Summarization:

1.Data Summarization:

The dataset comprises 4000 rows and 16 columns, including 15 input features which consists of 7 numerical features (day, age, duration, balance, campaign, pdays, previous) and 8 categorical features (education, job, marital, default, contact, housing, loan, poutcome). The output variable, representing the sale of the product, is categorical with two options: yes/no. The given dataset provides comprehensive information about customer interactions during a phone-based marketing campaign, encompassing diverse numerical and categorical features. There are no missing values.

Key statistics highlight customer age, balance, and call duration, laying the groundwork for in-depth analysis and model implementation and evaluation.

Features	Mean	Minimum	Maximum
Age	40	18	85
Balance	1339.32	-2082	34646
Duration	287.6	5	2653

2.Exploratory Data Analysis:

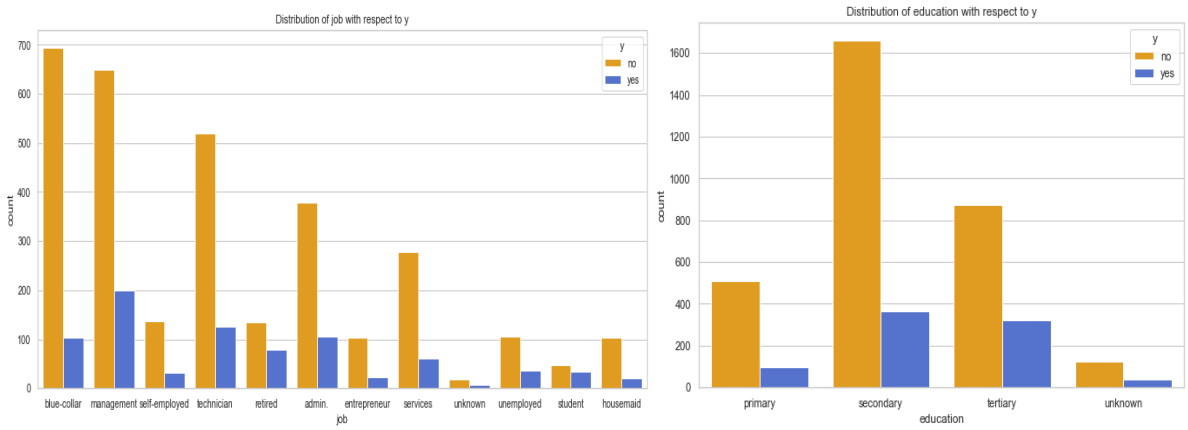


Fig.2.1

Distribution of job with respect to y: It can be observed from the above figure that certain occupational categories, such as individuals in management roles and technician, demonstrate a comparatively higher likelihood for subscription.

Education: It can be observed from Fig.2.1, the data distribution is notably skewed towards secondary education, resulting in a correspondingly high count of positive responses.

