

**AN6001 – AI & Big Data in Business**

**Credit Card Default Prediction with Machine Learning Models**

Zhu Wentao

G2200626G

Table of Contents

[Executive Summary 2](#_Toc116839843)

[1 Introduction 3](#_Toc116839844)

[1.1 Credit Risk and Default 3](#_Toc116839845)

[1.2 Machine Learning 3](#_Toc116839846)

[2 Data and Preprocessing 4](#_Toc116839847)

[2.1 Data Introduction 4](#_Toc116839848)

[2.2 Data Preprocessing 4](#_Toc116839849)

[3 Model Analysis 5](#_Toc116839850)

[3.1 Data Splitting 5](#_Toc116839851)

[3.2 Model Introduction 5](#_Toc116839852)

[3.3 Model Outcome 6](#_Toc116839853)

[3.4 Model Improvement 7](#_Toc116839854)

[3.5 Model Summary 7](#_Toc116839855)

[4 Conclusion 8](#_Toc116839856)

[Appendix 8](#_Toc116839857)

# Executive Summary

Commercial banks are exposed to all kinds of risks every day. Effective risk management and risk avoidance is quite essential for banks. Credit risk is a major risk in the operation of commercial banks' credit assets. It refers to the risk that the bank cannot recover all the loan principal and interest as scheduled due to the borrower's inability to repay the loan, which default happens. When credit risk, or default occurs, the bank will have bad debts, which will directly affect the bank's operating efficiency and asset quality. Therefore, commercial banks must keep alert and effectively strengthen their credit risk management to avoid defaults.

In the following report, I will use different machine learning models to predict if the bank customer will default or not. Artificial Intelligence (AI) and Machine Learning Algorithm are quickly developing these days. Generally, Artificial intelligence is the function of computer systems that mimic human cognitive functions, while Machine learning is a part of AI which analyzes and interprets the patterns and structure of historical data for the purpose of learning, reasoning, and decision-making. I have chosen to apply six machine learning models—Logistic Regression, Decision Tree, Random Forest, XGBoost, Neural Network and Autogluon—to predict whether the borrower will default or not based on selected features which has greater influence over default. To improve the performance of our models, techniques like oversampling, feature engineering are also used. The result suggest that machine learning algorithm can be effectively implemented to predict the likelihood of default. My models have practical value of improving the efficiency of the loan process for our bank by initially filtering out the potential borrowers who are more likely to pay back the loan. It can also serve as an advisor for our bank to avoid lending money to the borrowers who have higher risk of default.

Overall, it can be concluded that there are many opportunities to extend the scope and depth of this project in future research due to the seemingly endless capabilities of AI. My research demonstrates the concept of various techniques for default prediction using machine learning models. I believe that machine learning can bring many benefits to our bank and hope that my research will initiate the use of machine learning models in our bank to speed up and streamline more processes.

Key words: Credit Risk, Artificial intelligence and Machine Learning, Default Prediction

# 1 Introduction

## 1.1 Credit Risk and Default

Commercial banks are exposed to all kinds of risks every day. Effective risk management and risk avoidance is quite essential for banks. Although safety, liquidity and profitability are all objectives that banks pursue, there is always a trade-off.

Credit risk is a major risk in the operation of commercial banks' credit assets. Credit risk refers to the risk that the bank will not be able to recover all the loan principal and interest as scheduled due to the borrower's failure or inability to repay the loan after maturity. In this case, default happens. When a borrower defaults, the bank will incur financial losses because it fails to receive the principal and expected return.

Credit risk is rooted in the uncertainty of commodity economic activity itself. There are usually two influencing factors.

1. The contraction of economy. In this period, credit risk increases as profitability deteriorates overall and the likelihood that borrowers will not be able to make full and timely payments for various reasons increases.

2. Impact of special events on the company's operations. For example, market changes might result in product stagnation and poor liquidity, leading to the failure of repaying debts.

When credit risk occurs, the bank will have bad debts, which will directly affect the bank's operating efficiency and asset quality. Therefore, commercial banks must keep alert and effectively strengthen their credit risk management to avoid defaults.

## 1.2 Machine Learning

Artificial intelligence (AI) is the function of computer systems that mimic human cognitive functions (e.g., learning and problem solving). Through AI, computer systems use mathematics and logic to simulate the reasoning process which people use to learn new information and make decisions.

Machine learning is a part of AI. Machine learning analyzes and interprets the patterns and structure of data for the purpose of learning, reasoning, and decision-making without human interaction. In simple terms, machine learning supports users to give the computer algorithms a large amount of historical data, and then allows the computer to analyze the data, obtain patterns in it, and use the patterns to make predictions about unknown data, ultimately giving data-driven suggestions and decisions.

Machine learning has been widely used in computer vision, natural language processing, medical diagnosis, credit card fraud detection, securities market analysis, and other fields.

In this project, I will use different machine learning models to predict if the customer will default or not. If the model can do the prediction task well, it can then be used to give suggestions to the bank to avoid lending money to the customers who have high risk of default. The model can also improve the efficiency of loan process of the bank to initially filter out the potential customers who are more likely to pay back the loan.

# 2 Data and Preprocessing

## 2.1 Data Introduction

An official dataset from our bank about the borrower information and the outcome of each loan repayment was used in this project. The dataset contains 2000 historical loan data which collects the age and the amount of income and loan of the borrower, as well as a label representing whether default occurred.

## 2.2 Data Preprocessing

Python is one of the most popular programming languages and is used to do data exploration, cleaning, and model building in this project.

*2.2.1 Data Cleaning*

By using python functions, I found that there existed some NULL values in our dataset, indicating that we missed that value when recording the data. Additionally, some of the age values were below zero, which was abnormal. Considering that I already have enough amount of data, those data were directly deleted from our dataset.

*2.2.2 Data Visualization*

Next, I visualized the data to have a deeper look at the distribution and the relationship of each variable of the dataset.

*Histogram* shows the frequency (the amount of data) within different range. From Fig 1, we can see that the distribution of income and age is uniform (close amount in every range). The frequency decreases when the amount of loan increases. Most recorded data do not have default issue.

