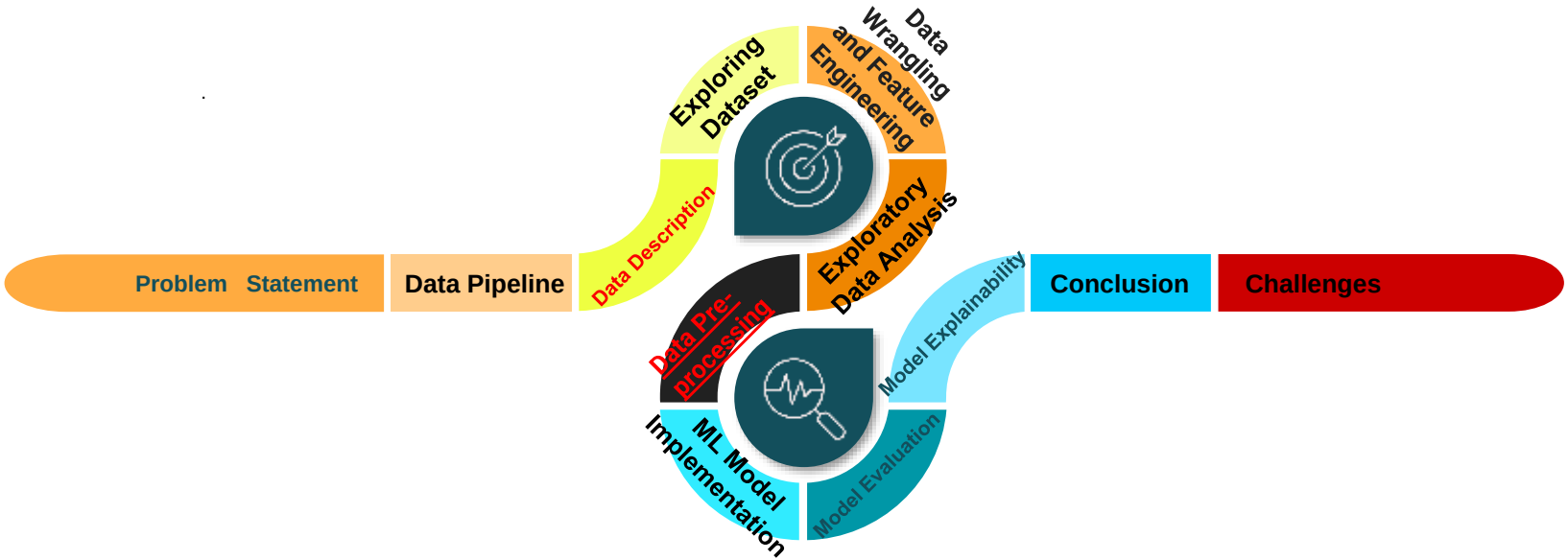
Capstone Project

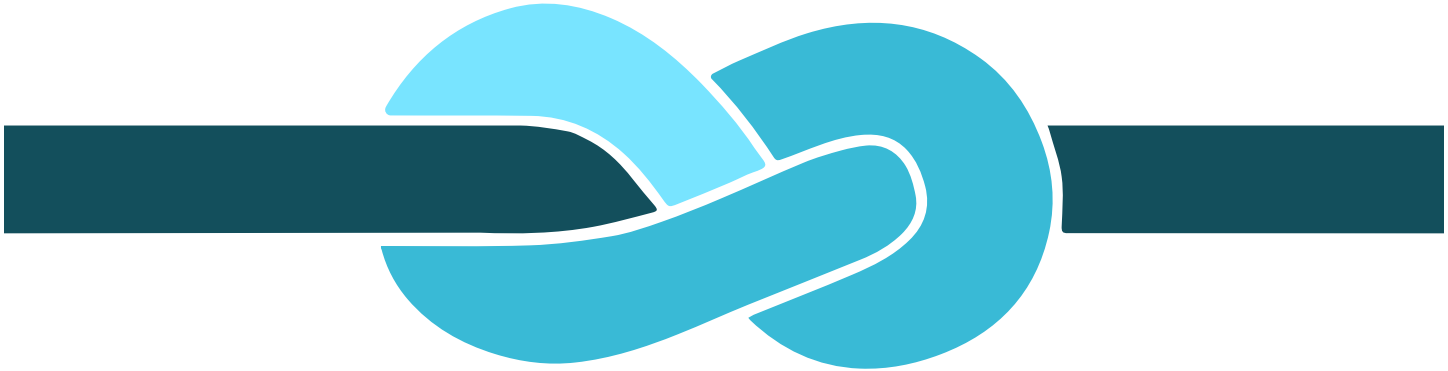
Bank Marketing Effectiveness Prediction

Ajit Sharad Mane
(ajitmane36@gmail.com)

Index

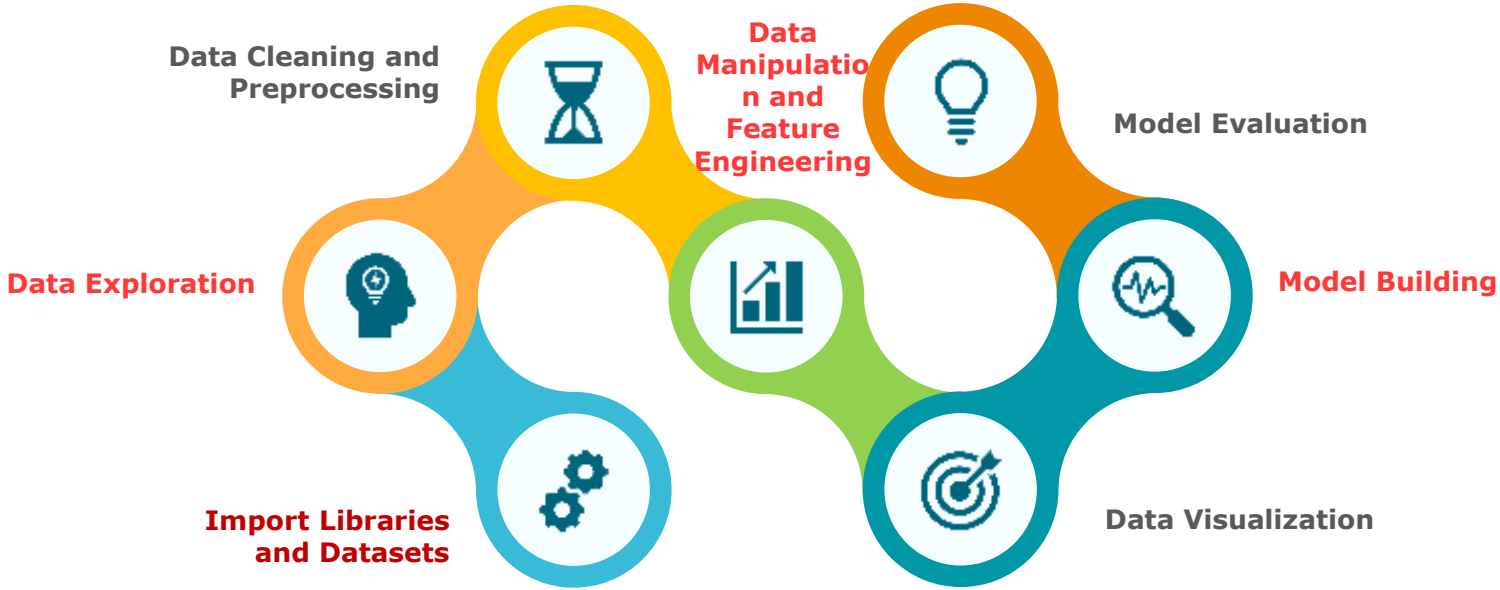

Problem Statement
Data Pipeline
Data Description
Exploring Dataset
Data Wrangling and Feature Engineering
Exploratory Data Analysis
Data Pre-processing
Model Explainability
Model Evaluation
ML Model Implementation
Conclusion
Challenges

Problem Statement



The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. The classification goal is to predict if the client will subscribe a term deposit (variable y).

Data Pipeline



Data Description

Sr.No.	Feature Name	Description
1	age	age of client
2	job	type of job
3	marital	marital status
4	education	qualification of client
5	default	has credit in default? housing: has housing loan? (categorical: 'no','yes','unknown')
6	loan	has personal loan?
7	contact	contact communication type
8	month	last contact month of year
9	day	last contact day of the month
10	duration	last contact duration, in seconds
11	campaign	number of contacts performed during this campaign and for this client
12	pdays	number of days that passed by after the client was last contacted from a previous campaign

Data Description

Sr.No.	Feature Name	Description
13	balance	Account balance of the client
14	housing	has housing loan? (categorical: 'no','yes','unknown')
15	poutcome	outcome of the previous marketing campaign
16	previous	number of contacts performed before this campaign and for this client (numeric)
17	y (target variable)	has the client subscribed a term deposit? (binary: 'yes','no')

Data Exploration

- Dataset having 45211 observations and 17 columns.In the dataset, there are object, float64, and int64 dtypes features present.

- Dataset having no duplicated values.

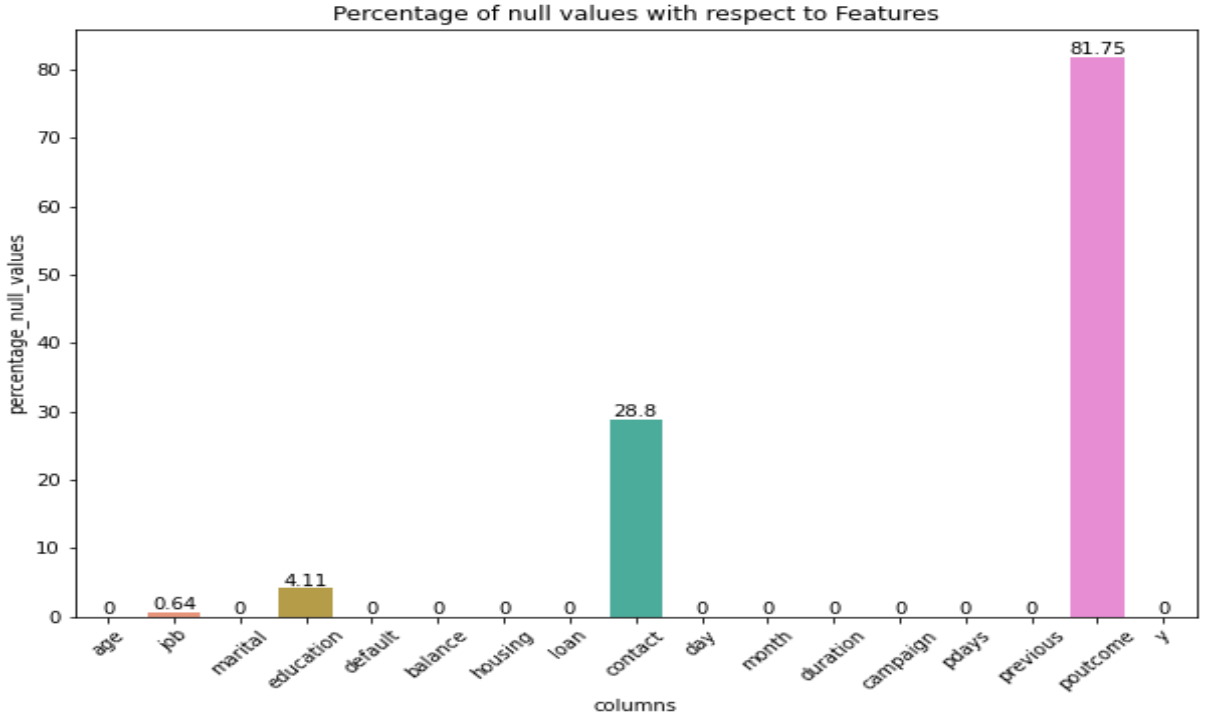
- job, marital, education, default, housing, loan, contact, month, poutcome, and y are among the 10 categorical variables in this dataset. There are 7 numerical variables in this dataset: age, balance, day, duration, campaign, pdays, and previous.

- The unknown values for features job, education, contact, and poutcome are 288; 1857; 13020; and 36959, respectively. Unknown values can be treated as null since they are not defined and can be taken out of features by treatment.

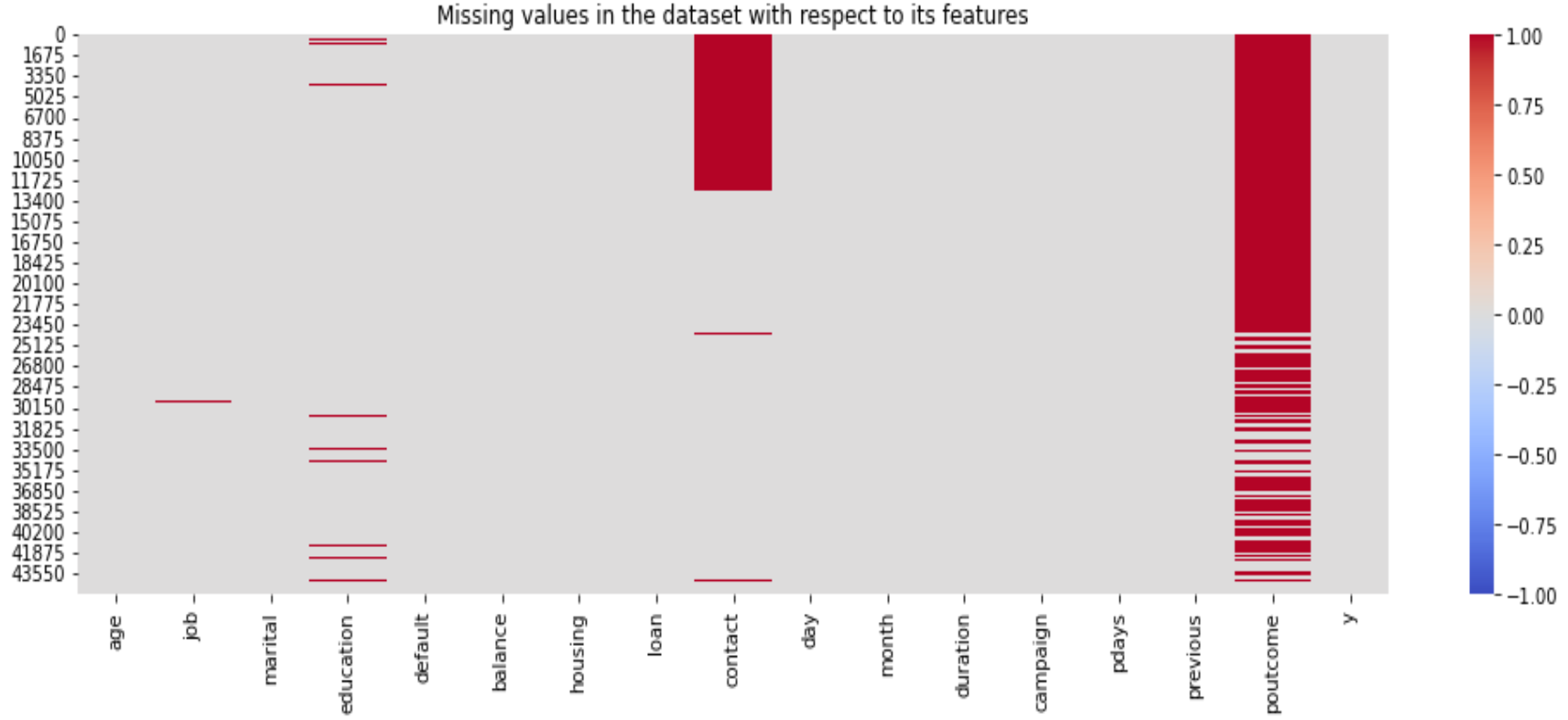
Data Wrangling and Feature Engineering

Handling Null Values

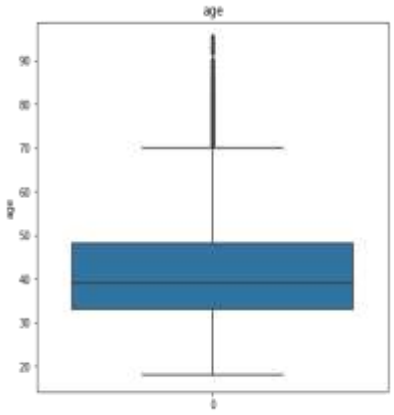
- **The dataset does not have any duplicates.**
- **The unknown values for features job, education, contact, and poutcome are 288; 1857; 13020; and 36959, respectively. Unknown values can be treated as null since they are not defined and can be taken out of features by treatment**