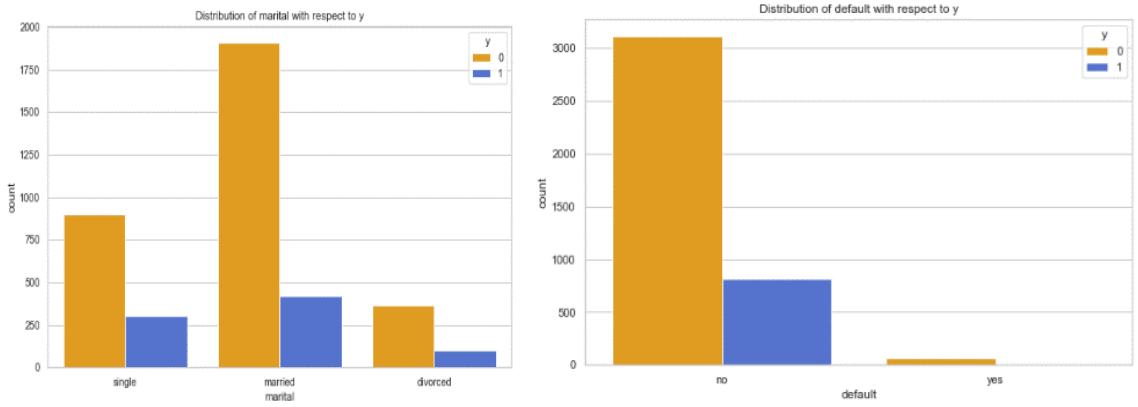


Fig.2.2

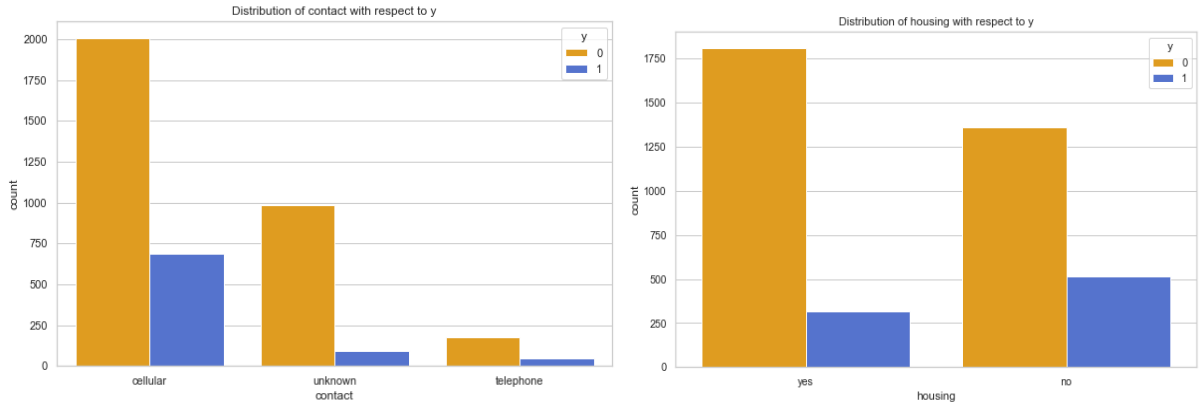


Fig.2.3

Distribution of marital status with respect to y: From the figure 2.2, it becomes evident that individuals who are divorced exhibit a notably lower occurrence of sales compared to those who are married or single. The married and single individuals are more likely to make a purchase.

Distribution of default and contact with respect to y: Fig2.2 and Fig.2.3 illustrates a significant trend where individuals not defaulting on credit payments have the highest count for depositing. The bar chart of contact communication types reveals a notable trend where a higher subscription rate is associated with the "cellular" communication category.

