The Calling: Unveiling Telemarketing Insights for Bank Long-Term Deposits

A portfolio project created by Jonatas Vieira

Dataset and Business problem

- **Data Set Information:** The data pertains to the direct marketing efforts of a Portuguese bank. These campaigns relied on phone calls, sometimes necessitating multiple contacts with the same client to determine whether they would subscribe to the bank's term deposit product ('yes') or not ('no').
- **Problem Definition**: The goal is to use all this information to predict whether someone will end up saving money with the bank. This helps the bank decide which feature to focus to get more money from the customers.

Data Source

This dataset describes the results of Portugal bank marketing campaigns. The campaigns primarily involved direct phone calls, where clients were offered the opportunity to place a term deposit with the bank.

If a client agreed to place a deposit after all marketing efforts, the target variable is marked as 'yes'; otherwise, it is marked as 'no'.

Records

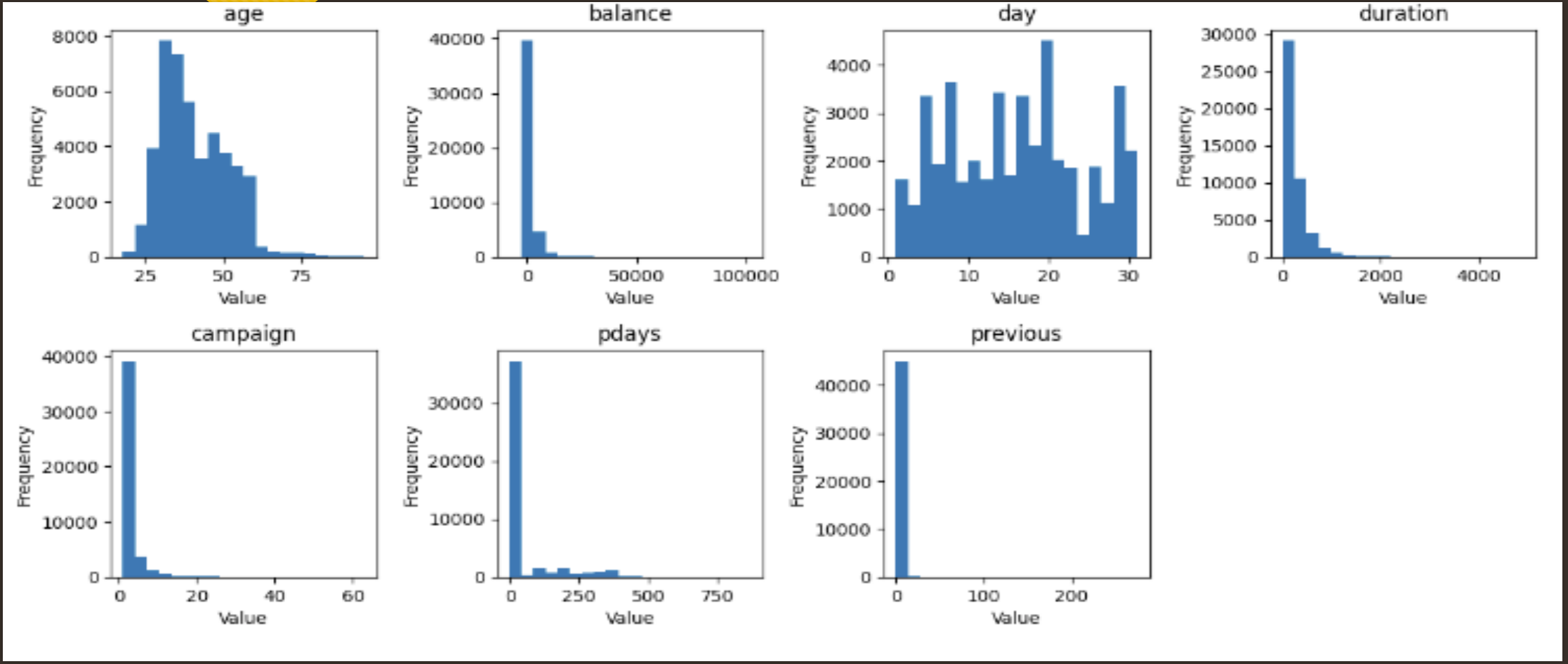
45211

Features

17

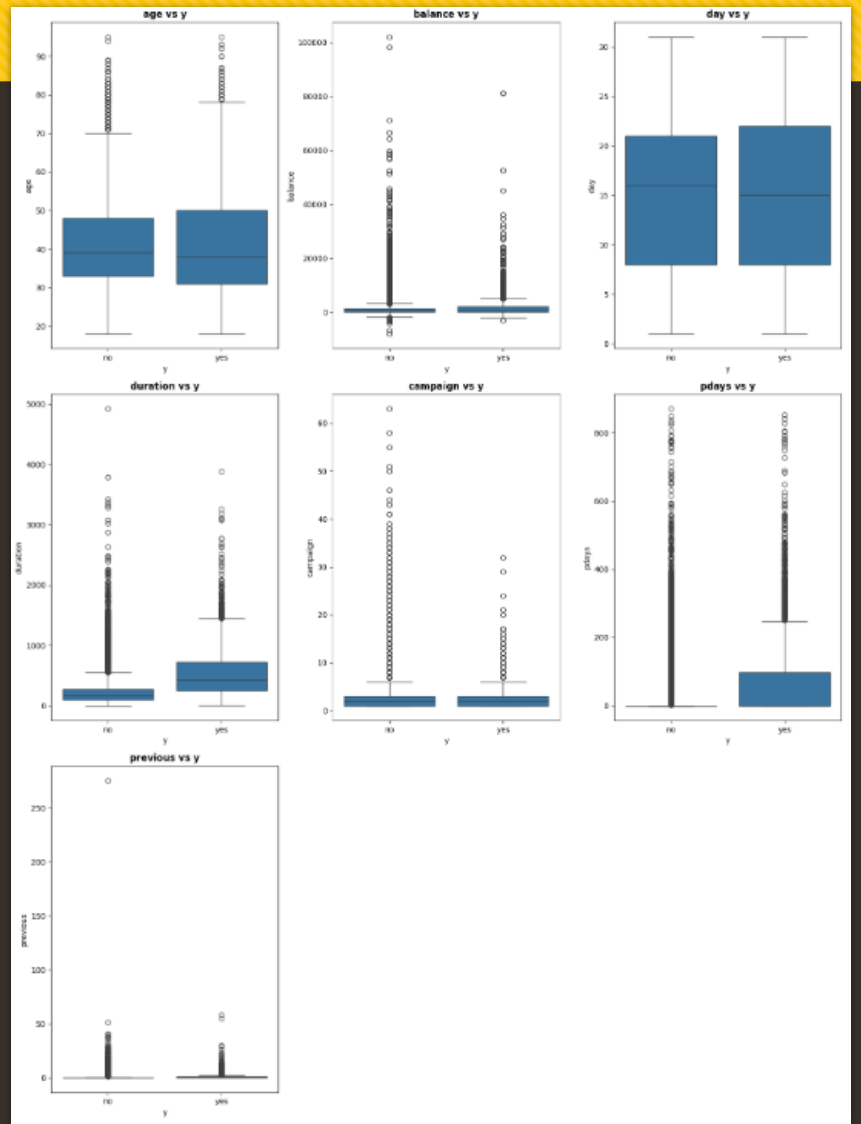
Exploratory Data Analysis

Univariate Analysis – Numerical Var.



Univariate Analysis – Categorical Var.

Bivariate Analysis – Numerical Var.



Conclusions:

1. The longer it takes to persuade a customer, the higher the likelihood of the customer subscribing to a term deposit.

2. The younger demographic (under the age of 50-60) has a greater likelihood of subscribing to a term deposit.

Bivariate Analysis – Categorical Var.

Conclusions:

1. Cellular communication tends to be more effective in convincing customers to subscribe to a term deposit.

2. In general, managers are significantly more inclined to subscribe to a term deposit compared to individuals in other fields.

3. Single customers show a greater propensity to utilize term deposits, even though their numbers are lower than those of married customers.

Bivariate Analysis – Categorical Var.

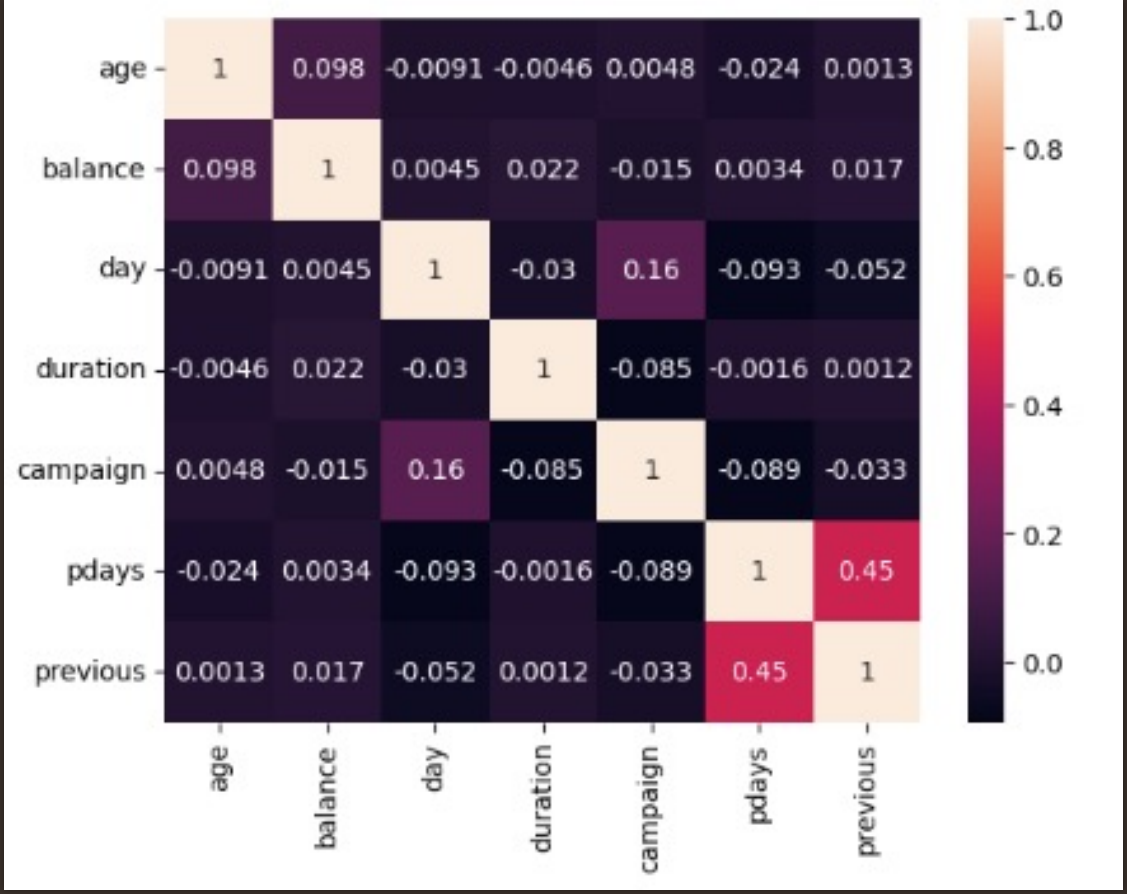
Conclusions:

4. Customers without housing loans are more likely to opt for a term deposit.

5. Customers with a secondary education are significantly more inclined to utilize a term deposit.

6. Customers who have defaulted on credit payments typically do not have a term deposit subscription.

Correlation Matrix



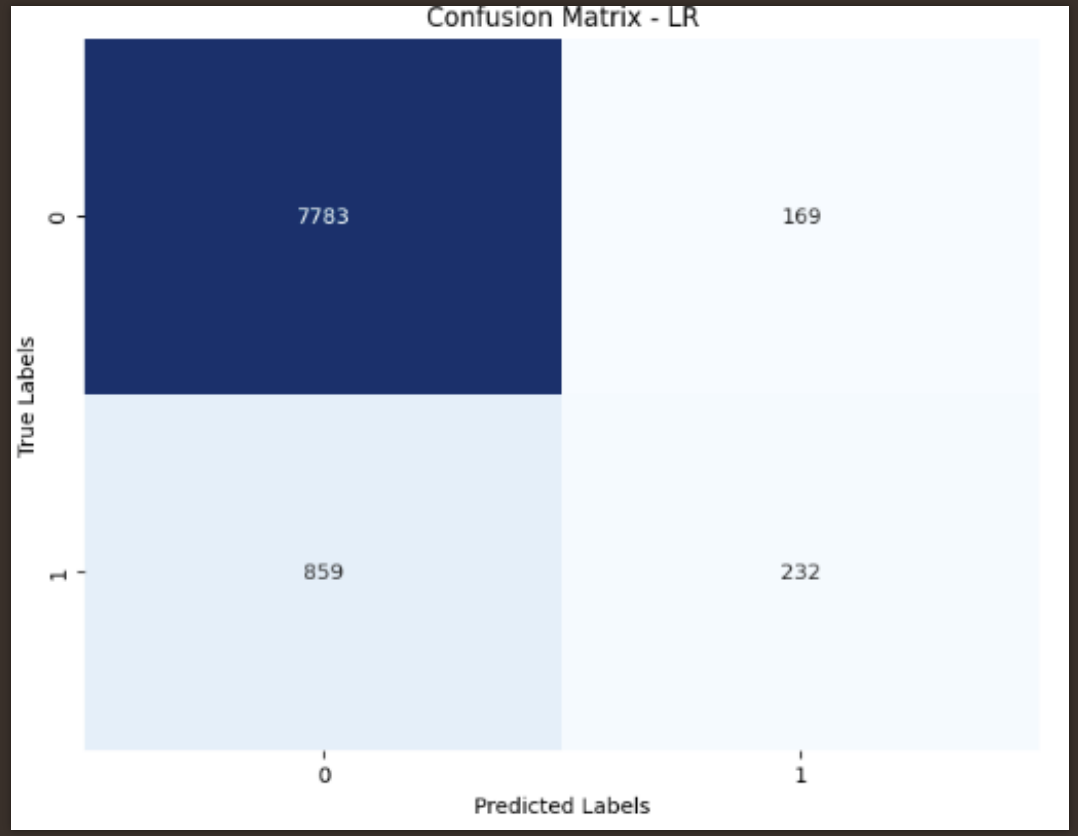
Conclusions:

Among numerical variables, there are no strong correlations, but *pdays* and *previous* show a moderate positive correlation of 0.45.

However, this relationship is not strong enough to indicate a direct dependency between these variables. This means that although there is some association between them, other factors may also influence their dynamics.