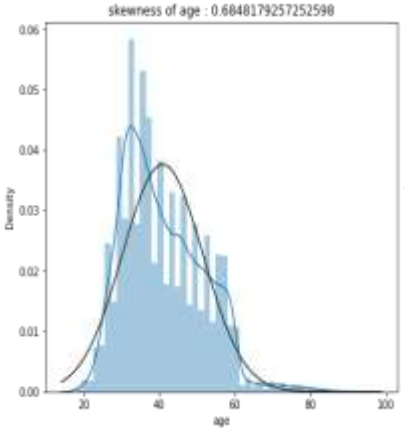


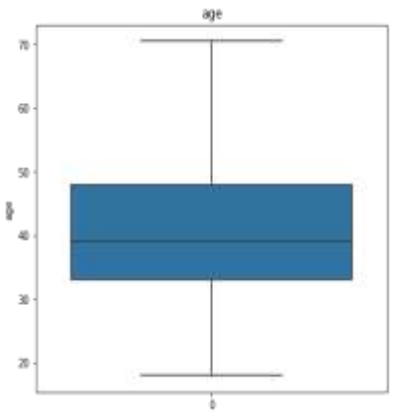
Handling Null Values

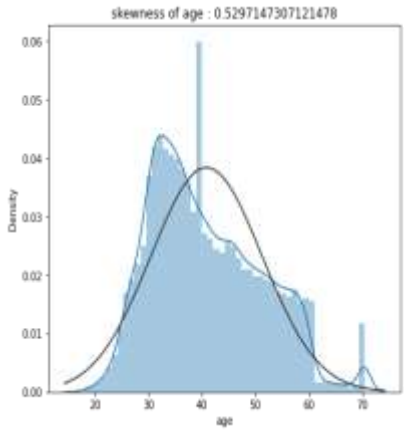


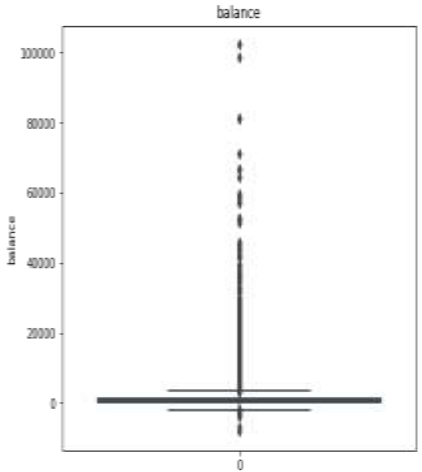
Handling Outliers

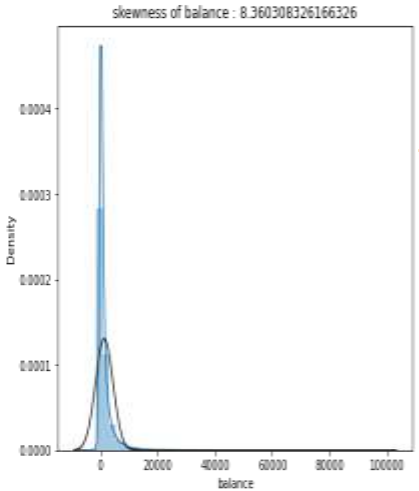


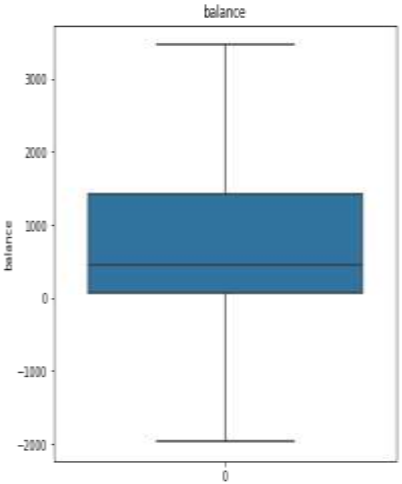


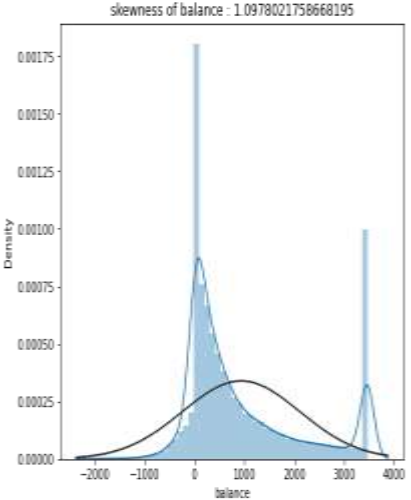




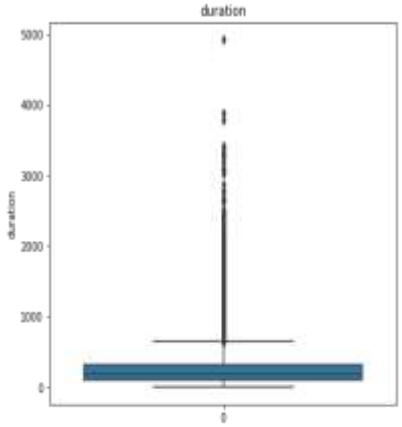


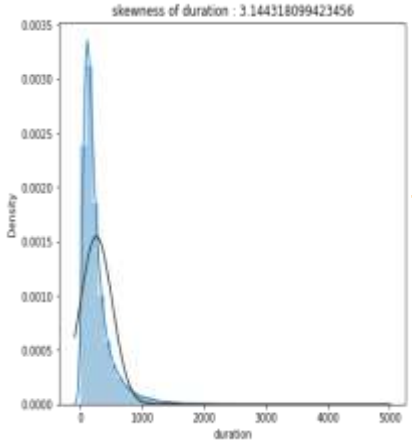


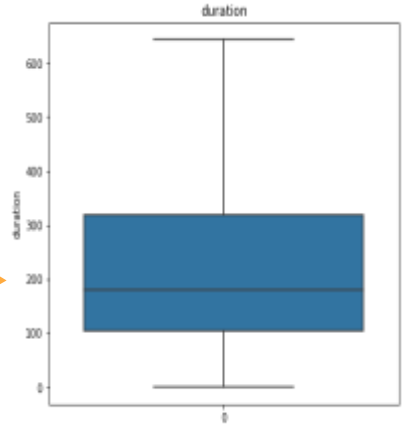


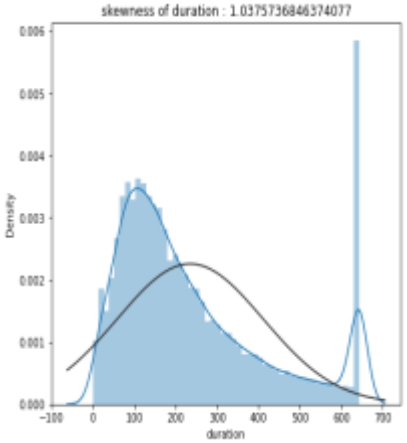


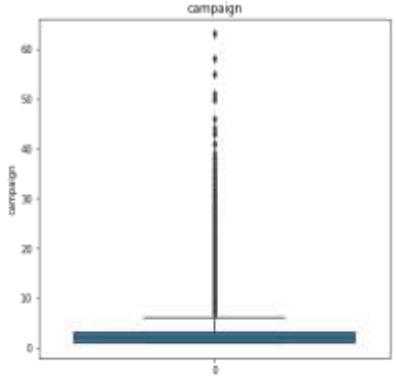
Handling Outliers

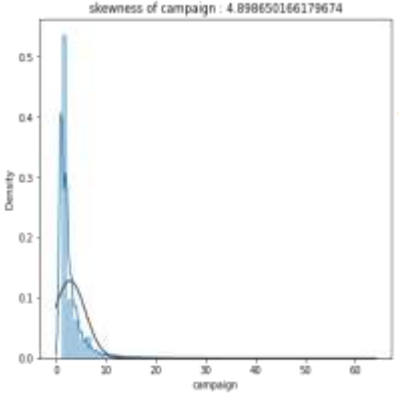


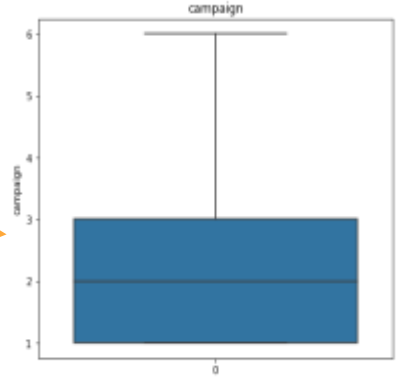


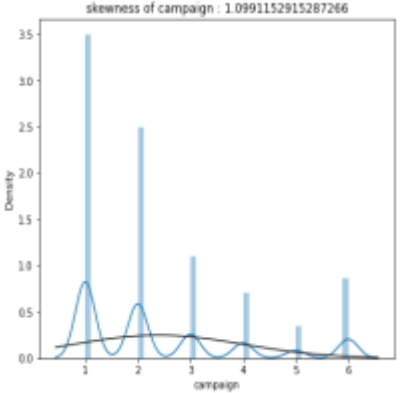






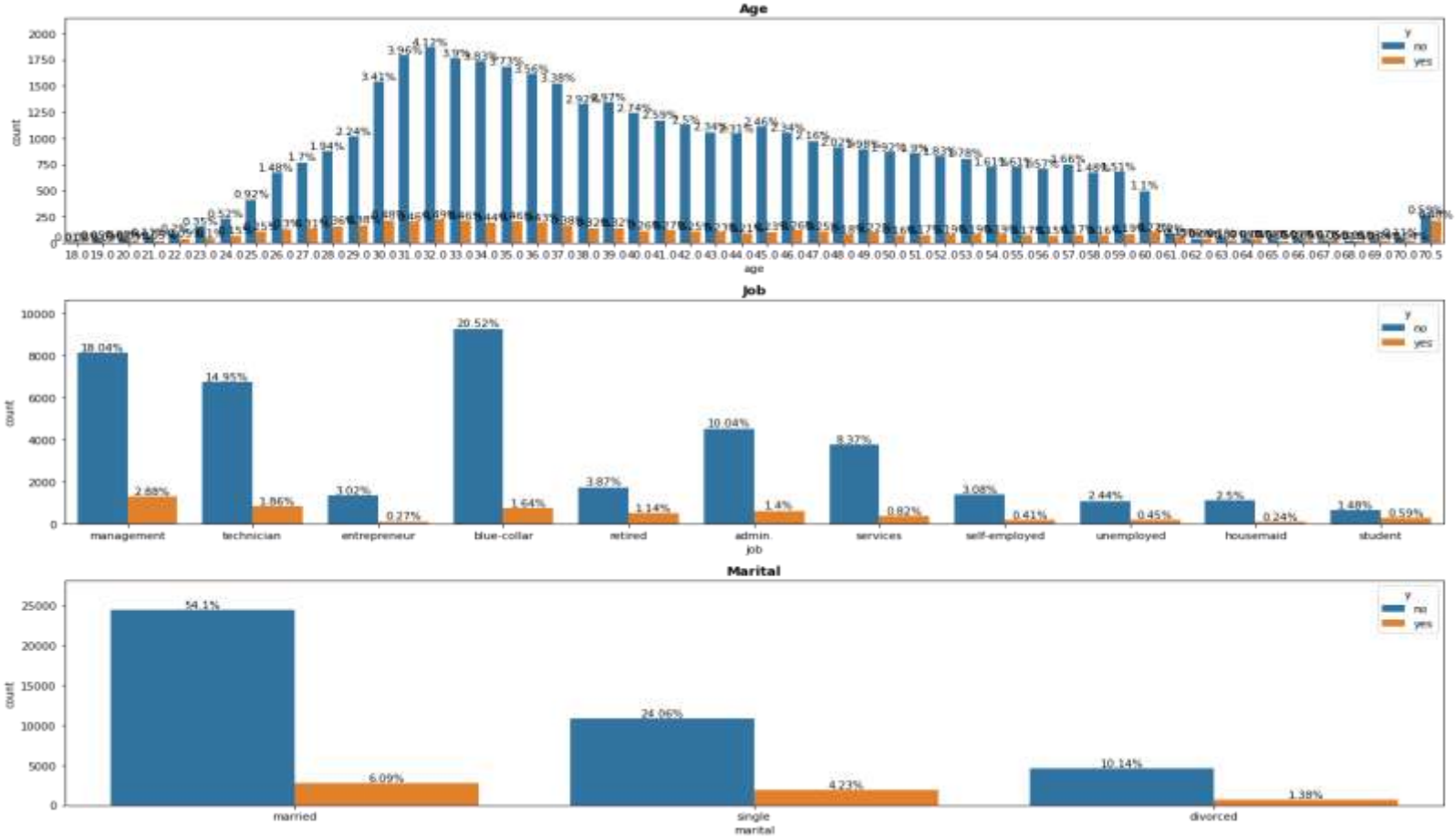






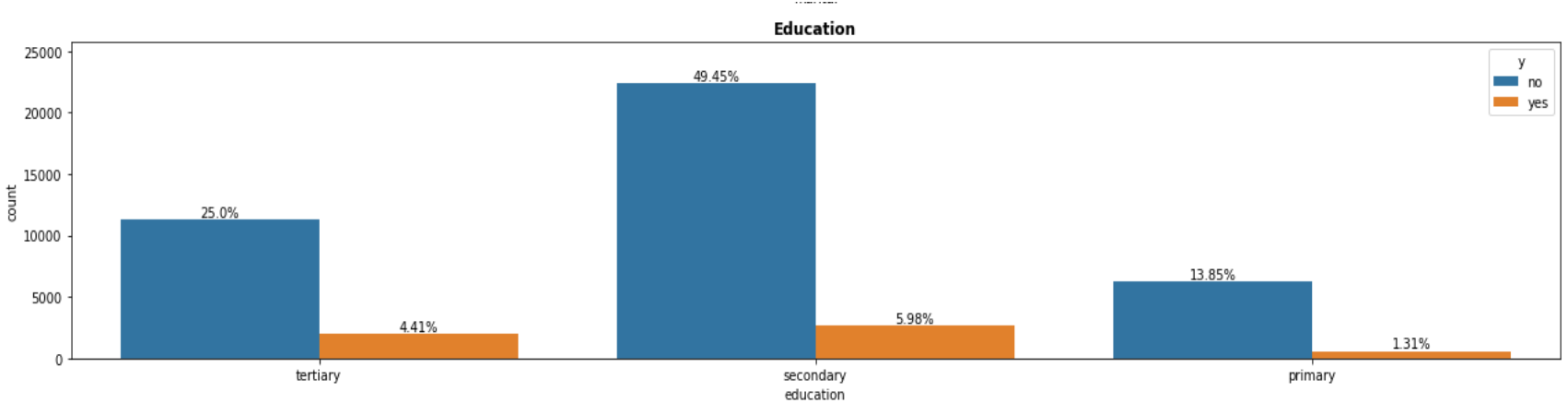
Exploratory Data Analysis

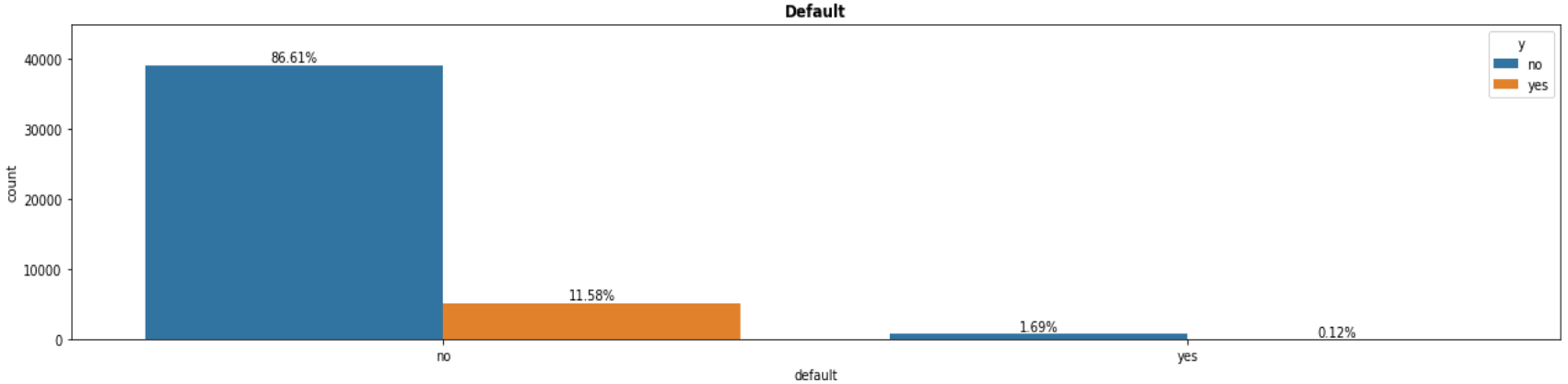
Univariate Analysis



Exploratory Data Analysis

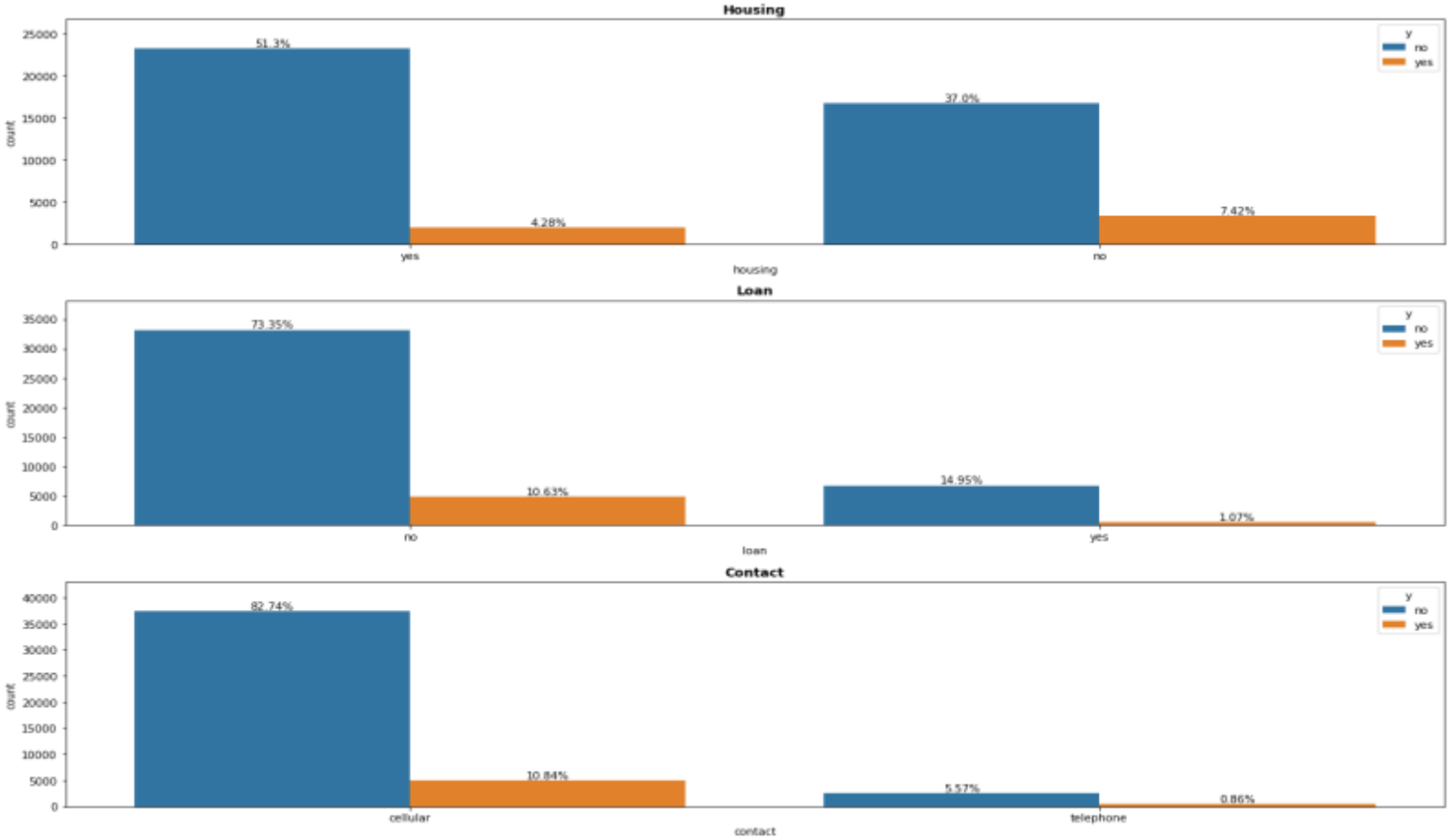
Univariate Analysis





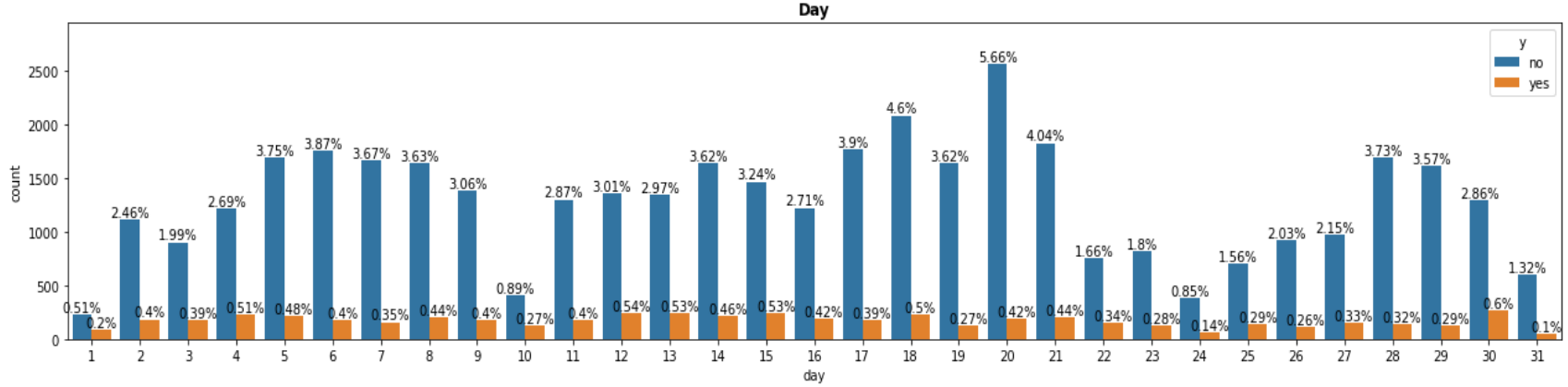
Exploratory Data Analysis

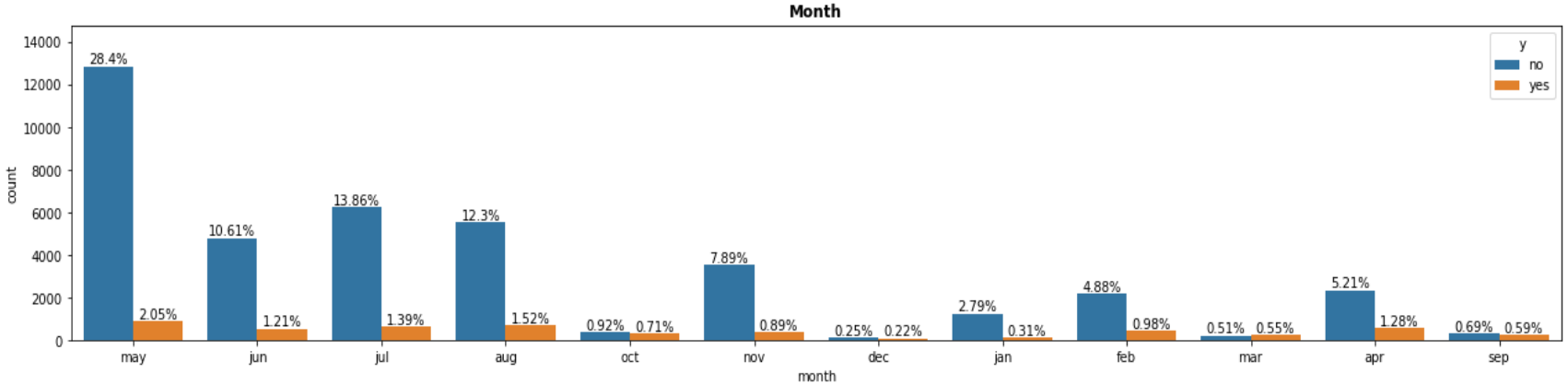
Univariate Analysis



Exploratory Data Analysis

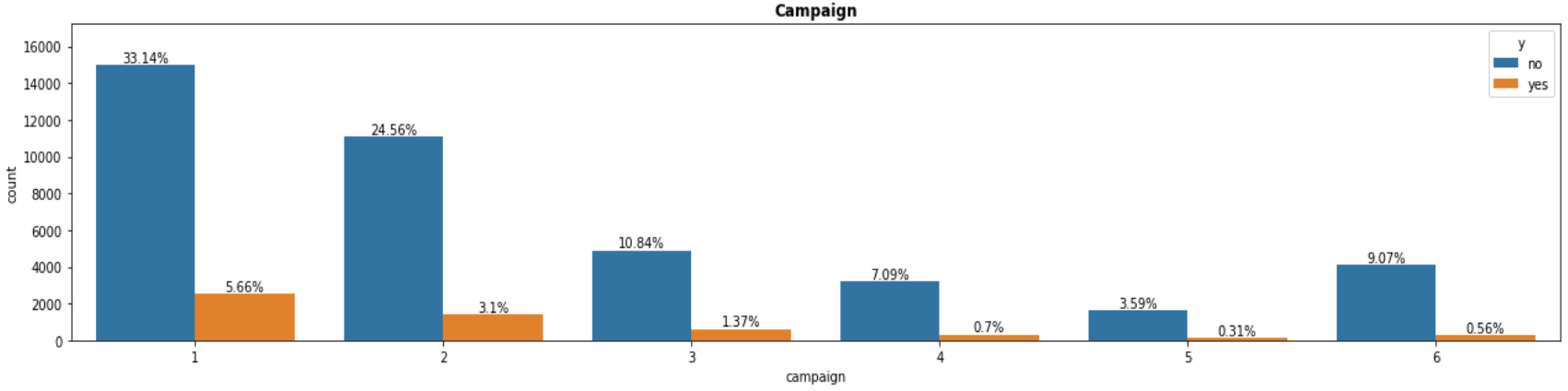
Univariate Analysis

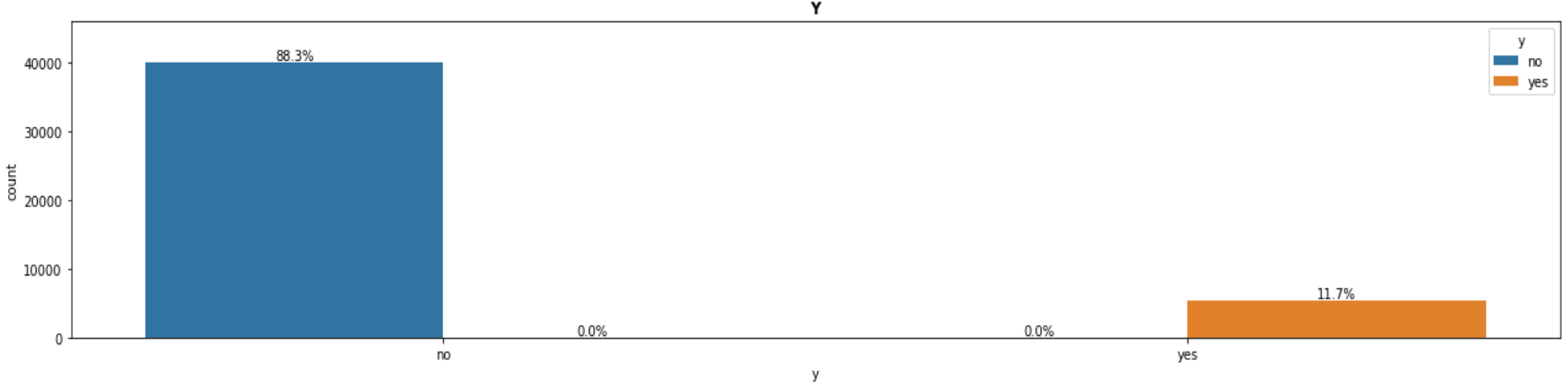




Exploratory Data Analysis

Univariate Analysis