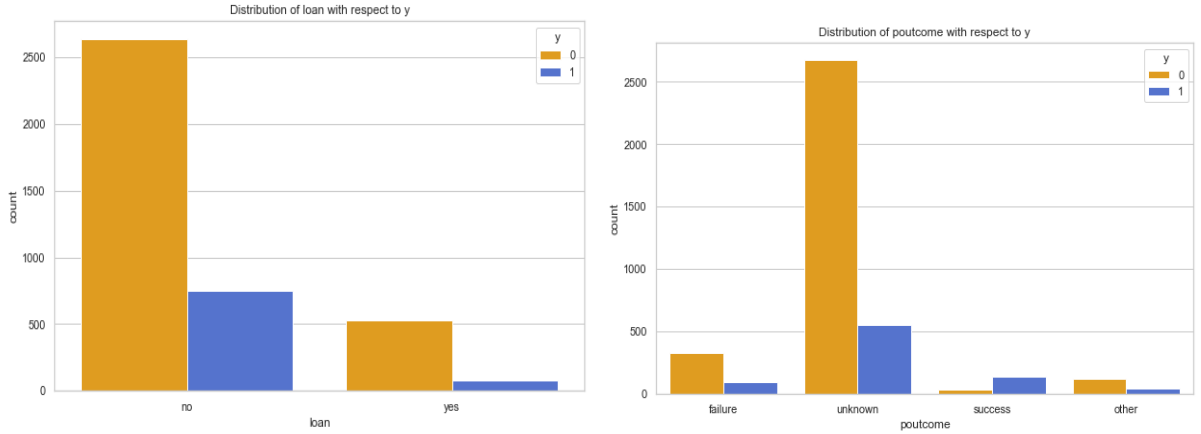


Fig.2.3

Housing and personal loan: Individuals without housing and personal loans exhibit the highest frequency of making deposits.

Distribution of poutcome with respect to y:It can be observed from the Fig2.3,the outcome of previous sales attempts i.e poutcome reveals that the majority of responses fall into the 'unknown' category, indicating a lack of information regarding previous campaign outcomes.

