILL_AMT_X* to verify the correlation among them, the scatterplots are presented below:

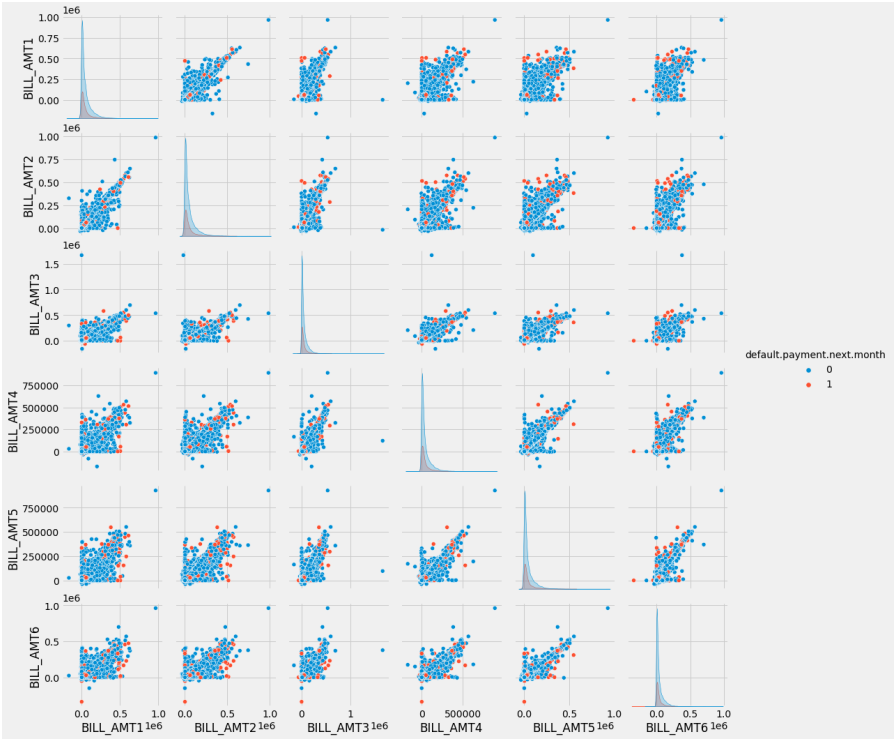

Fig.11 Correlation between *BILL_AMT_X*

According to the above plot we can see that most of the *Bill_AMT_X* have strongly positive correlation between each other. In this case, we do not expect to include all those variables. Or principal component analysis should be considered.

Fig.12 Correlation between *PAY_X*

The correlations between *PAY_AMT_X* are weak. We can say that the *PAY_AMT_X* explained different information. It can be inferred that all those variables need to be included in the model. In the feature engineering part, we reinforced that belief.

3 Feature Engineering

Feature engineering can produce new features, the goal is to simplify and speed up data transformations, at the same time enhancing model accuracy (Patel, 2021); which is important preparation for machine learning.

In the original data set, there are 23 features in total, and many of them describe the same concept but for different time periods. In this case, principal component analysis is being considered. For some categorical features, also be reproduced by different methods.

3.1 Categorical Variables

For categorical variables, sex, education, and marriage been keeping in original. Rest of those categorical variables (*AGE*, *PAY_0* - *PAY_6*) have been reproduced. The changes to the columns are listed as below:

Table.3 Changes of categorical Variables in Feature Engineering

Age Group	PAY_0 - PAY_6
21 - 30 (young): denoted as 1 in dataset	-1 = pay duly
30 – 45 (middle age): denoted as 2 in dataset	1 = payment delay for one month
45 – 60 (pre-old): denoted as 3 in dataset	2 = payment delay for two months
60 – 79 (old): denoted as 4 in dataset	

These changes are mainly conducted by grouping the ages and turning six *PAY_X* variables into one single variable. The changes in *AGE* have changed this continuous variable into a series of categories to support further classification. The changes in past payment variables have reduced the size of features and thus reduced the dimension. The sum of those columns produces a new feature named *Sumpay*. Thus, the interpretation of this new variable (*Sumpay*) should be: the larger *Sumpay* is, the higher probability of default payment.