Machine Learning Models

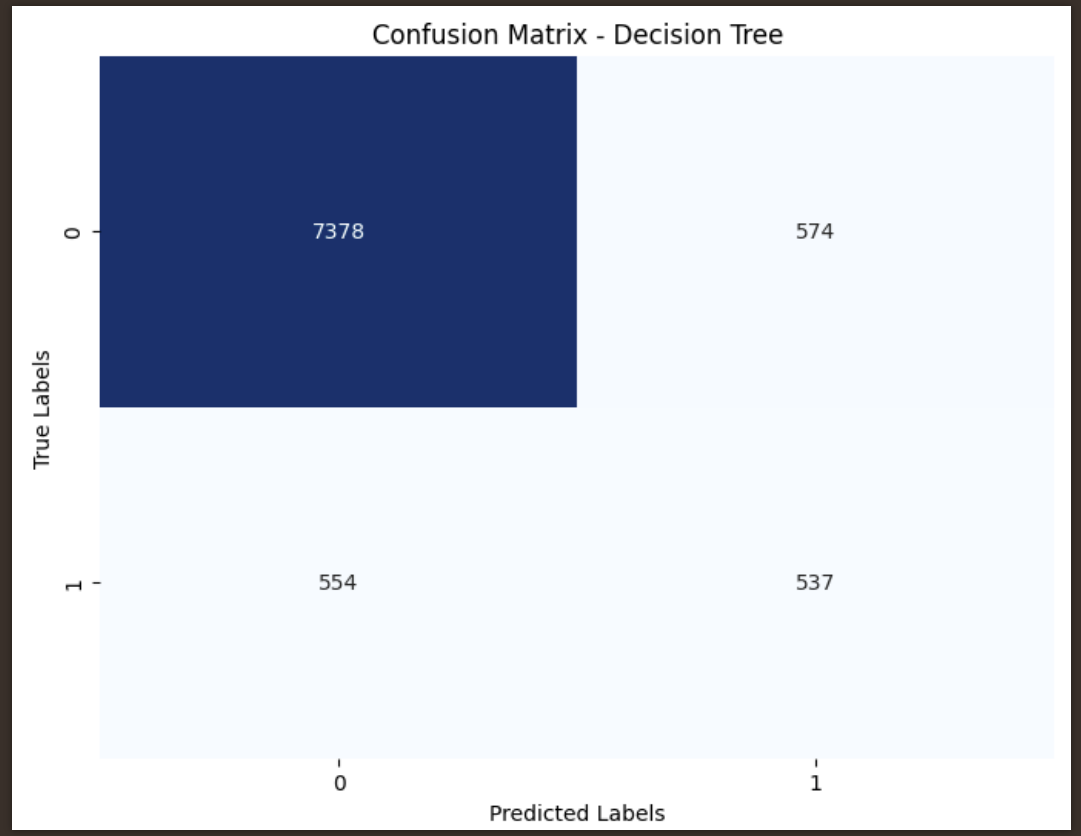
Logistic Regression – Confusion Matrix



Conclusions:

Logistic Regression Results:
True Positive = 7783
True Negative = 232.
Also, the Logistic Regression model achieved and accuracy of 88,63%

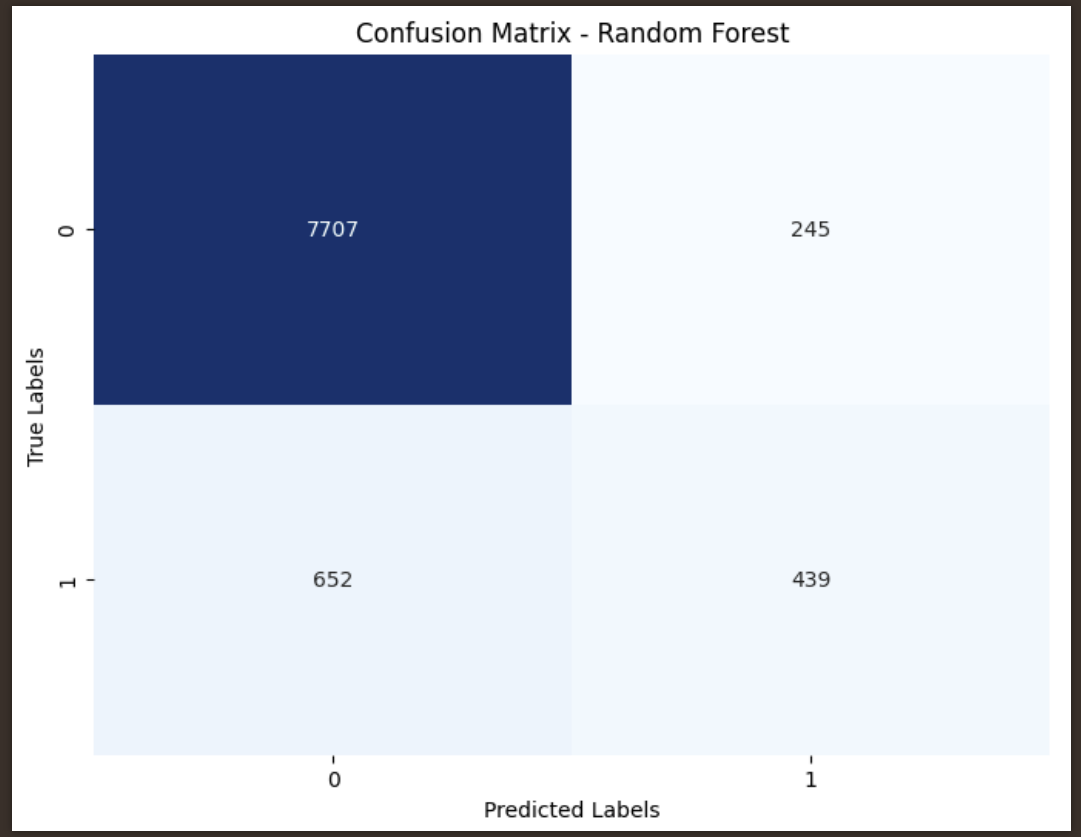
Decision Tree – Confusion Matrix



Conclusions:

Decision Tree Results:
True Positive = 7378
True Negative = 537.
The decision Tree model scored and accuracy of 87,65%.