Univariate Analysis

Observations :

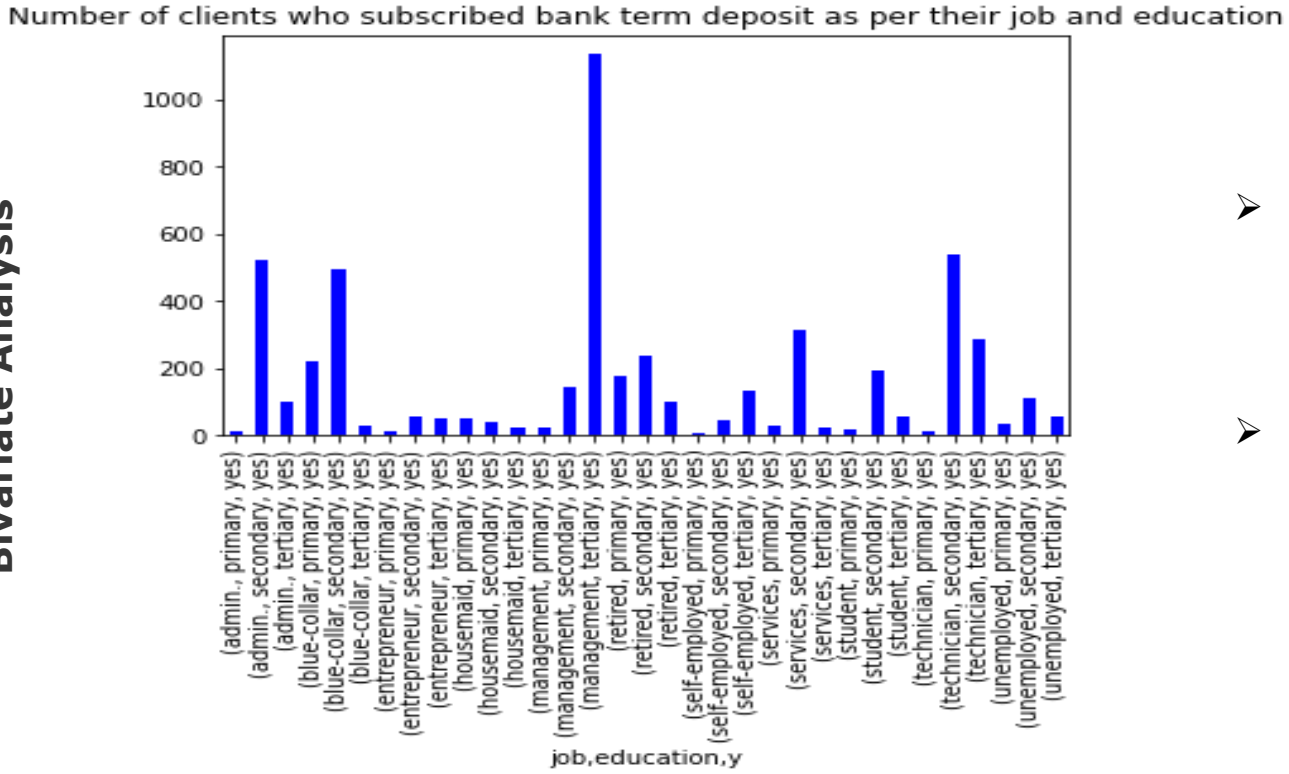
- The average client is between the ages of 25 and 60, but the majority of bank term deposits are made by clients between the ages of 30 and 36.
- Most clients with blue-collar jobs do not subscribe to bank term deposits (20.52%), but most clients with managerial jobs do (2.88%).
- Most of the clients are married. Clients who are married are the most likely to subscribe to term deposits, and they are also the least likely to subscribe to term deposits.
- Most of the clients are married. Clients who are married are the most likely to subscribe to term deposits, and divorced clients are less likely to subscribe to term deposits.
- Clients who are more educated than the primary are more likely to sign up for a term deposit.
- Most of the clients who subscribed to term deposits have no credit in default.
- The majority of clients who have signed up for a term deposit do not have any housing loan.
- If a client has a housing loan, there is a 51% chance that they will not subscribe to a term deposit.
- Clients are more likely to subscribe to the term deposit if they do not have any personal loans.
- If the client has a personal loan, there is a greater chance that they will not subscribe to a term deposit.