3.2 Numerical Variables

For numerical variables, *PAY_AMT1* to *PAY_AMT6* have been kept in original. *BILL_AMT1* to *BILL_AMT6* have been reproduced. We use PCA to deal with *BILL_AMT1* to *BILL_AMT6*. The reason why we don't apply PCA to *PAY_AMT1* to *PAY_AMT6* is that PCA is not a good choice to reduce dimension for these six variables. The reasons are as follows.

3.2.1 *PAY_AMT1 - PAY_AMT6*

These six variables represent amounts of previous payment from April 2005 to September 2005. We tried applying the principal component analysis (PCA) to those six variables, however, the result indicated not well. The total variance explained by the first five principal components is roughly 80%, despite the original data only having six variables. Therefore, PCA is not a good choice to reduce dimension for these six variables.

3.2.2 *BILL_AMT1 - BILL_AMT6*

In addition, we applied PCA into these six variables which represent the customers' bill statements from April 2005 to September 2005. We tried doing PCA to these six variables and the results stated excellent. The sample correlation matrix of these six variables are as follows.

Table.4 Sample Correlation Matrix of *BILL_AMTX*

	Bill Amount 1	Bill Amount 2	Bill Amount 3	Bill Amount 4	Bill Amount 5	Bill Amount 6
Bill Amount 1	1.00	0.95	0.89	0.86	0.83	0.81
Bill Amount 2	0.95	1.00	0.92	0.89	0.86	0.84
Bill Amount 3	0.89	0.92	1.00	0.92	0.89	0.86
Bill Amount 4	0.86	0.89	0.92	1.00	0.95	0.91
Bill Amount 5	0.83	0.86	0.89	0.95	1.00	0.95
Bill Amount 6	0.81	0.84	0.86	0.91	0.95	1.00

We use a sample correlation matrix to perform PCA. The eigenvalues and corresponding eigenvectors are as follows (Table.5).

Table.5 Eigenvalues and Corresponding Eigenvectors

Eigenvalues	5.441	0.298	0.113	0.068	0.043	0.038
Eigenvectors	-0.401	0.550	-0.427	0.162	-0.572	-0.038
	-0.409	0.451	-0.100	-0.030	0.784	0.051
	-0.411	0.144	0.680	-0.532	-0.223	-0.126
	-0.414	-0.222	0.366	0.632	-0.051	0.493
	-0.411	-0.418	-0.091	0.239	0.073	-0.765
	-0.402	-0.499	-0.451	-0.483	-0.024	0.389

From the table above, we can see that:

- The first PC is - 0.401* Bill Amount 1 - 0.409* Bill Amount 2 - 0.411 * Bill Amount 3 - 0.414 * Bill Amount 4 - 0.411* Bill Amount 5 - 0.402* Bill Amount 6, which is basically a straight average of the clients' previous six months' amount of bill statement.

- The second PC is 0.550 * Bill Amount 1 + 0.451 * Bill Amount 2 + 0.144 * Bill Amount 3 - 0.222 * Bill Amount 4 - 0.418 * Bill Amount 5 - 0.499 * Bill Amount 6, which is the difference between clients' average amount of bill statements from July 2005 to September 2005 and clients' average amount of bill statements from April 2005 to June 2005.