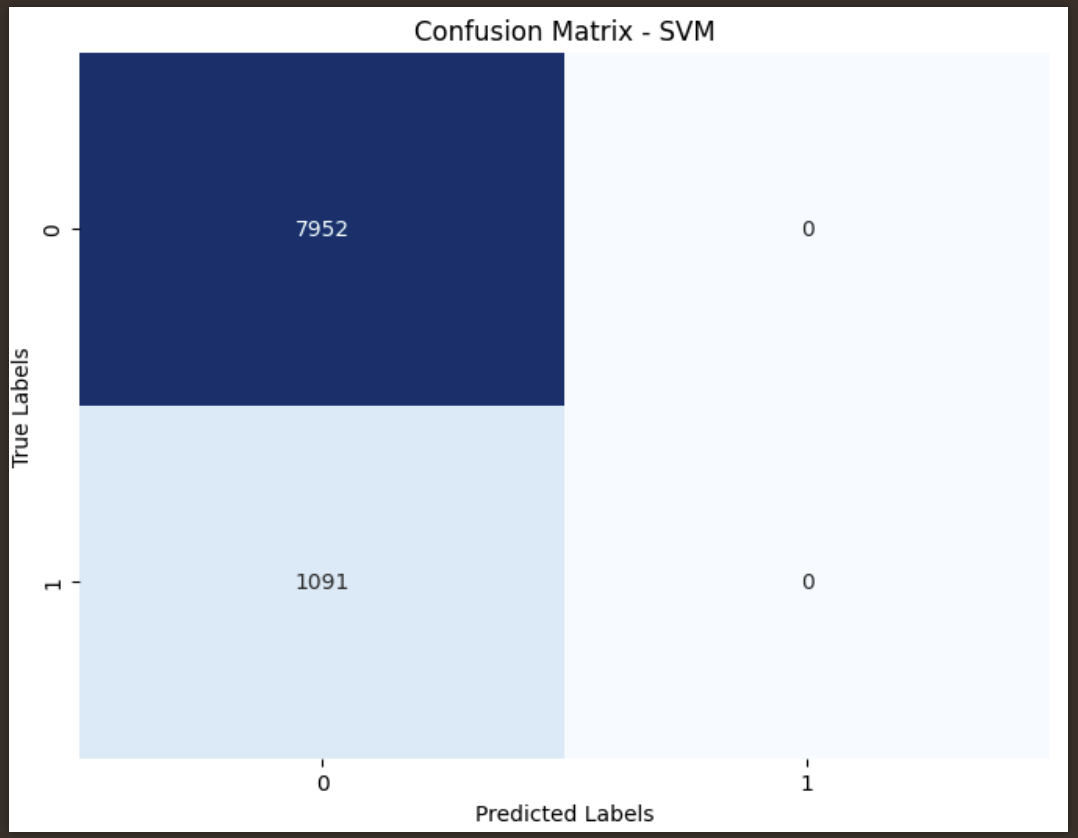
Random Forest– Confusion Matrix



Conclusions:

Random Forest Results:
True Positive = 7707
True Negative = 439.
The Random Forest scored and accuracy of 90,15%.

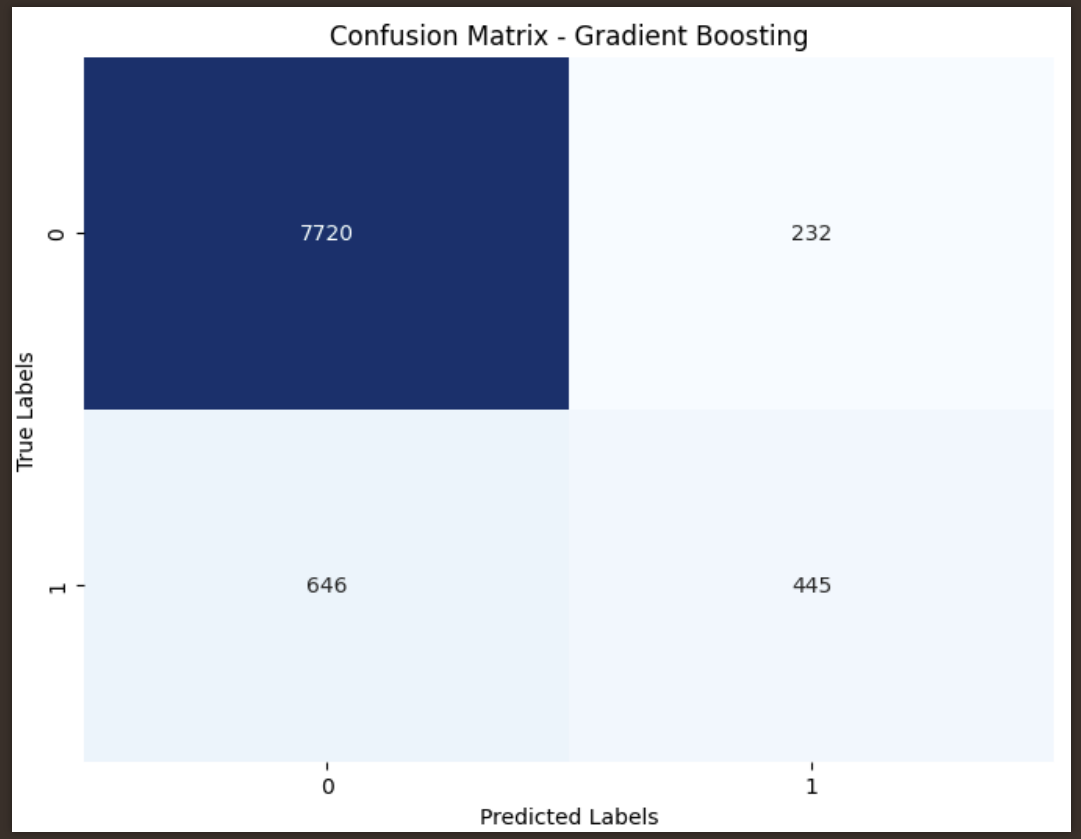
SVM (Support Vector Machine) – Confusion Matrix



Conclusions:

Support Vector Machine Results:
True Positive = 7952
True Negative = 0.
The Support Vector Machine scored and accuracy of 87,93%.

