FBA Coursework (20549904)

Analysis and Model Development for Targeted Marketing of N/LAB Platinum Deposit in the Banking Sector

As N/LAB Enterprises enter the banking sector with their "N/LAB Platinum Deposit," we have formulated a strategy to market the new financial product. We have used the historical data from a similar product to create a model predicting which people will likely be interested in subscribing to the product. The goal is to avoid calling too many people who won't be interested, saving time and money. The Chief Data Officer wants us to be cost-effective and focus on potential customers. We have used three different models- Naïve Bayes, Decision Tree, and Random Forest classifiers for this purpose. Out of them, Random Forest has proved to be the best. The report ends with practical suggestions based on the model's findings to help N/LAB Enterprises reach interested customers without wasting resources.

A. Summarization:

Dataset information: The dataset comprises 4,000 records from a phone call-based marketing campaign, featuring 16 columns of information. There are 15 input features which include 7 numeric and 8 categorical variables.

Numeric Variables	Age, balance, day, duration, campaign, pdays, previous
Categorical Variables	Job, marital, education, default, housing, loan, contact, poutcome

There is one output variable 'y'. To enhance clarity, we have opted to rename it as **'subscribed'**.

There are no missing values in the dataset.

Exploratory data Analysis:

1. Age: The mean age is 41.116 and all the values range between 18 to 92. The majority of customers who have subscribed to the new product are between the ages of 30 to 50. The median age here is 39.5 years.
2. Balance: The mean balance is 1342.14 and the majority of people have a balance of less than 5,000.
3. Duration: Mean duration is 297.71 seconds and the duration data is highly right skewed.

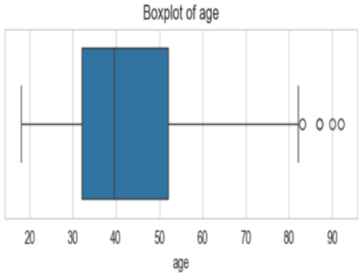


Fig 1.

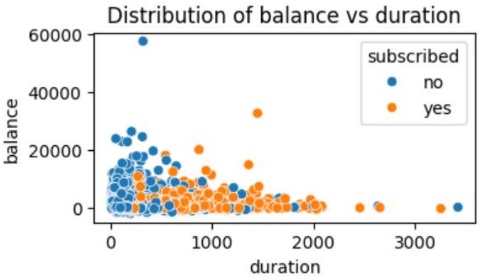


Fig 2.

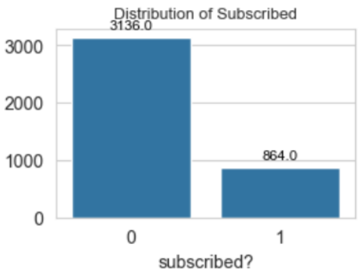


Fig 3.

4. Job: There are 12 different job values. The people who work in management are the ones who are most likely to subscribe to the new product.
5. Marital: It is evident from Fig.4 'marital', that a divorced person is very less likely to subscribe to the product as compared to married and single.

6. Education: The majority of subscribers possess either a secondary or tertiary level of education. Approximately 6% of individuals who subscribed have undisclosed information regarding their educational background.
7. Default: Approximately 98.8% of those who opted for a subscription do not exhibit any defaults. Conversely, the remaining 1% who subscribed despite having defaults have an average age of approximately 35. The interactions with this 1% group have, on average, lasted for 611.4 seconds, surpassing the time spent with individuals who did not default and still subscribed, which amounted to 569.47 seconds.
8. Housing: Individuals without housing loan are more inclined to subscribe; however, about 37.6% of individuals with a housing loan have still opted for a subscription, maintaining a balance of 1610.62.
9. Loan: Approximately 91% of subscribers do not have a personal loan, while the remaining 9% who have a loan and still opted for a subscription maintain an average balance of 1035.15 pounds.

   Furthermore, 39 individuals who possess both housing and personal loans have chosen to subscribe. They maintain an average balance of £1259.05 and have an average age of 39.69 years. Notably, the duration of the calls made to these individuals with dual loans is the highest, reaching 743 seconds.
10. Poutcome: There are 4 labels here - failure, success, unknown, and other. The poutcome value for 80% of data in the dataset is labelled as ‘unknown’. Consequently, it can be concluded that the available data for this particular attribute is notably insufficient.
11. Cellular: There are 3 ways of contact available: cellular, telephone, and an unidentified ‘unknown’ method. Individuals contacted via cell phone exhibited the highest likelihood of subscribing to the product.