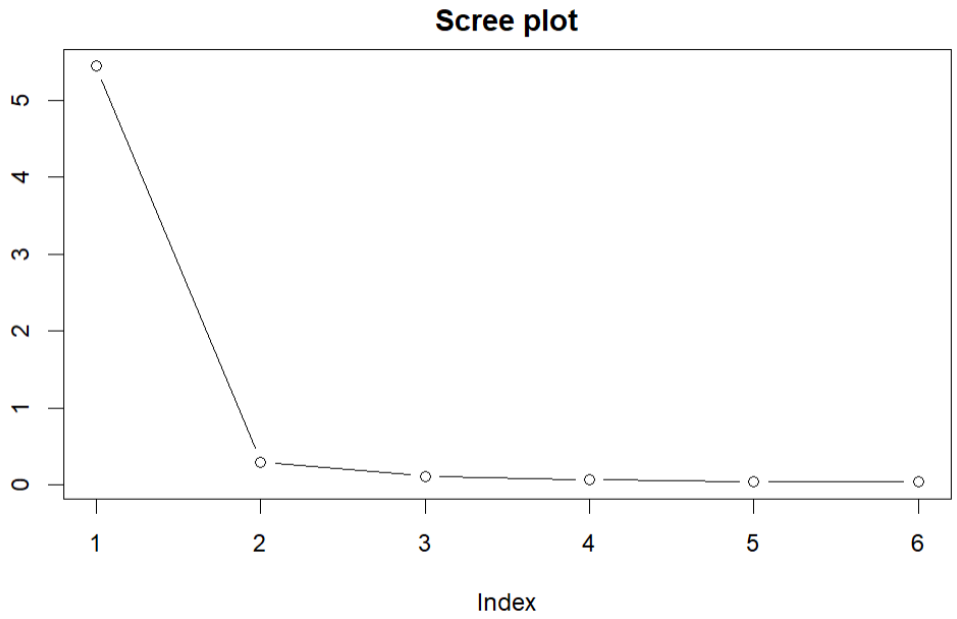

Fig.13 Scree Plot

Since the first two PCs account for 96% of the total variance, which indicates PCA is a good choice to reduce dimension for these six variables. Above is the scree-plot. The scree plot shows the eigenvalues on the y-axis and the number of factors on the x-axis with always a downward curve. The point where the slope of the curve is clearly leveling off at index equals to two which indicates two factors that should be generated in the analysis.

Therefore, as evidenced by the covariance matrix and scree plot, we finally reduced the dimension of these six variables and kept the total variance of the data by utilizing PCA, which was optimized for further analysis of the data set.

4 EDA after Feature Engineering

After feature engineering, we can explore more about the relationship between response variable and explanatory variables.

4.1 Numerical Variables

Table.6 Summary of PCA Variables

	Bill_pca1	Bill_pca2
MEAN	**-2.659000e-10**	**2.974000e-10**
STD	**2.332610e+00**	**5.458417e-01**
MIN	**-3.966195e+00**	**-6.787384e+00**
25%	**-1.473567e+00**	**-1.630588e-01**
50%	**–8.645734e-01**	**-4.536477e-02**
75%	**4.436441e-01**	**9.790291e-02**
MAX	**3.183719e+01**	**7.093761e+00**

4.2 Categorical Variables

4.2.1 Sumpay

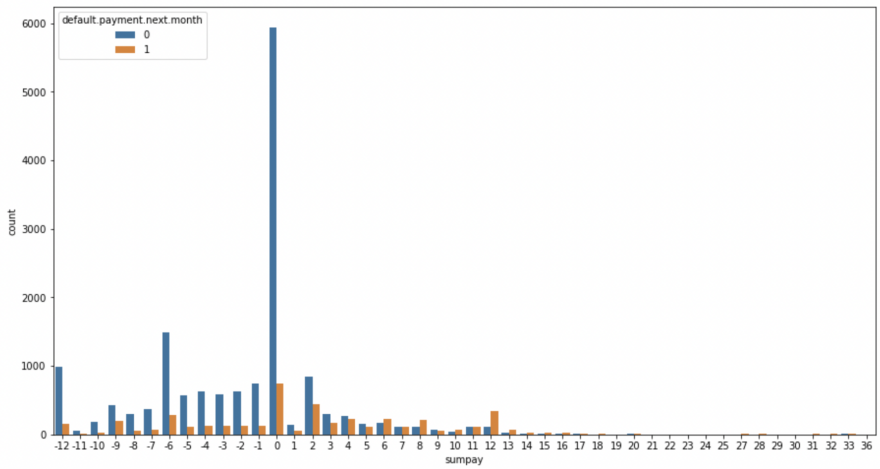

Fig.14 Histogram of *Sumpay* (1: default payment; 0: pay ontime)

According to the histogram, we found out that the proportion of default payment is smaller when the *Sumpay* is negative, with the increasing of *Sumpay*, the proportion of default payment increases.

And we check the conclusion above by calculating proportions in the training data set. For the full training data, the proportion of default payment = 1 is 22.79%, while *Sumpay* >= 4, the proportion of default payment = 1 is 57.54%, which is relatively high.

