**REPORT ON PREDICTING TERM DEPOSIT SUBSCRIPTIONS USING MACHINE LEARNING**

**-ANTONY JOSHUA A**

**INTRODUCTION**

A term deposit is a type of financial product offered by banks and other financial institutions. When a term deposit is opened, an individual agrees to deposit a certain amount of money for a fixed period, known as the "term."

**Data Set Explanation:**

The train dataset contains details such as age, job type, marital status, and call specifics, along with the target variable indicating subscription status. There are 31647 rows and 18 columns.

The test dataset contains details similar to train data but without the target variable. There are 13564 rows and 17 columns.

**METHODOLOGY**

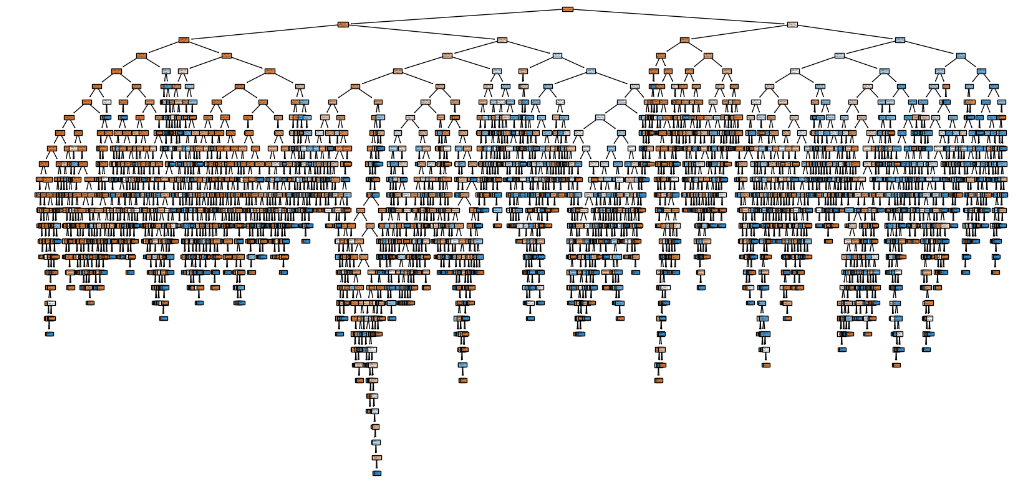
In order to provide insights on optimizing telephonic marketing campaigns by identifying clients most likely to subscribe to term deposits and also to reduce marketing costs by focusing resources on high-probability customers four different models were trained and build.

1. **BINARY LOGISTIC REGRESSION:** Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio- level independent variables. Dichotomous or dummy variables are usually coded 1, indicating "success" or "yes," and 0, indicating "failure" or "no." Here the target variable is the subscription status indicating “yes’” or “no”. Hence, we can use this model.
2. **DECISION TREE CLASSIFIER:** Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.

A decision node has two or more branches each representing values for the attribute tested.

* **Terminal node** represents a decision on the numerical target.
* The top most decision node in a tree which corresponds to the best predictor called **root node**. Decision trees can handle both categorical and numerical data.

For our data, since the dataset is big, the tree built is as follows:



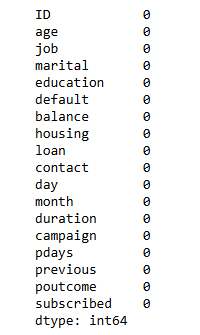
1. **RANDOM FOREST:** Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the "bagging" method. The general ides of the bagging method are that a combination of learning models increases the overall result. Random forest adds additional randomness to the model, while growing the trees, instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.
2. **GRADIENT BOOSTING:** Gradient Boosting is an ensemble method, meaning it combines multiple models (usually decision trees) to form a stronger predictor. The idea is that a collection of weak models can work together to create a robust model. The individual models in Gradient Boosting are often simple models known as "weak learners," typically decision trees with limited depth. These models alone might not be very accurate, but when combined, they produce strong predictive power.

