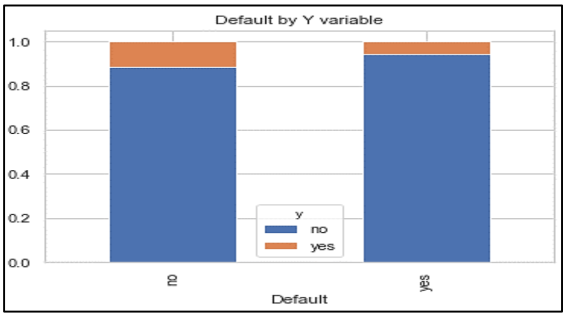
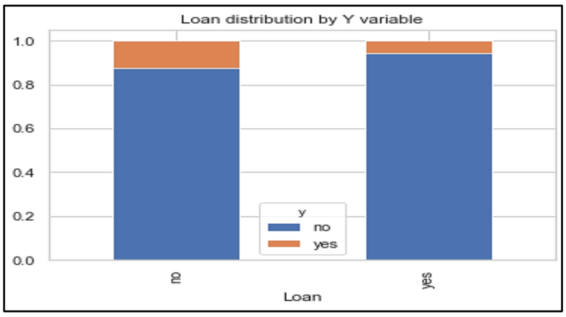
Section A: Summarization

In section A, statistical analysing of variables is accomplished to achieve relationship between each input features and the output variable *'y'*. In addition, examining relation of each variable are conducted. As a result, input variable *'default', 'loan', 'housing', 'duration'* and *'poutcome'* shows highly related to variable *'y'*. On the other hand, variable *'age', 'job', 'marital', 'education'* and *'balance'* determined as not significant variables. Lastly, variable *'contact', 'day', 'campaign', 'pdays'* and *'previous'* requires more analysing to figure out the impact on variable *'y'*.

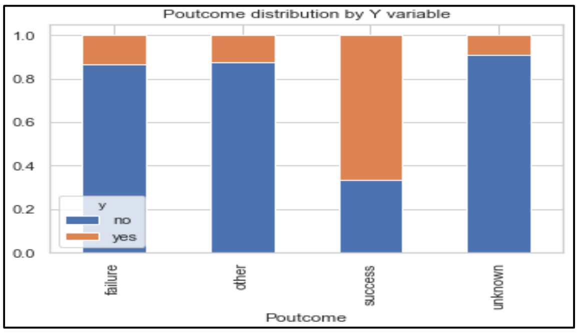
Statistical analysis is divided into two parts; 1) analysis of seven numerical variables and 2) analysis of eight categorical variables. However, in the reports only five variables which estimated as a significant variable are described



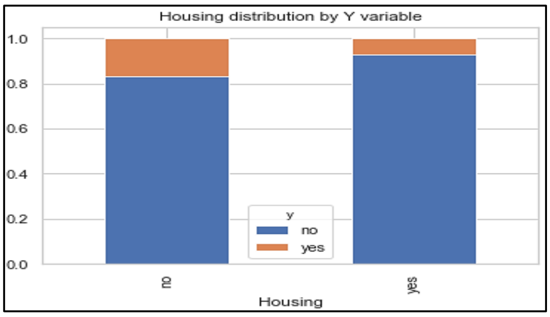
Input feature *'default'* and output variable *'y'* shows highly related relationship. People who have credit they are defaulting on indicates lower number of people who are likely to buy the 'N/LAB platinum' product.



Input feature *'loan'* and output variable *'y'* is highly related. This is because, people who has a loan shows lower number of people than who does not have a loan, in terms of purchasing the 'N/LAB platinum' product.