Then, in order to show whether default payment depends on *Sumpy* or not, we introduce the Chi-square test. Null hypothesis: *Sumpay* is independent of default payment. First, change the *Sumpay* into a categorical variable in two classes by dividing at 4, then count the observed value. Finally, using functions in python do the test.

Table.7 Change of *Sumpay* from Numerical to Categorical Variable

Sumpay	Result
>= 4	1
< 4	2

Table.8 Table of observed value for Chi-square Test

Default Payment next Month	Result	Total
0	2	14161
1	2	2822
1	1	1736
1	1	1281

According to the Chi-square test results, the p-value is smaller than 0.05, rejecting the null hypothesis, so we conclude that the *Sumpay* is dependent on the default payment.

Hence, the new variable *Sumpay* is important for the prediction of Y.

4.2.2 Age

In the original data set, age is considered as the continuous variable. Using the age after scaled to plot with dependent variable default payment. The plot indicates that there is no significant relation between those two variables. We do want to keep as much information as we can, so age has been redesigned as a categorical variable. The histogram of age after feature engineering (Fig.17) shows that most customers are between 21 to 45.

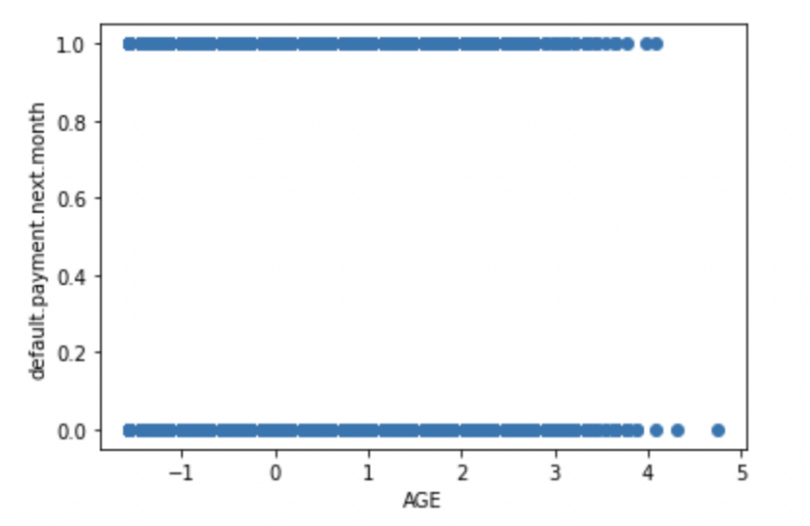

Fig.15 Scatter plot of Standard Age before Feature Engineering

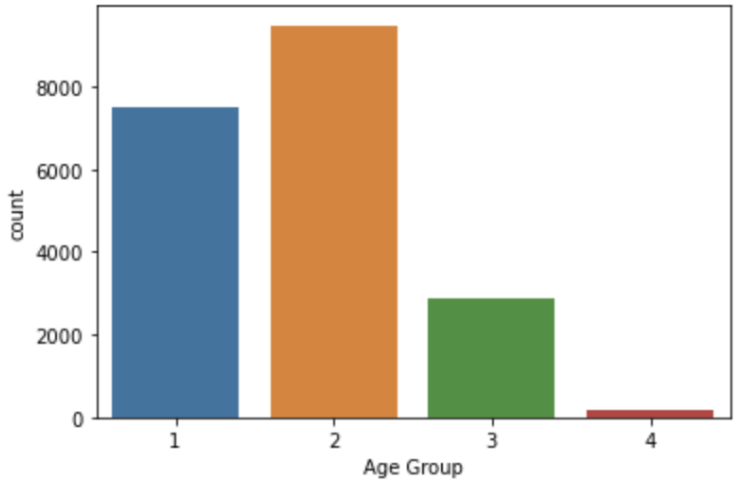

Fig.16 Histogram of Age after Feature Engineering

5 Model Construction

After feature analysis and feature engineering, five algorithms were introduced in this part, K-Nearest Neighbors, Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting Classifier with Random Forest. We examined accuracy, precision score, recall score, and F1 scores at first. High recall and precision is what we expected since we do care about false negatives as we do the prediction of default population. Moreover, F1 Score computed the average of Precision and Recall. In our case, the cost of false positives and false negatives are very different, and we have an imbalanced dataset. Accuracy is not a good metric to use when the dataset is imbalanced. Thus, we decide to choose F1 score as our final measurement of model performance (Korstanje, 2021).