**EVALUATION METRICS**

* **Accuracy:** The accuracy is the proportion of correctly classified cases. It can be calculated by adding the number of true positives (TP) and true negatives (TN) and dividing by the total number of cases (TP + TN + FP + FN).
* **Sensitivity/ Recall:** Sensitivity is the proportion of positive cases that were correctly identified (true positive rate). It is calculated by dividing the number of true positives (TP) by the total number of positive cases (TP + FN).
* **Specificity:** Specificity is the proportion of negative cases that were correctly identified (true negative rate). It is calculated by dividing the number of true negatives (TN) by the total number of negative cases (TN + FP)
* **F1-Score:** Harmonic mean of precision and recall, providing a single metric for model evaluation.
* **ROC curve:** An ROC curve is a visual tool that illustrates the performance of a binary classification model. It plots the true positive rate (TPR) on the y-axis against the false positive rate (FPR) on the x-axis.
* **Accuracy versus area under the ROC curve (AUC-ROC):** While accuracy gives a single value representing overall correctness, AUC-ROC provides an aggregate measure of performance across all classification thresholds. It measures the area under the ROC curve, indicating the model's ability to discriminate between positive and negative classes across all possible thresholds. A higher AUC-ROC value generally suggests better overall performance.

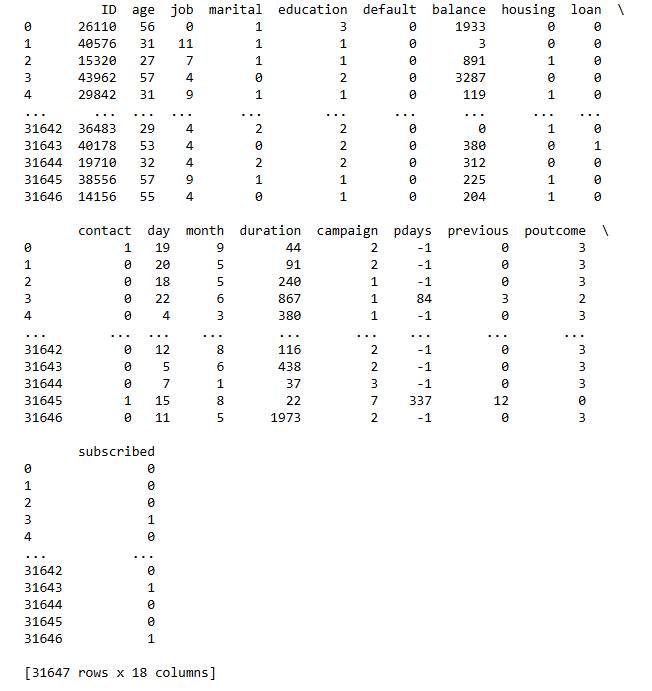
**MODEL PERFORMANCE**

**Data Preprocessing Script:** Using **PYTHON**, various libraries were imported. The train dataset is imported. The missing values were checked and there were no missing values.

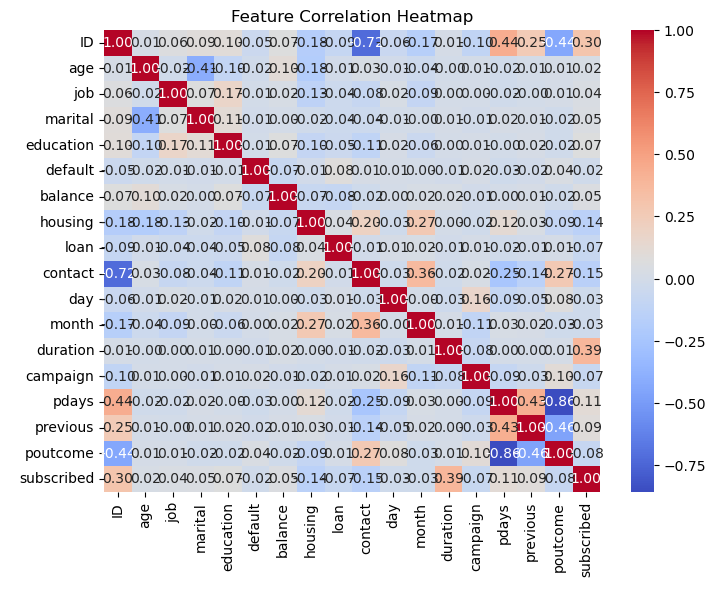


The datatypes were checked and there were categorical variables like job, Marital, education etc..

Using **Label Encoder**, the variables were transformed. Also, in order to make accurate predictions, **Standard scaler** was used to standardize the variables.

