# **Final Report**

Term Deposit Prediction

Class: IST\_707

Suihin Wong(Henry)

Introduction

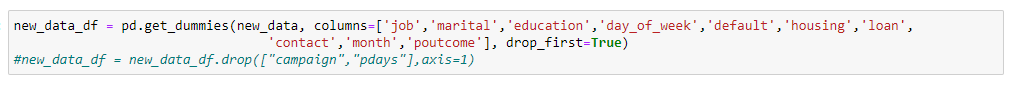
Bank is a financial institution that provides credit,debit and investment products to help their consumer or business customers succeed in their financial life. Term deposit (CD) is a fixed term investment product in which the customers sign an agreement with the bank to put a certain amount of money for a certain period of time in the bank, and for return, the bank will guarantee a fixed interest rate of return after the agreement ends. Usually, the interest rate of term deposit is lower than other investment products. However, term deposit offers a guarantee of return which a lot of customers who don’t like risk would purchase for investment.

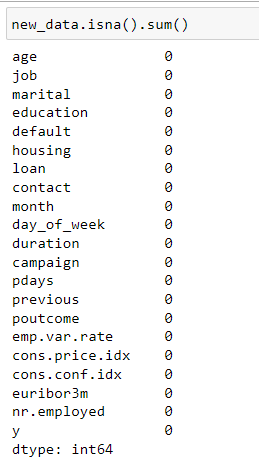
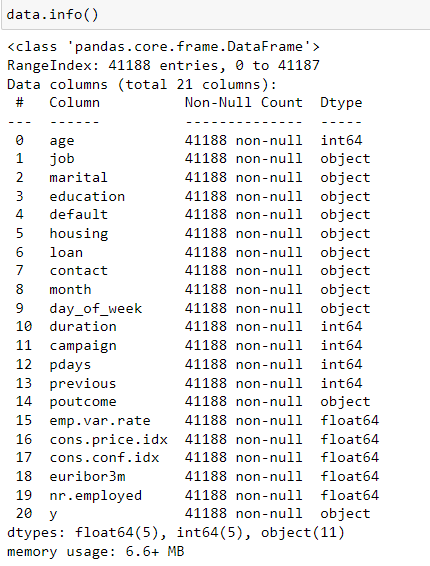
This report presents the findings for direct marketing campaigns conducted by a Portuguese banking institution. The findings are supported by data related to the phone calls made during the marketing campaigns. The goal of this data report is to predict if the client will subscribe to a term deposit and to increase revenue for the bank institution by targeting potential customers.

Furthermore, the goal of this report is to help the bank institution to prepare for their next telephone market campaign and increase the performance by generating more sales or saving cost. The bank would like to understand how to attract more customers, provide more sales and target potential customers.

# The Data

The data based on the phone calls was collected in Kaggle. The data set is related to a calling marketing campaign of a Portuguese banking financial institution. The dataset included 41,188 rows and 20 columns. The data set doesn’t include any missing values. Moreover, the category columns were converted into a numeric column with a one-hot encoding method. The data set is imbalanced with 89% of no and 11% of yes in the target variable. Resample method would be introduced in the model session.