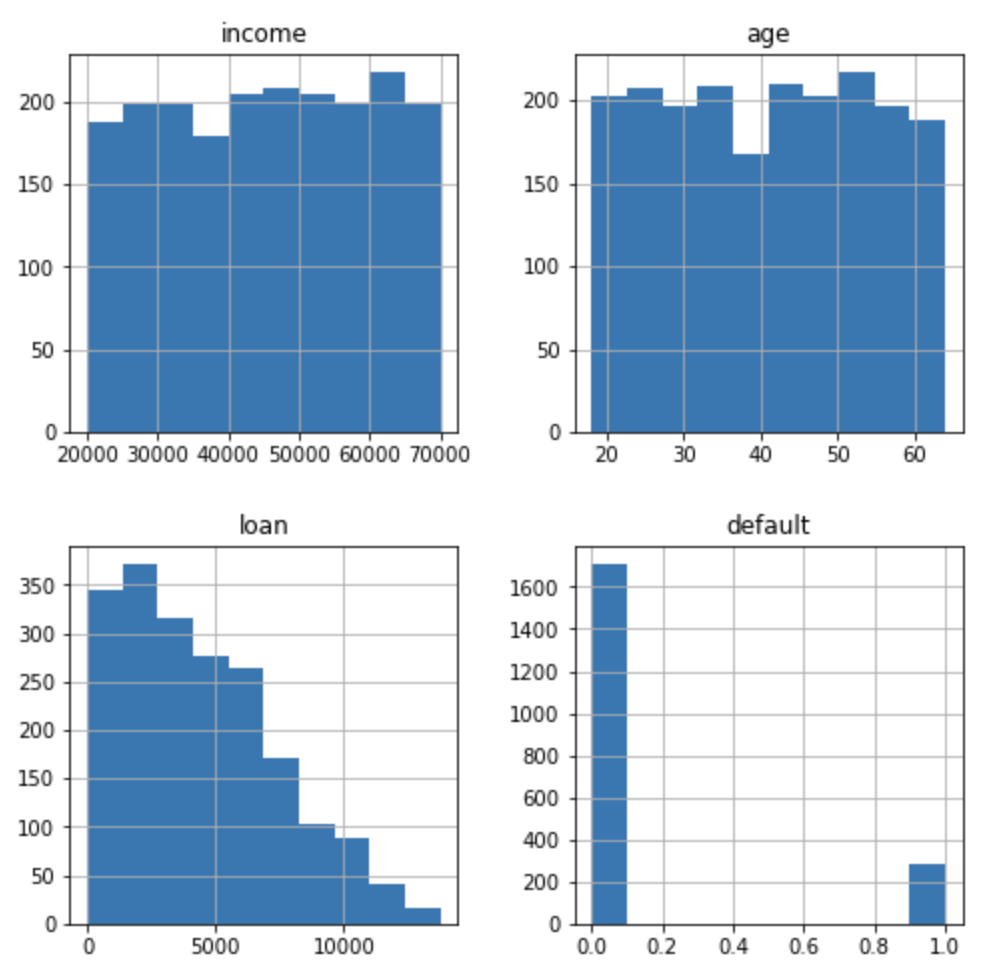


Fig 1. Histogram of each variable

*Heat map* shows the correlation between each variable. From Fig 2, we can conclude that whether default will occur or not is strongly related to the amount of the loan. On the other hand, age and default have negative correlation, meaning that the younger the borrower is, the more likely default will happen.



Fig 2. Correlation between variables

# 3 Model Analysis

## 3.1 Data Splitting

To enable our models to perform better when predicting default, I firstly spilt our datasets into training set and test set. The training set serves as “observed” data which contains most of the information from the original dataset to train our models while the test set is used to identify if the model is capable of deriving correct predictions given “unobserved” data. In general, a good model should perform well in both train set and test set. Here, I used 75% of the original dataset as training set, and the rest as test set.

However, I noticed that in the training set, the number of borrowers with and without default is imbalanced, with more than 1200 borrowers not defaulting and only 212 participants defaulted. This may result in issues when training our models and thus, I have used oversampling to deal with the imbalance data. By using this method, the program will automatically generate new data for the data with smaller size to let its number same as the larger-sized data.

On the other hand, I found that the variance of income and loan were large. Thus, to make the model perform better, I used z-score to change each of these two columns into a standardize normal distribution. This technical method will help decrease the difference of each income and loan data.

## 3.2 Model Introduction

The column ‘default’ only has value of 0 and 1, so we are trying to solve this binary classification problem in this project. Machine learning models such as logistic regression, decision tree, random forest, XGBoost and neural network will be used.

*Logistic Regression* is one of the most basic machine learning models which is a basic method for classification problems. The idea of this model is to find a linear formular for the variables and put the formula into the sigmoid function, and finally find the best parameters and threshold to make the prediction.

*Decision Tree* is a common machine learning technique which predicts the value of a specified target by inferring rules from the features of a dataset. Decision tree is a very simple way to classify data. We can simply imagine the model as a tree of questions that must be answered in sequence to yield a predicted classification.

*Random forest* is a better version of decision tree which avoids overfitting problem (will be explained later). Random Forest constructs multiple decision trees at training time. Every individual tree in this random forest will make a prediction and the prediction with the greatest number of votes becomes the final prediction of the Random Forest model.

*XGBoost* is an advanced machine learning model. Unlike a single decision tree, XGBoost is an ensemble algorithm whose implementation relies on the continuous iteration of weak classifiers. In other words, it will continuously add new trees to learn from the residual values. Generally, XGBoost will outperform the previous models.

*Neural network* is a method of deep learning in machine learning. An artificial neural network is just like a human brain, each layer is made up of neurons. These neurons are connected to each other, passing data, and eventually giving predictions for various problems.

## 3.3 Model Outcome

I will first introduce some indicators and terms to analyze the model result.