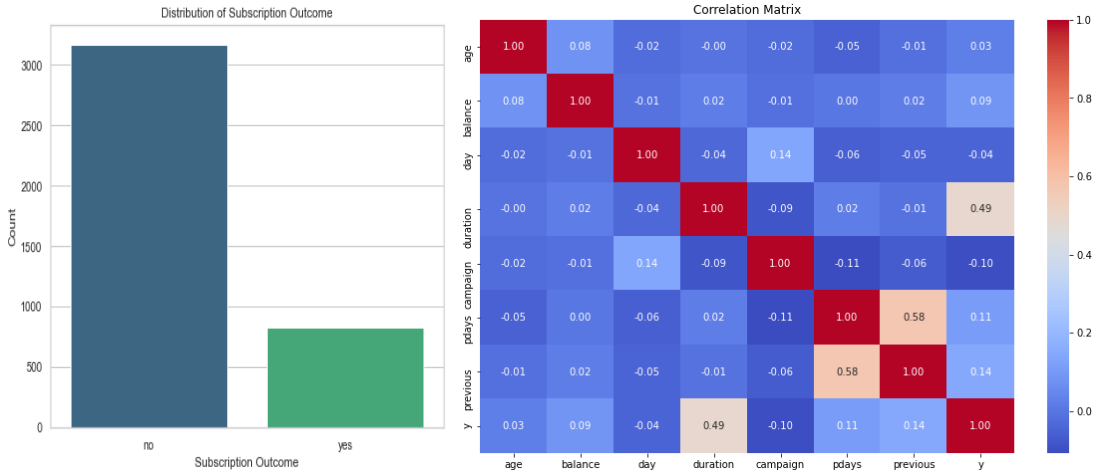


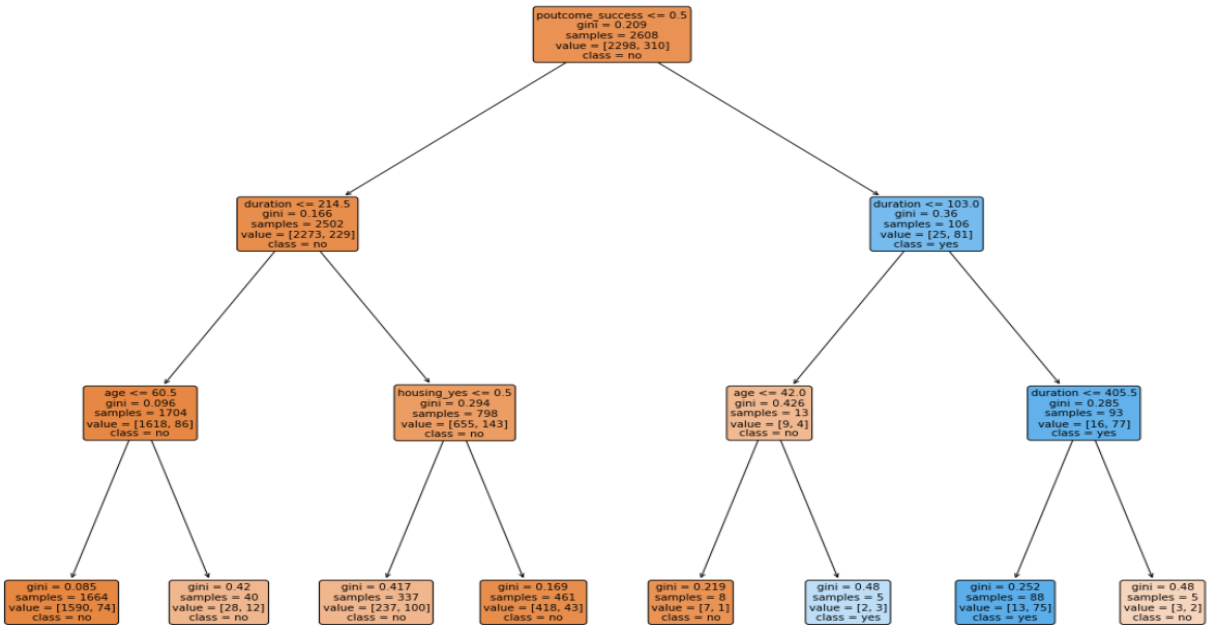
Fig.2.4

Subscription Outcome (output feature): The visual representation above clearly indicates a significant imbalance in the dataset. The category representing non-resulted sales comprises 3,172 instances, constituting 79.3% of the data. In contrast, the category indicating resulted sales has a count of 828, representing only 20.7% of the total 4,000 rows in the dataset.

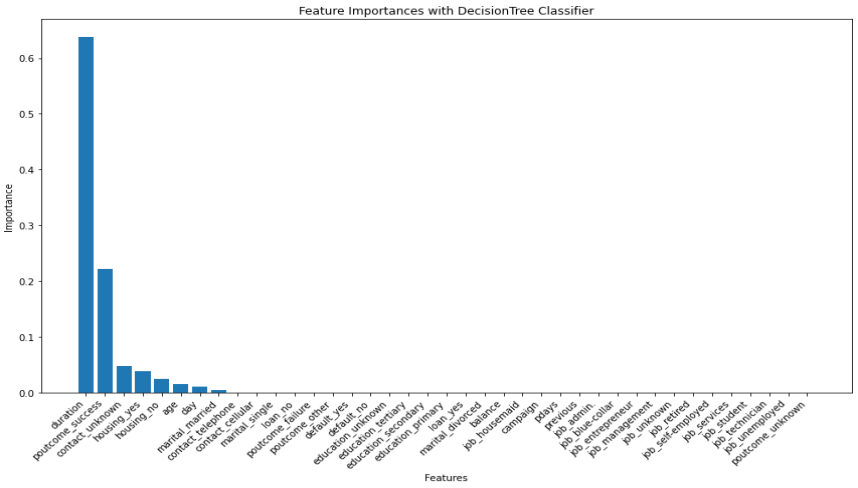
From Fig.2.4, The 'duration' variable represents the duration of the call, and it is highly correlated with the output variable 'y' (whether the call resulted in a sale or not).

Exploration:

Decision Trees are a supervised learning technique that can be used to classify data. They work by iteratively splitting the data set based on the input features, resulting in a hierarchical tree structure that aids decision-making.Decision trees prove invaluable for solving classification problems like identifying potential prospects who will buy this new product. The tree comprises of nodes representing decision points based on input features, branches denoting possible outcomes, and leaves indicating the final classification or decision.



The above figure is the application of decision tree on our dataset, it employs variable selection at each step to form subgroups that aim for maximum purity using the attributes.



