SUPERVISED MACHINE LEARNING (CLASSIFICATION)

PROJECT ON

Credit Card Default Prediction

Name of the participants:

CHARAN

Data science trainees,
AlmaBetter, Bangalore

Table of Contents

Abstract:

Aiming at the problem that the credit card default data of a financial institution is unbalanced, which leads to unsatisfactory prediction results, this project proposes a prediction model based on various supervised, classification algorithms. It mostly focuses on enhancing the classifier performance for credit card default prediction. Different machine learning models are also employed to obtain efficient results. We developed the hypothesis of whether developed models using different machine learning techniques are significantly the same or different and whether resampling techniques significantly improves the performance of the proposed models. One-way Analysis of Variance is a hypothesis-testing technique, used to test the significance of the results. The split method is utilized to validate the results in which data has split into training and test sets. The results using our proposed methods significantly improve the accuracy of Taiwan clients credit dataset.

1. Problem Statement

This project is aimed at predicting the case of customers' default payments in Taiwan. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients. We have to evaluate which customers will default on their credit card payments.

Financial threats are displaying a trend about the credit risk of commercial banks as the incredible improvement in the financial industry has arisen. In this way, one of the biggest threats faced by commercial banks is the risk prediction of credit clients.

To analyze and predict the above given database, the current project is developed. This project is an attempt to identify credit card customers who are more likely to default in the coming month

2. Introduction

Credit-Card Default Definition – When a customer applies for and receives a credit card, it becomes a huge responsibility for the customer as well as the credit card issuing company. The credit card company evaluates the customer's credit worthiness and gives him/her a line of credit that they feel the customer can be responsible for. While most people will use their card to make purchases and then diligently make payments on what they charge, there are some people who, for one reason or another, do not keep up on their payments and eventually go into credit card default. Credit card default is the term used to describe what happens when a credit card user makes purchases by charging them to their credit card and then they do not pay their bill. It can occur when one payment is more than 30 days past due, which may raise your interest rate. Most of the time, the term default is used informally when the credit card payment is more than 60 days past due. A default has a negative impact on the credit report and most likely leads to higher interest rates on future borrowing.

In recent years, the credit card issuers are facing the cash and credit card debt crisis as they have been over-issuing cash and credit cards to unqualified applicants, in order to increase their market share. At the same time, most cardholders, irrespective of their repayment ability, overused credit cards for consumption and accumulated heavy credit and cash–card debts. The crisis is an omen for the blow to consumer finance confidence and it is a big challenge for both banks and cardholders.

This project is an attempt to identify credit card customers who are more likely to default in the coming month by using customer's personal and financial information like credit line, age, repayment and delinquency history for the past 6 months to predict the probability of the particular customer to become default next month. Many statistical and data mining techniques will be used to build a binary predictive model.

3. Understanding The Problem Statement

This project is an attempt to identify credit card customers who are more likely to default in the coming month. A lot of credit card issuing companies are working on predictive models which would help them predict the payment status of the customer ahead of time using the customer's credit score, credit history, payment history and other factors.

This project is aimed at using customer's personal and financial information like credit line, age, repayment and delinquency history for the past 6 months to predict the probability of the particular customer to become default next month. Many statistical and data mining techniques will be used to build a binary predictive model. If the credit card issuing companies can effectively predict the imminent default of customers beforehand, it will help them to pursue targeted customers and take calculated efforts to avoid the default, to overcome future losses efficiently.

The data, in any sense, does not directly reveal the identity of any individual or provide information that could be decrypted to connect to an individual. In this project, the plan is to predict the probability of credit-card holders to go default in the next month by using payment data from October 2015 to March 2016. Among the total 30,000 observations,To determine the binary variable – default payment in April 2016 (Yes = 1, No = 0), as the response variable.

Some of the key attributes consisting those variables which are used for this project are listed below:

Attributes and their Description

This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. This study reviewed the literature and used the following 23 variables as explanatory variables:

- 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. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 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; . . .; X17 = amount of bill statement in April, 2005.
- X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

4. Data Pre-processing

Before jumping to the exploration stage, we need to perform basic data pre-processing steps like null value imputation and removal of unwanted data.