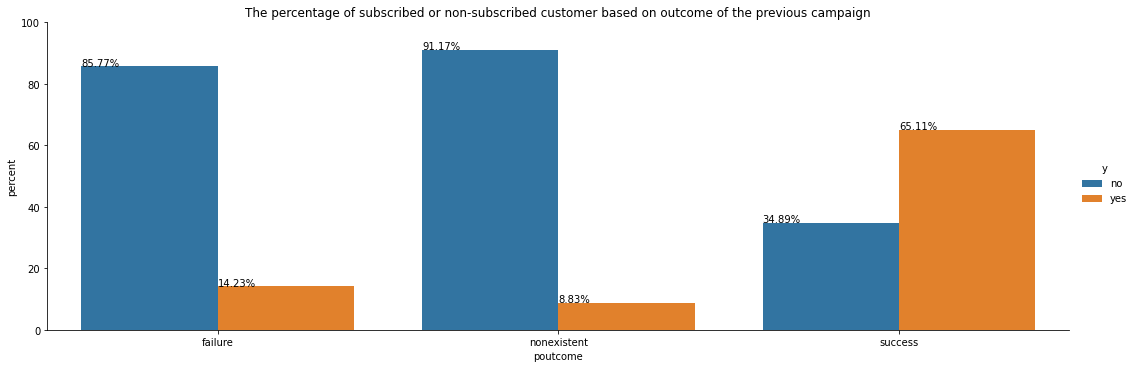


# EDA

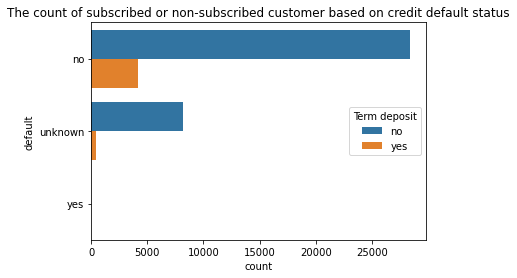
Exploratory Data Analysis is a process to investigate on a dataset to identify any trends and patterns in the data to help answer any business related questions with visualization and mathematical methods. Below graph shows descriptive statistics for numeric variables in the data.

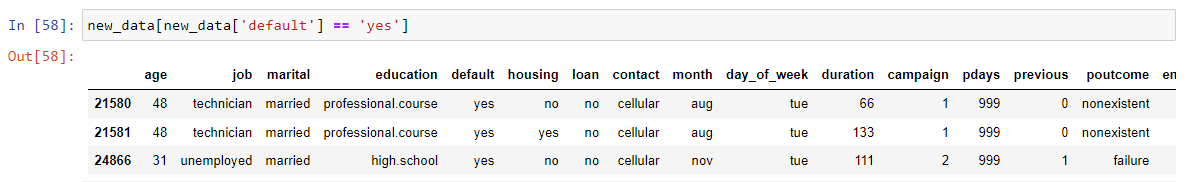
****

Based on the outcome from the previous campaign, there are around 65% of previous term deposit subscribed customers that subscribe in the current campaign. There are only around 9% of non-existent customers from the previous campaign which subscribe to the current campaign. Moreover, around 15% of customers did not subscribe from the previous term deposit but subscribe to the current campaign.

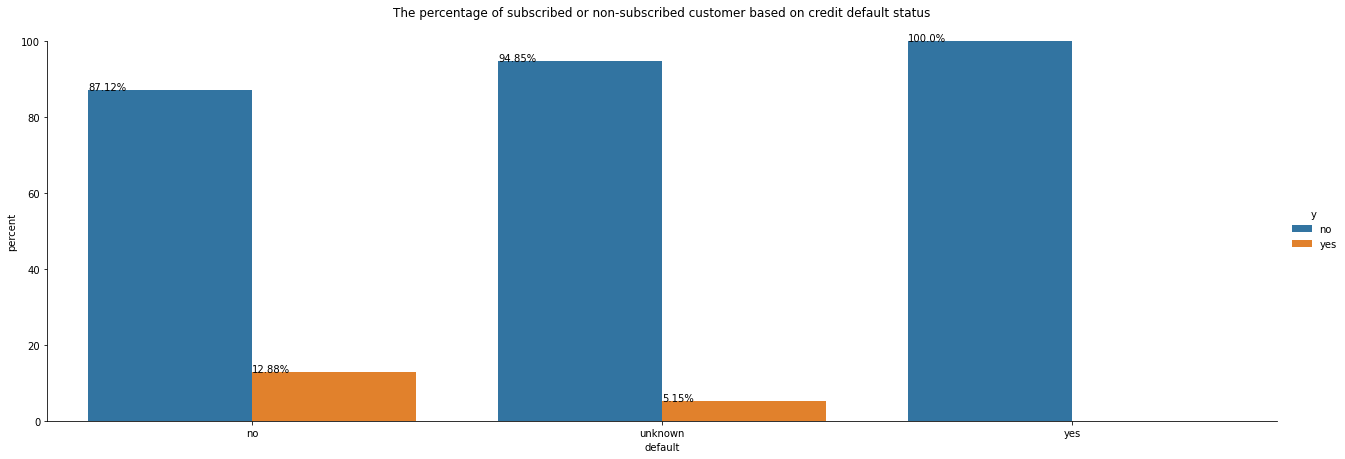
****

Credit default means people fail to make payment to credit products. Credit default status is a critical part in term deposit because people would not spend money on saving if they don’t have enough funds to make their regular payment. Most of the customers in this market campaign are not credit default. The second graph shows there are only 3 customers who are in credit default and around 8,000 customers are unknown in their credit default status.