The above figure depicts the most influential feature is 'duration,' with a high importance score of 0.638 followed by 'poutcome_success' and 'contact_unknown. This suggests that the duration of the marketing call plays a significant role in determining whether a customer will subscribe to the product. Feature importance in models involves ranking input features based on their significance in making predictions.The greater the magnitude of a characteristic's weight, the more necessary that feature is for identifying the target. And, if a feature's weight is close to zero, the associated feature can normally be disregarded or eliminated for improved model performance.These scores provide insights into the data and model, aiding in the identification of the most and least influential features. Understanding feature importance helps streamline predictive models by allowing the exclusion of less impactful features in the prediction process.

From our dataset it can been observed that duration is the feature which is most relevant to the target and the least important features are the job and poutcome_unknown(result of trying to sell the individual something on a previous campaign which is categorically defined as unknown). These insights can guide in emphasizing the importance of call duration and the outcome of previous marketing attempts in predicting customer subscriptions.

Model Evaluation:

The performance measures of the point predictor model are taken as the baseline measure. The three classification models used are Random Forests, Decision Trees and Naive Bayes Classifier.

1.Performance measures of the Point Model Predictor (Baseline Accuracy=0.79)

Output	Precision	Recall	F1-score
No	0.79	1.00	0.88
Yes	0.00	0.00	0.00

Due to the imbalanced nature of the dataset, the precision, recall, and F1-score values for the point model are all zero.

2.Performance measures of the Naive Bayes Classifier Model (Accuracy-0.78)

Output	Precision	Recall	F1-score
No	0.88	0.83	0.86
Yes	0.47	0.56	0.51

3.Performance measures of the Decision Tree Classifier Model (Accuracy-0.81)

Output	Precision	Recall	F1-score
No	0.87	0.90	0.89
Yes	0.56	0.49	0.52

4.Performance measures of the Random Forest Classifier Model (Accuracy-0.85)

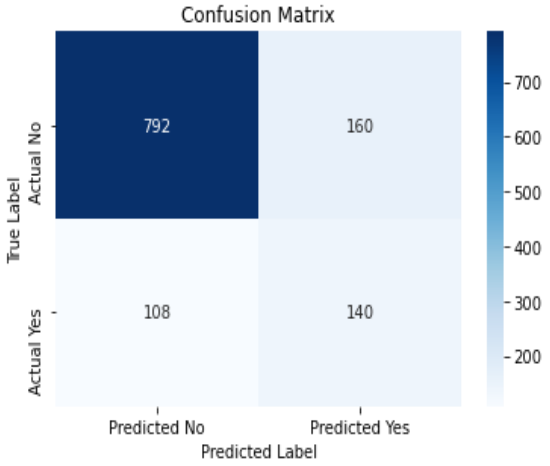
Output	Precision	Recall	F1-score
No	0.88	0.94	0.91
Yes	0.69	0.52	0.59

Naïve Bayes Classifier: It is "naive" in the sense that it ignores feature interactions and instead predicts the impact of each characteristic on the target separately. It is highly fast, efficient, and unexpectedly effective due to its simplicity. A probabilistic classifier based on the Bayes theorem is called the Naive Bayes classifier. Importing the model and using the Naïve bayes classifier is shown below:

from sklearn.naive_bayes import GaussianNB

nb_classifier = GaussianNB(var_smoothing=1e-9)

In this model the parameter used is with its default value.



The recall value for Naive Bayes Classifier is 0.83

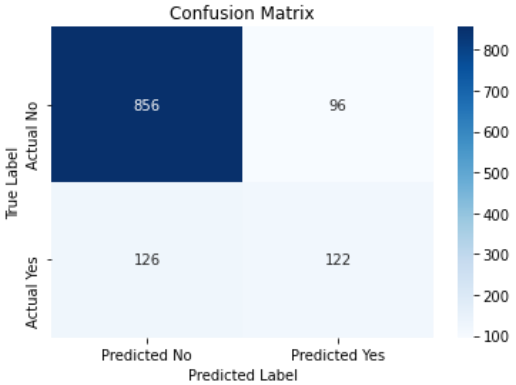
Decision Trees Classifier: A machine learning algorithm that is supervised and used for classification tasks, with the purpose of predicting instances' categorical class labels based on their feature values. It can handle both categorical and numerical data and requires less data processing during each iteration than other models such as Naive Bayes. In this classifier, the following parameterizations are used as shown below:

decision_tree = DecisionTreeClassifier(max_depth=20, min_samples_leaf=7,random_state=42)