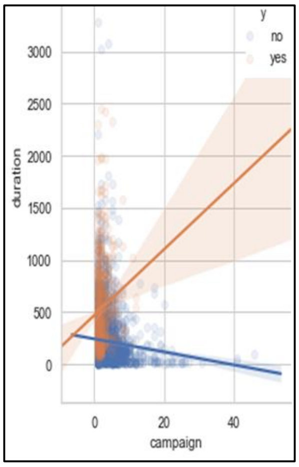


Input feature *'poutcome'* and output variable *'y'* demonstrates meaningful relationship, especially success category shows high number of people who are likely to invest on 'N/LAB platinum' product.

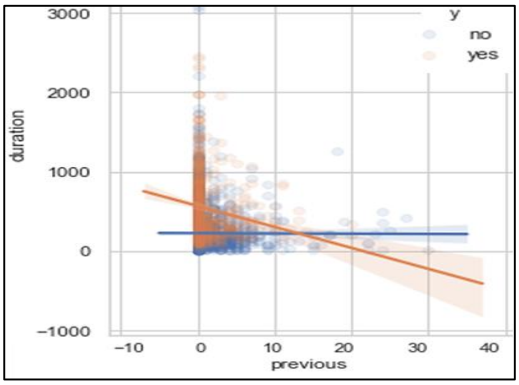


Input feature 'housing' and output variable 'y' points to substantially related relation. People who has taken out a housing loan tend to not interested in our product.

Last variable is *'duration'*. In this variable *'duration'* analysis, relation of each numerical variable are conducted. After comparing variable *'duration'* with other numerical variables, I found out meaningful relationship with variable *'campaign'* and *'previous'*.



Variable 'campaign' and 'duration' shows negative correlation. The more number of contacts performed during the campaign, the less last contact happened.



This negative correlation features also happened between variable *'previous'* and *'duration'*. The more number of contacts performed before the campaign, the less last contact happened.

However, these numerical variables require more analysis to identify whether it is a significant variable to output variable *'y'*. And this analysis will be continued in Decision Trees.

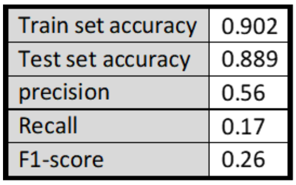
Section B: Exploration

Since statistical analysis in section A, five variables founded out as a significant variable to output variable 'y'. However, variable *'contact', 'day', 'campaign', 'pdays'* and *'previous'* need to be figure out whether it is significant variable or not, through putting as an input variable in decision tree model. All the six Decision Trees models contain variable *'default', 'loan', 'housing', 'duration'* and *'poutcome'*, however, in five models, it additionally includes each a single potential variable from above and the other one would be standard model with only contains five variables. To find out the result, comparing three criteria is essential; 1) difference of training set accuracy and test set accuracy, 2) precision and 3) f1-score. Moreover, all the models including Decision Trees in Section C, K-Nearest Neighbours, Logistic Regression and Random Forests model in Section D used equal four split data set; *'x_train', 'y_train', 'x_test'* and *'y_test'*. The size of test set is 25 per cent and stratified by variable *'y'*.

	Five variables	**+ variable 'contact'**	**+ variable *'day'***	**+ variable *'campaign'***	**+ variable *'pdays'***	**+ variable *'previous'***
Train set accuracy	0.969	0.976	0.996	0.985	0.971	0.971
Test set accuracy	0.868	0.864	0.850	0.854	0.868	0.867
Precision	0.38	0.37	0.38	0.38	0.34	0.38
F1-score	0.40	0.38	0.37	0.38	0.37	0.39

Since all the decision tree model shows high accuracy on training set and relatively low accuracy on test set, overfitting problem require to be dealt with. Moreover, models with an additional single potential variable indicates more differences between train set accuracy and test set accuracy than standard model. Precision is another criterion to consider. This is because, during the business problem, major consideration is expense of pointless calls to individuals and to reduce these cost, thus it is must to highly regard about false positive in confusion matrices and precision score. Standard model (five variable model) shows highest precision and f1-score which is blend of precision and recall. As a result, standard model with variable *'default', 'loan', 'housing', 'duration'* and *'poutcome'* is the optimal model and these five variables is defining as 'x' data frame, input features.