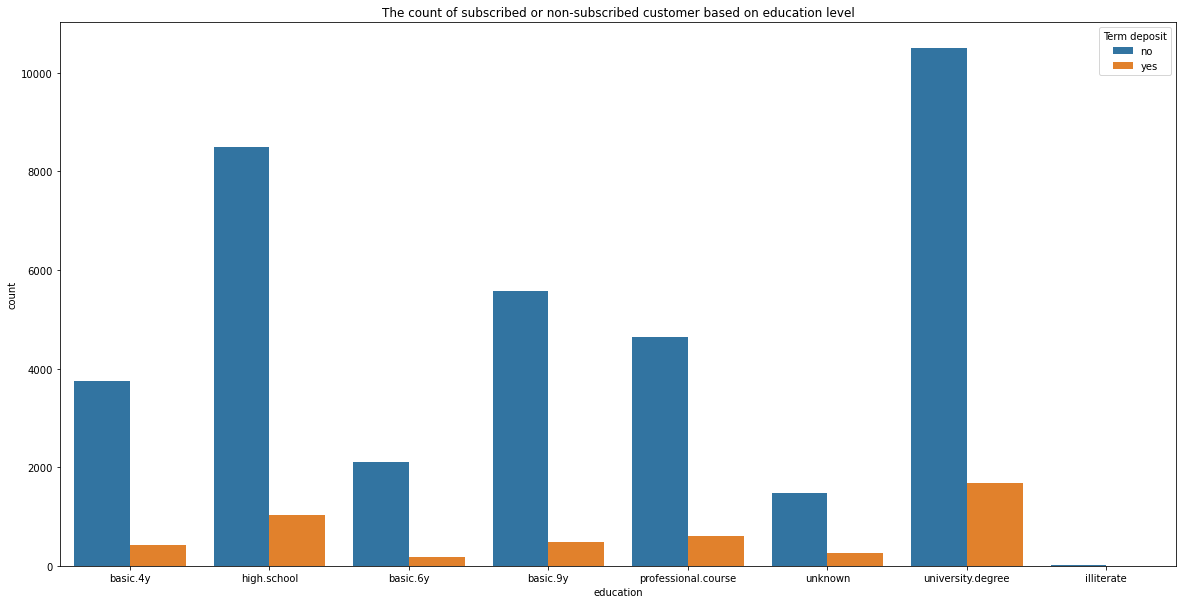
****

****

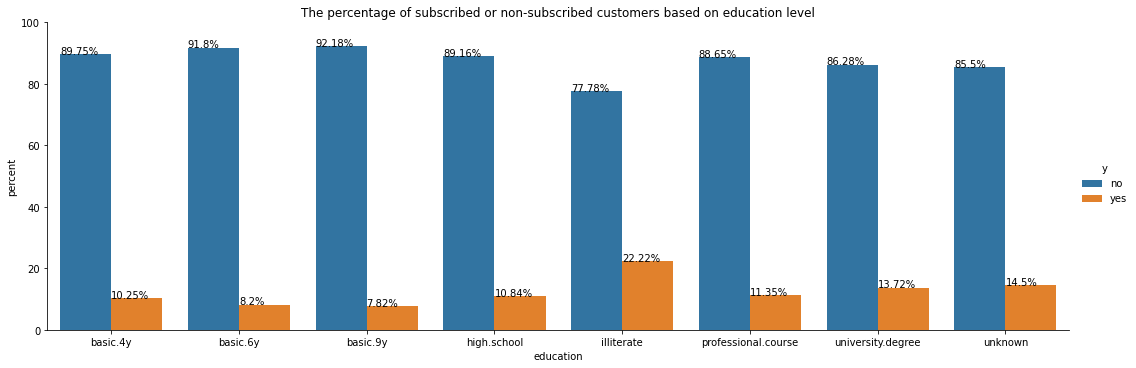
Furthermore, the percentage graph shows the accept rate in each status in credit default. Around 13% of credit default customers subscribe to term deposit which matches the assumption that credit default customers would have less chance to accept term deposit. Moreover, there are only 3 customers who are credit default and all of them did not subscribe to term deposit. However, more data are needed to conclude the finding to better represent the population.