5.1 KNN Classification

KNN is one of the most popular and simple algorithms in machine learning, which can be used to predict classification problems. The idea behind KNN algorithm is we assume that similar things are near to each other. An object is classified by vote of his K nearest neighbor (Bhutani, 2021). Basically, first we need to define the distance function we use that measures distance between samples and select appropriate K values. In our model, the number of neighbors is equal to five, meaning that the class of the new data is voted by his five closest neighbors.

Table.9 Table of KNN Classification

Accuracy	**0.769833**		
classification	**precision**	**recall**	**f1-score**
0	**0.81**	**0.91**	**0.86**
1	**0.50**	**0.29**	**0.37**

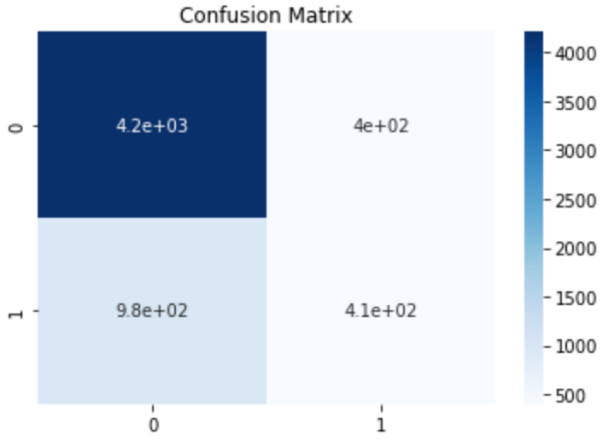

Fig.17 Parameters of KNN Classification Model

The accuracy rate is 0.7698 which means there are a high 76.98% of correct predictions over total predictions in the k nearest neighbor classification model. The precision score in the default population is 0.81 and in non-default population 0.50 that represents the ratio of correctly predicted positive observations to the total predicted positive observations. Besides, recall measures how good a model is when it correctly predicts positive classes. Logistic regression model measures recall as 0.91 in the non-default population and 0.29 in the default prediction. The F1 score in the default population is 0.86 which indicates both precision and recall are high, and in non-default population 0.37 which indicates both precision and recall are relatively low.

5.2 Logistic Classification

This method is a predictive algorithm using independent variables to predict the dependent variable which is a categorical variable. Logistic regression is a statistical model that uses the logistic function to model the conditional probability. For binary regression, we calculate the conditional probability of the dependent variable Y (the default of payment), given independent variables. It is the conditional probability of Y = 1 (default), given X, or conditional probability of Y = 0 (non-default), given X. To implement this method, we split the data set into training and testing data sets by 70:30, and then we fit the model by using the *sklearn* model function in python, and calculate the accuracy rate, precision, recall, and F1 score.

Table.10 Table of Logistic Classification

Accuracy	**0.786**

classification	precision	recall	F1-score
0	0.79	0.98	0.88
1	0.67	0.15	0.24

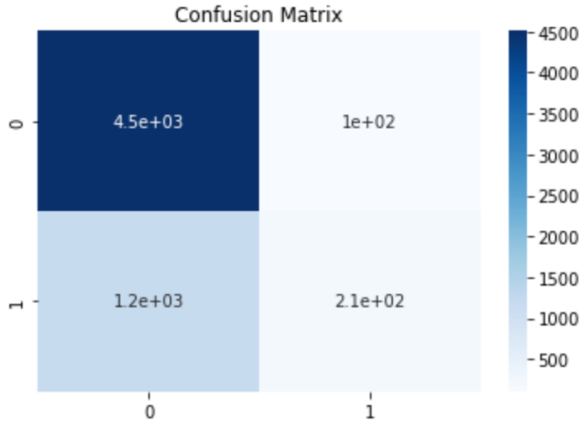

Fig.18 Parameters of Logistic Classification Model

There are 78.6% of correct predictions over total predictions in the logistic regression model. The precision score in the default population is 0.79 and in non-default population 0.67 that indicates a high ratio of correctly predicted positive observations to the total predicted positive observations. Logistic regression model measures recall as 0.98 in the non-default population and 0.15 in the default prediction. The F1 score in the default population is 0.88 which indicates both precision and recall are high, and in non-default population 0.24 which indicates both precision and recall are relatively low.