The result of Decision Trees after applying RandomizedSearchCV to find optimal hyper - parameter is below. For major feature of optimal Decision Trees, 'Gini' is used for criterion, max depth is 76, max features are 'auto', minimum samples leaf is 8 and minimum sample split is 10. Overfitting concern is diminished because the difference of train and test set accuracy is small and

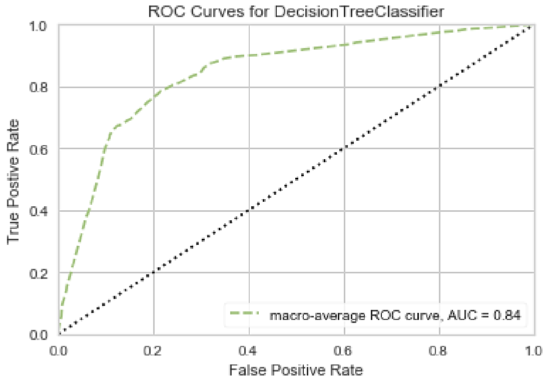


Train set accuracy	0.902
Test set accuracy	0.889
precision	0.56
Recall	0.17
F1-score	0.26

both accuracy is high in each set. Moreover, precision in this model shows 56 per cent. From confusion matrix, receiver operating characteristic (ROC) curve can be created.

Optimal Decision Trees

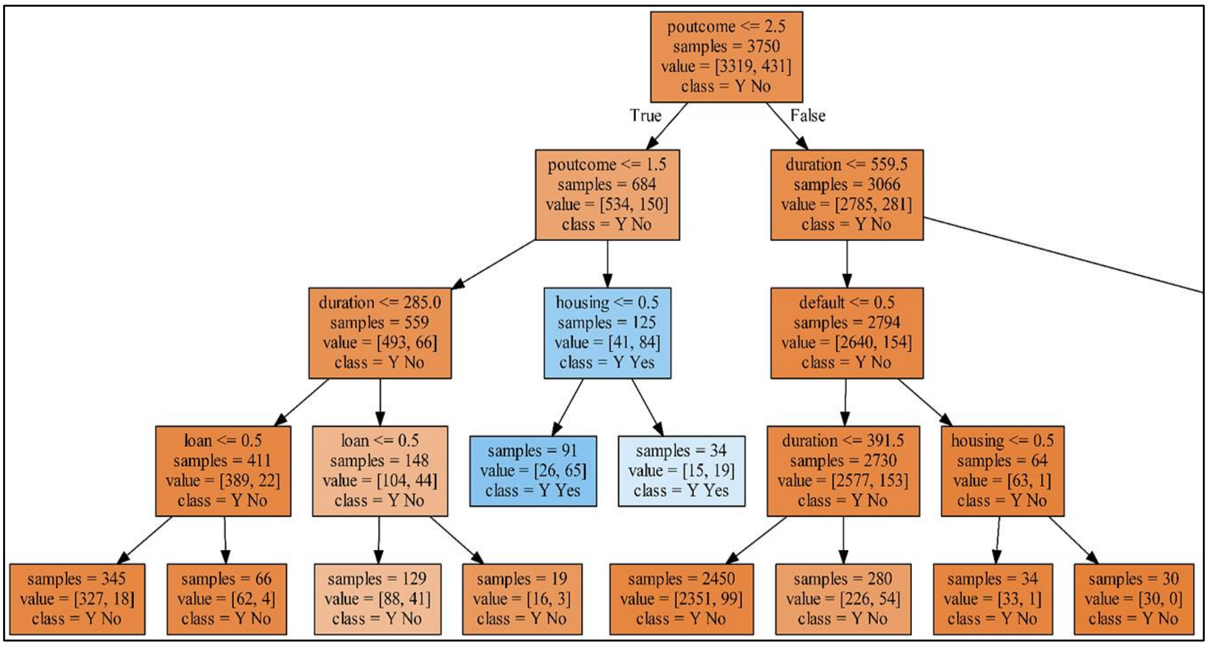
N = 1,250	Predict Y: YES	Predict Y: NO
Actual: Yes	24	120
Actual: No	19	1087