Univariate Analysis

Observations :

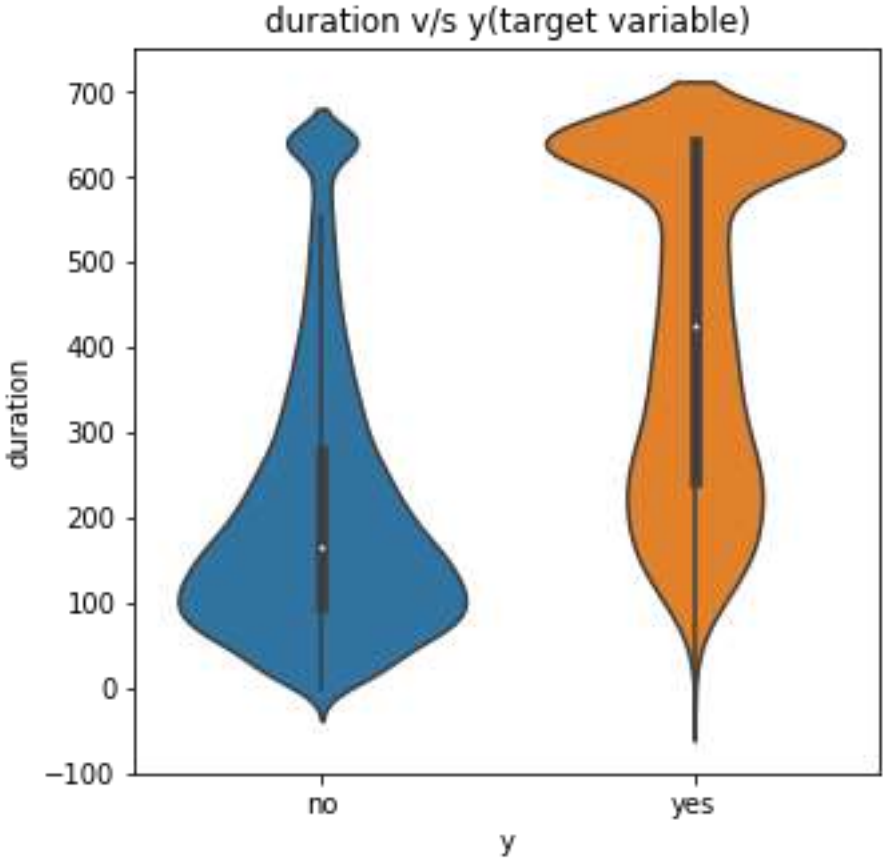
- The clients who were contacted with celluler are mostly subscribed to term deposits.
- Less than one percent of total clients contacted per day subscribe to term deposits.
- In May, June, July, August, and April, more than 1 percentage of clients subscribed to the term deposit, but other than this month, less than 1 percentage of clients subscribed to the term deposit.
- In June, July, August, and April, more than 1 percentage of clients subscribed to the term deposit, but other than this month, less than 1 percentage of clients subscribed to the term deposit. May's subscriber rate is more than double that of the other months of the year, a difference of more than 2 percentage.
- No one has signed up for term deposit if they have received more than three phone calls. Less than three times contacted clients who signed up for term deposits.
- Only 11.7% of total clients sign up for term deposits, which means that there is an 88.3% chance that clients will not subscribe to term deposits.

Bivariate Analysis



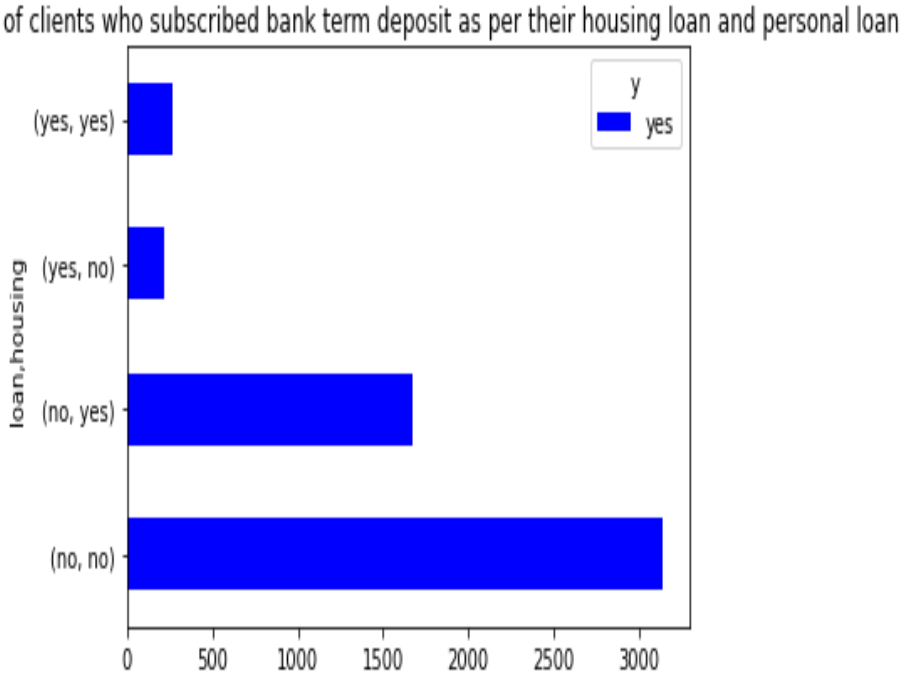
- Most clients who have management-related jobs and a tertiary degree have subscribed to the term deposits.
- Customers with a secondary education are the second most likely to subscribe to term deposits.

Bivariate Analysis



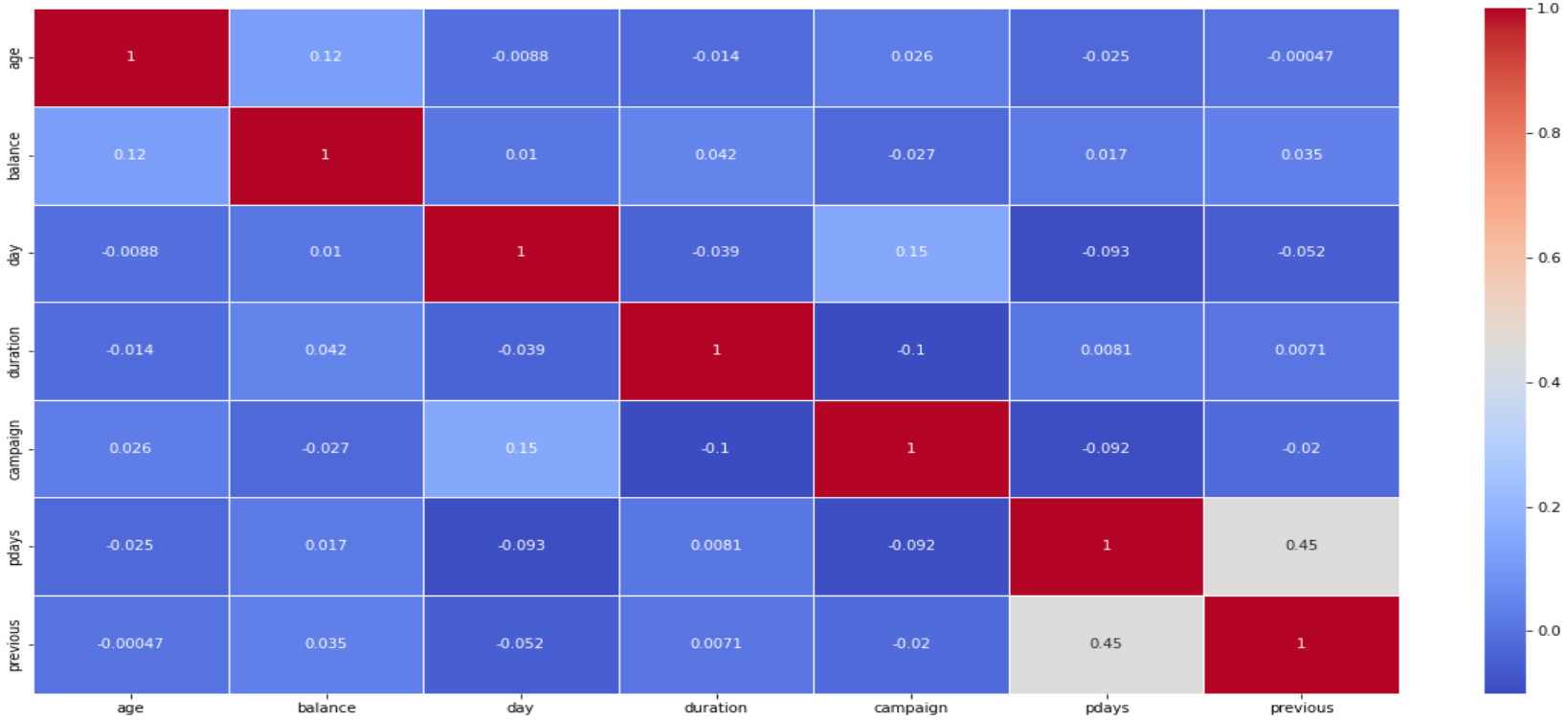
- Clients are more likely to subscribe to term deposits if they spend more time on the phone.
- Average of 400 seconds required to convey clients' intent to subscribe and make a term deposit

Bivariate Analysis



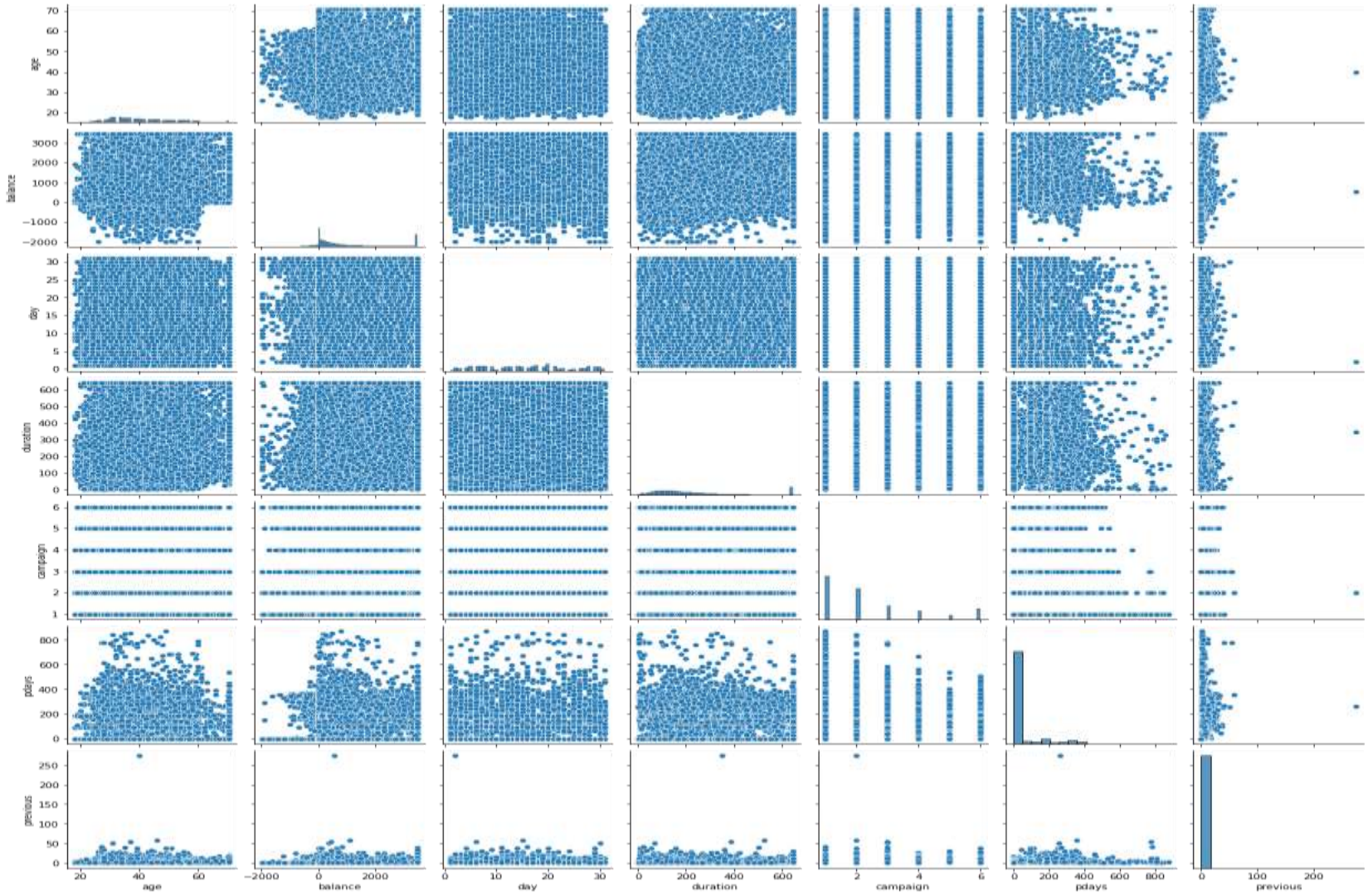
- A customer is more likely to sign up for a term deposit if he is entirely debt-free.
- Customers are less likely to choose a term deposit if they already have both types of loans.

Multivariate Analysis



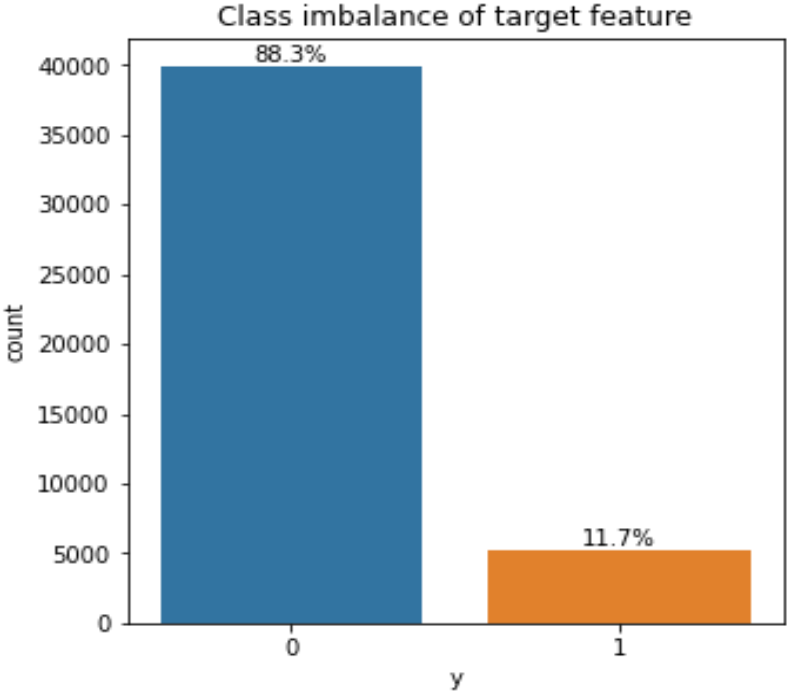
- There is no correlation between any independent variables.

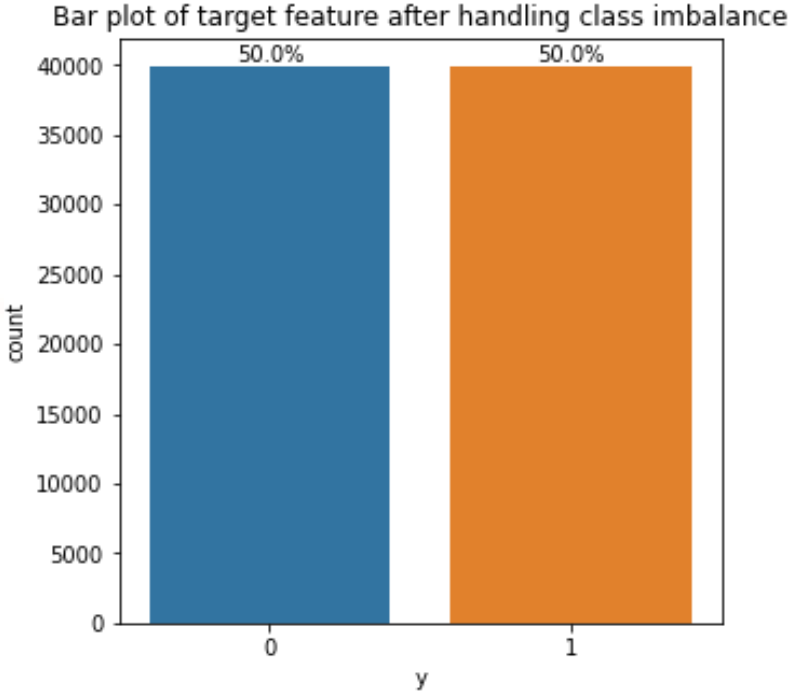
Multivariate Analysis



Data Pre-processing

Handling Imbalanced Dataset





- We clearly detect a class imbalance beacause discovered that the number of clients who subscribed to term deposits is 11.7% lower than the number of clients who did not (88.3%).
- Class imbalance handled successfully using the Synthetic Minority Oversampling Technique (SMOTE).