4.1. Libraries Used

Pandas ,Numpy , Matplolib, Seaborn,Sklearn

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.pipeline import Pipeline, FeatureUnion
from sklearn.impute import SimpleImputer
imputer = SimpleImputer(missing_values=np.nan, strategy='mean')
from sklearn.base import BaseEstimator, TransformerMixin
from sklearn.svm import LinearSVC
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split, cross_val_score,
GridSearchCV, cross_val_predict
from sklearn.metrics import classification_report,
accuracy_score,confusion_matrix,precision_score, recall_score, roc_curve,
roc_auc_score
from sklearn.decomposition import PCA
from sklearn.manifold import LocallyLinearEmbedding
from sklearn.ensemble import BaggingClassifier, RandomForestClassifier,
```

VotingClassifier

The dataset contains 30000 rows and 25 columns. But we only need information about important columns. and dropping other unwanted columns.

4.2. Handling Null Values

There are a few null values in the dataset. The major part of the dataset is null values which triggers many stumbling blockages in processing the necessary columns, therefore it is required to identify the columns containing lump sum null values and get dropped. So, let's drop these null values columns which contain more than 80 percent of null values.

5. PREPARING DATA FOR EXPLORATORY DATA ANALYSIS(EDA)

5.1. data loading

We have loaded the data into our drive and then finally loaded it to google colab notebook using pandas library.

5.2. Renaming the columns

This method is useful because it lets you modify a column heading without having to create a new column.

5.3. Removing unwanted columns from dataset

5.4. Renaming the columns

One way of renaming the columns in a Pandas dataframe is by using the rename() function. This method is quite useful when we need to rename some selected columns because we need to specify information only for the columns which are to be renamed.