*Accuracy*: how many times the machine learning model was correct overall

*Precision*: how good the model is at predicting a specific category (default happens)

*Recall*: how many times the model was able to detect a specific category

*Overfit*: good performance on training set, while performing poorly on test set

\*There are other ways used to analyze the performance of a model such as ROC plot, precision recall curve. I will append those plots in Appendix, but in this section, I will only focus on the previous mentioned indicators.

Table 1 shows the performance of each model. I noticed that even though all the models achieve good performance on the training set (above 90% for all indicators), most of them perform poorly on the test set. Despite the high accuracy, the large difference between precision and recall indicates that our models are overfitting. Specifically, high recall and low precision means that our models are so sensitive that they are very likely to predict default happening.

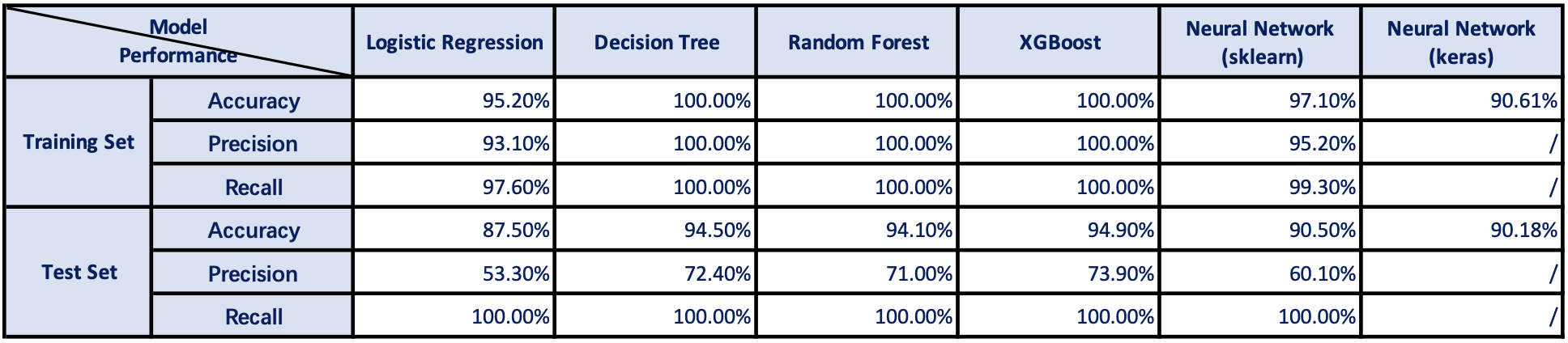
****

Table 1. Model Performance

## 3.4 Model Improvement

Since the performance of the models were not well, I tried to improve the model by changing the features selected and data wrangling. This process is also called feature engineering.

1) There might exist some non-linear relationship between independent variables and the target variable (default). Hence, I also added features such as the squared value, cubic value, square root value of age, income, and loan.

2) Based on intuition, the ratio of loan/income is also a key factor deciding whether default will occur. Hence, I also added this feature to the dataset.

Similarly, I split the dataset to training set and test set and used overfitting to deal with imbalanced training data. However, this time I did not use z-score to standardize the dataset. It is because income and loan are both positive values and the size of the numbers is of comparative significance.

As mentioned in 3.2 section, *XGBoost* is an advanced machine learning model which generally outperform other models, I will apply this model to the new dataset. Additionally, I used *Autogluon* to build a new model. *Autogluon* is based on auto machine learning structure. It can automatically conduct feature engineering, choose appropriate machine learning models, and train the dataset. I did not use neural network because its hyperparameters (number of layers and neurons in each layer) are hard to choose.

Table 2 shows the performance of the improved models. The feature engineering process significantly improves the performance of the models. Both models (*XGBoost* and *Autogluon*) achieved very high accuracy, precision and recall value. Thus, considering my improved models have excellent performance, they can be used to help our bank predicting default issue.

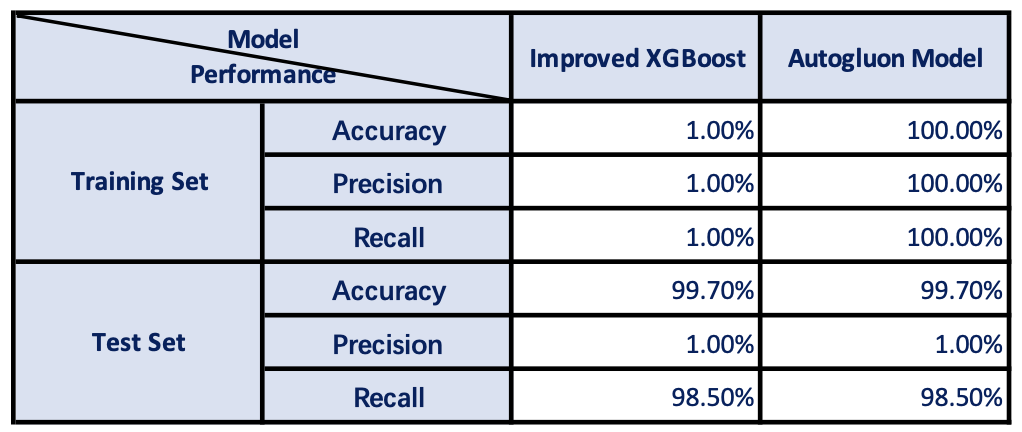


Table 2. Improved Model Performance

## 3.5 Model Summary

This section will qualitatively explain the pros and cons of all models used in this report.

For *logistic regression, decision tree, random forest model*, they have the advantage that those models are simple and training speed is fast. Yet, for complex dataset, those models are more likely to overfit and have poor performance.

For *XGBoost* and *Neural Network*, they have the advantage of having better performance when given complex dataset. Yet, these two models have many parameters, and it is time-consuming to find the best parameters for each problem. It is also hard to explain the structure of *Neural Network*, and the training process of neural network is usually time-consuming.

For *Autogluon* model: this model can automatically conduct the entire process of feature engineering, choosing appropriate machine learning model and training the dataset. Thus, applying this model is quite simple (we simply need to give the dataset). Yet, despite its excellent performance, it is also hard to explain the structure of the model.