Machine Learning Model Implementation

[1] Logistic_Regression

[2] Decision_Tree

[3] Random_Forest

[4]Gradient_Boosting_Machine

[5] XGBoost

[6] K_Nearest_Neighbor

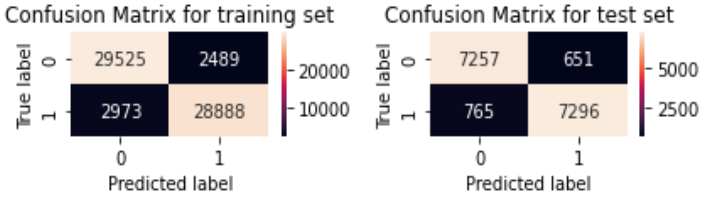
[7] Naive_Bayes

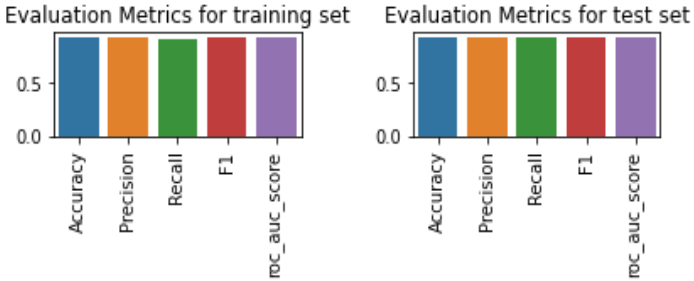
[8] Support_Vector_Machine

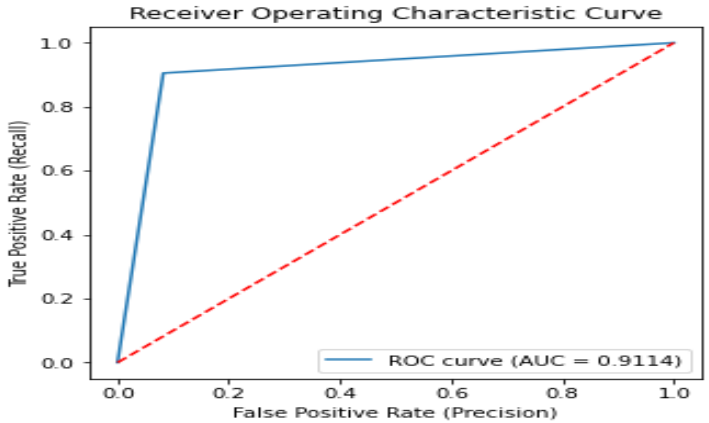
[9] Artificial Neural Network

[1] Logistic Regression

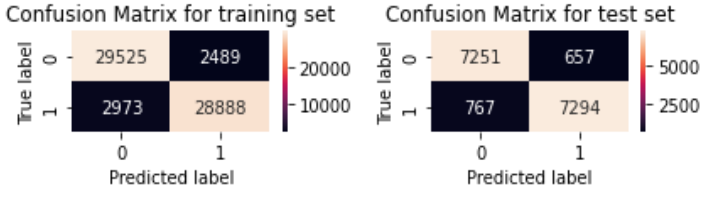
Train/Test Split results

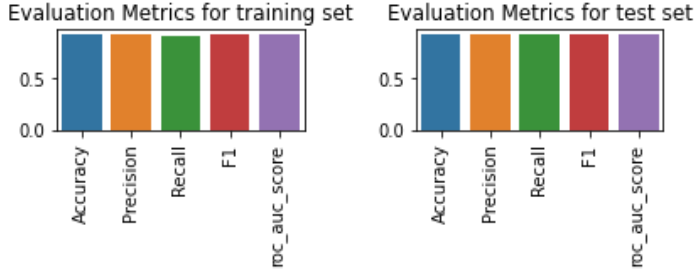


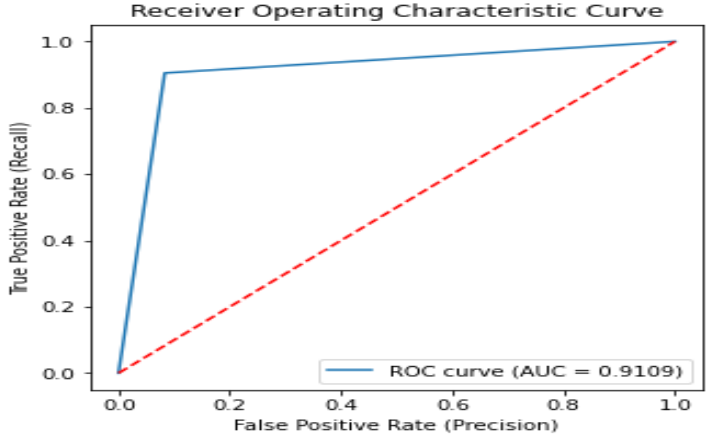




Cross-Validation results

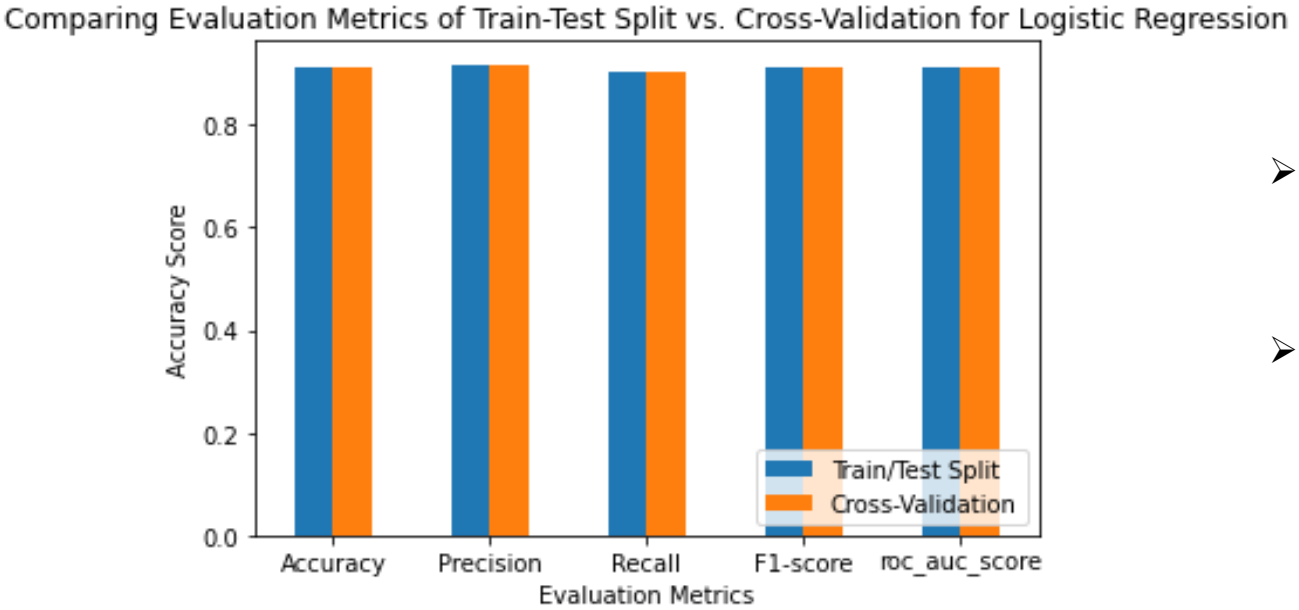






[1] Logistic Regression

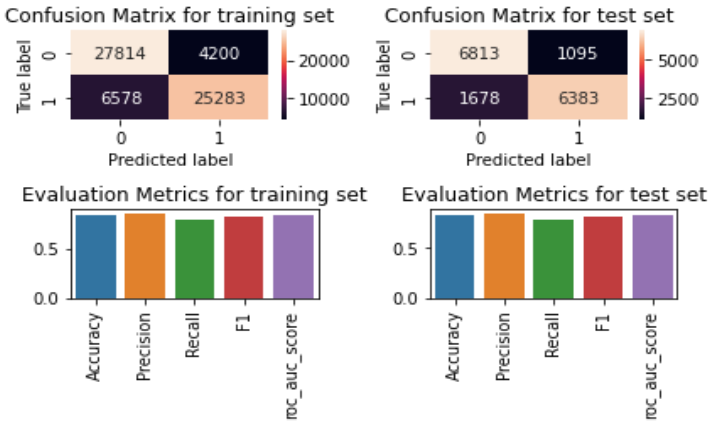
Comparing Evaluation Metrics of Train/Test Split and Cross-Validation

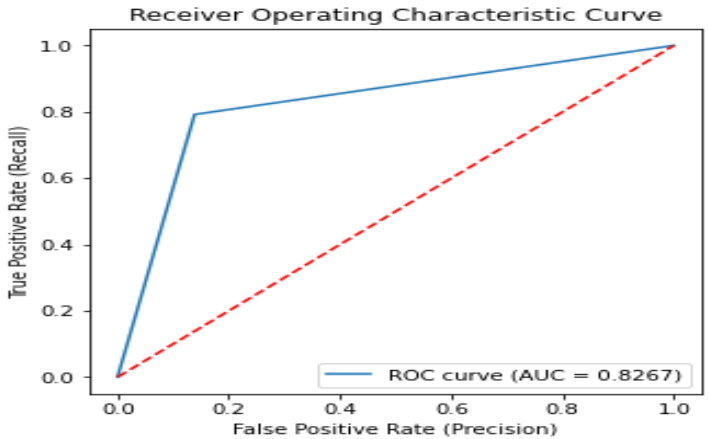


- We saw no improvement in the model after training with cross-validation.
- We got 0.91 % of accuracy in logistics regression model.

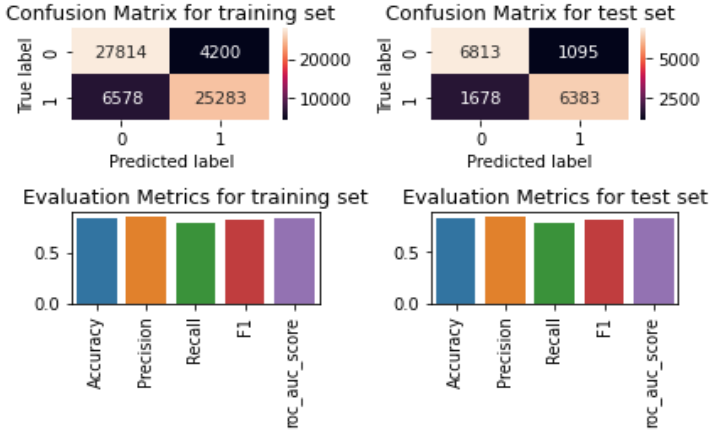
[2] Decision Tree

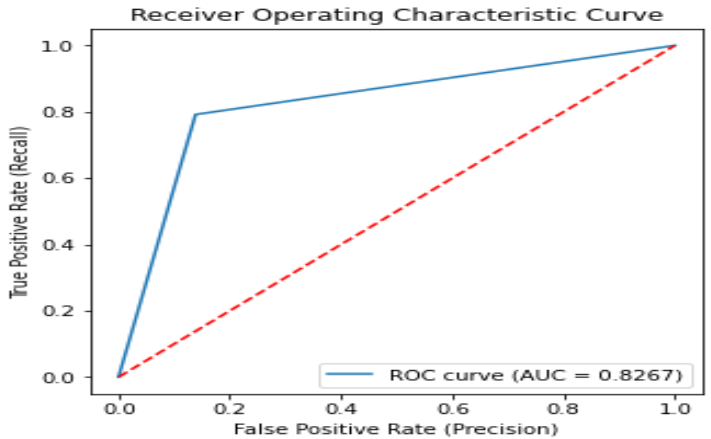
Train/Test Split results





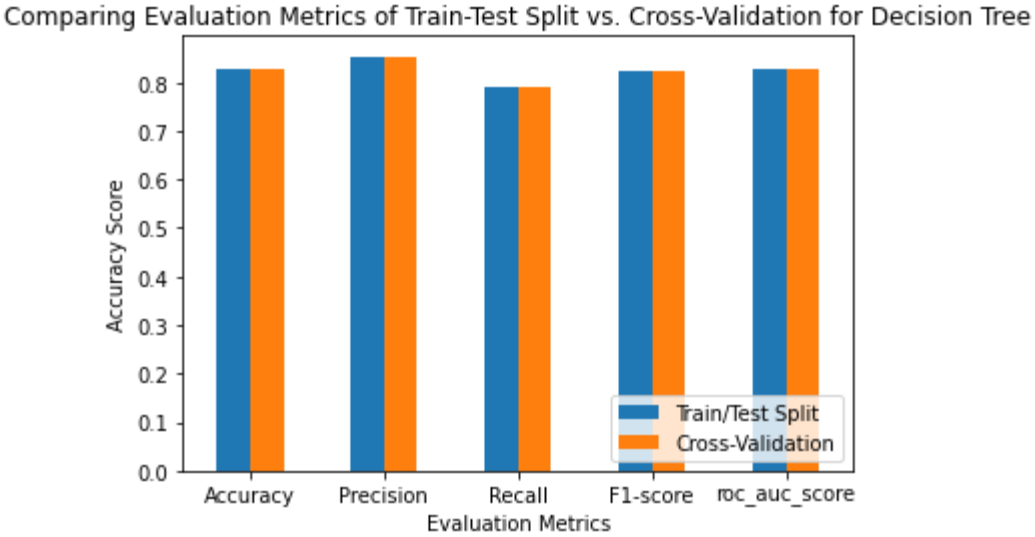
Cross-Validation results





[2] Decision Tree

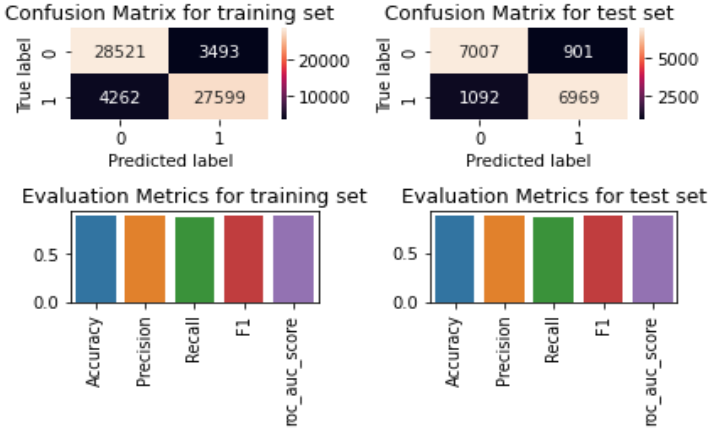
Comparing Evaluation Metrics of Train/Test Split and Cross-Validation

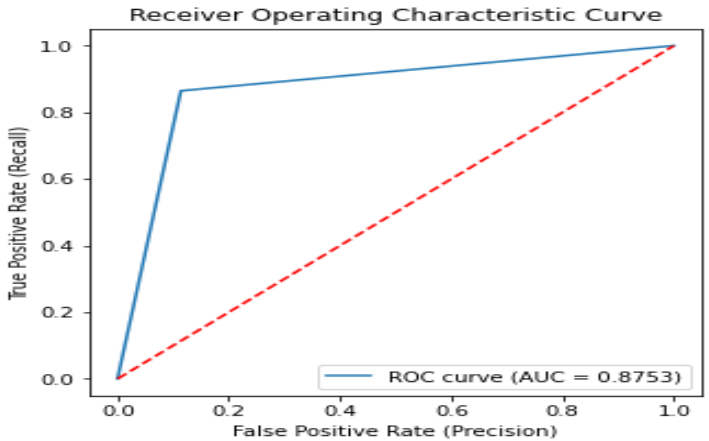


- No improvement seen in the model after training with cross-validation.
- We got 0.82 % of accuracy in Decision Tree model.