6. EXPLORATORY DATA ANALYSIS:

6.1Shape :

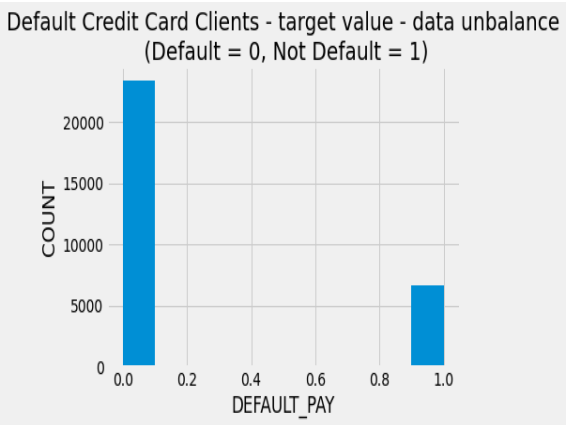
The shape attribute tells us a number of observations and variables we have in the data set. It is used to check the dimension of data. The given data set has more than 30000 observations and 23 variables in the data set.

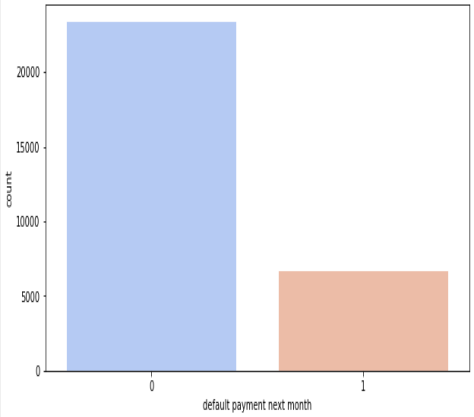
6.2. Describe :

It computes a summary of statistics pertaining to the DataFrame columns. This function gives the mean, std and IQR values. And, function excludes the character columns and gives a summary about numeric columns.

6.3. Analysis of different variables

6.3.1. Analyzing the default pay

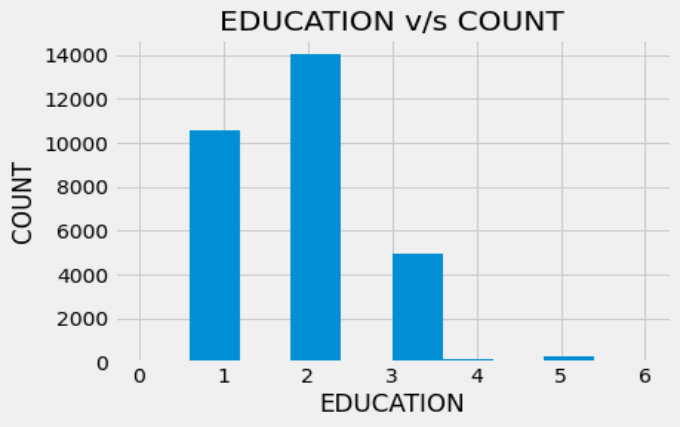


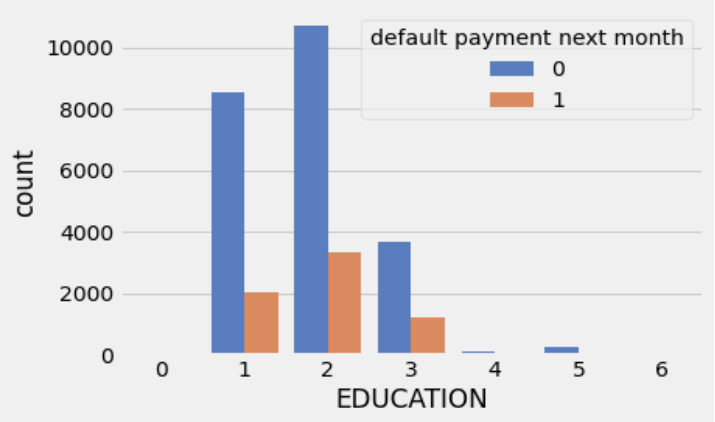


From the above graph on x-axis 0 indicates as not a default payment and 1 indicates the default payment.From this we can say that for more customers there are no default payments for next month.

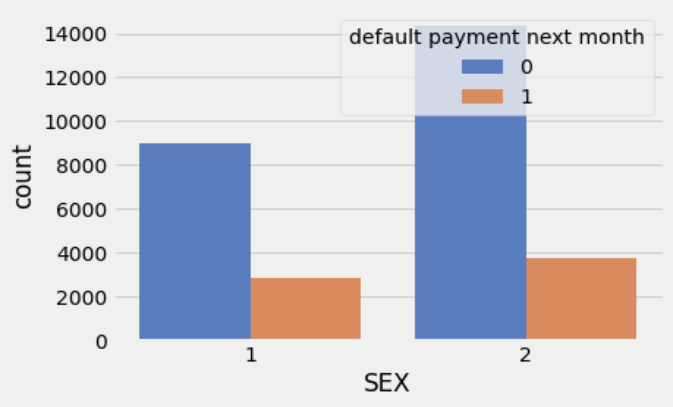
score of the customer. Using this variable will help the model predict defaults more effectively

6.3.3. Analyzing the customers based on their Education



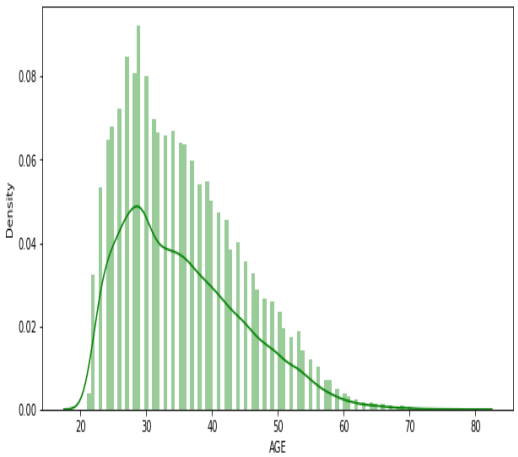


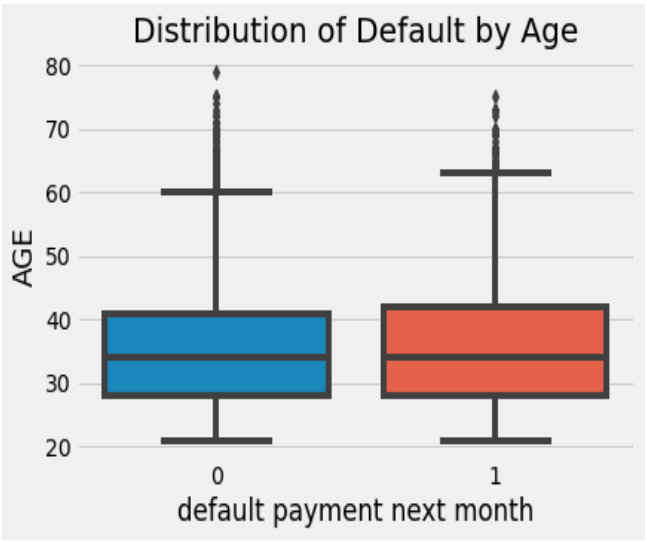
Education (1 = graduate school; 2 = university; 3 = high school; 4 = others) It might be useful to see whether the education level of the customer is in any way related to his/her probability of default. The distribution of defaults based on education level will be an interesting chart to look at. From the above we can say that most of the people are university educated followed by graduated schoolMore number of credit holders are university students followed by Graduates and then High school students.



Gender (1 = male; 2 = female) It might be useful to see whether the gender of the customer is in any way related to his/her probability of default.

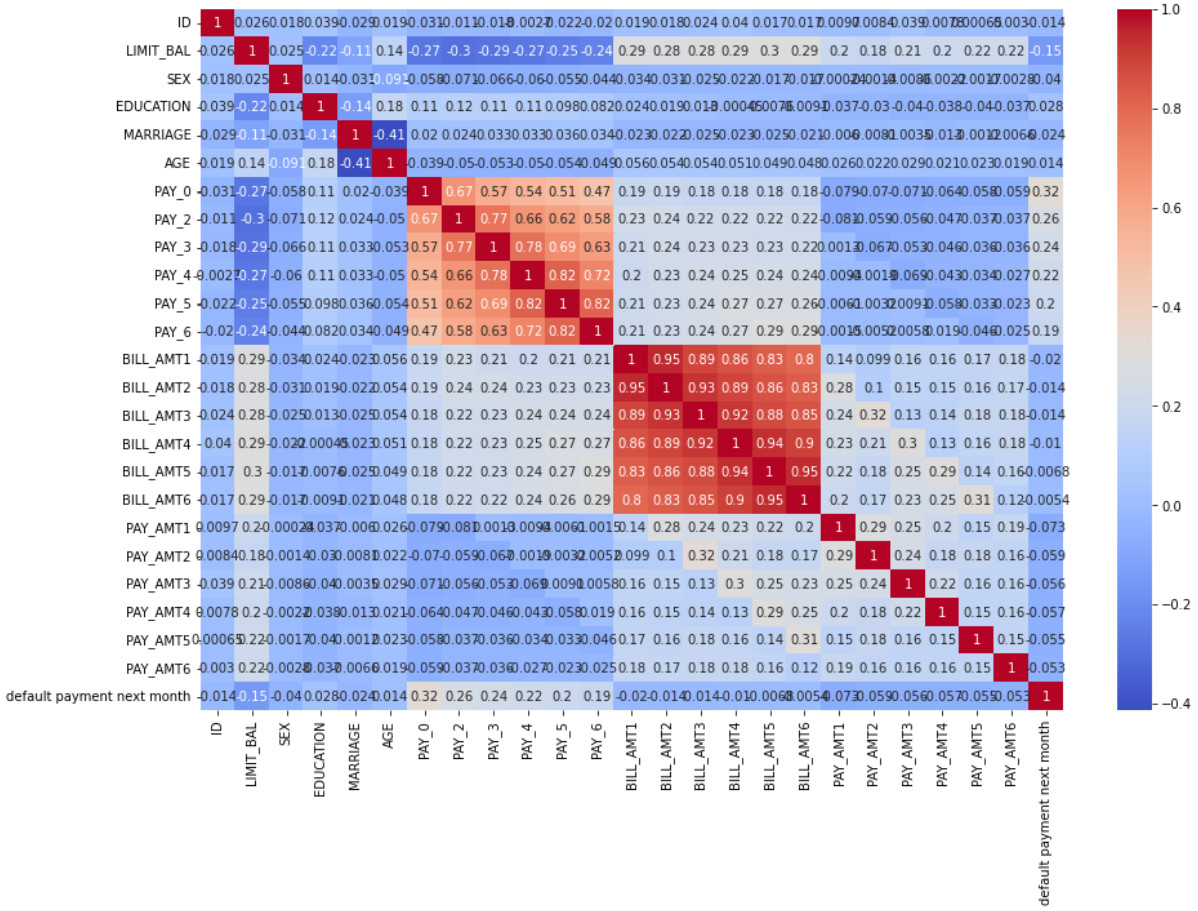
6.3.8. Analyzing the age of the customers