False positive rate (1 – Specificity) is on X axis and true positive rate (Recall) is on Y axis. This ROC curve shows a degree of round shape and area under the curve (AUC) is 0.84. However, the size of this model is too massive to transform to the graph, downsizing is demand.

Limiting max depth optimization parameter to five, other parameter also changed. Downsized decision trees' optimal hyper parameter is Gini for criterion, max depth is four, max features is 'sqrt', minimum sample leaf is seven and minimum samples split is two. This model also has less probability of overfitting issue and high accuracies in each data set. Moreover, the rate of precision is higher than previous model.

Train set accuracy	0.901
Test set accuracy	0.896
precision	0.60
Recall	0.28
F1-score	0.39

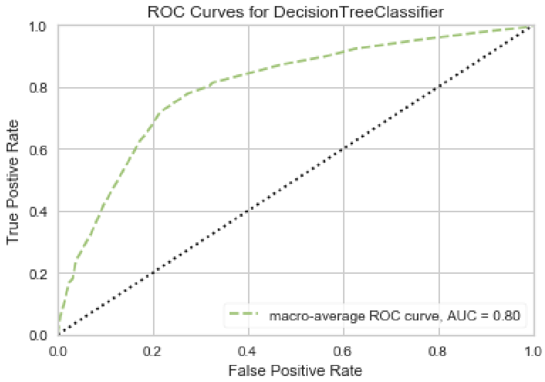


This is the two-thirds part of downsized optimal Decision Trees which has 14 leaf nodes. Root node is using variable *'poutcome'* as a first criterion. Very first internal nodes from the root node take variable *'poutcome'* and *'duration'* as a criterion. Lastly, 14 leaf nodes are divided as class Y as Yes or Y as No. Interpreting the two blue colour box in the graph, left one shows 91 samples in class Y as Yes and right one shows 34 samples in class Y as Yes. Does who categorised as 'success' in variable 'poutcome' and categorised as 'no', 'yes' or 'unknown' in variable 'housings' are classified as who is likely to purchased 'N/LAB Platinum Deposit'. In other words, no matter having house loan or not, who have experience of buying product on a previous campaign is predicted as potential buyer.

Based on confusion matrix in below, ROC curve and AUC can be illustrated.

Downsized Optimal Decision Trees

N = 1,250	Predict Y: YES	Predict Y: NO
Actual: Yes	41	103
Actual: No	27	1079



In this downsized optimal Decision Trees, ROC curve is not typically round compare to optimal decision tree above and AUC is lower as 0.80. However, when comparing precision score, downsized optimal Decision Trees shows more accurate performance, thus downsized optimal Decision Trees is better model than previous Decision Trees.

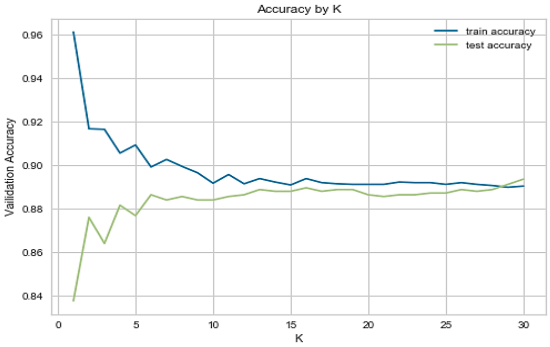
Section C: Model Evaluation

In section C, K-Nearest Neighbours (KNN), Logistic Regression and Random Forests classification is chosen to test their effectiveness in modelling with five variables; *'default', 'loan', 'housing', 'duration'* and *'poutcome'* against an output variable *'y'*. The main reason for chosen the same five parameters in each modelling is to compare each model's evaluation and performance in the same condition. Moreover, these five variables are decided from analysing 15 input features by statistical analysis and Decision Trees. From statistical analysis, five input features founded out as an important variables and another five variables as a potential variable. However, after analysing by Decision Trees, only five variables discovered as a significant variables considering the difference between train and test set accuracy, precision and f1-score.