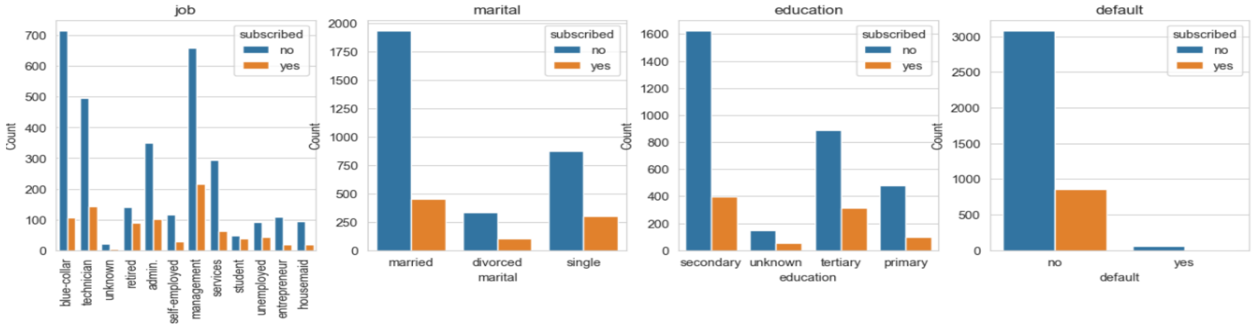


Fig 4.

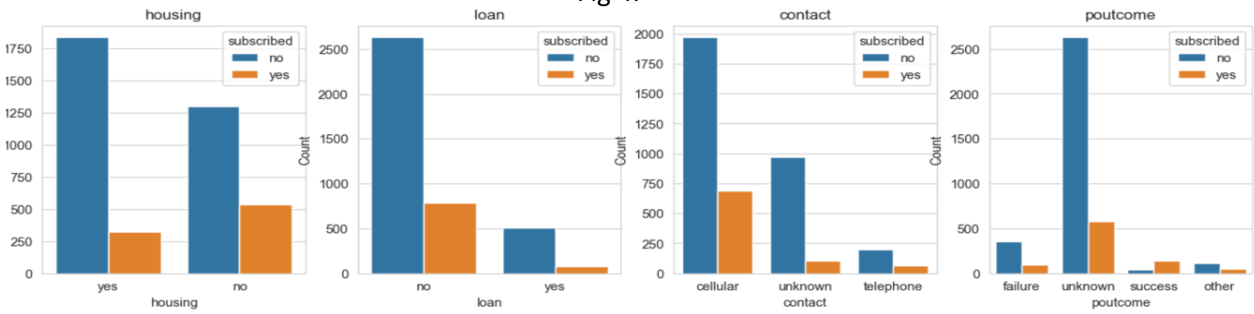


Fig 5.

As seen in Fig 3, the dataset is highly unbalanced with only 864 people subscribing to the new product while the remaining 3136 people do not.

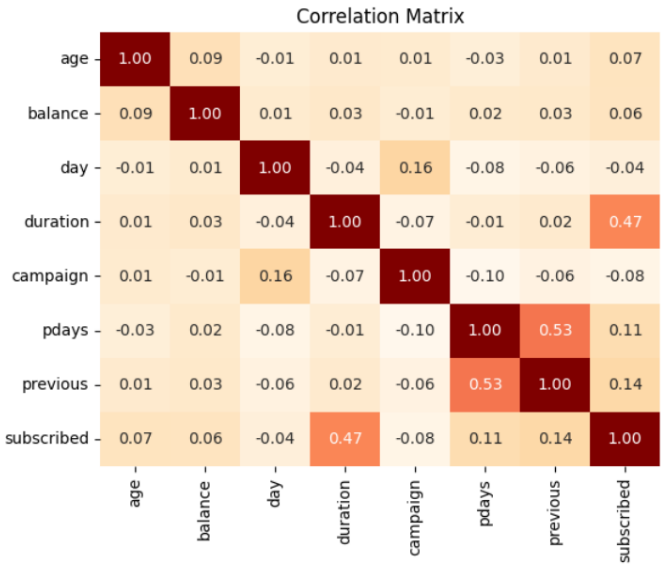


Fig 6

As per the correlation matrix, Fig 6, there is a high correlation between the output variable- ‘subscribed’ and ‘duration’ as compared to other variables.