The following parameters and their influence on decision trees:

max_depth=20 (the decision tree will stop when any of the leaf node has a depth of 20 or greater)

min_samples_leaf=7 (the minimum number of samples required to be present in a leaf node is 7 or greater)



The recall value for Decision tree Classifier is 0.90

Random Forest Classifier: During training, Random Forest builds numerous decision trees and combines their outputs to create a robust ensemble learning classifier that can also be used for supervised learning tasks such as classification and regression. Unlike individual decision trees prone to overfitting, Random Forest mitigates this risk by leveraging an ensemble approach, ensuring adaptability to new data. Notably, random forest works well with missing values, a valuable trait in real-world datasets often plagued by incomplete information. It builds several decision trees during training and the decision of the majority tree is taken as the output but unlike decision trees, the possible disadvantage is overfitting which can be solved by choosing an appropriate number of trees.
rf_clf = RandomForestClassifier()

mean_rf_cv_score = np.mean(cross_val_score(rf_clf, X_train, y_train, cv=5))

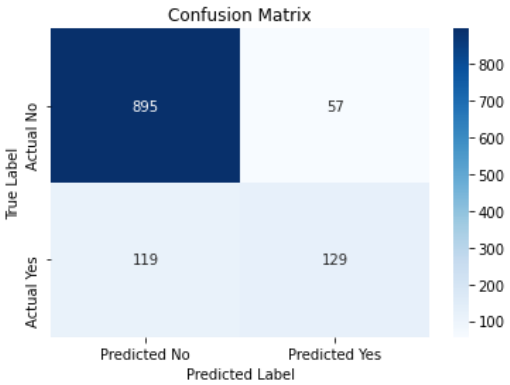
print(f"Mean Cross Validation Score for Random Forest Classifier: {mean_rf_cv_score :.2%}")

The following parameters are used in random forest:

rf_param_grid = { 'n_estimators': [10, 30, 100],'criterion': ['gini', 'entropy'],'max_depth': [None, 2, 6, 10], 'min_samples_split': [5, 10], 'min_samples_leaf': [3, 6]}

The below parameterizations are obtained by intitiating the GridSearch and finding optimal parameters for the classifier:

Optimal Parameters: {'criterion': 'gini', 'max_depth': None, 'min_samples_leaf': 3, 'min_samples_split ': 10, 'n_estimators': 100}
The recall value for Random Forests Classifier is 0.94



To assess and identify the most effective classifier model, various evaluation metrics are employed, selected based on the specific criteria relevant to the problem being addressed. Using the confusion matrix as the base of the evaluation model.

Observation from the confusion matrix: The confusion matrix's interpretation requires a thorough comprehension of its four major components: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). These elements serve as the foundation for calculating critical performance metrics.

1. Accuracy: This statistic assesses the model's overall accuracy, calculated as the sum of TP and TN divided by the total number of instances. It provides an overall evaluation of the model's predictive power.

2. Precision: It is the proportion of correct positive predictions. It is calculated as TP divided by the total of TP and FP. When minimizing false positives is a top goal, precision is necessary.

3. Recall: It calculates the proportion of true positive cases accurately classified by the model. It is computed by dividing TP by the total of TP and FN. In scenarios where minimizing false negatives is vital, such as in the N/LAB dataset which is to minimize the business costs for customers that are not interested in the product, achieving a higher recall value is crucial.