# 4 Conclusion

From section 1.1, we learn the concept, cause and impact of credit risk and default. Credit risk will greatly affect the bank's operating efficiency and asset quality. Hence, commercial banks must keep alert and effectively strengthen their credit risk management to avoid defaults. In this project, my goal is to help our bank to predict the likelihood of a borrower having default. Several machine learning models are used. The final performance of the improved models is excellent with accuracy more than 99%.

Therefore, I recommend our bank to use my models to decide whether to accept the loan application of a customer or not. My models can improve the efficiency of the loan process of the bank by initially filtering out the potential borrowers who are more likely to pay back the loan. It can also serve as an advisor for our bank to avoid lending money to the borrowers who have higher risk of default.

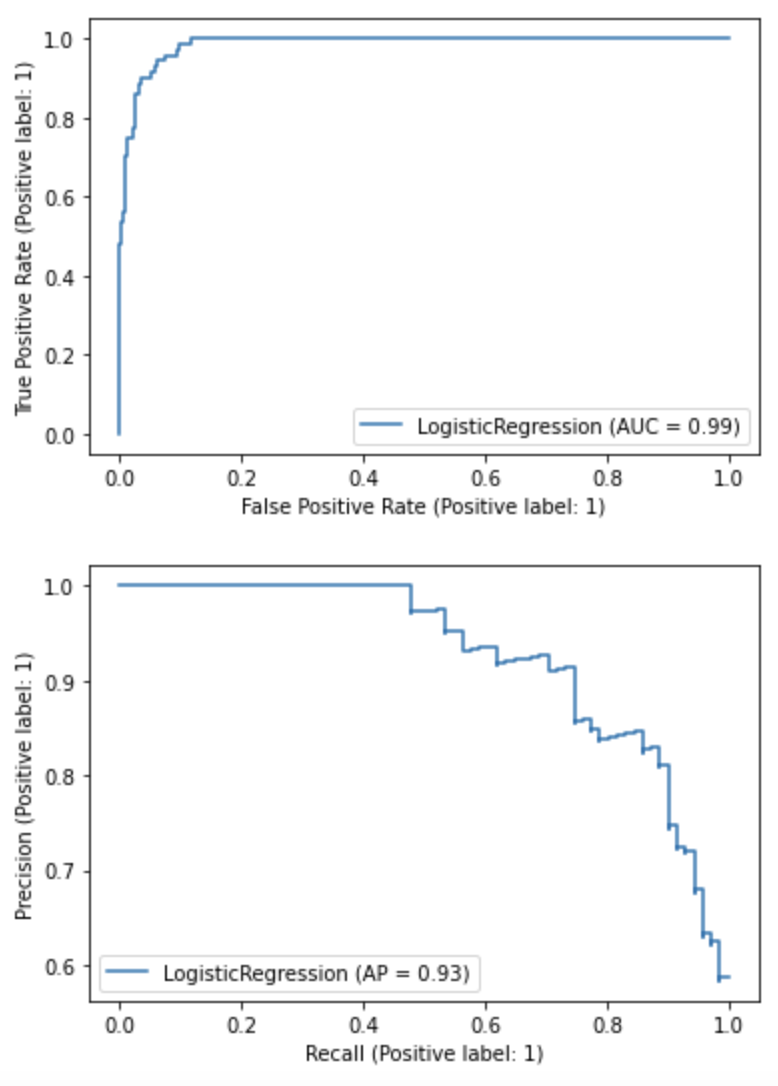
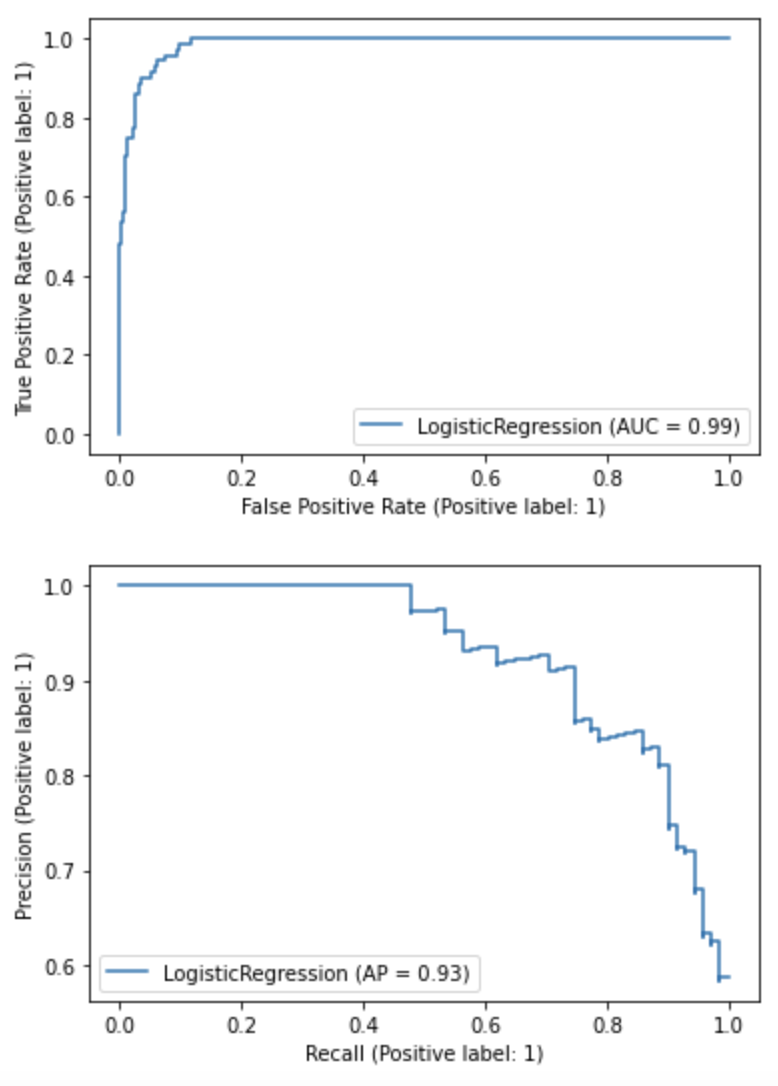
In future study, a more complex dataset can be used to train the model. Here, I only used the dataset that contains the information of age, income, and loan amount of the borrower. There are other factors that will influence whether default will occur such as the marriage status, education level, occupation, and credit history of the borrower. In this case, we will also use more advanced machine learning models and other techniques like Random Search, Bayesian Optimization to find the best parameters for each model to achieve better performance on the prediction result.