B. Data Exploration:

Decision tree is a non-parametric supervised learning algorithm that is widely used for the classification of data. It provides an intuitive and qualitative ranking of features based on their contribution in reducing the uncertainty for the target variable.

On executing the decision tree algorithm (with max_depth=4, criterion=’entropy’ as hyperparameters) on the dataset, the below tree was formed:

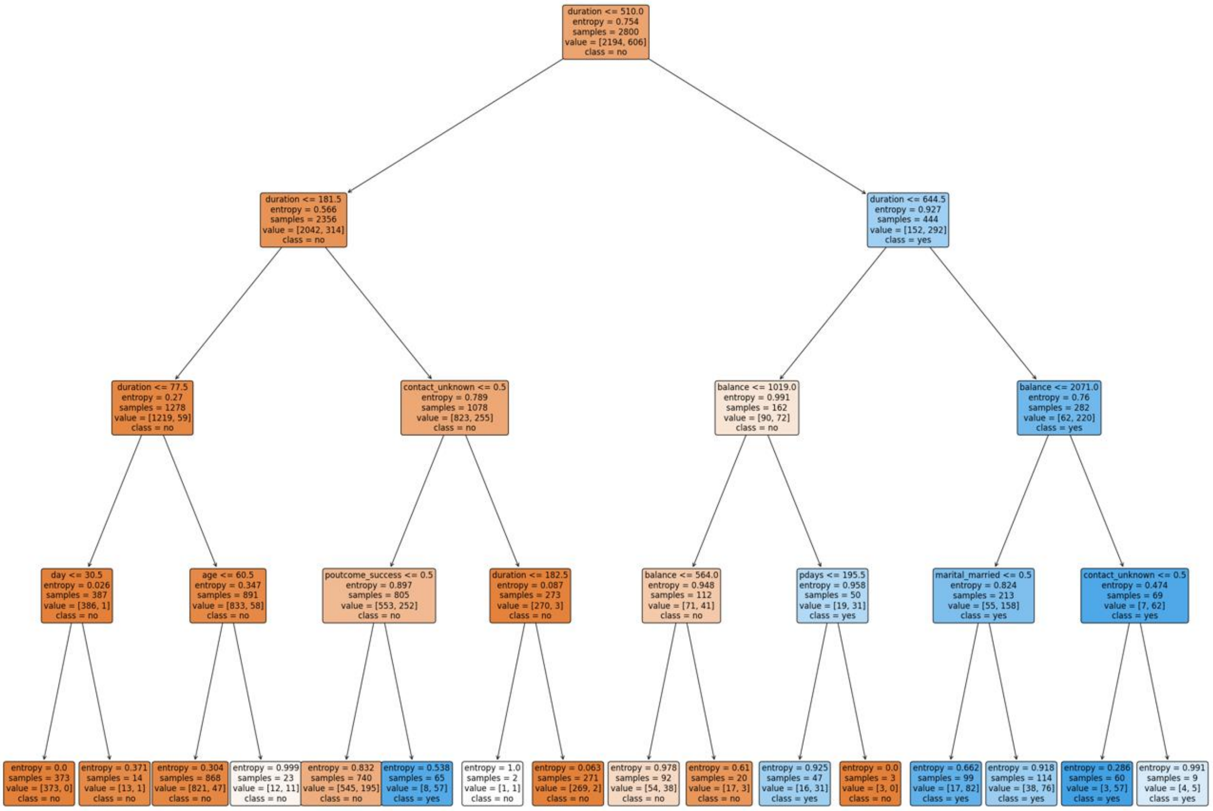

duration <= 510.0 entropy = 0.754 samples = 2800 value = [2194, 606] class = no
duration <= 181.5 entropy = 0.566 samples = 2356 value = [2042, 314] class = no
duration <= 644.5 entropy = 0.927 samples = 444 value = [152, 292] class = yes
duration <= 77.5 entropy = 0.27 samples = 1278 value = [1219, 59] class = no
contact_unknown <= 0.5 entropy = 0.789 samples = 1078 value = [823, 255] class = no
balance <= 1019.0 entropy = 0.991 samples = 162 value = [90, 72] class = no
balance <= 2071.0 entropy = 0.76 samples = 282 value = [62, 220] class = yes
day <= 30.5 entropy = 0.026 samples = 387 value = [386, 1] class = no
age <= 60.5 entropy = 0.347 samples = 891 value = [833, 58] class = no
poutcome_success <= 0.5 entropy = 0.897 samples = 805 value = [553, 252] class = no
duration <= 182.5 entropy = 0.087 samples = 273 value = [270, 3] class = no
balance <= 564.0 entropy = 0.948 samples = 112 value = [71, 41] class = no
pdays <= 195.5 entropy = 0.958 samples = 50 value = [19, 31] class = yes
marital_married <= 0.5 entropy = 0.824 samples = 213 value = [55, 158] class = yes
contact_unknown <= 0.5 entropy = 0.474 samples = 69 value = [7, 62] class = yes
entropy = 0.0 samples = 373 value = [373, 0] class = no
entropy = 0.371 samples = 14 value = [13, 1] class = no
entropy = 0.304 samples = 868 value = [821, 47] class = no
entropy = 0.999 samples = 23 value = [12, 11] class = no
entropy = 0.832 samples = 740 value = [545, 195] class = no