4.F1 Score: This balanced statistic takes into account both false positives and false negatives. It is calculated as the harmonic mean of precision and recall.

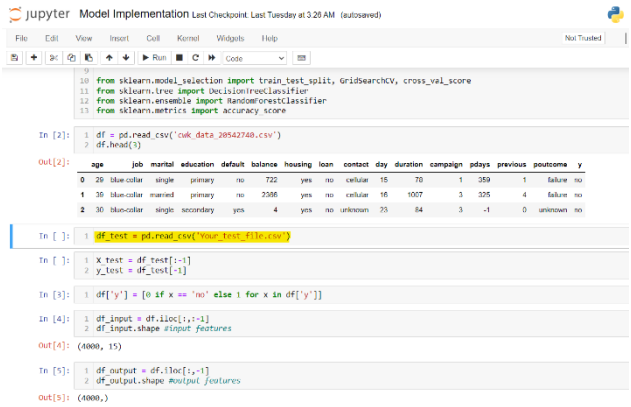
Final Assessment:

In the context of our business objectives, the classifier achieving the highest recall is the winning classifier.
The ranking of models based on their recall scores is as follows:
Random Forest Classifier > Decision Tree Classifier > Naive Bayes Classifier.

The Random Forest Classifier demonstrates the highest recall with a score of 0.94, followed by the Decision Tree Classifier with a recall of 0.90, and the Naive Bayes Classifier with a recall of 0.83. Consequently, the Random Forest Classifier, being the top performer in terms of recall, will be implemented as the preferred model.

Model Implementation:

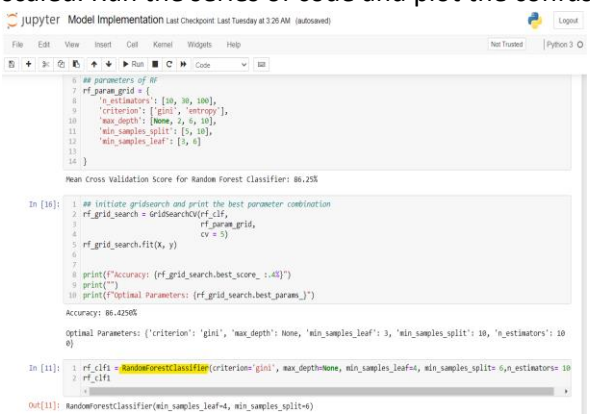
The best performance model that we use to predict the output label of N/LAB is Random forest classification model. This classifier must then be trained against the whole historical data set ready for deployment. The model will classify the test dataset into two class categories (YES/NO)



Firstly, we will input the test dataset:

df_test = pd.read_csv('Your_test_file.csv') ,where the file should be in '.csv' format.
Then we encode the data using both label encoder and one hot encoder and then we will scale the encoded data.

After this we will fit the Random Forest and whole historical data set we have that is encoded and scaled. Run the series of code and plot the confusion matrix as a heatmap.



Business Case Recommendations:

Insights from input features:

Duration: This feature is identified as highly important, emphasizing its role in predicting customer subscriptions. Duration of the marketing call is crucial. Longer durations positively correlate with a higher likelihood of subscription.

Job: It is observed that people in the management job category have the highest likelihood of subscription. Therefore, job category can also be focused on during the next campaign for better classification results.

Age: The analysis reveals a heightened probability of individuals making a deposit when contacted through marketing calls falls within the age bracket of 20 to 50. Consequently, a more concentrated effort should be directed towards engaging with this dynamic and responsive age group.

Marital: The focus should be more on married and single customers. They are more likely to make a purchase compared to divorced individuals, with more likeliness to deposit is from married customers.

Loan: Individuals without housing and personal loans have the highest likelihood of subscription.

Default: Target non-defaulters as they have a good credit history, as well as they exhibit a higher count for depositing.

Education: Prioritize customers with secondary education, as this category shows a higher count of positive responses.