To compare each model's performance, three estimations is going to use; 1) the difference of train and test set accuracy, 2) precision and 3) f1-score. The difference of train and test set accuracy provide the possibility of overfitting and how the model fits on the data. Secondly, precision is most important criterion, because the CEO emphasize of the expense of pointless calls. Thus, to reduce the cost, maximizing precision is demand. Lastly, f1-score is used to perceive the balance of precision and recall.

C-1. K-Nearest Neighbours (KNN)

KNN classification has great strengths of simple approach and intuitive understanding, as well as, flexible decision boundaries. Major purpose of using KNN modelling is to investigate without using any parameters and glance overall situation. To find out the optimal K in KNN, train and test set accuracy is tested by range of K between 1 to 30 and as K increases. The higher the K, the lower accuracy in train set and the higher accuracy in test set, however, after nearly K is 10, two dataset's accuracy becomes similar.



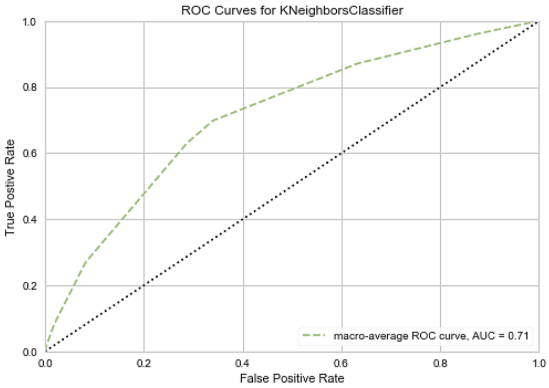
To find out the optimal KNN model, GridSearchCV is used and as a result, the best parameter is 'auto' in algorithm, leaf size is 1, weights are 'uniform' and K is four. After tuning

Train set accuracy	0.907
Test set accuracy	0.883
precision	0.47
Recall	0.12
F1-score	0.20

parameters, the difference of train and test set accuracy is pretty low that overfitting problem is out of concern, but precision is lower than Decision Trees and recall and f1-score is low. Based on confusion matrix, ROC curve can be illustrated.

Optimal K-Nearest Neighbours

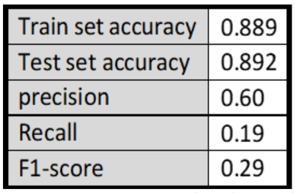
N = 1,250	Predict Y: YES	Predict Y: NO
Actual: Yes	18	126
Actual: No	20	1086



KNN's ROC curve is not rounded as Decision Trees and AUC score lower than Decision Trees. Therefore, downsized optimal Decision Trees is better model than KNN.

C-2. Logistic Regression

Next classification model is Logistic Regression which has a nice probabilistic interpretation and strength of regularizing to avoid overfitting. To find out optimal parameters, GridSearchCV is used. The optimal parameter C is 0.026… which is the inverse of regularization strength in Logistic Regression.

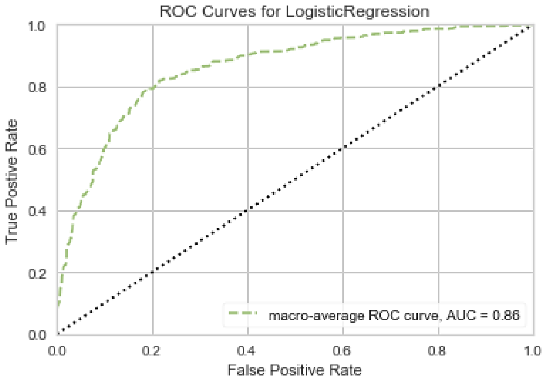


Train set accuracy	0.889
Test set accuracy	0.892
precision	0.60
Recall	0.19
F1-score	0.29

Both data set of accuracy is high and the difference gap of two score is nearly 0.003, however, test set accuracy is higher than train set which is significant error. Even though, precision is 0.60 modification is demand. Based on confusion matrix, ROC curve can be illustrated.