entropy = 0.538 samples = 65 value = [8, 57] class = yes
entropy = 1.0 samples = 2 value = [1, 1] class = no
entropy = 0.063 samples = 271 value = [269, 2] class = no
entropy = 0.978 samples = 92 value = [54, 38] class = no
entropy = 0.61 samples = 20 value = [17, 3] class = no
entropy = 0.925 samples = 47 value = [16, 31] class = yes
entropy = 0.0 samples = 3 value = [3, 0] class = no
entropy = 0.662 samples = 99 value = [17, 82] class = yes
entropy = 0.918 samples = 114 value = [38, 76] class = yes
entropy = 0.286 samples = 60 value = [3, 57] class = yes
entropy = 0.991 samples = 9 value = [4, 5] class = yes

Fig 7

The decision tree splits by choosing the optimal feature, resulting in nodes that are either pure or have lower impurity. Assessing the significance of features helps in identifying the input variables with the most substantial influence on the final decision. Additionally, it reveals redundant or less informative features that can be eliminated to enhance model efficiency. This process is essential for analyzing the impact of different input variables on the target variable.

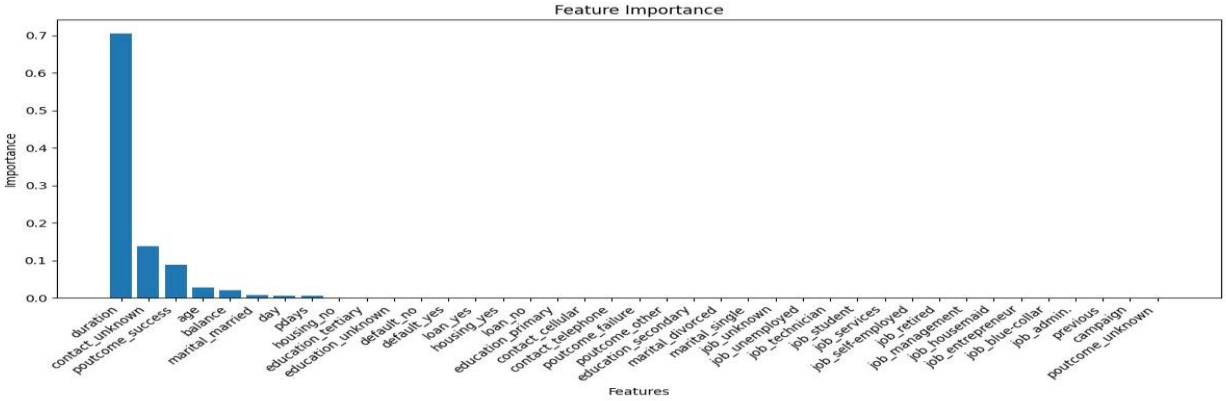


Fig 8

Feature analysis:

1. Top Features:
   - "Duration" is identified as the most critical feature, with the highest score of 0.706.
   - "Contact_unknown" and "poutcome_success" also hold notable importance, with scores of 0.139 and 0.088, respectively.
   - "age" , "balance", "marital_married", "day" and "pdays" have non-zero scores, indicating their influence on the model. The customer's age, balance, and marital status influence their possibility of subscribing to the new product.
2. Less Relevant features:
   Other features, such as job categories, previous interactions, and demographic factors, receive scores of 0, indicating minimal impact on the decision-making process according to the decision tree.