5.3 Decision Tree Classification

Decision tree is a model that divides data into different branches according to the characteristics of the judgment. After several branches, the model can form a tree-like model of decisions and arrive at consequences. And we can adjust some parameters such as the max depth or max features of the tree to improve the accuracy.

Table.11 Table of Decision Tree Classification

Accuracy	0.701166		
classification	precision	recall	F1 score
0	0.81	0.80	0.81

1	**0.37**	**0.39**	**0.38**

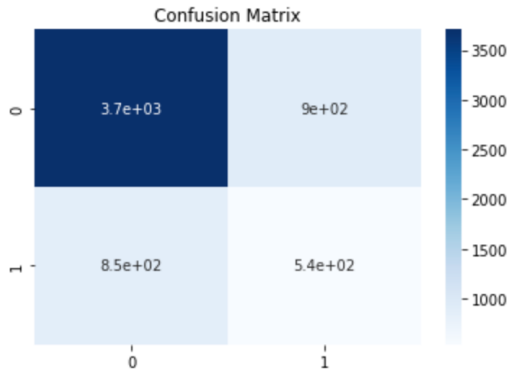

Fig.19 Parameters of Decision Tree Classification Model

There are 70.1% of correct predictions over total predictions in the decision tree classification model. The precision score in the default population is 0.81 and in the non-default population 0.37 that indicates a high ratio of correctly predicted positive observations in default population but low ratio in non-default population. Logistic regression model measures recall as 0.80 in the non-default population and 0.39 in the default prediction. The F1 score in the default population is 0.81 which indicates both precision and recall are high, and in non-default population 0.38 which indicates both precision and recall are relatively low.

5.4 Random Forest Classification

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. We tried applying a random forest method to our data. The results are as follows.

Table.12 Table of Random Forest Classification

Accuracy	**0.7921666**		
classification	**precision**	**recall**	**F1 score**
0	**0.82**	**0.94**	**0.87**
1	**0.61**	**0.29**	**0.39**

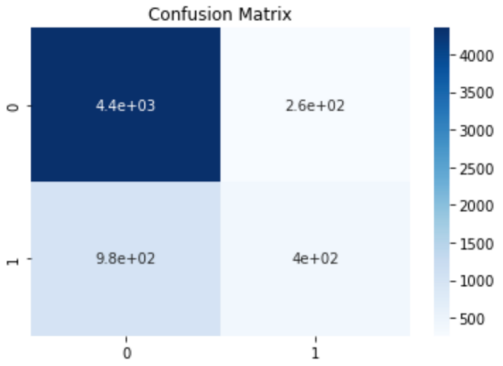

Fig.20 Parameters of Random Forest Classification Model

There are 79.2% of correct predictions over total predictions in the random forest classification model. The precision score in the default population is 0.82 and in the non-default population 0.61 that indicates a high ratio of correctly predicted positive observations in the default population but low ratio in the non-default population. Logistic regression model measures recall as 0.94 in the non-default population and 0.29 in the default prediction. The F1 score in the default population is 0.87 which indicates both precision and recall are high, and in non-default population 0.39 which indicates both precision and recall are relatively low.

5.5 Gradient Boosting Classifier with Random Forest

Instead of learning a single weak classifier, now we are learning many weak classifiers. The final boosting classifier will be a linear combination of the votes of the different classifiers weighted by their strength. The boosting classifier is effective and simple to implement. There are two most popular types of boosting which are Adaboost and Gradient Boosting. Gradient boosting is present here.