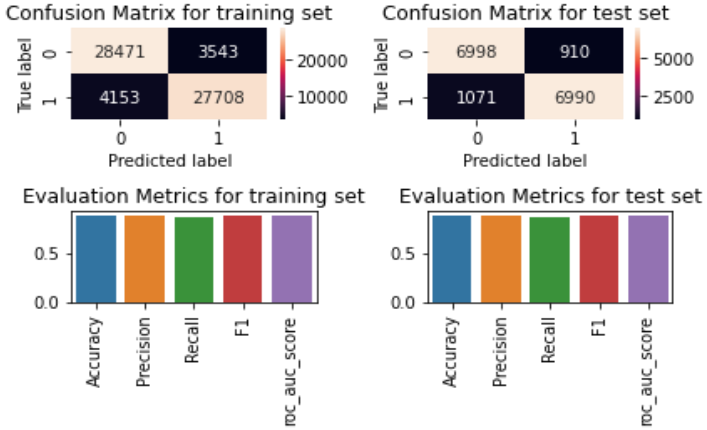
[3] Random Forest

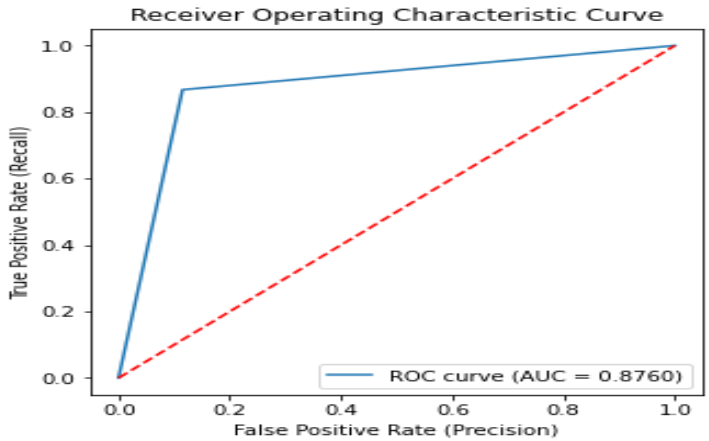
Train/Test Split results





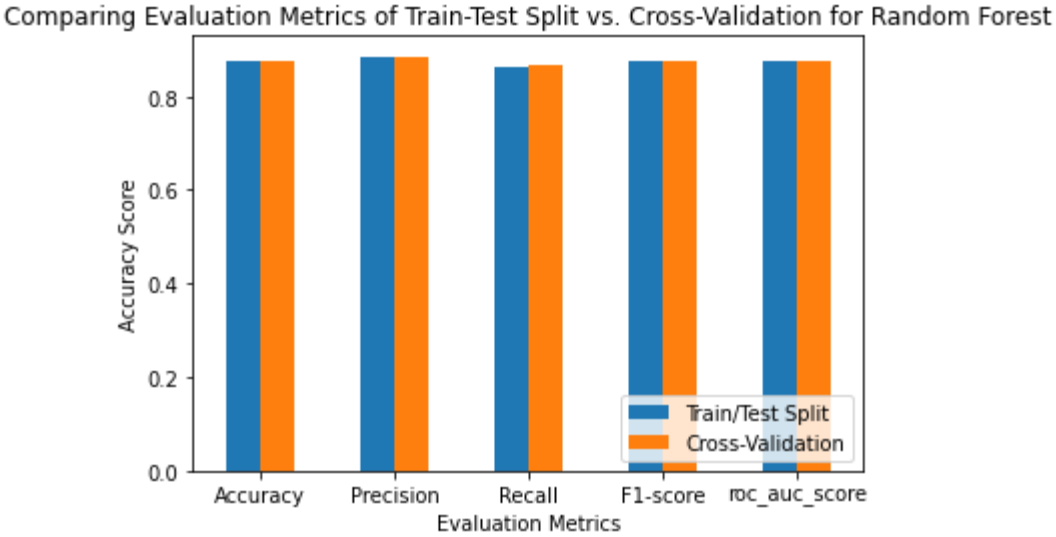
Cross-Validation results





[3] Random Forest

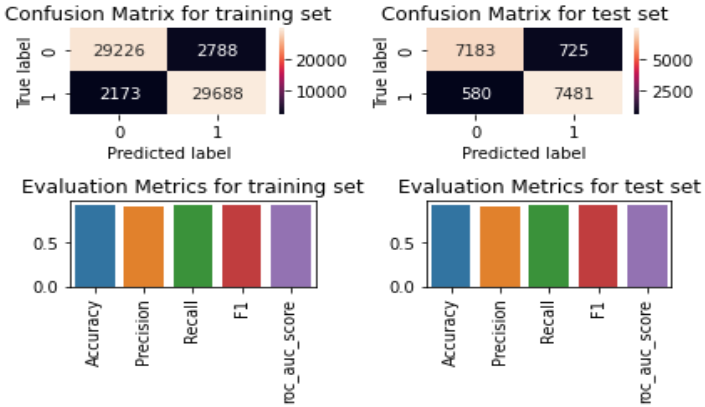
Comparing Evaluation Metrics of Train/Test Split and Cross-Validation

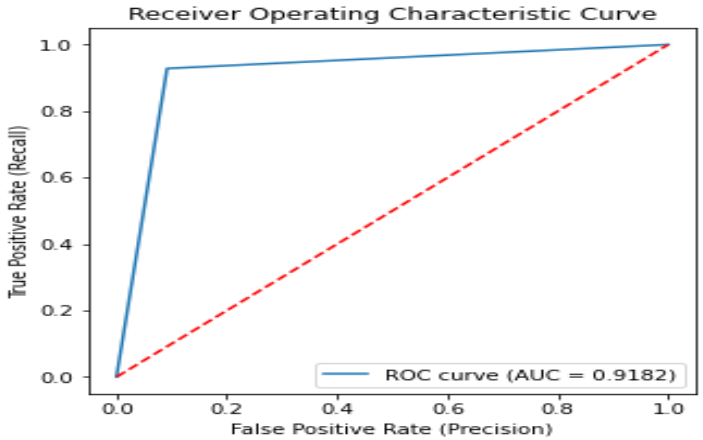


- We saw slightly improvement in the model after training with cross-validation.
- Mejor improvement seen in reacall.
- We got 0.87 % of accuracy using train_test_split and 0.87 % of accuracy using cross validation in Random Forest model.

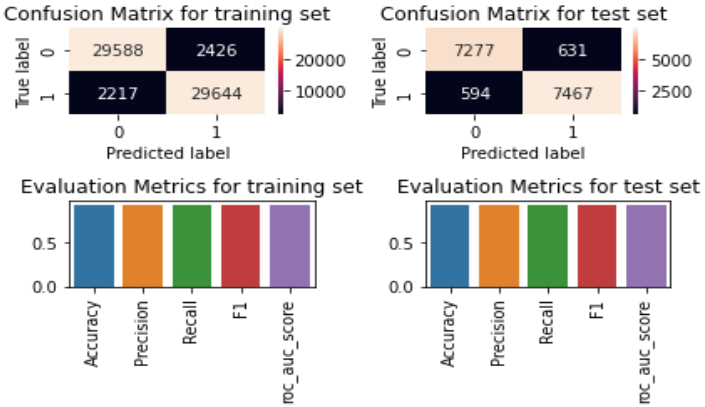
[4] Gradient Boosting Machine

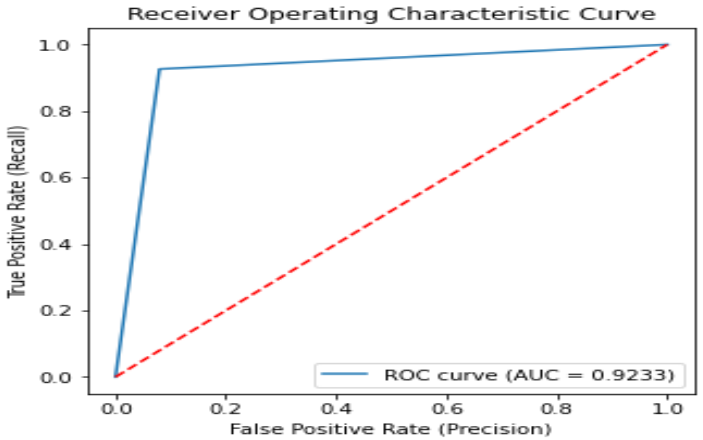
Train/Test Split results





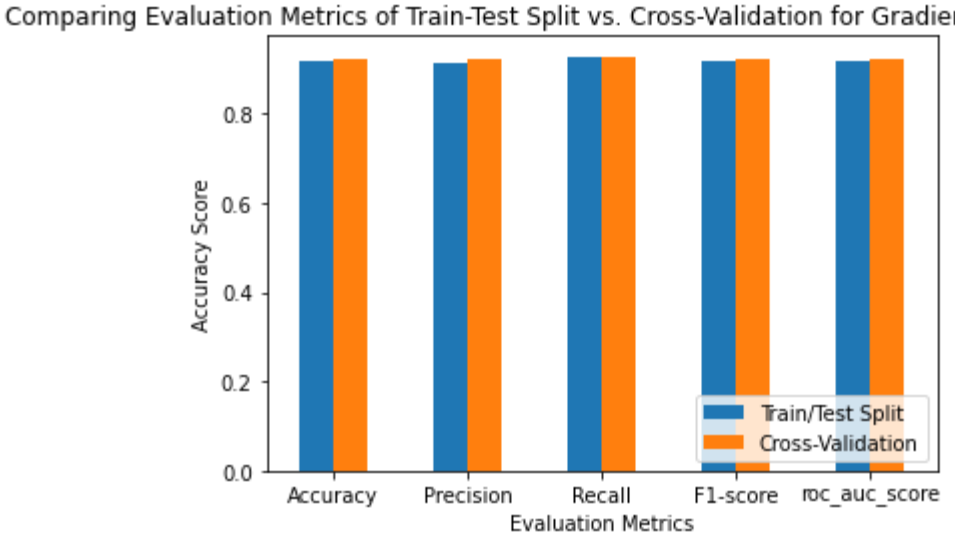
Cross-Validation results





[4] Gradient Boosting Machine

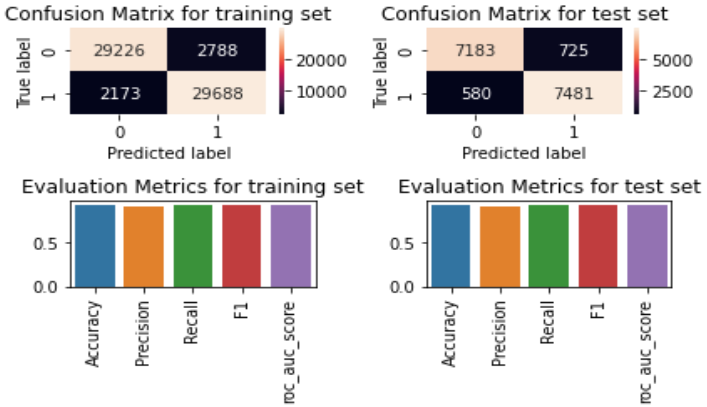
Comparing Evaluation Metrics of Train/Test Split and Cross-Validation

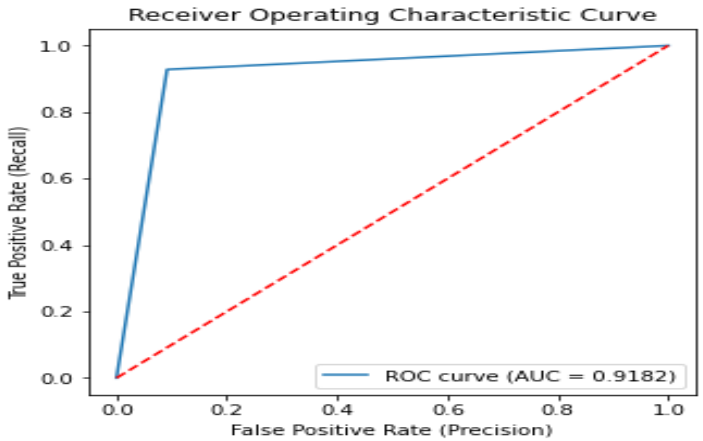


- Got slightly improvement in the model after training with cross-validation.
- All evaluation metrics improved except for recall.
- We got 0.91% of accuracy using train_test_split and 0.92 % of accuracy using cross validation in Gradient Boosting Machine model

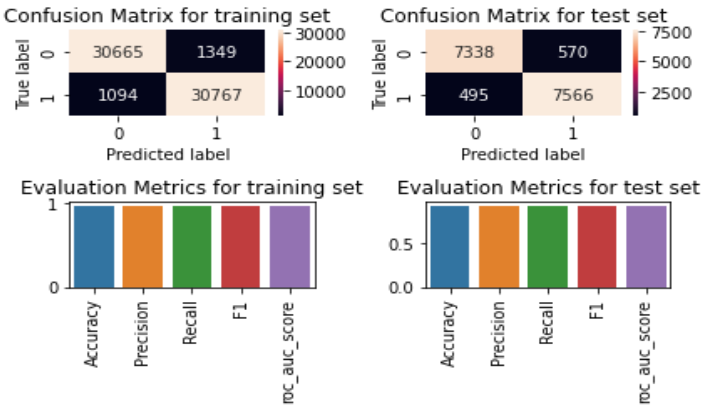
[5] XGBoost

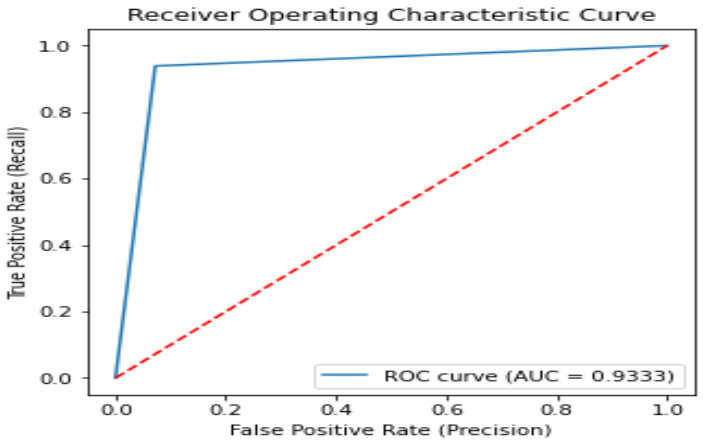
Train/Test Split results





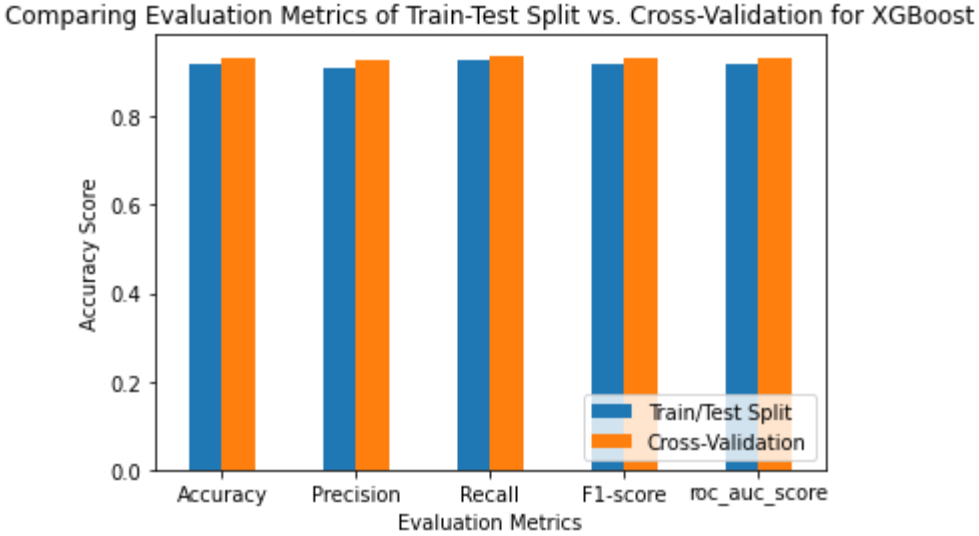
Cross-Validation results





[5] XGBoost

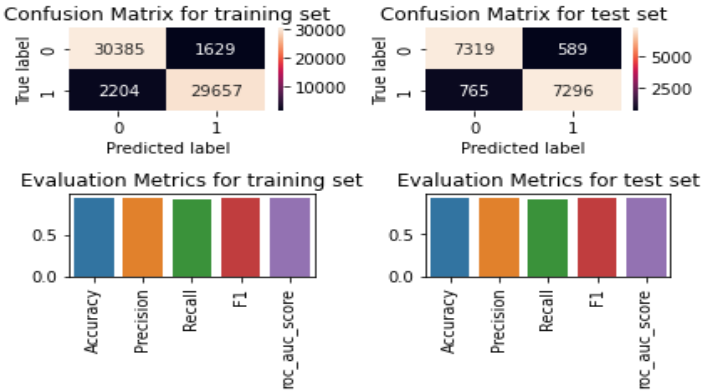
Comparing Evaluation Metrics of Train/Test Split and Cross-Validation

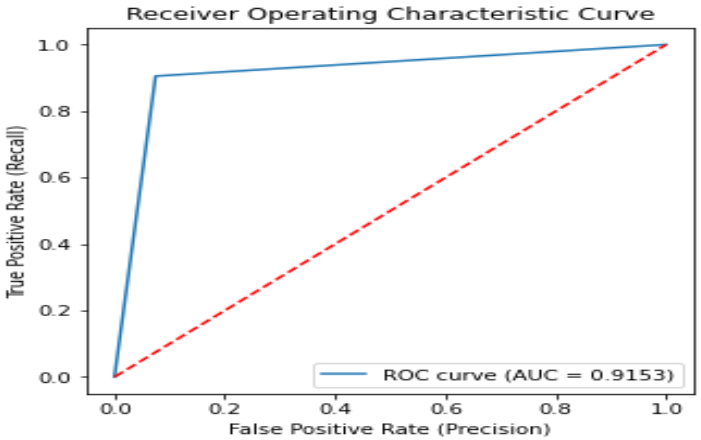


- We found improvement in the model after training model using cross-validation.
- More accuracy seen in model trained using cross validation.
- We got 0.91 % of accuracy using train_test_split and 0.93 % of accuracy using cross validation in XGBoost model.
- Major improvement found in precision which is 0.94 from 0.90.