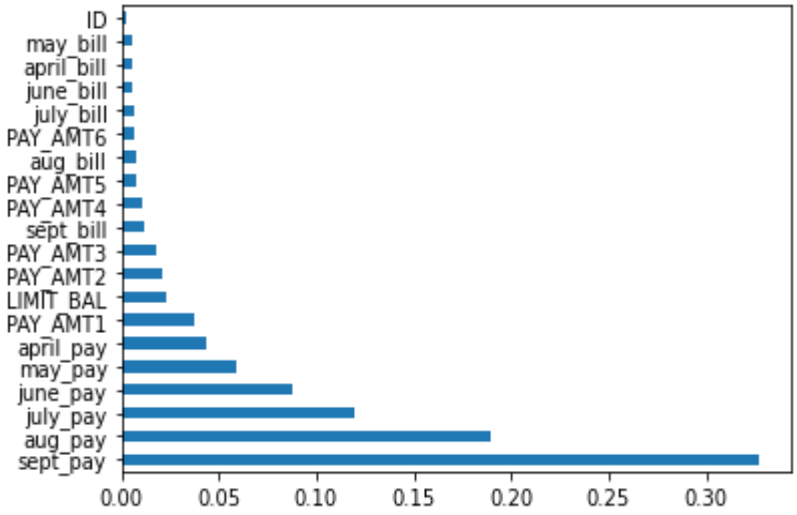
Age (year) It might be useful to see whether the gender of the customer is in any way related to his/her probability of default. From the above graphs we can say that most of the customers are from the age group of 25-30 years.

CORRELATIONS:



The above figure is an indication about the changes between two variables. We can plot correlation matrices to know which variable is having a high or low correlation in respect to another variable.

7. Feature Importance



8. Training the models

8.1.Fitting different models

For modeling we tried various classification algorithms like:

➢ Logistic Regression

Logistic Regression using LLE

➢SVC

➢SVC with LLE

Ensemble Learning

➢Bagging Classifier

➢Voting Classifier

8.1. Algorithms:

Logistic Regression:

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, True or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic regression equation:

$$p = 1/1 + e^{-(\beta 0 + \beta 1 X1 + \beta 2 X2 ..... \beta n Xn)}$$

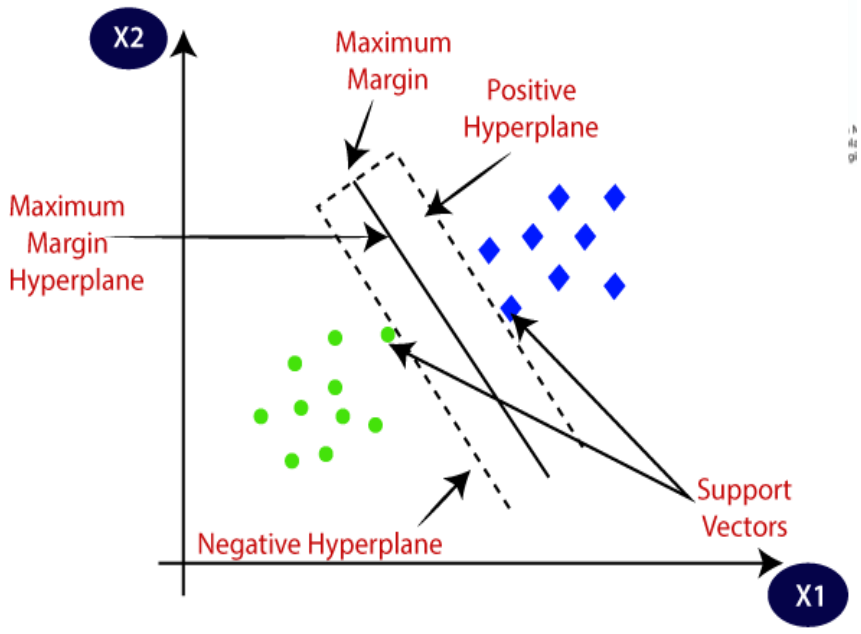
Logistic Regression using LLE:

SVM:

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence the algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:

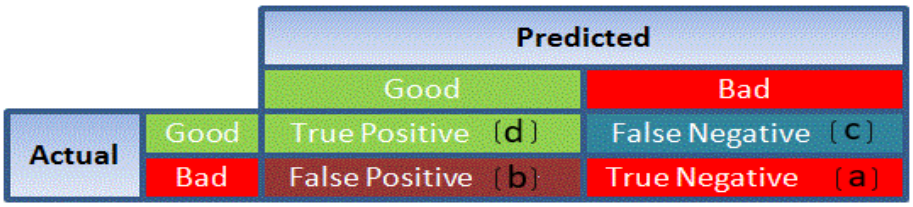


Ensemble learning is a general meta approach to machine learning that seeks better predictive performance by combining the predictions from multiple models.The three main classes of ensemble learning methods are **bagging**, **stacking**, and **boosting**, and it is important to both have a detailed understanding of each method and to consider them on your predictive modeling project.

9. Evaluating the model

After the model is built, if we see that the difference in the values of the predicted and actual data is not much, it is considered to be a good model and can be used to make future predictions.