In this project, I have only presented one possibility which our bank could apply machine learning tools which is default prediction, but the potential of machine and deep learning models is vast and there are still many ways to broaden the scope and depth of my study. For example, credit card fraud detection and securities market analysis are also the popular fields where machine learning can be applied to. In all, I believe that machine learning can bring many benefits to our bank and hope that my research will initiate the use of machine learning models in our bank to speed up and streamline more processes, and help our bank generate more profit and wealth.

# Appendix

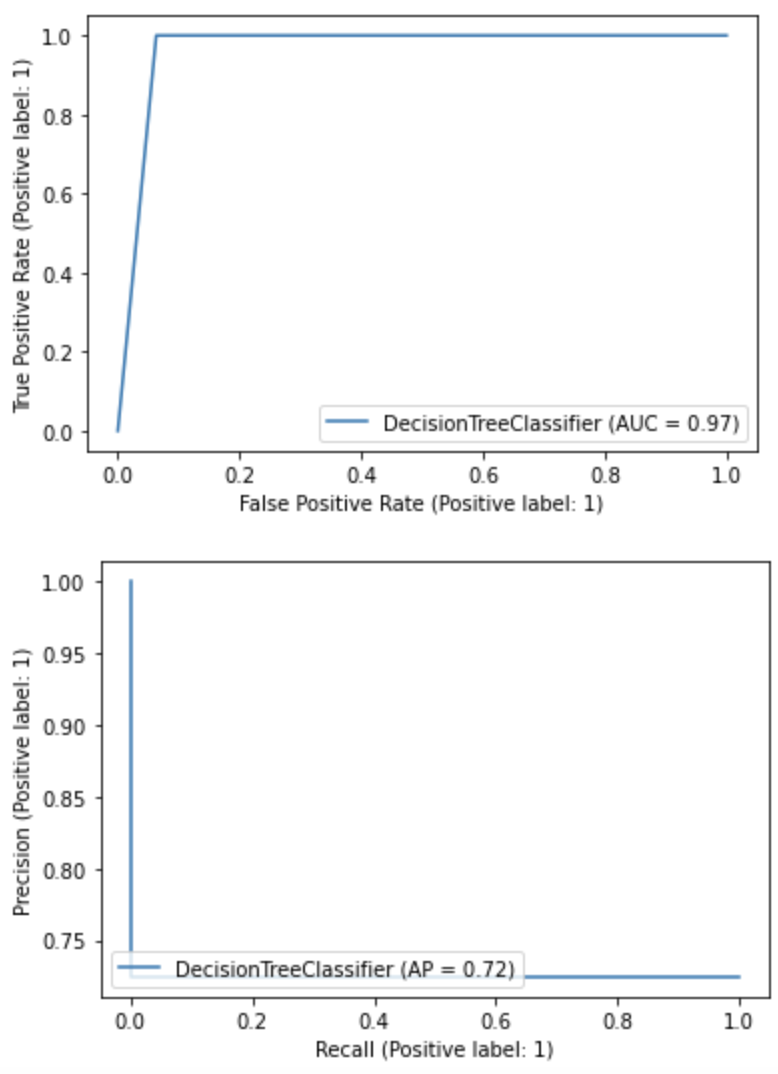
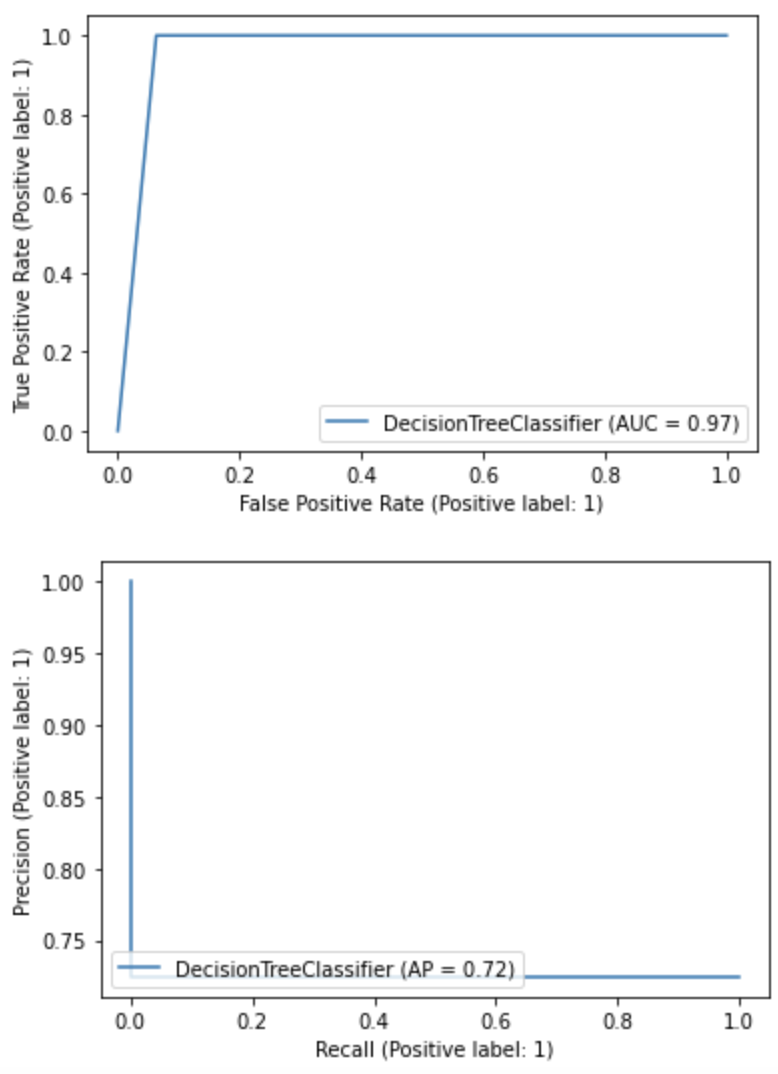
This section shows the ROC plot and precision recall curve of each model. The ROC plot reflects the sensitivity and specificity of the model. The curve which is more close to top left part of the plot has better performance. Precision Recall curve shows the balance between precision and recall. Generally, higher recall leads to lower precision. Thus, the PR curve which is more close to top right part of the plot has better performance.