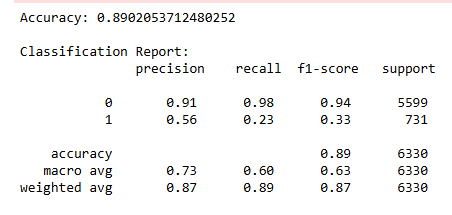


**Feature Engineering Script:** Feature importance is a technique used to identify and quantify the contribution of each feature in a dataset to the predictive performance of a machine learning model. By understanding which features are most influential, the most relevant data is found by potentially improving model accuracy. Using this in decision tree and random forest, to improve accuracy of the model. Heatmap is as follows:

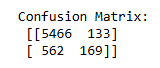


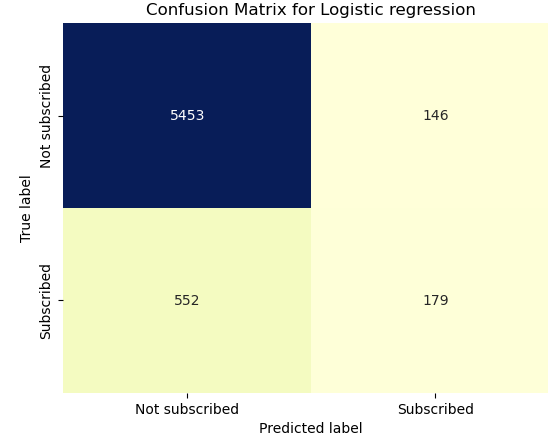
**Model Training Script:**

1. **Binary logistic regression**: The train dataset is split into train and test dataset in 80:20 ratio. The classification report is as follows:

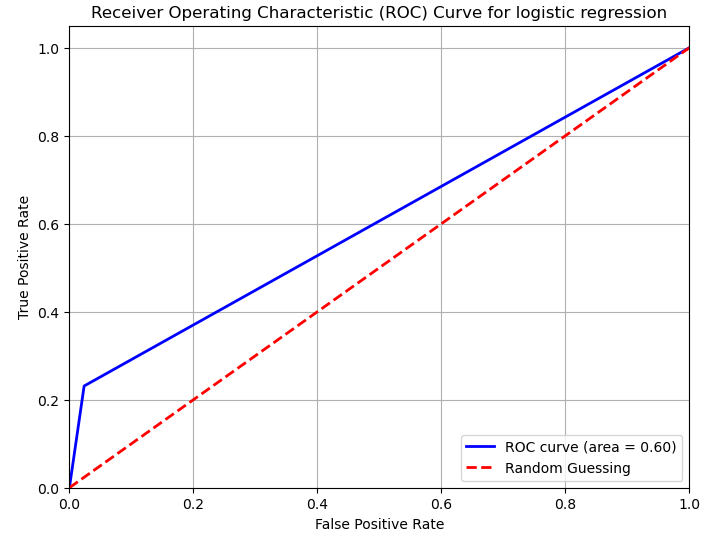


Confusion matrix:





**ROC:**



**Cross-validation:** Cross-validation is a crucial technique in machine learning and statistical modeling used to evaluate the performance of a model and to ensure that it generalizes well to unseen data. It involves dividing the data into subsets, training the model on some subsets, and testing it on others. This helps prevent overfitting and gives a more accurate estimate of the model’s performance.

Cross-validation scores: [0.88262243 0.88135861 0.88276189 0.88212988 0.88244588]

Mean cross-validation score: 0.8823

**Hyperparameter tuning:** Hyperparameter tuning is the process of finding the optimal set of hyperparameters for a machine learning model.

* **Grid search:** Grid Search exhaustively searches through a manually specified subset of the hyperparameter space.

Best Parameters: {'C': 0.01, 'penalty': 'l2', 'solver': 'saga'}

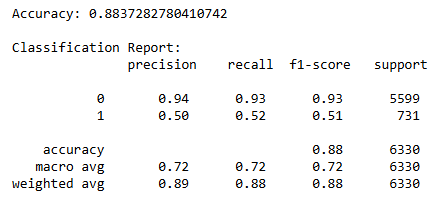
Best Score: 0.8820159903072785

* **Random search:** Random Search randomly samples combinations of hyperparameters from a specified distribution for a fixed number of iterations.