****

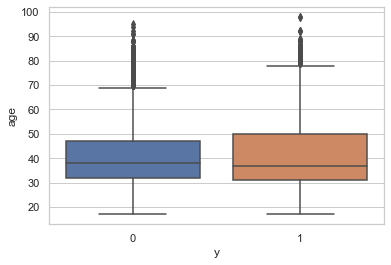
Education level is one of the most important elements to consider saving. People with higher education would have a higher salary so they would have a higher chance to put it in savings or investment. The graph below shows the customer count for different education levels. University education levels have the highest total count and illiterate education levels have the least count. Illiterate education level does not give too much information on the analysis due to the size of the sample.

****

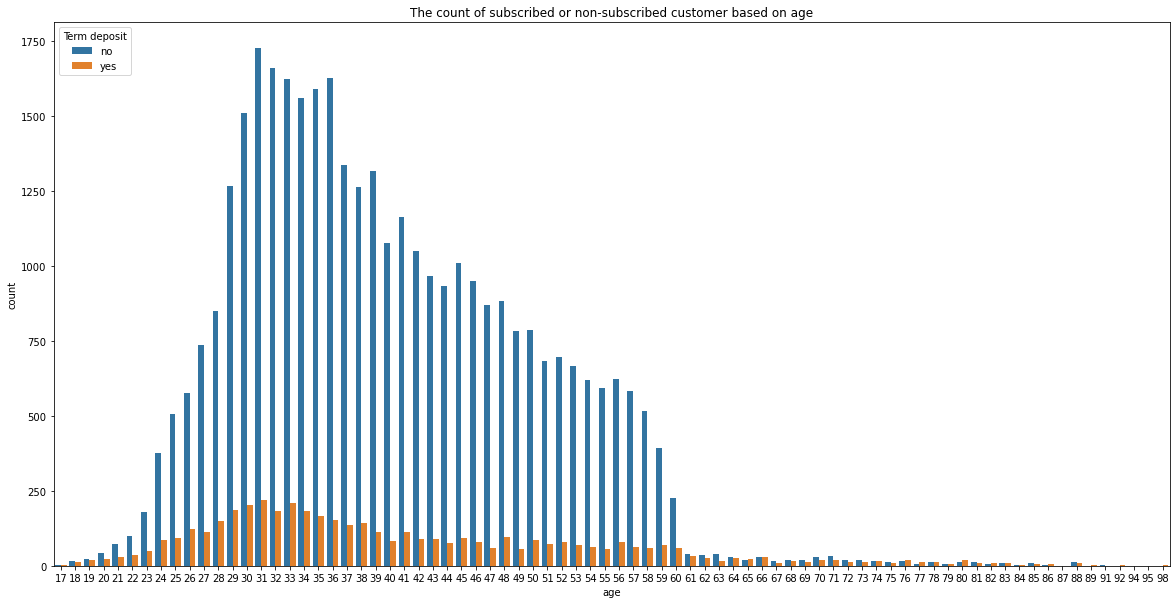
Furthermore, in the education level percentage graph, basic 4,6,9 years and high school education levels have an acceptance rate from 7% to 11%. However, professional, university and unknown education have slightly higher acceptance rates from 11% to 15%.

****

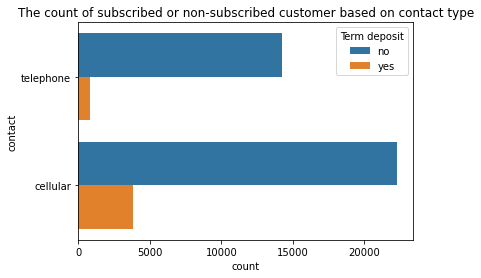
Boxplot below shows the IQR, max and min of the two groups based on their age and subscription status of term deposit. The group of customers that subscribed to term deposit has a longer range of age.