*Logistic Regression*

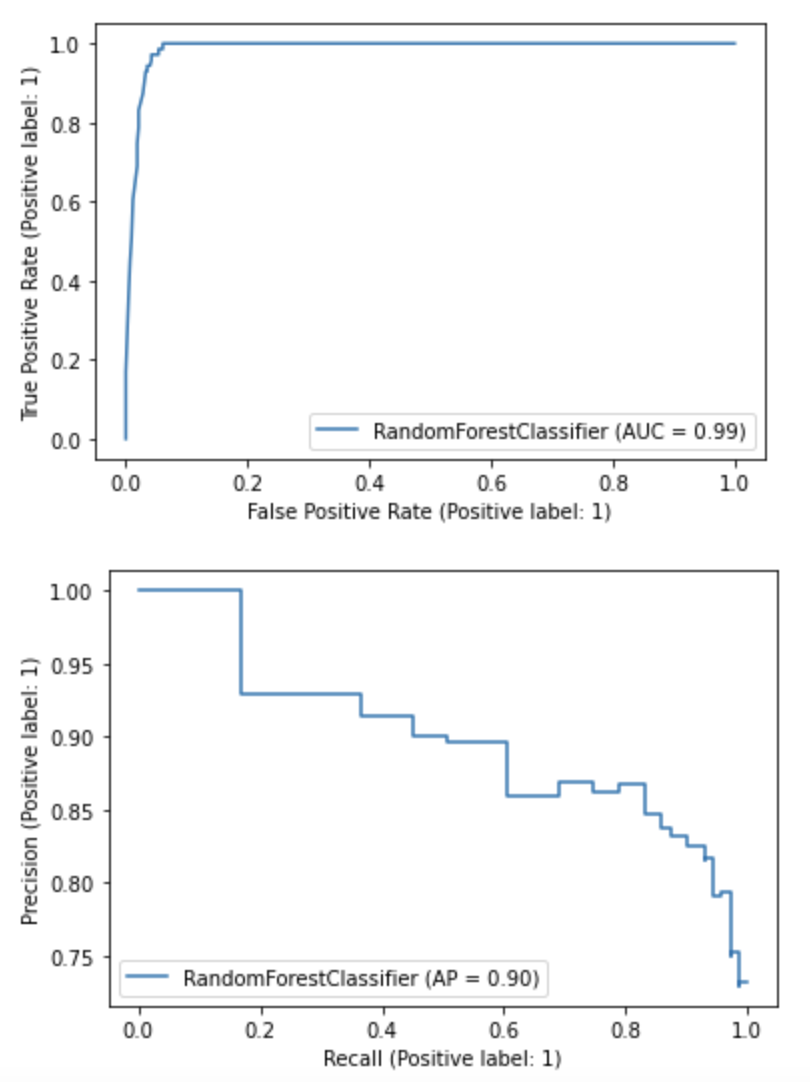
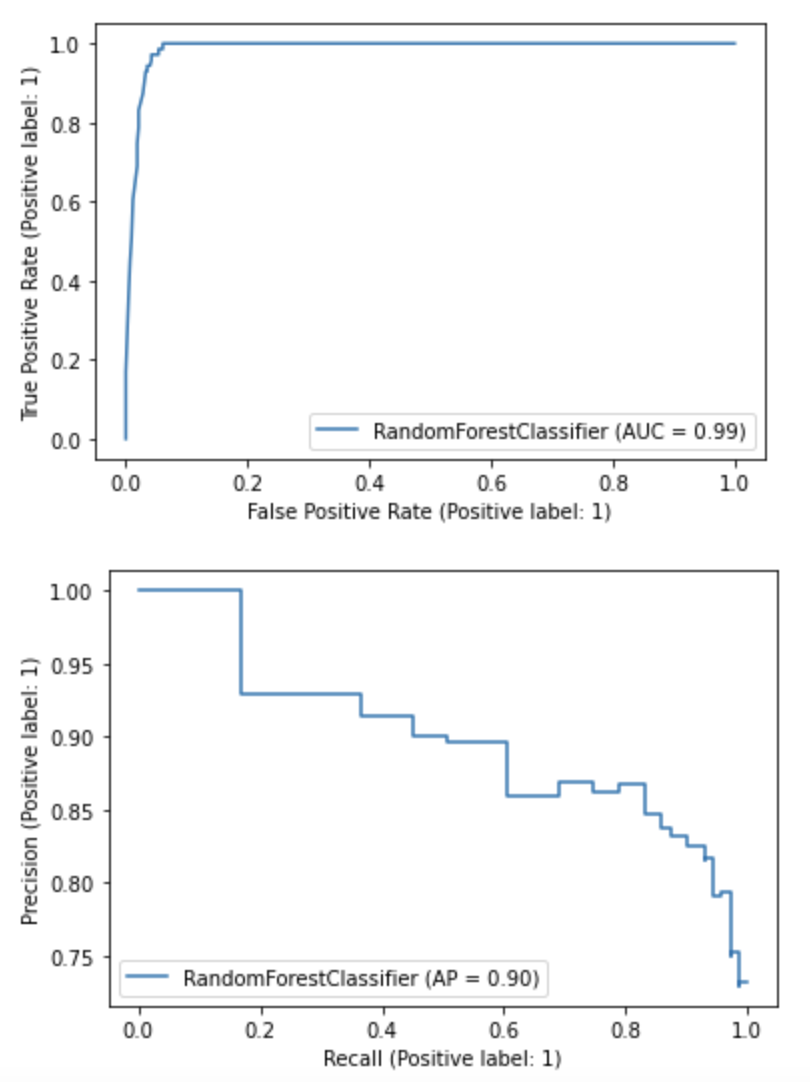
ROC and PR curve of Logistic Regression