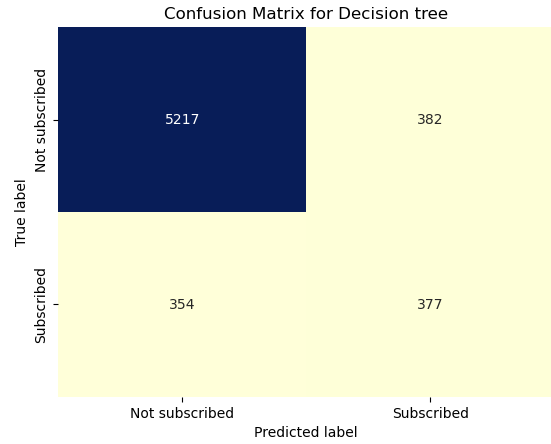
Best Parameters: {'C': 2.9926542674407717, 'penalty': 'l1', 'solver': 'liblinear'}

Best Score: 0.8989215349471852

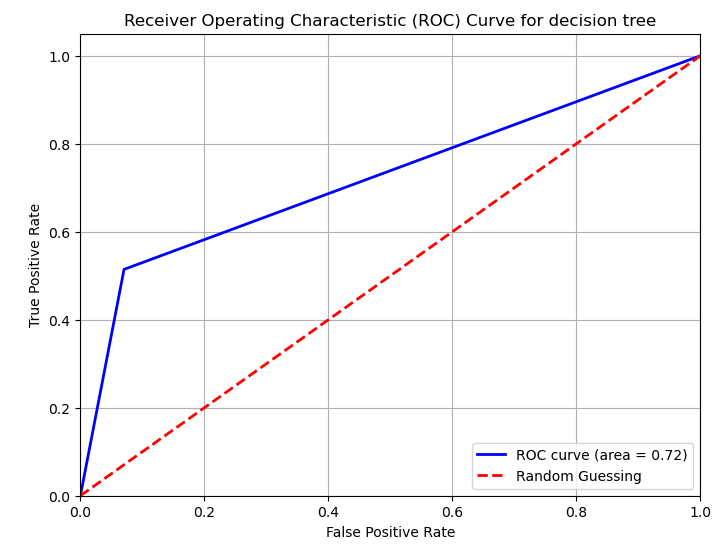
1. **Decision tree classifier:** The classification report is as follows:



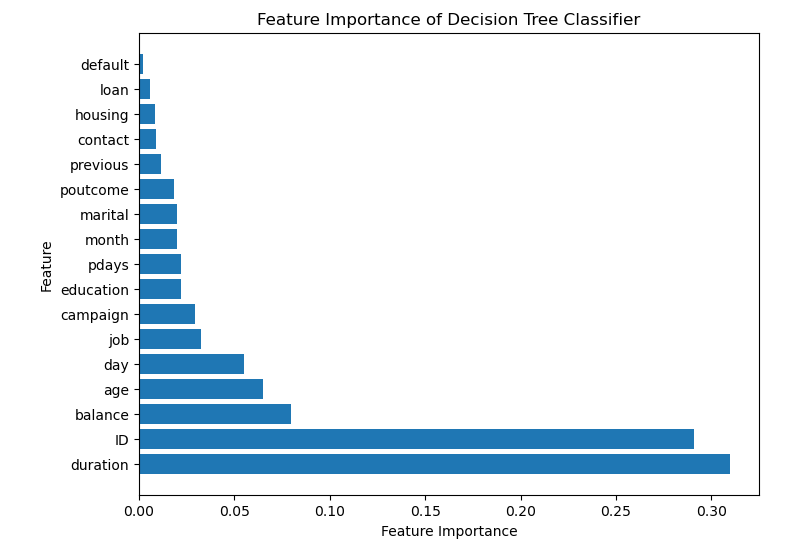
Confusion matrix:



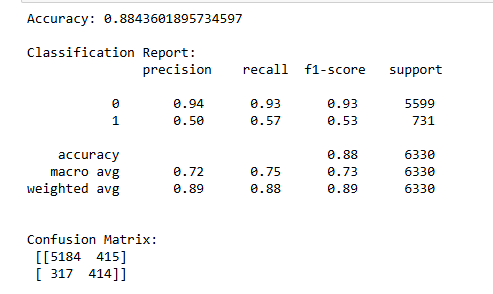
ROC:



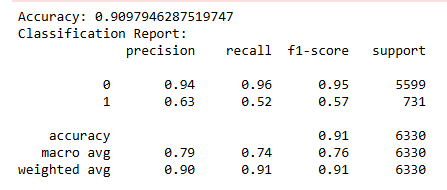
**Feature importance:** This shows that **duration, balance, age** are some important variables. So we take interaction terms 'duration\_campaign', 'job\_education', 'housing\_loan'. Also some aggregated statistics of 'age\_mean\_duration', 'poutcome\_sum\_previous'.



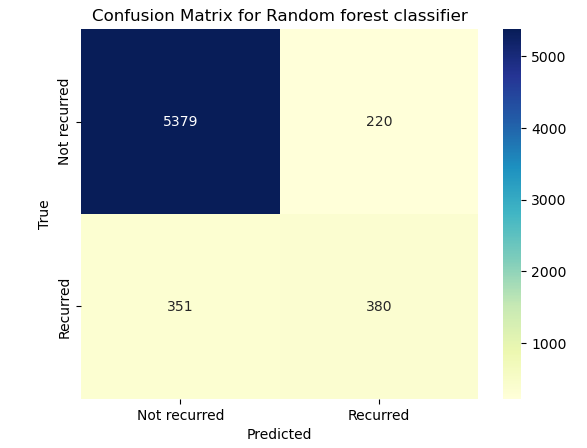
After adding interaction terms,



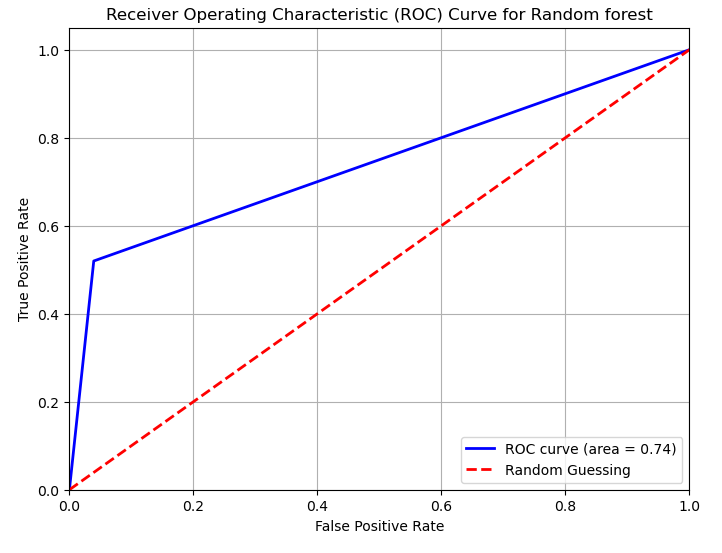
1. **Random forest:** The classification report is as follows:



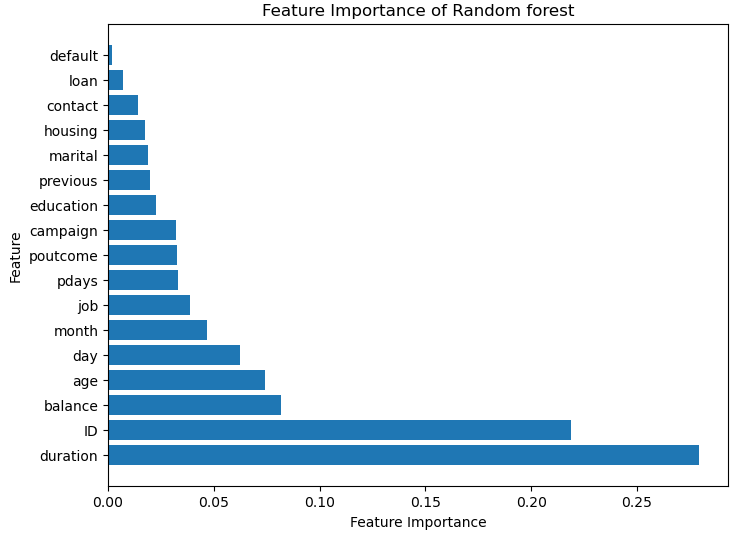
Confusion matrix:



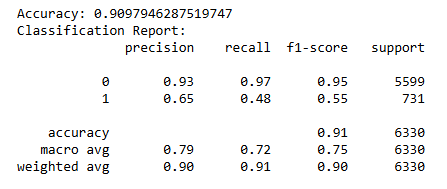
ROC:



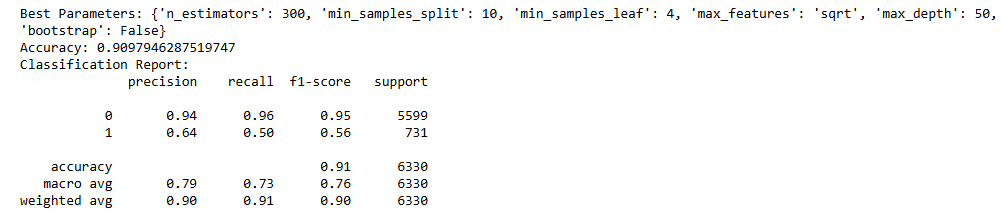
Feature importance shows the same duration as the most important variable. So, adding some variables like 'duration\_balance', 'job\_education', 'age\_mean\_duration' to the dataset.



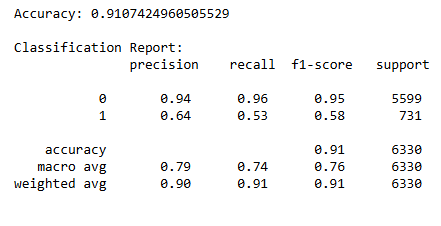
After adding those variables:



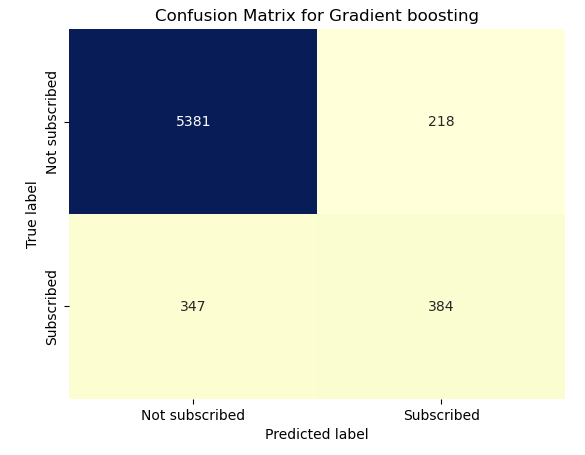
Random search hyperparameter tuning was imposed and gave results as:



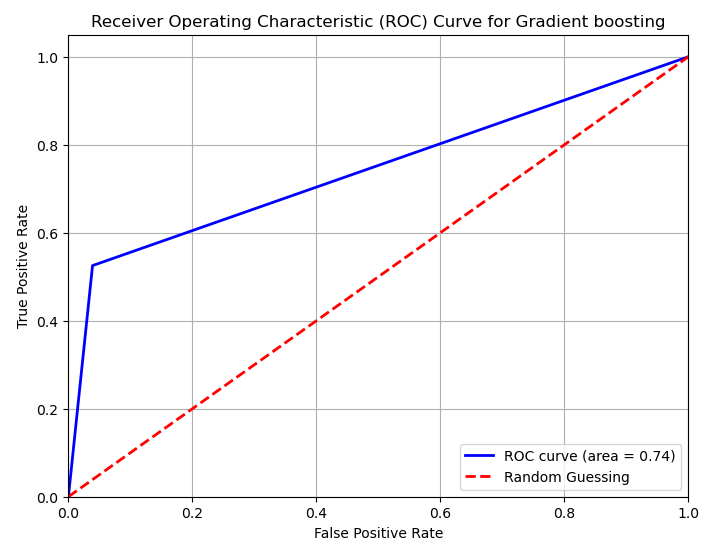
1. **Gradient boosting:** The classification report is as:



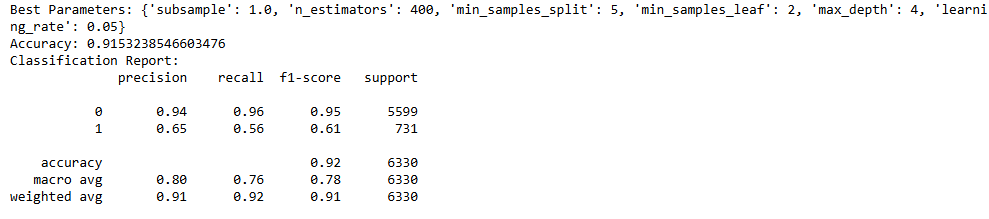
Confusion matrix:



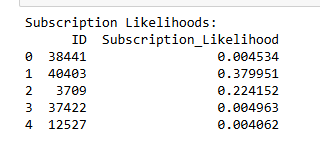
ROC:

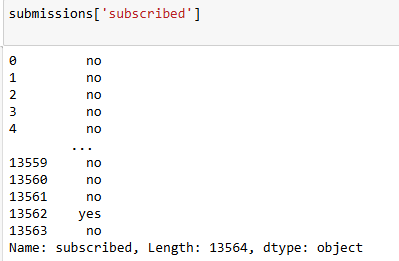


Random search:

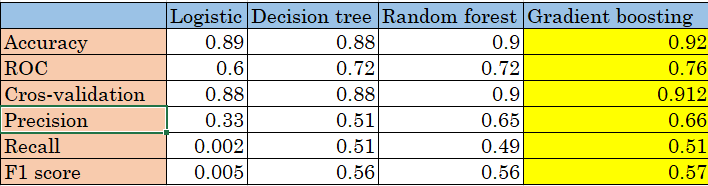


**TEST DATASET:** The same preprocessing was done on this dataset and for each ‘id’ the subscription and likelihood was predicted using the best model and saved as an excel file.





**VALIDATION OF BEST MODEL**



Overall, **Gradient Boosting** appears to be the best performing algorithm, with the highest scores in accuracy, cross-validation, precision, recall, and F1 score.

**EVALUATION REPORT:**

* **Accuracy** measures the overall correctness of the model, indicating the proportion of correctly classified instances. Gradient Boosting has the highest accuracy of 0.92, followed by Random Forest at 0.9, then Logistic Regression at 0.89, and Decision Tree at 0.88.
* **ROC (Receiver Operating Characteristic)** is a curve that shows the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity). The area under the ROC curve (AUC) is a good measure of the overall performance of the model. All algorithms have a similar ROC, with Gradient Boosting slightly better at 0.76, followed by Random Forest and Decision Tree at 0.72, and Logistic Regression at 0.6.
* **Cross-validation** measures the model's ability to generalize to unseen data by splitting the data into multiple folds and training the model on different combinations of folds. Gradient Boosting has the highest cross-validation score of 0.912, followed by Random Forest and Decision Tree at 0.9, and Logistic Regression at 0.88.
* **Precision** measures the proportion of correctly predicted positive instances out of all instances predicted as positive. Gradient Boosting has the highest precision of 0.66, followed by Random Forest at 0.65, then Decision Tree at 0.51, and Logistic Regression at 0.33.
* **Recall** measures the proportion of correctly predicted positive instances out of all actual positive instances. Gradient Boosting and Decision Tree have the highest recall of 0.51, followed by Random Forest at 0.49 and Logistic Regression at 0.002.
* **F1 score** is the harmonic mean of precision and recall, providing a balanced measure of the model's performance. Gradient Boosting has the highest F1 score of 0.57, followed by Decision Tree at 0.56, then Random Forest at 0.56, and Logistic Regression at 0.005.

**RESULTS:**

The best-performing model to predict the likelihood of new clients subscribing to term deposits is **Gradient boosting.**

The final model achieved an **accuracy of 91.2%** on the validation dataset.

Precision, Recall, and F1-Score values were 66%, 51% and 0.56 respectively, indicating the model's effectiveness in predicting term deposit subscriptions.

Gradient boosting is considered the best because it has the highest accuracy (0.92), highest cross-validation score (0.912) and a decent F1 score (0.57).

While it has a lower recall than decision tree, the other metrics make it a more robust model overall, indicating a good balance between precision and recall.

Hence, it is clear from definition of Gradient Boosting that it is an ensemble method, meaning it combines multiple models (usually decision trees) to form a stronger predictor and also it is a collection of weak models that can work together to create a robust model. This is the reason it gave a high accuracy.

**Thank You**