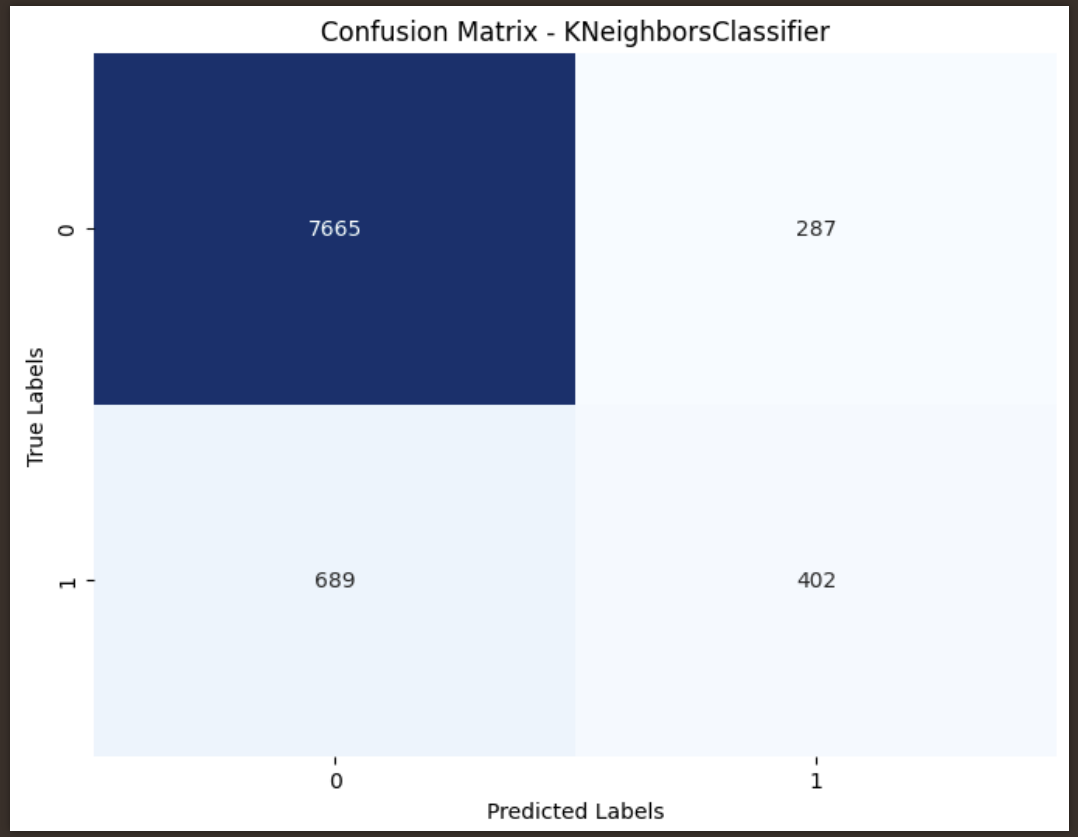
Gradient Boosting– Confusion Matrix



Conclusions:

Gradient Boosting Results:
True Positive = 7720
True Negative = 445.
The Gradient Boosting scored an excellent accuracy of 90,29%.

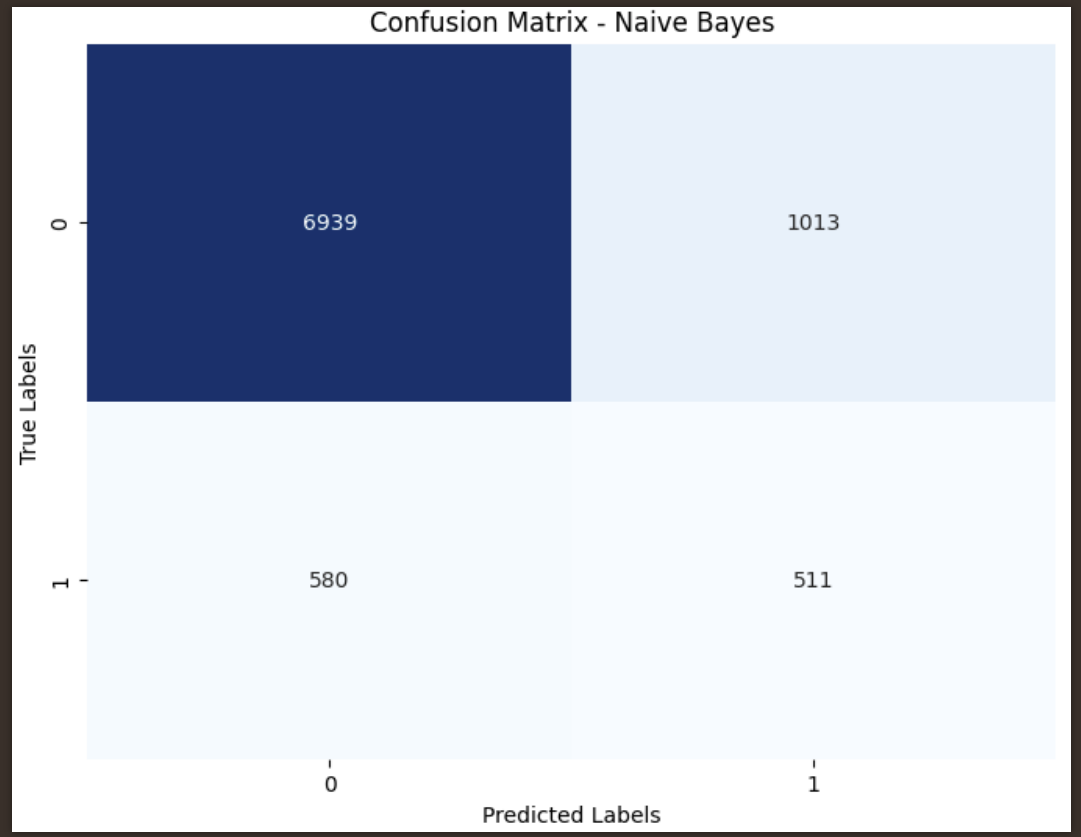
KNN (K-nearest Neighbors) – Confusion Matrix