*Decision Tree*

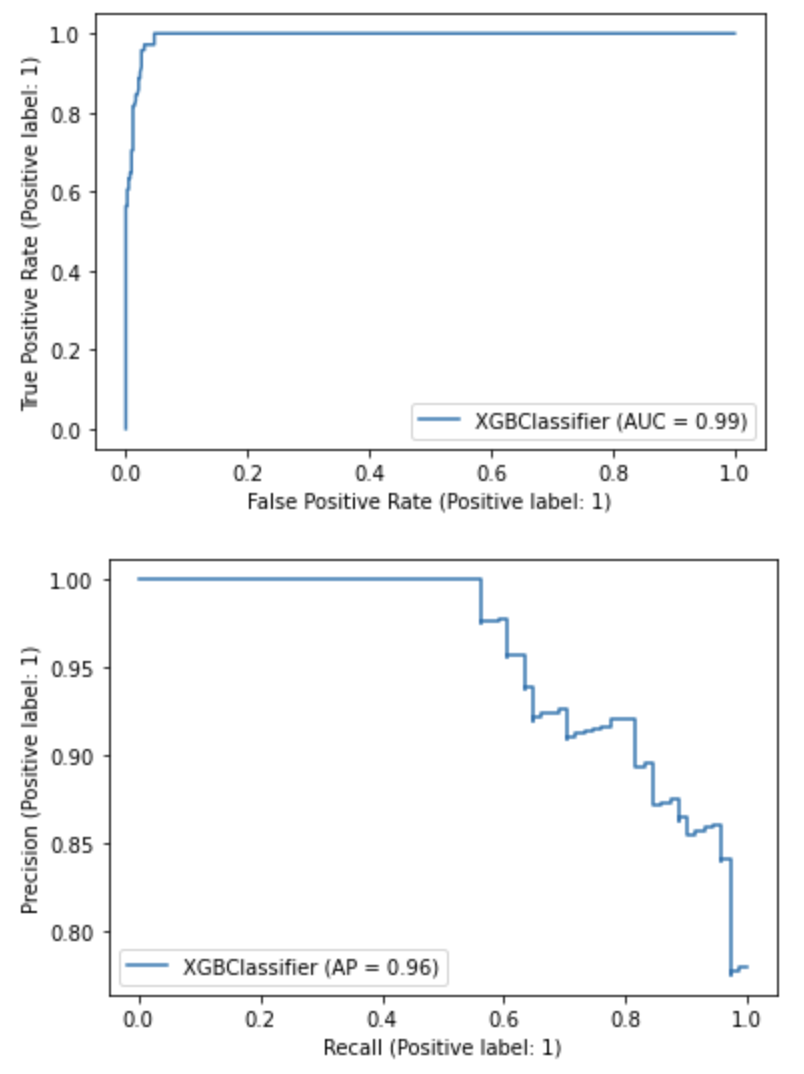
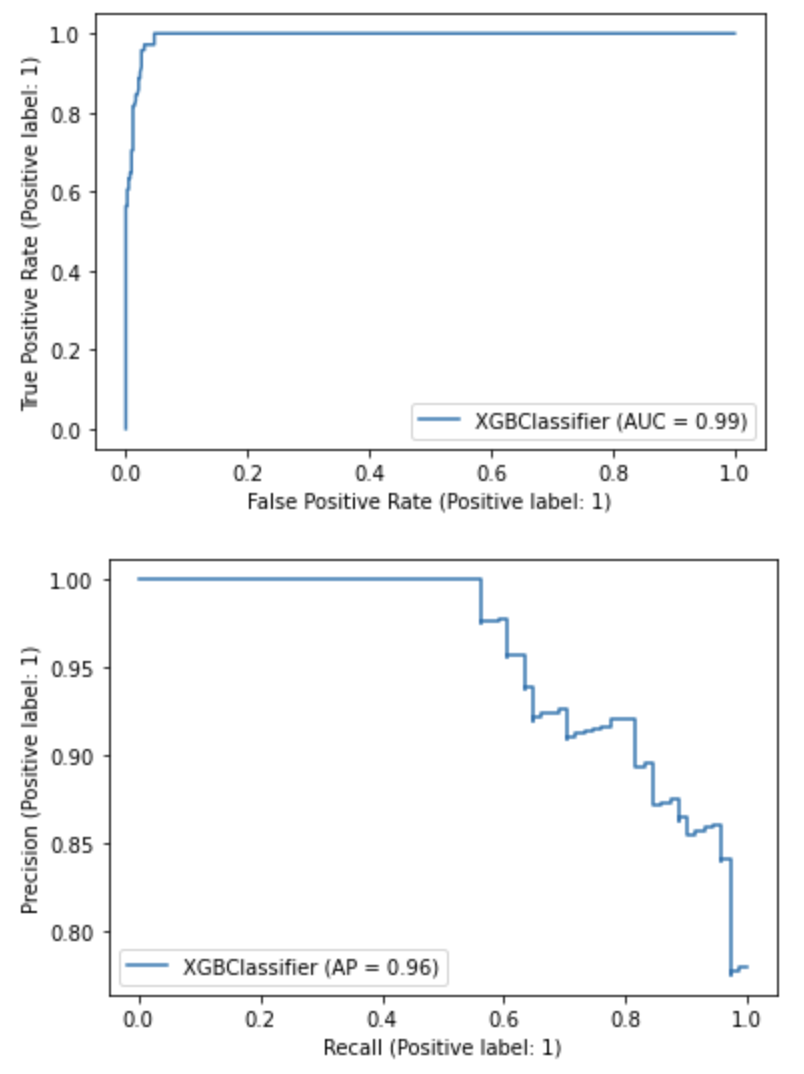
ROC and PR curve of Decision Tree