1. **Confusion Matrix:** It is nothing but a tabular representation of Actual vs Predicted values. This helps us to find the accuracy of the model and avoid overfitting. This is how it looks like:

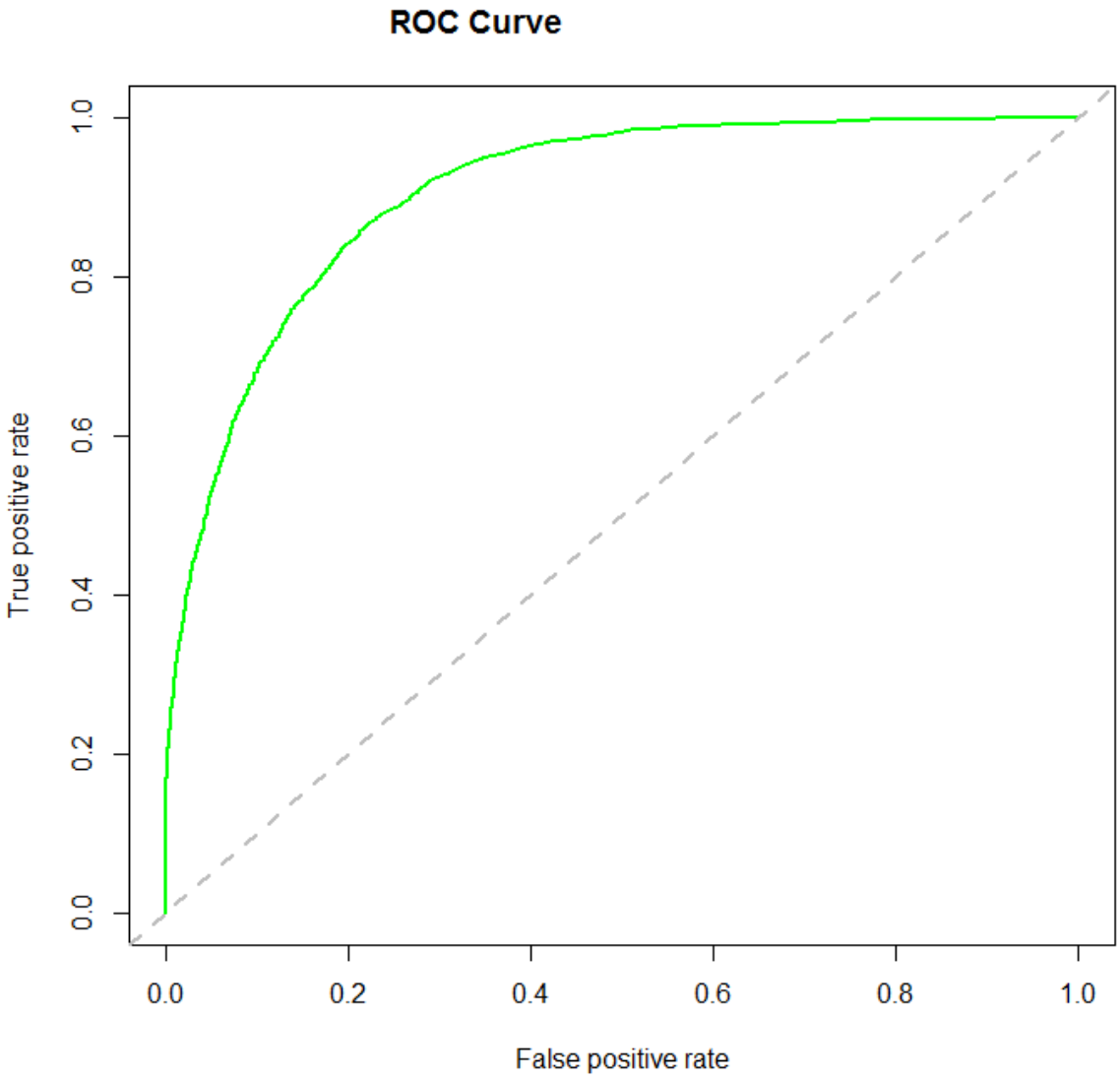


calculate the accuracy of your model with:

$$\frac{\text{True Positive + True Negatives}}{\text{True Positive + True Negatives + False Positives + False Negatives}}$$

2. **ROC Curve:** Receiver Operating Characteristic(ROC) summarizes the model's performance by evaluating the trade-offs between true positive rate (sensitivity) and false positive rate(1- specificity). For plotting ROC, it is advisable to assume p > 0.5 since we are more concerned about success rate. ROC summarizes the predictive power for all possible values of p > 0.5. The area under curve (AUC), referred to as index of accuracy(A) or concordance index, is a perfect performance **m**etric for ROC curve. Higher the area under the curve, better the prediction power of the model. Below is a sample ROC curve. The ROC of a perfect predictive model has TP equals 1 and FP equals

0. This curve will touch the top left corner of the graph.



3.

10. Conclusion:

The objective of this project is to train various supervised learning algorithms to predict the client's behavior in paying off the credit card balance. In classification problems,

an imbalanced dataset is also crucial to enhance the performance of the model, so different resampling techniques were also used to balance the dataset. We first investigated the datasets by using exploratory data analysis techniques, including data normalization. We started with the logistic regression model, then compared the results with traditional machine learning-based models. Then K-means SMOTE resampling method on Taiwan client's credit dataset.