Optimal Logistic Regression

N = 1,250	Predict Y: YES	Predict Y: NO
Actual: Yes	27	117
Actual: No	18	1088

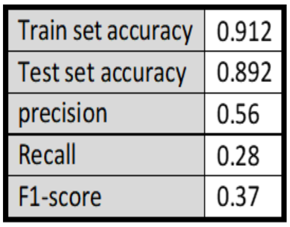


ROC curve shows is rounded shape and AUC score is 0.86 which is higher than Decision Trees and KNN.

C-3. Random Forests

Last classification model is Random Forests. Generally, Random Forests perform better than Decision Trees and it is powerful model when it comes to handling large data sets with higher dimension. After tuning hyper parameters using RandomizedSearchCV, the optimal parameter is 216 estimators, 'auto' in max features, 35 max depths, 18 minimum samples split and six minimum samples leaf.

The train set accuracy is very high in Random Forest and the number of difference with test set accuracy is only 0.020, thus the possibility of overfitting is low. Precision score is 0.56 in Random Forest which is



Train set accuracy	0.912
Test set accuracy	0.892
precision	0.56
Recall	0.28
F1-score	0.37

higher than K-Nearest Neighbours and Logistic Regression.

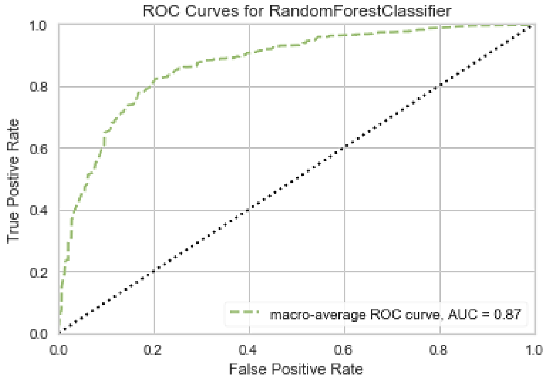
Variable *'duration'*	0.6720
Variable *'poutcome'*	0.2113
Variable *'housing'*	0.0927
Variable *'loan'*	0.0231
Variable *'default'*	0.0008

From optimal Random Forests, importance of each variable can be found. Variable *'duration'* indicates as most significant variable in input features and variable *'default'* shows least important input features in the optimal Random Forests.

Based on confusion matrix, ROC curve can be illustrated.

Optimal Random Forests

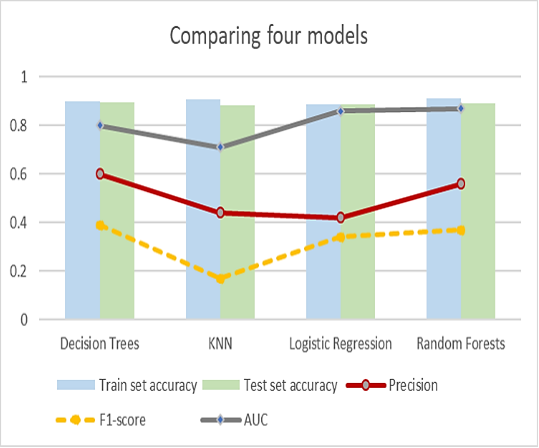
N = 1,250	Predict Y: YES	Predict Y: NO
Actual: Yes	41	103
Actual: No	32	1074



Random Forests' ROC curve shows rounded shape and AUC score is 0.87 which is higher than any other classifiers.