*Random Forest*

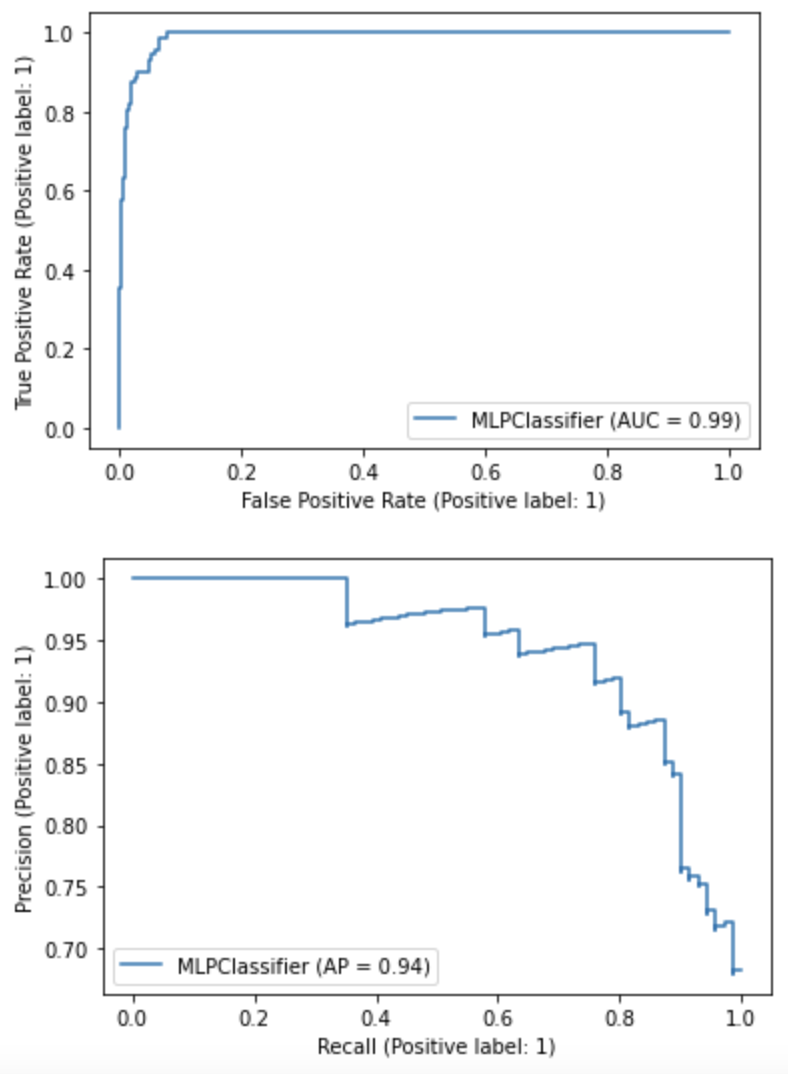
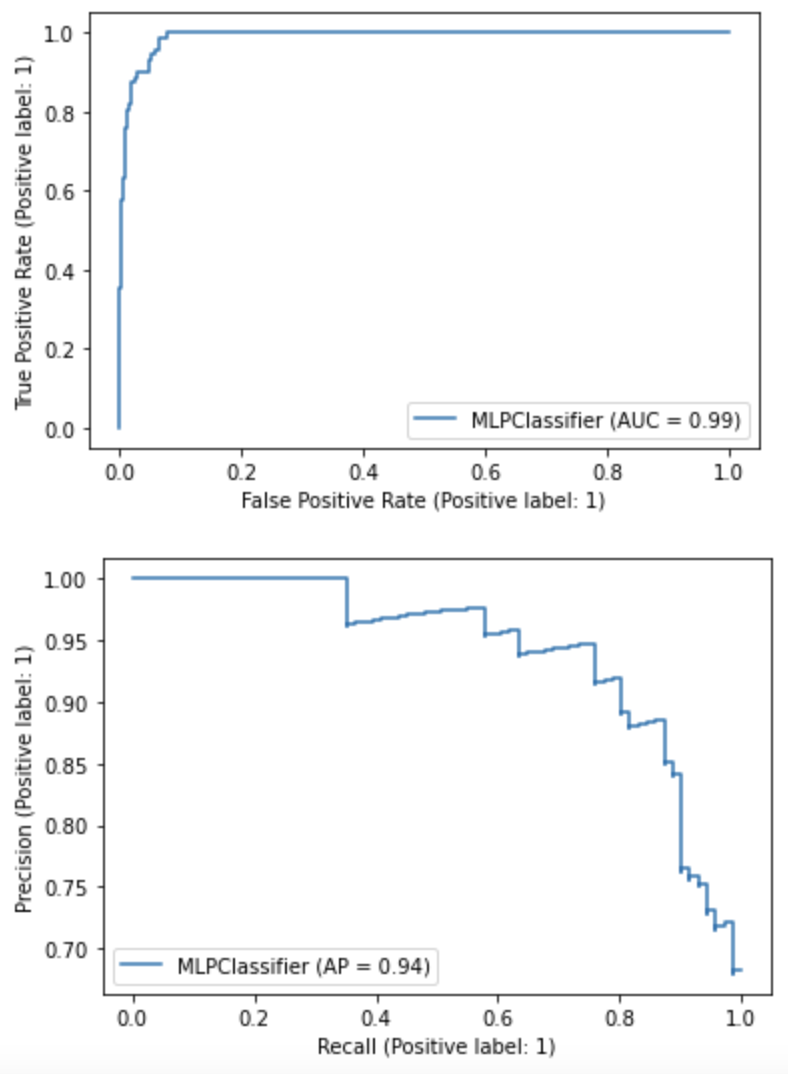
ROC and PR curve of Random Forest

*XGBoost*



ROC and PR curve of XGBoost

*Neural Network*

ROC and PR curve of Neural Network