In the end, the proposed method has also been deployed on the web to assist the different stakeholders. Therefore, when the financial institution considers issuing the client a credit card, the institution needs to check the payment history of that person because the decision on whether to pay on duly or owe the bill on a specific month usually relates to the previous payment history. For instance, if a person owes numerous bills already, he or she is likely to delay the payment of the current month unless this person gets a windfall so that the total arrears can be paid off. Besides the payment history, it is also imperative to look at the applicants' credit limit of their current credit cards. This is a result of a virtuous circle: people who pay on duly tend to have better credit scores, so the banks prefer to increase these people's credit lines by taking less risk. As a result, if a potential client already has a credit card with a high credit limit line, this person is unlikely to fail to pay the full amount owed in the future. Although the financial institution often collects clients' personal information such as age, educational level, and marital status when people apply for credit cards, this information also affects the default behavior. In other words, the financial institution should equally consider their potential clients who are men or women, obtain bachelor degrees or master degrees, single or married when deciding whether to approve their credit card/loan applications. We tried our best to make a thorough analysis, and there are still a few possible improvements that may require longer-term action. For the boosting models, only the GBDT method was trained, but various variants of boosting techniques may also be utilized in the future. The financial market changes rapidly every day, and people's economic status and performance are affected by the market all the time. So, if more economic indicators are added to the dataset, this will lead to a more generic model.