Table.13 Table of Gradient Boosting Classification

Accuracy	**0.7983333**		
classification	**precision**	**recall**	**F1 score**
0	**0.82**	**0.95**	**0.88**
1	**0.64**	**0.30**	**0.41**

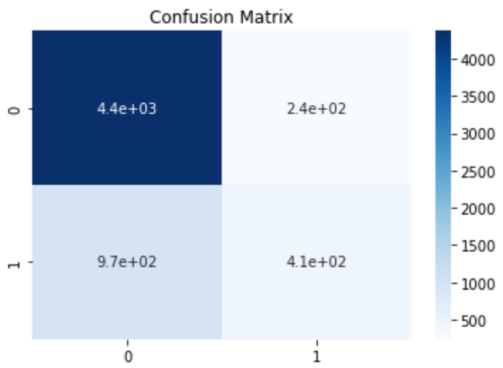

Fig.21 Parameters of Gradient Boosting Classifier with Random Forest Model

There are 79.8% of correct predictions over total predictions in the random forest classification model. The precision score in the default population is 0.82 and in the non-default population 0.64 that indicates a high ratio of correctly predicted positive observations in the default population but low ratio in the non-default population. Logistic regression model measures recall as 0.95 in the non-default population and 0.30 in the default prediction. The F1 score in the default population is 0.88 which indicates both precision and recall are high, and in non-default population 0.41 which indicates both precision and recall are relatively low.

6 Model Comparison

6.1 Precision, Recall, F1 score

Table.14 Precision and Recall among different models

Model	F1 score	Precision	Recall
KNN	**0.37**	**0.5**	**0.29**
Logistic Regression	**0.24**	**0.67**	**0.15**
Decision Tree	**0.38**	**0.37**	**0.39**
Random Forest	**0.39**	**0.61**	**0.29**
Gradient Boosting	**0.41**	**0.64**	**0.30**

6.2 Validation Curve

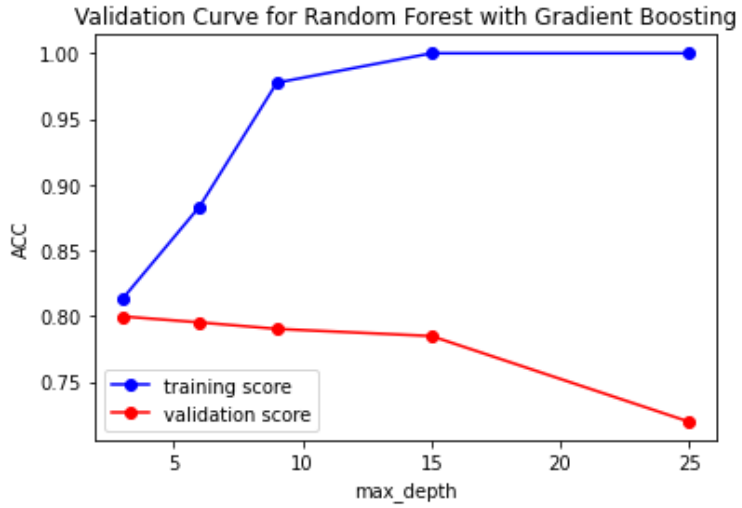

Fig.22 Validation Curve for Random Forest with Gradient Boosting

That cross validation is a procedure used to avoid overfitting and estimate the skill of the model on test data. According to the plot, with the increase of max depth, the validation score increases at first and then decreases. The Accuracy of the model is the average of the accuracy of each fold during the cross validation. We can conclude that when the max depth equals one, the accuracy of the model is the largest.

6.3 Sort the Importance Features

Sumpay	**0.689089**
Bill_pca1	**0.081166**
Bill_pca2	**0.041057**
PAY_AMT1	**0.033244**
LIMIT_BAL	**0.031943**

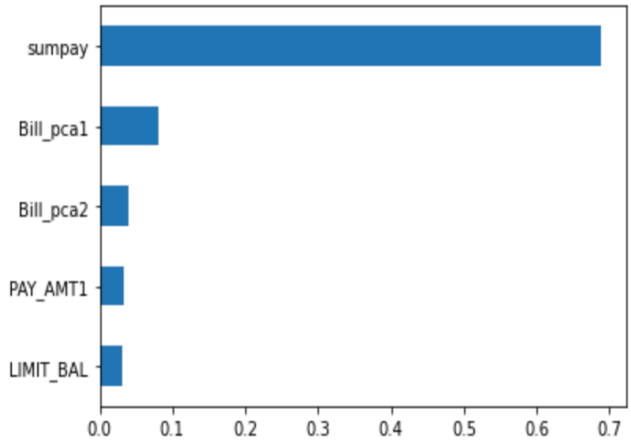

Fig.23 Values with Different Features

By calculating the importance of features, the above graph shows that the variable *Sumpay* has the largest value compared to other variables and the importance of the variable *Sumpay* is approximately equal to the 0.7 in the gradient boosting with random forest classification. Therefore, we can conclude that the *Sumpay* is an extremely important feature in the final model.