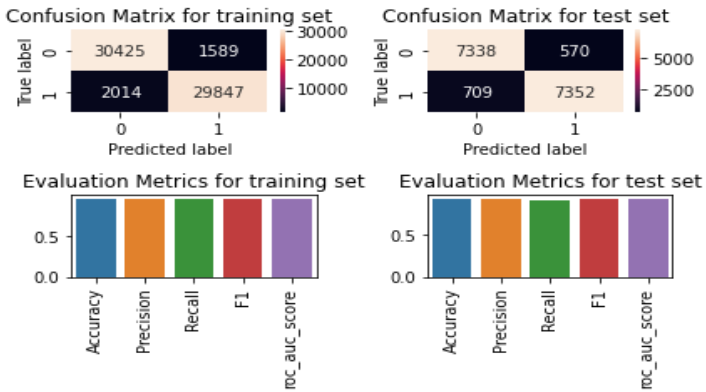
[6] K-Nearest Neighbor(KNN)

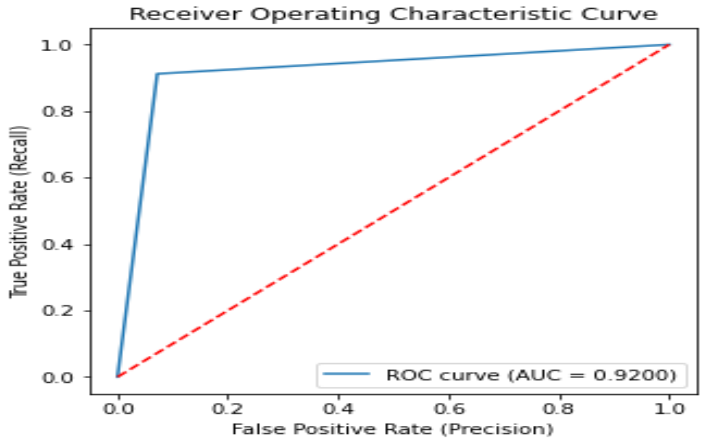
Train/Test Split results





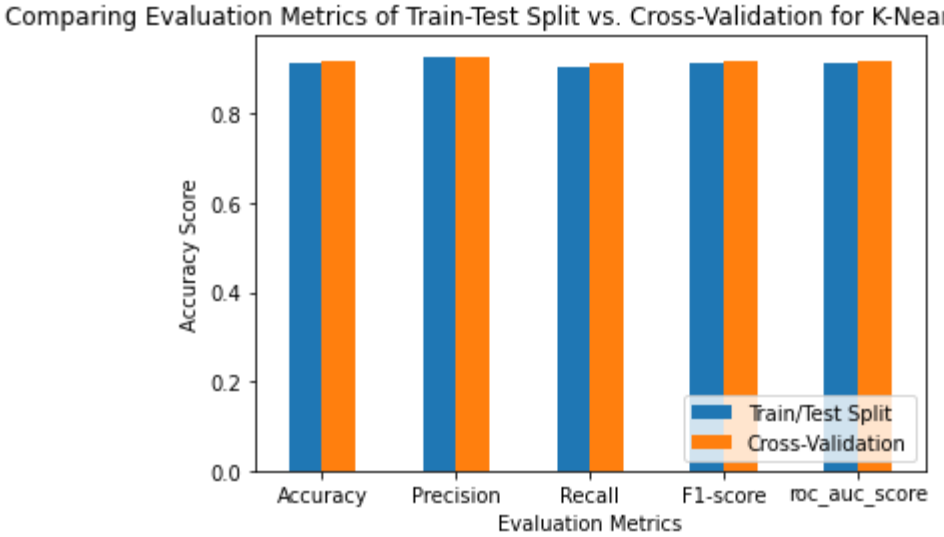
Cross-Validation results





[6] K-Nearest Neighbor(KNN)

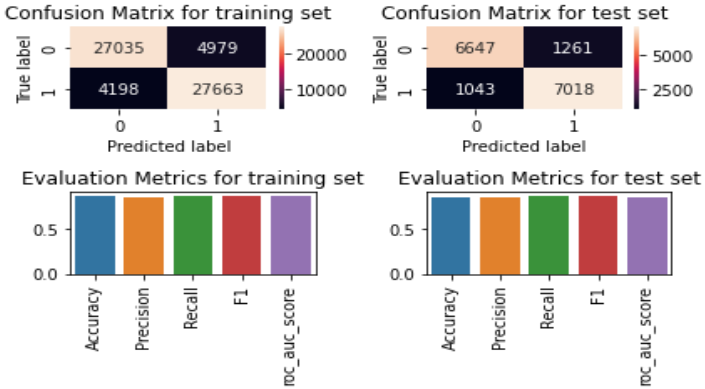
Comparing Evaluation Metrics of Train/Test Split and Cross-Validation

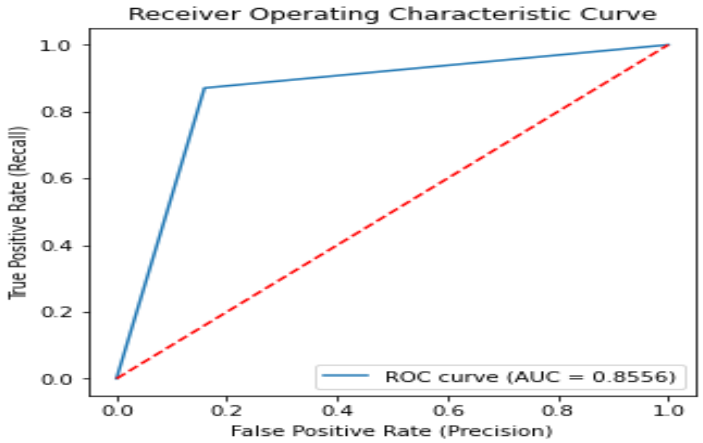


- We found slightly improvement in the model after training model using cross-validation.
- We got improvement in F1-score and recall using cross validation.
- We got improved deree of roc_auc_score in the model using cross validation over train_test_split.

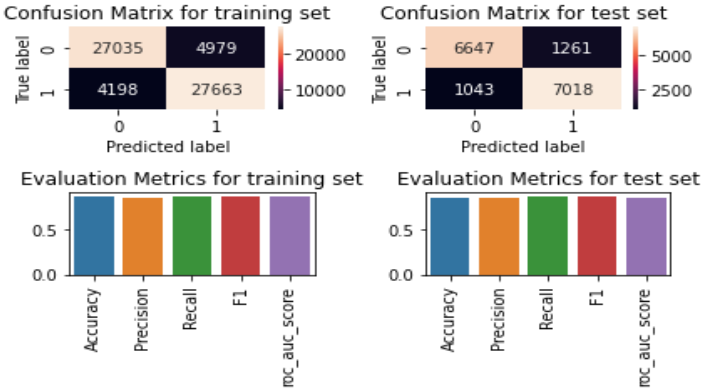
[7] Naive Bayes

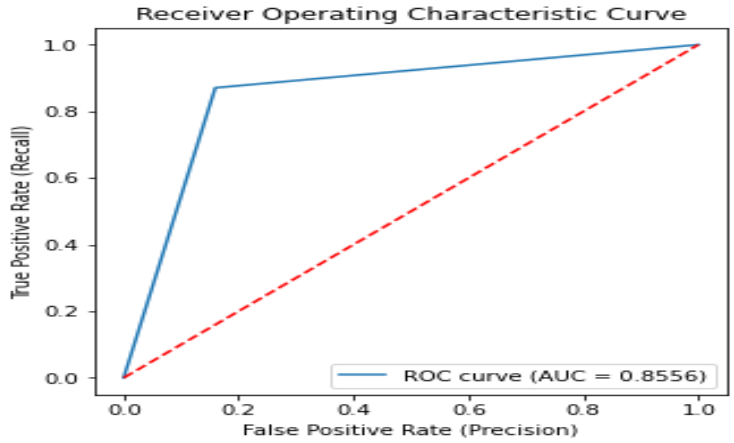
Train/Test Split results





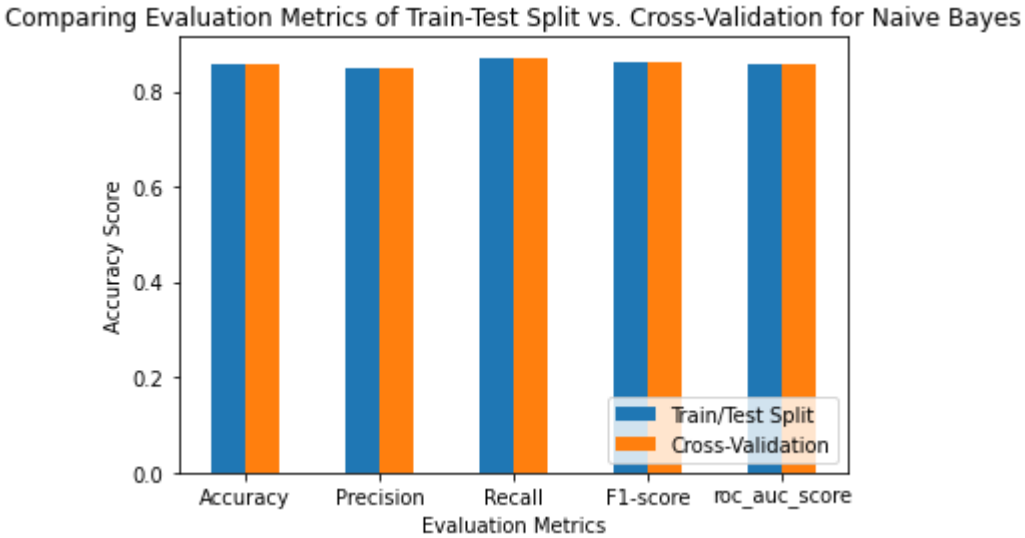
Cross-Validation results





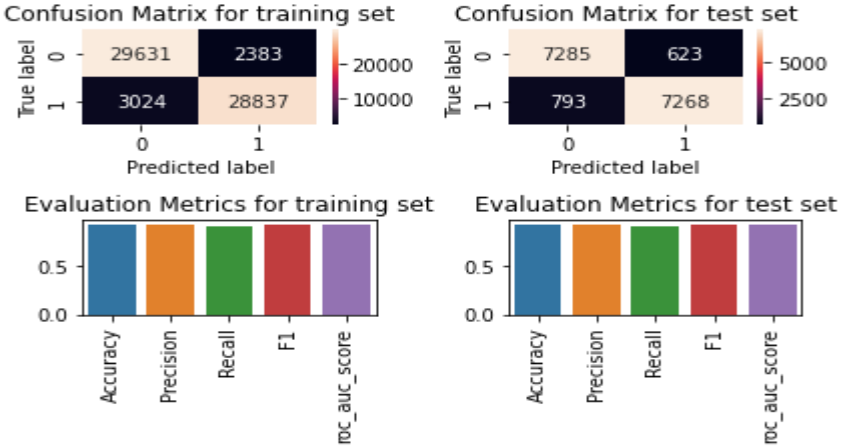
[7] Naive Bayes

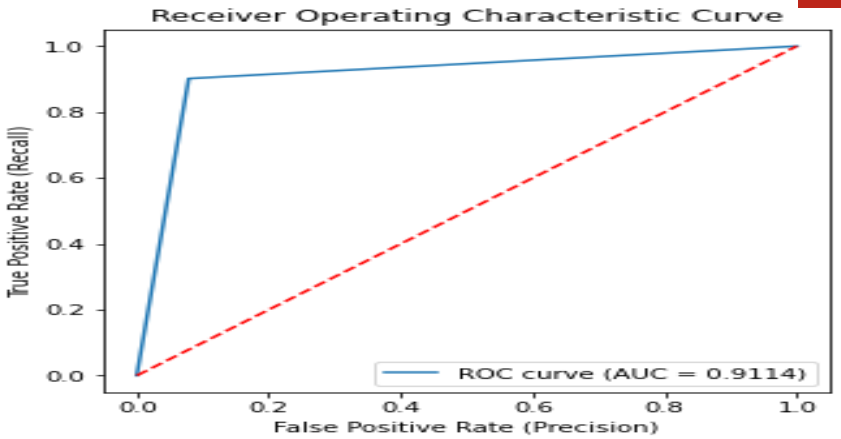
Comparing Evaluation Metrics of Train/Test Split and Cross-Validation

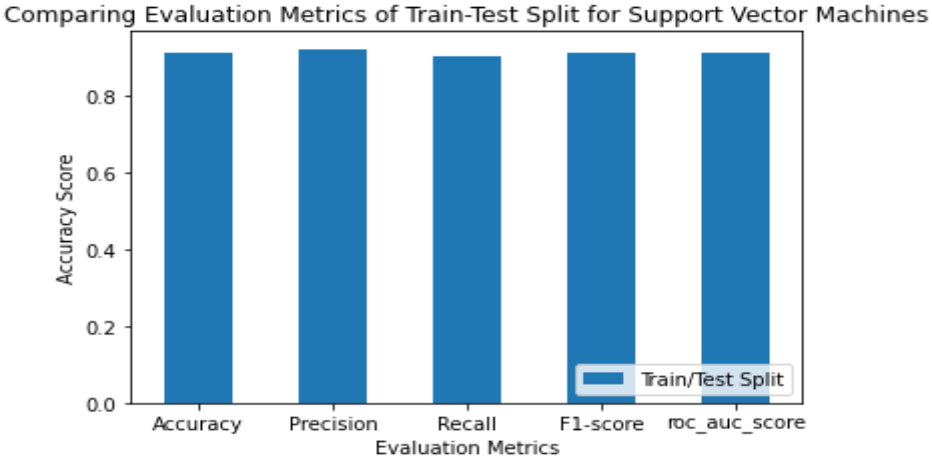


- We not found any improvement in the model after training model using cross-validation.
- We got 0.85 % of accuracy using train_test_split and 0.85 % of accuracy using cross validation in Naive Baye model.

[8] Support Vector Machine

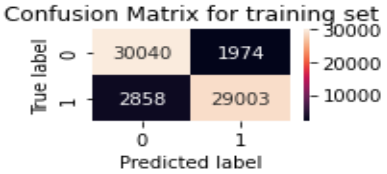


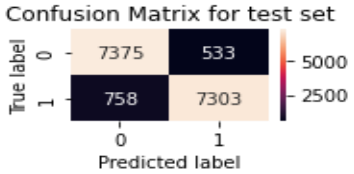


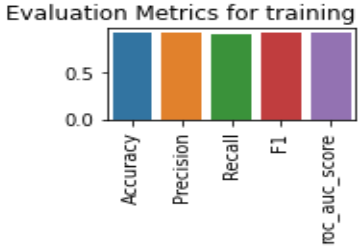


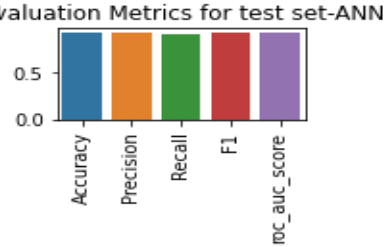
- SVM take long time to train using cross validation, so we not performed cross validation in SVM model, we skipped it.
- We got 0.91 % of accuracy using train_test_split in Support Vector Machines model.

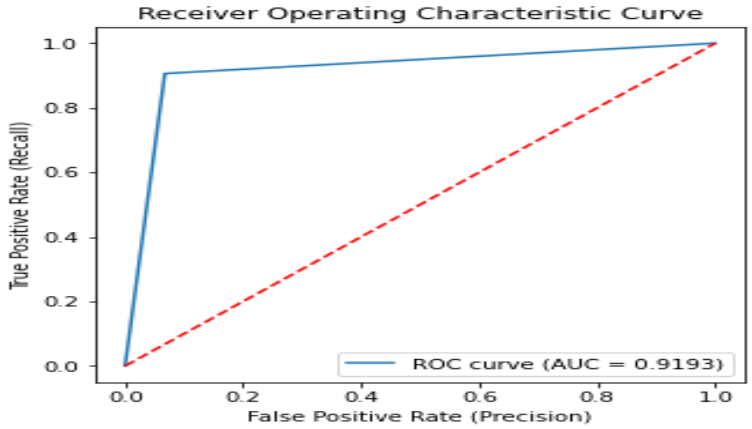
[9] Artificial Neural Networks - ANNs

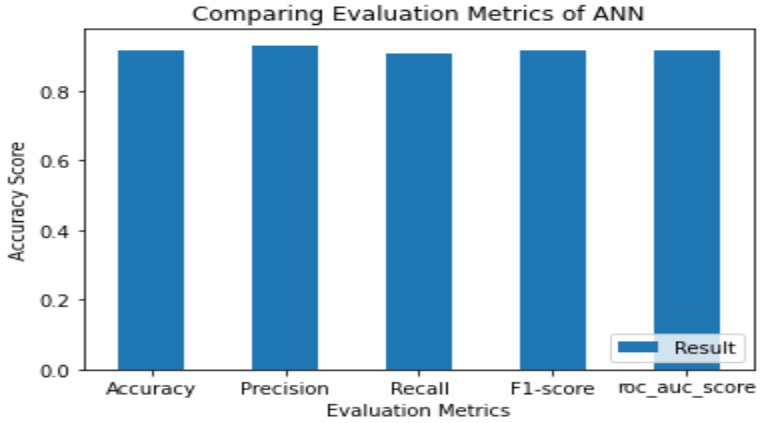










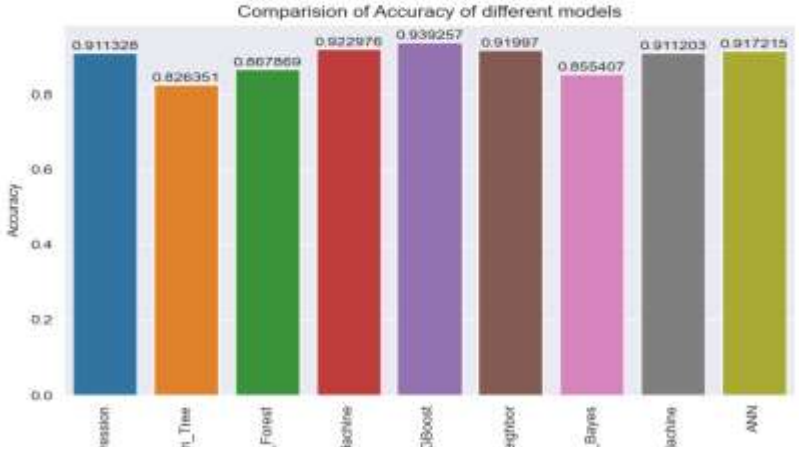


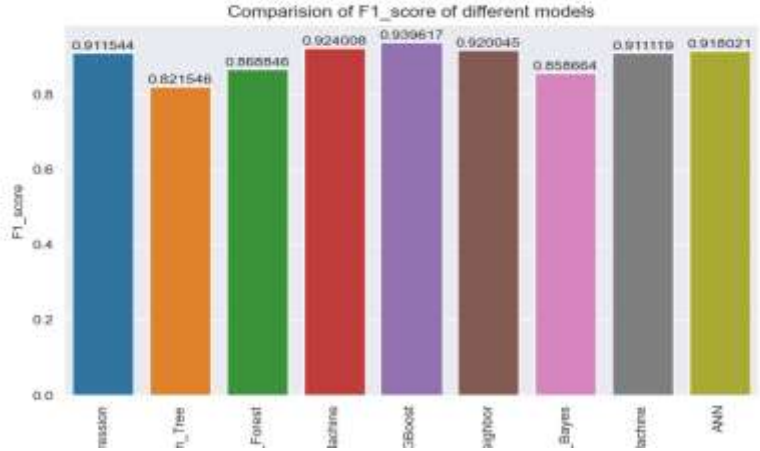
- We got 0.91 % accuracy and 0.91 F1-score in Artificial Neural Networks.