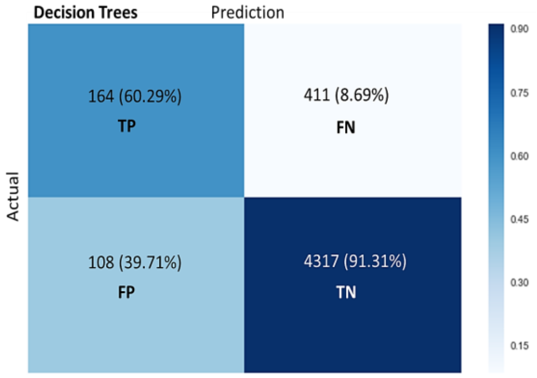
C-4. Comparing classification models

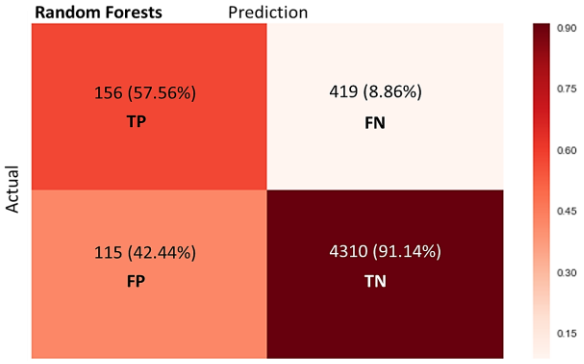
Classified historical dataset using Decision Trees, KNN, Logistic Regression and Random Forests. Since all the classifier model shows high degree of train set accuracy and test dataset accuracy, more detail comparison is demand to choose a 'winning classifier'.

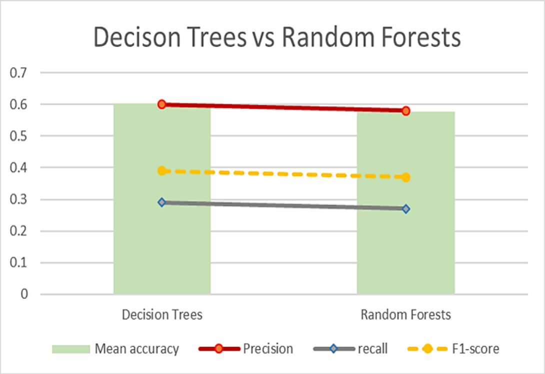


All of model shows high score of train and test dataset accuracy, which means that all the models have low possibility of having overfitting problem and having a high degree of fitting to historical data. When comparing precision score, downsized Decision Trees and Random Forests shows high performance.

To compare more specifically, folds and cross validation approach of creating dataset would be used instead of splitting. The result of each confusion matrices of heat map is below. Comparing FP rate, Decision Trees has lower rate than Random Forests.







As a result, both models shows mean accuracy score. However, Decision Trees' precision is higher than Random Forests. Therefore, a winning classifier would be Decision Trees.

Section D: Final Assessment

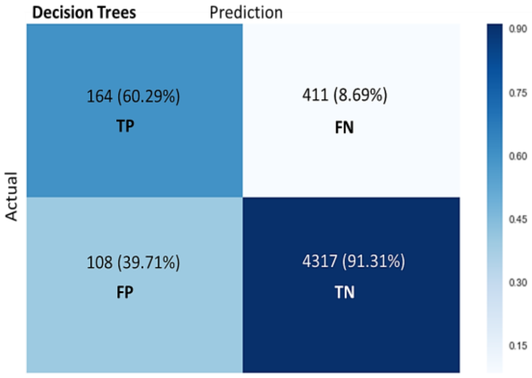
In previous section comparing four classification models; 1) Decision Trees, 2) K-Nearest Neighbours, 3) Logistic Regression and 4) Random Forest and decided that Random Forest was the winning classifier. Three performance strategies is used to comparing each models; 1) the difference of train and test set accuracy, 2) precision and 3) f1-score. Most importantly, precision score is significant standard to compare, this is because the CEO of N/LAB enterprises is focusing on the expense of fruitless calls to customers who are not interested in N/LAB Platinum Deposit. This cost could be minimised when precision score is high. The formula of precision is below:

True Positive / True Positive + False Positive

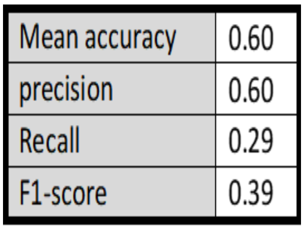
This measurement is used when false calls are costly and this is what the CEO is emphasised to diminish. Moreover, in generally, the lower the False Positive and False Negative value in confusion matrix is, the better model.

Section E: Model Implementation

In this section, folds and cross validation is used instead of creating train and test sets. After folds and cross validation, same parameter as downsized optimal Decision Trees in Section C is used on modelling process. These optimal parameters were discovered by using RandomSearchCV. The optimal parameter is 'gini' for criterion, max depth is four, max features is 'sqrt', minimum sample leaf is seven and minimum samples split is two.



164 (60.29%) TP	411 (8.69%) FN
108 (39.71%) FP	4317 (91.31%) TN



Mean accuracy	0.60
precision	0.60
Recall	0.29
F1-score	0.39

From above confusion matrix heat map, mean accuracy is 0.60 and precision is 0.60.

Final Decision Trees is ready to deployment. The recipient requires to follow six steps from loading the data set to evaluation. Brief instruction for the recipient is given below.

Step 1. Load in the new data

Step 2. Encoding categorical input features

Step 3. Setting features (X) and label (Y)

Step 4. Setting up folds and cross validation

Step 5. Model prediction

- Detail about the model: This Decision Trees used 'gini' for criterion, four max depths, 'sqrt' max features, seven minimum sample leaf and two minimum samples split. These hyper parameter is founded out by using RandomSearchCV.

- Result: After run the code, mean accuracy, confusion matrix, and classification report including precision, recall and f1-score will be present. Precision is the most significant score to understand the performance of this model, the higher is the better.

Step 6. Evaluation using the confusion matrix

Section F:
Business Case Recommendations

This business case is mainly about identifying customer target group who is willing to purchase 'N/LAB Platinum Deposit'. The CEO of N/LAB enterprises strongly mentioned that the expense of pointless calls to individuals must be avoided, thus the rate of False Positive rate in confusion matrix need to be minimize. The prior data is already validating enough to use in analysis because N/LAB enterprises have taken over banking operation and this company had an experience of launching a similar product before. The available dataset is composed of total 16 variables; 15 input features and one output variable. Through statistical analysis and Decision Trees, five variables, *'default', 'loan', 'housing', 'duration'* and *'poutcome',* discovered as significant variables which has a strong relationship with output variable *'y'*. Based on these five variables, four classification model including Decision Trees, K-Nearest Neighbours, Logistic Regression and Random Forests are assessed their effectiveness. As a result, Decision Trees became the most effective classification model. As training set accuracy scored 0.901 and test set accuracy is 0.896, the gap of difference is nearly 0.005, which indicates low possibility of overfitting issues. Moreover, precision rate is 0.60 which is the highest score of all models. With this model, N/LAB banking would figure out the target group who have are willing to invest on 'N/LAB Platinum Deposit'.

From the Decision Trees, there are two groups who has high possibility to purchased 'N/LAB Platinum Deposit'. One is who have experience of buying product on a previous campaign regardless of house loan statue. The other group is whose last contact duration is between 694 seconds and 806 second and categorised in 'success', 'failure' or 'unknown' in variable *'poutcome'*. To recapitulate, the company should select the target group who has experience of buying product on a previous campaign and customer whose last contact duration is between 694 seconds and 806 second.