After exploring, manipulating and experimenting with different models on the credit card default data set we have obtained a maximum accuracy of 82% to determine whether a person defaults on their credit card or not. Ideally we would have been able to increase this accuracy by trying out various ways of preprocessing the data, utilizing dimensionality reduction, fine-tuning the models' hyperparameters, and applying ensemble learning.

First we derived new features from the data set. Since this resulted in about 100 features we explored dimensionality reduction. Using the second form of reduction, LLE, we ran our initial classification models, Logistic Regression and Support Vector Classifier with the original training set and the reduced set. Seeing no improvement with the set produced in LLE we continued using our original training set.

Next we fine-tuned the better performing model, Logistic Regression, with GridSearchCV as another attempt to improve the model. After using GridSearchCV no significant improvements were seen.

This led us to trying ensemble learning to see if our overall accuracy could be improved by combining various models' predictions. Once again there did not seem to be a significant improvement in both the Bagging Classifier and Voting Classifier when compared to our Logistic Regression model.

In the end, looking at the accuracy scores of each model was not enough information to choose which model performed best when trying to predict whether or not a person would default on their credit card. Above one can clearly see how close all these accuracy scores fall. In fact, by just seeing the accuracy scores the Voting Classifier seems to perform the best. But if looked at performance via ROC curves the Logistic Regression model seems to perform the best and the Voting Classifier does significantly worse.

For future work, we think it would be interesting to develop more complex models, such as implementing a neural network and seeing if there could be a better performance of an 82% accuracy score since all our previous techniques did not seem to affect the accuracy.