Examining the key features (fig 8), namely 'duration' and 'Contact_unknown,' we can deduce that client communication holds the most significant influence on the outcome. In this context, 'duration' refers to the time spent on a call with the client during the last contact in seconds, while 'Contact_unknown' indicates the method of communication, such as cellular, telephone, or unknown.

Other variables associated with communication-related information include 'previous' and 'pdays.' It is known that a 'pdays' value of '-1' indicates that the client was never contacted. Consequently, when a client was never contacted, the 'previous' contacts value becomes zero, and the influence of communication type ('contact_unknown') becomes more prominent in predicting the outcome as the duration value as well by default becomes '0'.

C. Model Evaluation:

Three classification models, namely Naïve Bayes, Decision Tree, and Random Forest, were evaluated for their efficacy in modeling our historical training dataset in comparison to a point predictor benchmark determined using the dummy classifier.

1. Dummy Classifier: [Accuracy: 78%]

Performance of Point Predictor (Dummy Classifier):

Subscribed?	precision	recall	f1-score
No	0.78	1.00	0.88
Yes	0.00	0.00	0.00

Since the dataset is unbalanced, the baseline classifier tends to predict only the majority class (label ‘No’) to improve accuracy leading precision, recall, and f1-score values to be 0 here. Thus, it is a biased and highly misleading classifier.

2. Naïve Bayes Classifier: [Accuracy: 77.83%]

Naïve Bayes is a probabilistic classifier based on the Bayes theorem. It assumes independence of the input features. However, in our case, as the dataset is unbalanced, it can lead to biased predictions favouring the majority class. Class imbalance can impact the model's ability to detect patterns in the minority class, affecting its overall predictive performance.

The model is trained using the default parameters.

Performance of Naive Bayes Classifier:

Subscribed?	precision	recall	f1-score
No	0.89	0.82	0.85
Yes	0.49	0.61	0.54

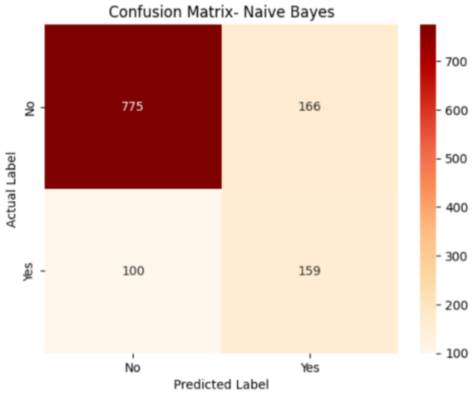


Fig 9

Here, the recall value of 82% for the label ‘No’ indicates that the classifier accurately identified 82% of people who opted to not subscribe to the product.

Additionally, for the label ‘Yes’, the classifier was able to correctly identify 61% of all people who subscribed to the product among all the people who genuinely subscribed.

3. Decision Tree Classifier: [Accuracy: 80.16%]

Decision Tree being is a supervised learning algorithm and it provides an interpretable representation of the decision-making process by identifying the features that influence the final decision. However, it is prone to overfitting. Thus, the hyper-parameters have been tuned to resolve this issue.

Using gridsearchcv, the best hyperparameters were selected. They were as:-

'criterion'	'entropy'	‘entropy’ is selected as criterion for measuring node impurity over ‘gini’.
'max_depth'	None	The max_depth being none means all nodes are expanded until we get pure nodes or until all leaves contain less than min_samples_split samples.
'min_samples_leaf'	1	Minimum no. of samples required to be a leaf node.
'min_samples_split'	2	Minimum no. of samples needed to split an internal node.
'splitter'	'random'	The strategy used to split at every node

Performance of Decision Tree Classifier:

Subscribed?	precision	recall	f1-score
No	0.87	0.88	0.87
Yes	0.54	0.52	0.53

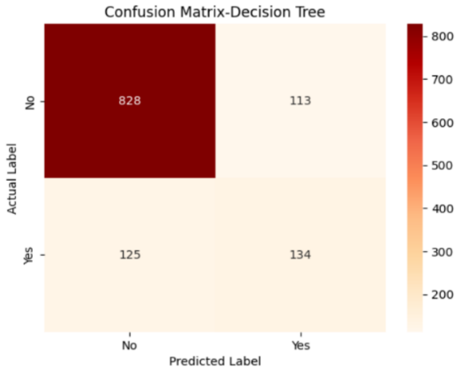


Fig 10

Here, the recall value for the label ‘No’ is 88% which is 6% more than that for Naïve Bayes classifier. This suggests that it is more efficient in correctly identifying the customers who opted not to subscribe to the product.

The recall value for the label ‘Yes’ is 52%, which is less than the Naïve Bayes classifier indicating that it correctly identified 52% of customers subscribing to the product among the genuine subscribers.

4. Random Forest: [Accuracy: 85.91%]

Random Forest is a collection of several decision trees. It resolves the challenge of overfitting by aggregating the predictions from every decision tree in the ensemble. Tuning the hyperparameters here helps to improve the performance of the classifier.

Using gridsearchcv, the best possible parameters for the model were identified. They are as follows:

'max_depth'	None	The max_depth being none means all nodes are expanded until we get pure nodes or until all leaves contain less than min_samples_split samples.
'max_features'	None	No. of features to consider for the best split
'min_samples_leaf'	2	Minimum no. of samples required to be a leaf node.
'min_samples_split'	10	Minimum no. of samples needed to split an internal node.
'n_estimators'	50	No. of trees in the forest

Performance of Random Forest Classifier:

Subscribed?	precision	recall	f1-score
No	0.89	0.93	0.91
Yes	0.71	0.59	0.64

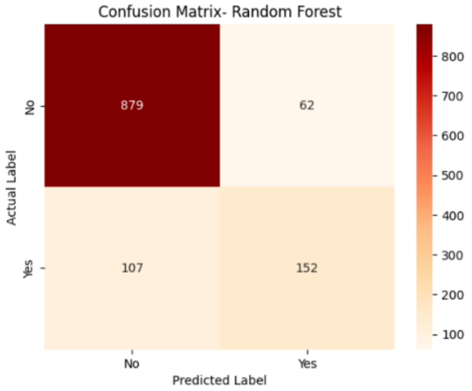


Fig 11

The recall value for the label ‘No’ here is 93% indicating that the Random forest classifier has been successful in identifying 93% of customers who genuinely opted not to subscribe to the product. The recall value here is the highest among all three classifiers.

Additionally, 59% recall for the label ‘Yes’ which is close to the one for naïve Bayes for the same label implies that the model was 59% successful in recognizing actual customers.

Identifying the best success measure to evaluate performance:

Using the confusion matrix, 4 performance measures were calculated- Accuracy, Precision, Recall, and F1-score.
Accuracy: It evaluates the overall correctness of the model. However, as our dataset is highly unbalanced, using accuracy can be deceptive.

Precision: It focuses on the accuracy of correctly predicting the target class.

Recall: It is the proportion of actual positive instances that the model identifies correctly. A higher recall value indicates a better ability of the model to minimize the false negatives. This scenario perfectly fits our requirement as per CDO's message of minimizing the business cost that may be incurred due to fruitless calls to individuals who are not at all interested in subscribing to the product. Therefore, we will use the recall performance metric.

F1-score: It is a harmonic mean of precision and recall values.

D. Final Assessment:

Model Accuracy comparison:

Random Forest (85.91%) > Decision Tree (80.16%) > Naïve Bayes (77.83%)

According to the confusion matrix, the recall value associated with the label "No," signifies customers who chose not to subscribe. It reflects the model's ability to correctly identify individuals who decided not to subscribe among the entire set of customers who refrained from subscribing to the product. This is most important for minimizing false negatives. Thus, a higher recall value is needed for this scenario.