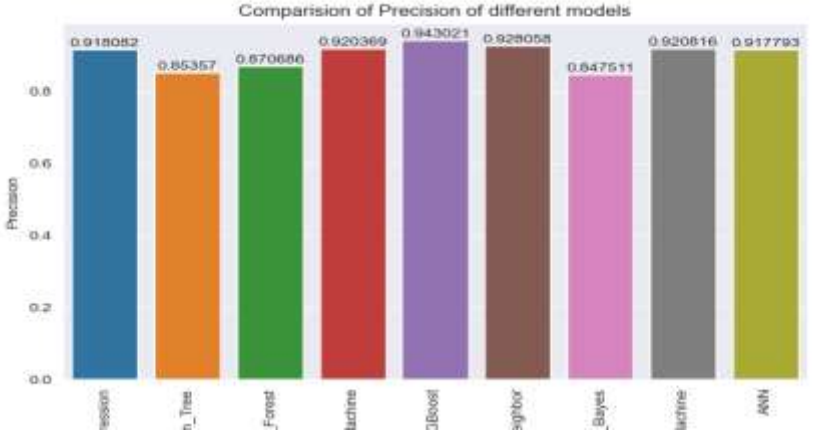
Model Evaluation

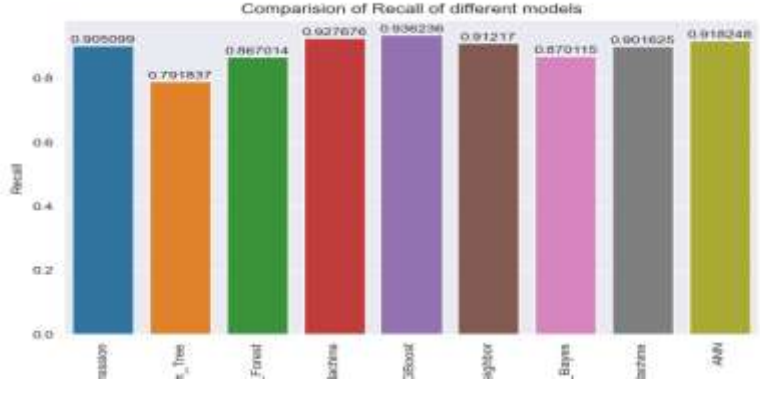
ML Model / Metrics	Logistic Regression	Decision Tree	Random Forest	Gradient Boosting Machine	XGBoost	K Nearest Neighbor	Naïve Bayes	Support Vector Machine	Artificial Neural Network
Accuracy	0.911328	0.826351	0.867869	0.922976	0.939257	0.919970	0.855407	0.911203	0.917215
Precision	0.918082	0.853570	0.870686	0.920369	0.943021	0.928058	0.847511	0.920816	0.917793
Recall	0.905099	0.791837	0.867014	0.927676	0.936236	0.912170	0.870115	0.901625	0.918248
F1 score	0.911544	0.821546	0.868846	0.924008	0.939617	0.920045	0.858664	0.911119	0.918021
Roc auc score	0.911388	0.826685	0.867877	0.922930	0.939287	0.920045	0.855265	0.911296	0.917205

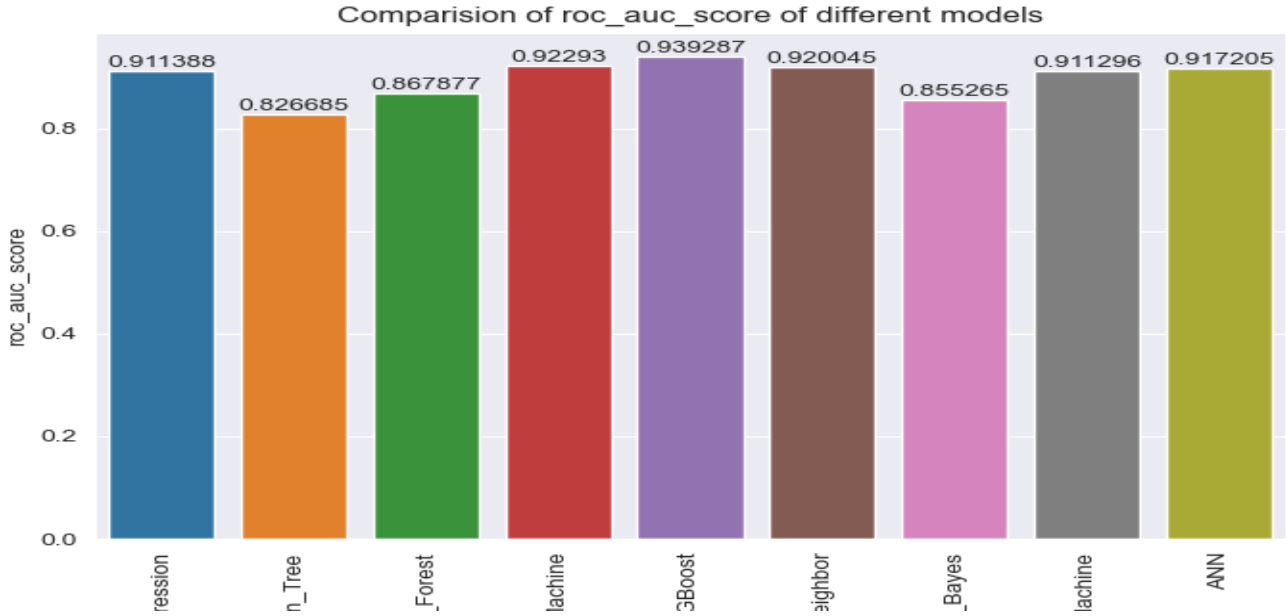
Comparison of Evaluation Metrics







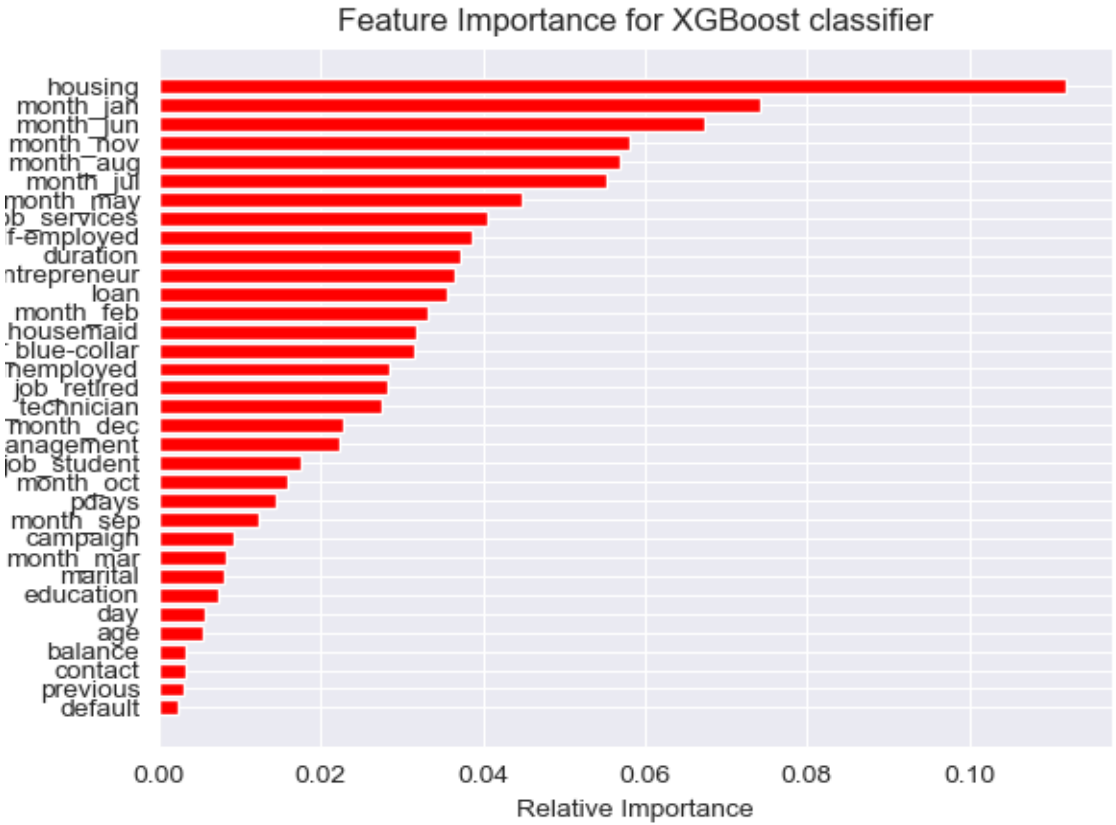




- ***Among all models, the XGBoost classification model has the highest accuracy, precision, recall, and F1-score. Also, XGBoost has a roc_auc_score of 0.93, which is very close to 1, which means the classifier is able to perfectly distinguish between classes.***

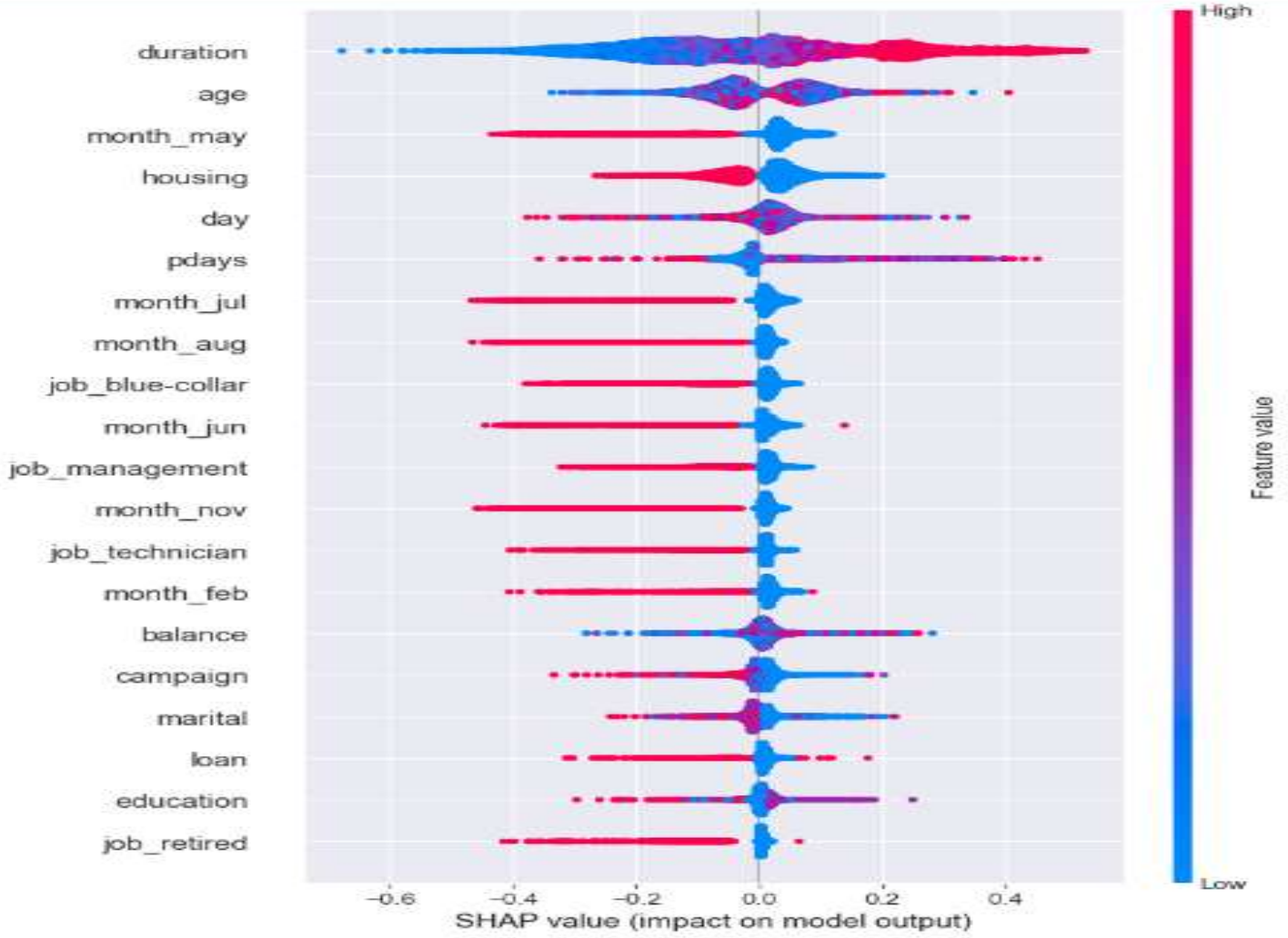
Model Explainability

Feature Importance :



- The presence or absence of a housing loan has a significant impact on the model output used to predict whether a client will subscribe to a term deposit or not.
- A higher feature importance score for features like housing, month_jan, and month_jun, month_nov, and month_aug indicates that those specific features have a greater influence on the model output used to predict whether or not a client will subscribe to a term deposit.
- Features such as day, age, balance contact, previous, and default are very less impact on the model.

Model Explainability



Observations :

- In descending order of their impact on the prediction of a model, the top five features are duration, age, month_may, housing, and day. We can get features from the shap summary plot in descending order of their impact on the prediction of a model.

- Higher values of features like month_may, housing, month_jul, month_jun, month_aug, blue-collar, management, month_feb, month_nov, job_technician, loan, campaign, education, and job_services have a negative impact on prediction, while lower values have a positive impact on the prediction.

- Higher values of feature duration have a positive impact, but lower values have a negative impact on prediction.

- Lower values of Age, day, and balance features have both positive and negative impacts on prediction.

- Overall, we can conclude that the model's prediction is positively impacted by lower values of the majority of the input features and negatively impacted by higher values of the majority of the input features.

Conclusion

- **The XGBoost classification model has the highest accuracy, precision, recall, and F1-score of all the models. Furthermore, XGBoost has a roc auc score of 0.93, which is very close to one, indicating that the classifier is perfectly capable of differentiating between classes.**

- **The XGBoost classification model trained using cross validation is the ideal model and well-trained for predicting whether the client will subscribe to a term deposit or not due to its high accuracy (0.93), precision (0.93), recall (0.93), F1 score (0.93), and rou auc score (0.93), which is close to one.**

Challenges

AI

- **Data Preprocessing:**

Data preprocessing is an essential step in any machine learning project, and we faced difficulties identifying and fixing errors in the data.

- **Feature Engineering:**

we faced difficulties choosing the right features, but it was difficult to determine which ones were most important.

- **Algorithm Selection:**

Choosing the right algorithm is critical to the success of the model. It was difficult to determine which algorithm would be most effective for a particular problem.

- **Model Training:**

Training the model requires a lot of resources and can take a long time. It was important to choose the right parameters for the model in order to achieve the best performance.

- **Model Evaluation:**

Evaluating the performance of the model is essential to determining how well it is performing. It was important to choose the right metrics for evaluating the model and to interpret the results correctly.

Thank You !