Conclusions:

Gradient Boosting Results:
True Positive = 7665
True Negative = 402.
The K Neighbors scored an accuracy of 89,20%.

Naïve Bayes – Confusion Matrix



Conclusions:

Naive Bayes Results:
True Positive = 6939
True Negative = 511.
The model scored an accuracy of 82,38%.

All Classifiers Report

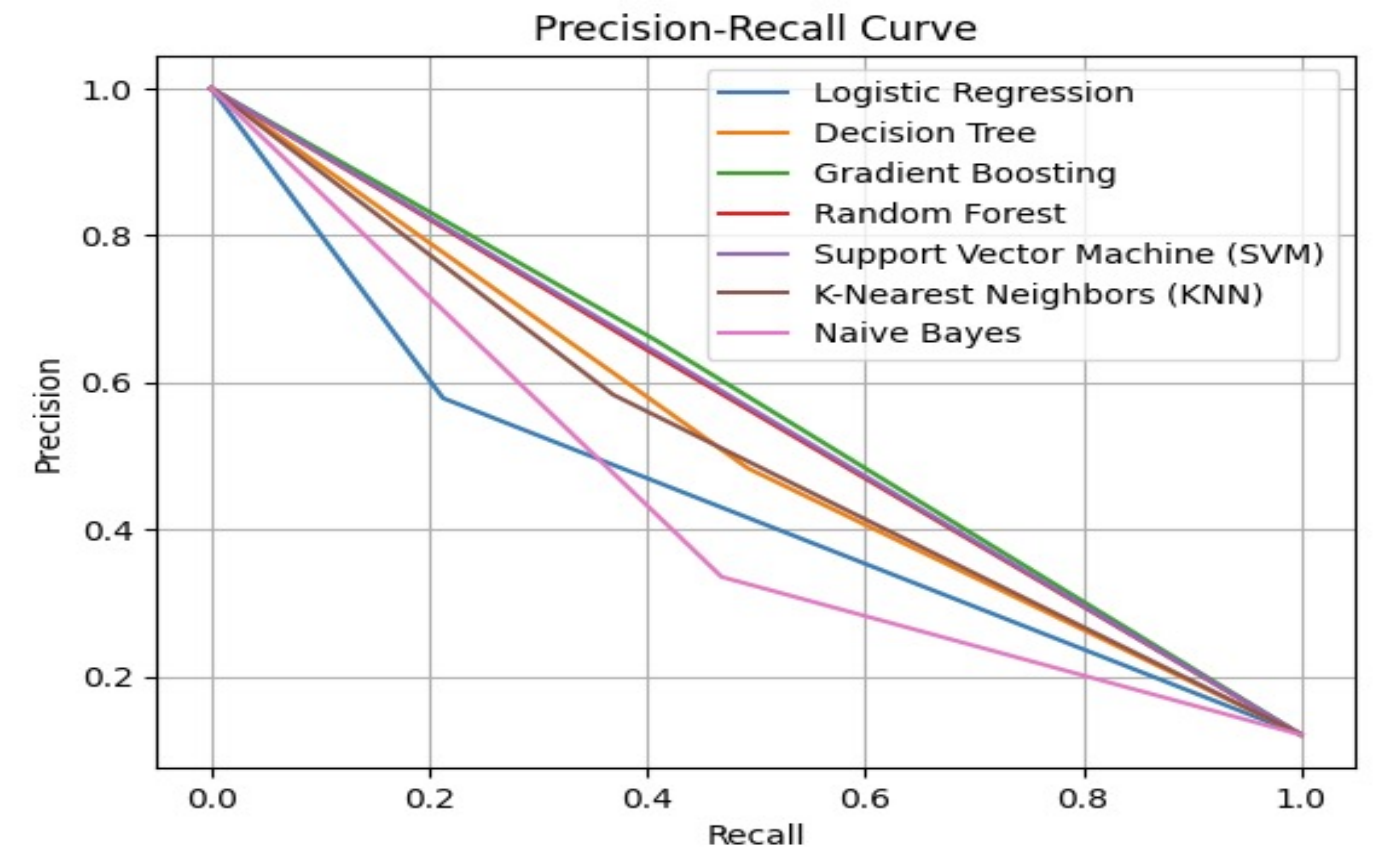
```
Logistic Regression:
Accuracy: 0.886
Precision: 0.579
Recall: 0.213
F1 Score: 0.311

Decision Tree:
Accuracy: 0.875
Precision: 0.483
Recall: 0.492
F1 Score: 0.488

Gradient Boosting:
Accuracy: 0.903
Precision: 0.657
Recall: 0.408
F1 Score: 0.503
```

```
Random Forest:
Accuracy: 0.901
Precision: 0.642
Recall: 0.402
F1 Score: 0.495

Support Vector Machine (SVM):
Accuracy: 0.879
Precision: 0.000
Recall: 0.000
F1 Score: 0.000

K-Nearest Neighbors (KNN):
Accuracy: 0.892
Precision: 0.583
Recall: 0.368
F1 Score: 0.452

Naive Bayes:
Accuracy: 0.824
Precision: 0.335
Recall: 0.468
F1 Score: 0.391
```