****

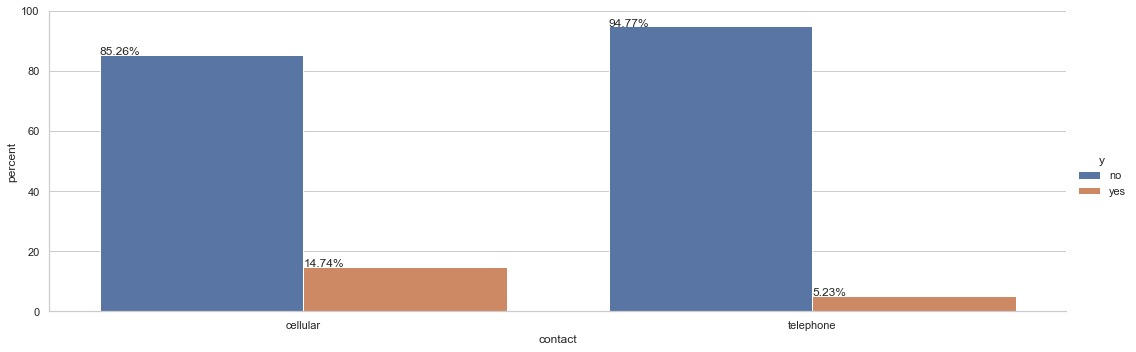
Below graph shows the total number of customers based on their age. Most of the customer’s age are from around 25 to 50. There are trends showing that there are perks at age 31 to accept term deposits in this data.

****

The below graph shows the total number of customers based on their contact type from the campaign. Customers with cell phones are more than customers with telephones. Moreover, customers with cell phones as contact type have higher acceptance rate than customers with telephones.

****

Below graph shows the percentage of both cell phone and telephone customer acceptance rate in term deposit. There are around 15% of customers who would subscribe to a term deposit when using a cell phone to answer the market campaign calls. There are only 5% of customers who would not subscribe to a term deposit when using a telephone to answer the market campaign calls.

****

# **Modeling**

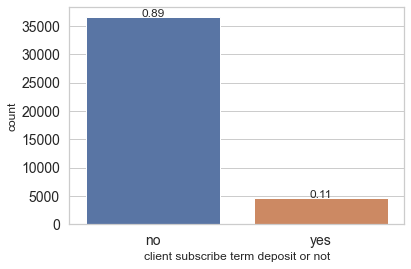
# **Data preparation**

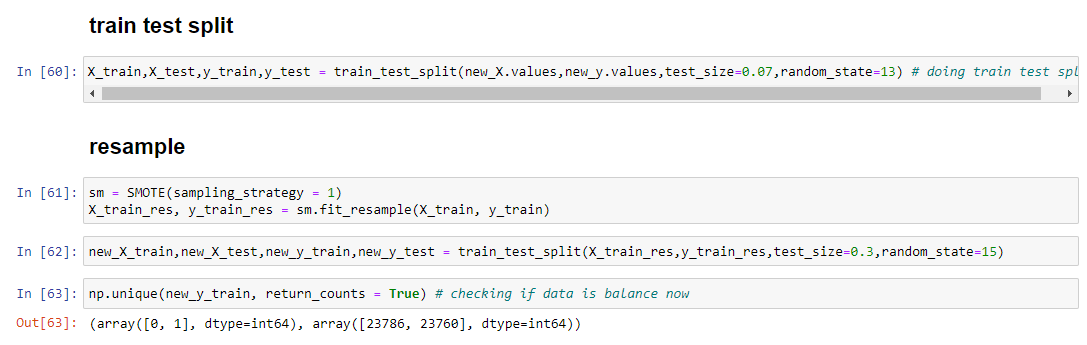
## Train test split

Perform a train test split method from the original data into 93% of the data in the train set and 7% in the test set due to a small data set. The purpose of setting 7% in the test set is to have a bigger train set to train the model for better performance if possible.

## Oversample

As mentioned in the beginning, this dataset is an imbalance data set. An Oversample method was applied to solve this problem. Using Smote function to re-sample the training data set to split 70% in the training data set and 30% in testing dataset for validation.



****

**Cost Benefit matrix**

To compare the performance with different classification models, Accuracy is one of the ways to evaluate models. However, Accuracy does not perform well in evaluating imbalanced cases. In this report, F1,precision and recall would be used to evaluate performance for different models. Cost benefit matrix is built on top of the confusion matrix. Based on the cost benefit matrix, the company would like to minimize false negative which means focus on higher recall.

|  |  |  |
| --- | --- | --- |
|  | Predicted yes | Predicted no |
| Actual yes | $800 | -$1000 |
| Actual no | -$50 | $0 |

# **Logistic Regression**

Logistic regression is one of the most popular classification algorithms to predict the probability of a binary outcome. Trained the model with using 1000 in max\_iter parameter, balanced in class\_weight and liblinear in solver. Max\_iter parameter is the maximum number of iterations for the solver parameter. Balanced class weight is to use the value in the target variable to automatically adjust the weight for classes. Liblinear solver is good in optimization for small datasets. Test validation in the new\_X\_test set with 87% in accuracy and the actual result is 85% accuracy. Both of the results are very close and the actual result is slightly worse than the test validation. In this Logistic model, the precision is 41% and recall is 88%. Based on the cost benefit matrix, this model can generate $152,650 revenue with 0.5 threshold.