6.4 ROC Curve for Model which Has the Best Performance

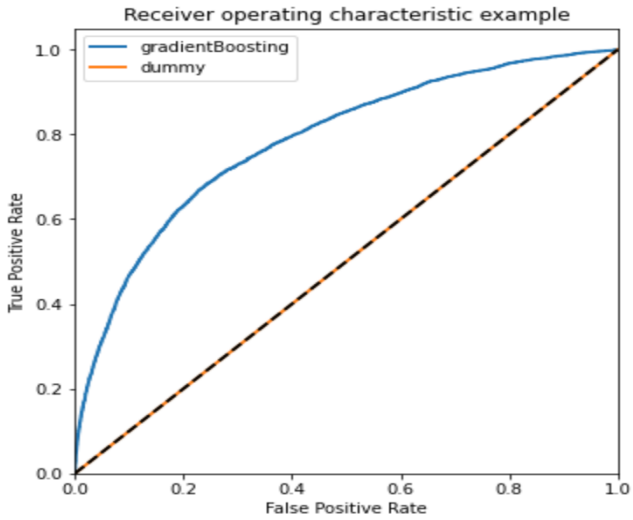

Fig.24 ROC Curve for Gradient Boosting Classification

One of the most important methods that evaluates the performance of the classification problem is the ROC curve. The ROC curve evaluates the relationship between true positive rate and false positive rate. The diagonal line is FPR = TPR in the plot. If TPR > FPR, the curve is closer to the top-left corner, the model is better, if TPR < FPR, the curve is closer to the bottom-right corner, the model is worse.

(Chan, n.d.). According to the plot, the curve is closer to the top-left corner, which means that the AUC is larger than 0.5. Therefore, it has good performance which remains both good specificity and minimizes the false positive rate.

All in all, after trying different models and comparing F1 scores of 5 models, we conclude that Gradient Boosting Classification has the best performance in terms of F1 score. The F1 score of this model is 0.41, the precision of this model is 0.64, recall is 0.30.

7 Summary

The main goal of this project is to analyze the data in order to predict whether credit card holders will default or not next month.

Firstly, based on our exploratory data analysis, we find that most categorical variables are not important features, like education level, gender. Among the most important features, we use PCA in order to reduce the dimension and also create new features during the feature engineering process. Then we fit our data with those new features to different machine learning models, finding Gradient Boosting Classification is the best classifier compared to others. For future considerations, we may be able to make improvements to get an even better F1 score, such as using other machine learning models like neural networks. A more informative source of data including other predictors such as income, crime records and other demographic factors can help make a more accurate prediction, although there may be difficulties ensuring data accessibility and consistency.

References

Chan, Carman (n.d.) *What is a ROC Curve and How to Interpret It,* Retrieved from https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/

Bhutani, Harman (Jan, 2021.). *K-Nearest Neighbors (KNN) for Machine Learning.* Retrieved from https://medium.com/analytics-vidhya/k-nearest-neighbors-knn-for-machine-learning-d266d7c43830#:~:text=An%20object%20is%20classified%20by,of%20that%20single%20nearest%20neighbor./

Korstanje, J. (2021, August 31). *The F1 score. Medium.* Retrieved from https://towardsdatascience.com/the-f1-score-bec2bbc38aa6

Patel, Harshil (Aug. 2020.). *What is Feature Engineering*. Retrieved from https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10#:~:text=It%20can%20produce%20new%20features,working%20with%20machine%20learning%20models.

Wang (2022.) .*Taiwan's Credit Card Crisis*. Sevenpillarsinstitute.org. (n.d.). Retrieved May 6, 2022, Retrieved from https://sevenpillarsinstitute.org/case-studies/taiwans-credit-card-crisis/

Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, 36(2), 2473-2480.