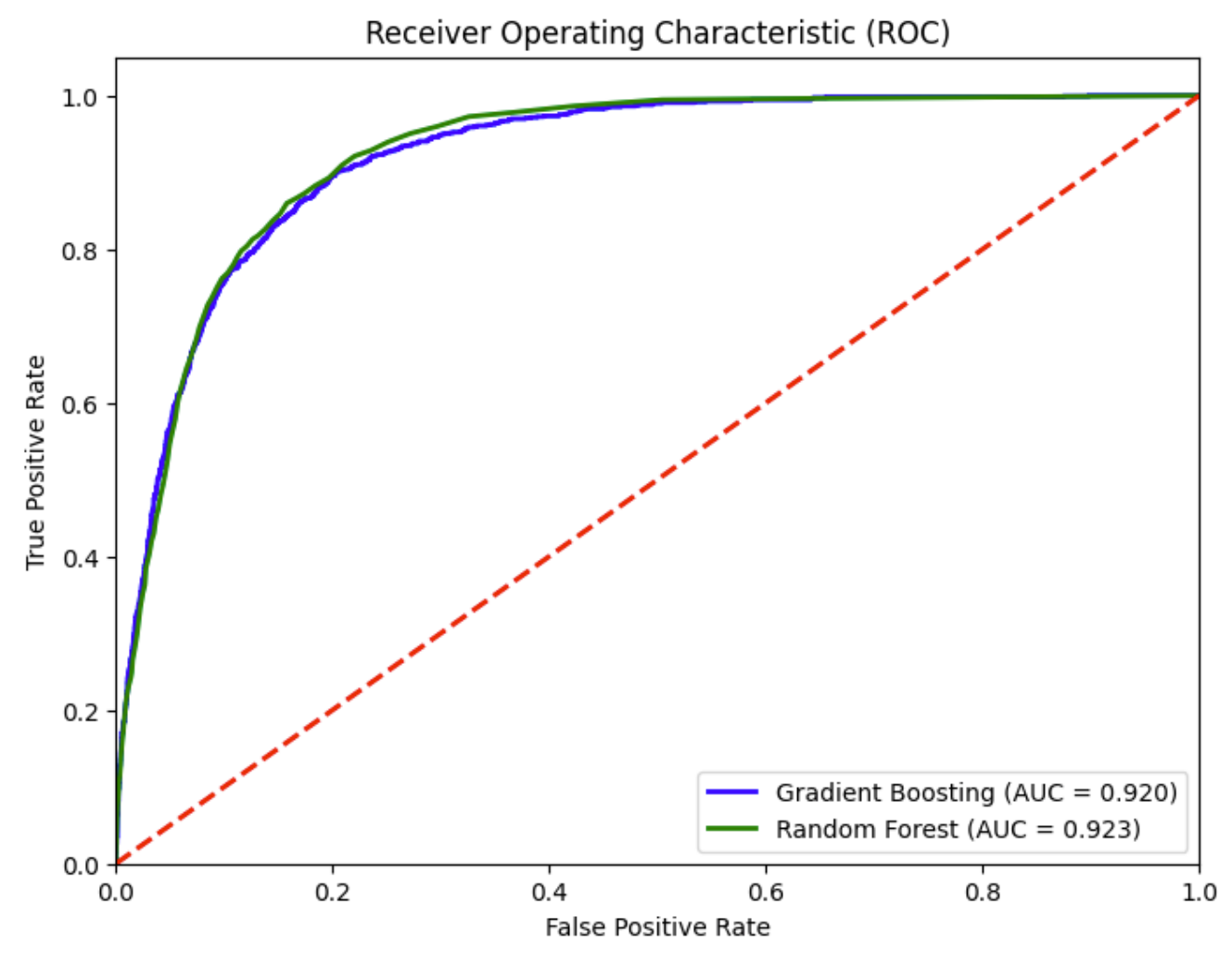
Conclusions:

The best models classifiers was Random Forest (acc = 0.901) and Gradient Boosting (acc = 0.903)

Precision-Recall Curve (All models)



ROC Curve (Best models)



Feature Importance

Gradient Boosting

Top 4 - Feauture importances - Gradient Boosting

1. `duration` - This feature holds the utmost significance, suggesting that the duration of the call greatly influences the outcome.
2. `month` - The last contact in the month is a crucial factor.
3. `poutcome` - The result of the previous marketing campaing is also very important.
4. `pdays` - The elapsed time since the client's last contact from a previous campaign is also an essential factor.

Random Forest

Top 4 - Feauture importances - Random Forest

1. `duration` - Likewise, the duration of the call remains as the primary predictor in the Random Forest model.
2. `balance` - A annual balance is also a very important feature that can influence the outcome using the random forest model.
3. `day` - The last contact of a day contribute to the model.
4. `month` - The last contact in the month have it importance.