Graphical user interface, text, application, email

Description automatically generated

# **Random Forest**

In the random forest model, trained with 5000 in n\_estimators, 6 in max\_depth and 3 in min\_samples\_split. The n\_estimators mean the number of trees in the model, the max\_depth is to limit the growth of trees in the model. Min\_samples\_split means the minimum number of samples for the tree to split. By adjusting these parameters can decrease overfitting and increase performance in the test score. The precision is 42% and recall is 70%. Based on the cost benefit matrix, this model can generate $65,350 revenue with 0.5 threshold.

Graphical user interface, text, application, email

Description automatically generated

# **Gradient Boosting**

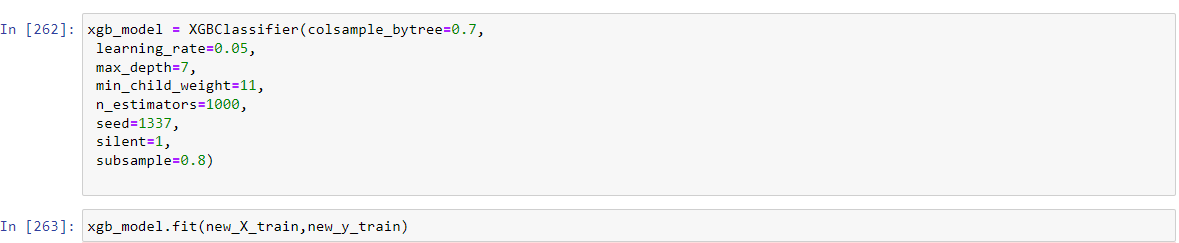
In the gradient boosting model, trained with 0.05 in learning\_rate. Learning rate means how fast the model can learn. With a low learning rate, the performance performs better. In this model, the precision is 49% and recall is 75%. Based on the cost benefit matrix, this model can generate $95,550 revenue with 0.5 threshold.

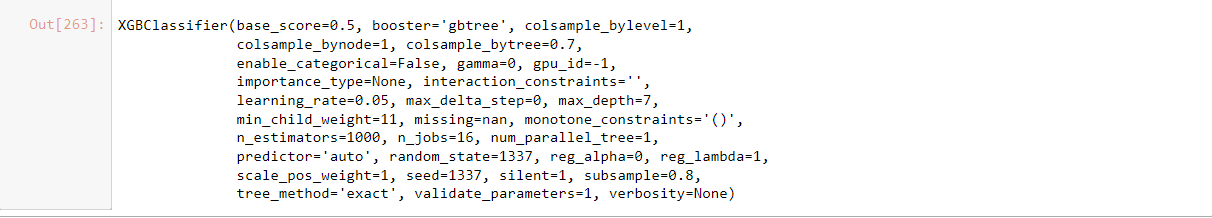
Graphical user interface, text, application, email

Description automatically generated

# **XGboost**

In the XGboost model, trained with 0.7 in colsample\_bytree, 0.05 in learning\_rate, 7 in max\_depth, 11 in min\_child\_weight, 1,000 in n\_estimators and 0.8 in subsample. Learning rate means how fast the model can learn, max\_depth, colsample\_bytree, min\_child\_weight and subsample are to tune to tree to prevent overfitting. In this model, the precision is 60% and recall is 54%. Based on the cost benefit matrix, this model lost $11,700 with 0.5 threshold. The false negative is highly weighted in the cost benefit matrix, so any model which does not have high recall would not perform well with the cost benefit matrix.

****

****

Graphical user interface, text, application, email

Description automatically generated

# **SMV**

In the SVM model, standardize the train and test data. After that, trained with three different kernels such as sigmoid, polynomial and Gaussian radial basis. First, rbf svm model generates a precision of 52% and recall of 67%. Second, the sigmoid svm model generates a precision of 31% and recall of 80%. Finally, the polynomial svm model generates a precision of 48% and recall of 71%. Based on the cost benefit matrix, the sigmoid model can generate the most revenue of $104,150 with 0.5 threshold.