Recall value comparison: **Random Forest (93%) > Decision Tree (88%) > Naïve Bayes (82%)**

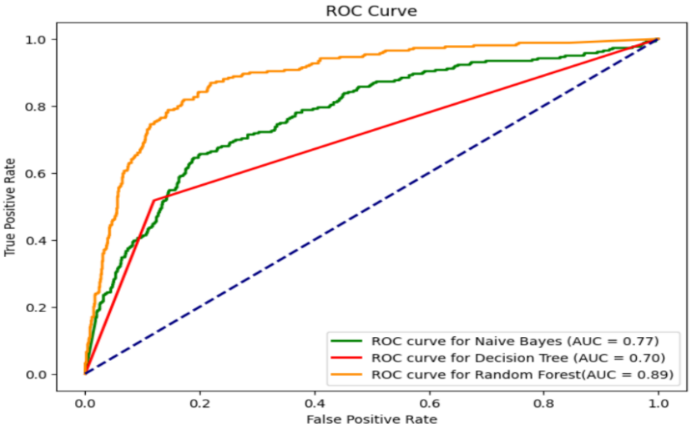


Fig 12

Fig 12, illustrates the ROC-AU curve plotted for the 3 classifiers. As per the area under the curve (AUC) value for all 3 classifiers, Random Forest still emerges as the best choice.

Therefore, considering all the analyses for the classifiers, it can be said that Random Forest is the best classifier for our dataset.

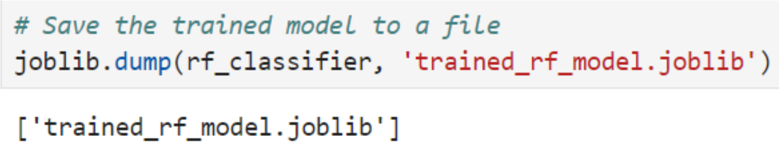
E. Model Implementation:

As Random Forest has emerged as our best classifier we are going to use it to predict whether the N/Lab customer will subscribe to the product or not. The classifier is trained using the entire historical dataset received. The hyperparameters used during training the classifier are the best parameters selected through gridsearchcv for the Random Forest model during model evaluation.

There are two files for file model execution:

1.Model training: Final_model_training.ipynb

Using this file, we will train the random forest model. Execute the series of code blocks and the trained model will be saved. This will be done using the joblib library. Thus we don't have to train the model repeatedly for testing the data.

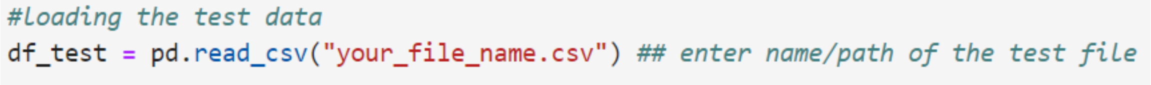


```
# Save the trained model to a file
joblib.dump(rf_classifier, 'trained_rf_model.joblib')
```

```
['trained_rf_model.joblib']
```

2.Testing the model: Model_Testing.ipynb

The trained_rf_model used for prediction is imported here using the joblib library. The test dataset can be loaded by adding the filename or file path as mentioned below:



```
#Loading the test data
df_test = pd.read_csv("your_file_name.csv") ## enter name/path of the test file
```

Note: The file should be in .csv format. Also, if the file path is as:

"C:\Users\ABC\OneDrive\Desktop\test_data1.csv", replace the backward slash with forward slash as "C:/Users/ABC/OneDrive/Desktop/test_data1.csv"

Later, execute the further steps to test the model accuracy, classification report, and confusion matrix.

Business Recommendations:

1. Focus on customer engagement duration:

The 'duration' is a pivotal feature for targeting customers. It is observed that longer durations are strongly correlated with the increased likelihood of product subscription. Thus, formulate strategies to extend customer engagement during phone calls for product marketing, emphasizing the potential benefits that customers stand to gain.

2. Use Cellphone Contact method:

It is observed that customers contacted on cell phones are more likely to subscribe. Hence, focus marketing efforts on this communication channel for contacting potential customers.

3. Target customers who have previously subscribed to such products:

'poutcome_success' is one of the important features that suggests that the customers who have subscribed to a similar product previously as per the dataset have a high chance of subscribing again. Thus, try targeting those customers.

4. Target the correct age group:

Target customers between the ages of 30 to 50. They are mostly likely to subscribe to the product.

5. Target married customers:

It is observed that customers who are married have their balance higher than the mean balance value and are more likely to subscribe. Focus on marketing messages tailored to resonate better with their preferences.