Graphical user interface, text, application, email

Description automatically generated

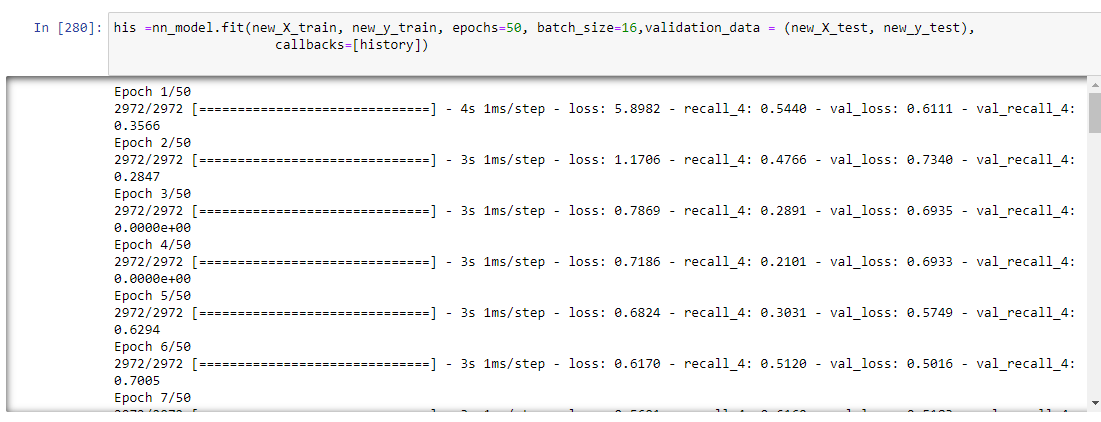
Graphical user interface, text, application, email

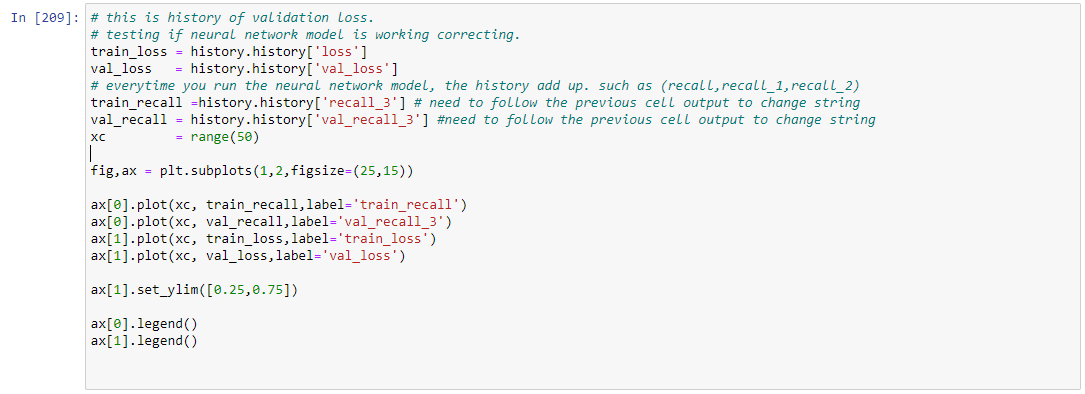
Description automatically generated

# **Neural Network**

In the Neural Network model, trained with five layers with activation function relu and at the end with one layer with activation function sigmoid. The learning\_rate is 0.0001 because the Neural Network learns very slowly. Used cross-entropy for the loss function because this is a binary classification case. In this model, the precision is 38% and recall is 93%. Based on the cost benefit matrix, this model can generate $179,300 revenue with 0.5 threshold. The second last validation graph shows the performance in both train and test to validate the model run correctly.

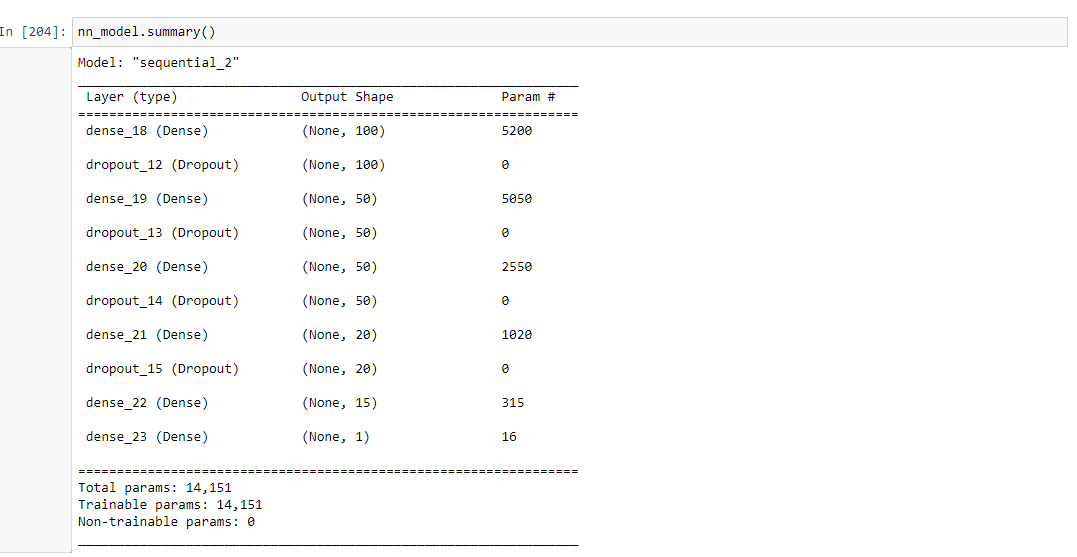
****

****

****

Graphical user interface, text, application

Description automatically generated

****

**Chart

Description automatically generated**

# **Result**

# **Logistic regression**

Accuracy: 85%, F1: 56%, Precision: 40%, Recall: 88%, Profit: $196000 with threshold .85

# **Random forest**

Accuracy: 92%, F1: 53%, Precision: 40%, Recall: 70%, Profit: $195350with threshold .7

# **Gradient boosting**

Accuracy: 93%, F1: 59%, Precision: 49%, Recall: 75%, Profit: $194850 with threshold .85

# **XGboost**

Accuracy: 92%, F1: 59%, Precision: 63%, Recall: 56%, Profit: $189450 with threshold .95

# **SVM**

**Kernel = rbf**

Accuracy: 90%, F1: 59%, Precision: 40%, Recall: 69%, Profit: $187850 with threshold .95

**Kernel = sigmoid**

Accuracy: 79%, F1: 45%, Precision: 32%, Recall: 79%, Profit: $154000 with threshold .65

**Kernel = polynomial**

Accuracy: 89%, F1: 58%, Precision: 48%, Recall: 71%, Profit: $180000 with threshold .95

# **Neural Network**

Accuracy: 85%, F1: 56%, Precision: 40%, Recall: 88%, Profit: $195750 with threshold .75

**ROC curve**

ROC curve with AUC is one way to measure the performance of classification models with different thresholds. However AUC-ROC curve is sensitive to imbalance class and provides misleading information. In this report, instead of ROC-curve, the profit curve is used to measure the performance for different models in different thresholds.

Chart, line chart

Description automatically generated

**Profit curve**

Neural network and Random Forest generate the highest profit with 0.7 threshold based on the cost benefit matrix. The other models are very close with different thresholds. However, the profit curve is based on the cost benefit matrix that was created to achieve our business goal. In the cost benefit matrix, the False negative is minus 1000 dollars in profit compared to false positive only minus 50 dollars in profit as a labor cost.

**Chart, line chart

Description automatically generated**

Text, email

Description automatically generated with medium confidence

# **Conclusion**

After the analysis in the market campaign, we identified that contact type was an important element to find out which customers are more likely to purchase a term deposit. Education level also had an impact on whether customers could purchase a term deposit. Higher education customers had more potential to subscribe to term deposits. Moreover, credit default customers would be less likely to subscribe to term deposits because they couldn’t have enough money to pay their debt.

Additionally, the company could send notice to customers to request an update on their contact type to put their cell phone as current contact type. attract higher education level customers by advertising low-rate student loans to build customer pipelines. Generally, helped customers to consolidate their debt. It could help customer retention and keep customer loyalty.

Moreover, the classification prediction model could generate around 190,000 in profit to the bank. The model could save labor cost by targeting the right customers to order term deposits, so employees could perform other tasks. However, the cost benefit matrix was only set for certain business goals. If the matrix changed, the model performance would vary and the best model would change.