CAPSTONE PROJECT-3

Bank Marketing Effectiveness Prediction

Presented by

Minal Kharbade
Deveshya Gupta

CONTENT

- Problem statement
- Data description
- Exploratory data analysis
- Model implementation
- Evaluation metrics
- Challenges
- Conclusion

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 DESCRIPTION

Dataset contains 17 features and 45211 observations.

Categorical Features

- Job -(housemaid,retired,self employed etc.)
- Marital- (single,married,divorced)
- Education -(primary,secondary,tertiary)
- Contact- (telephone,cellular,unknown)
- Month -(Jan,Feb,March etc.)
- Poutcome -(success,failure,other,unknown)
- Housing -(yes/no)
- Loan -(yes/no)
- Default -(yes/no)

Desired target

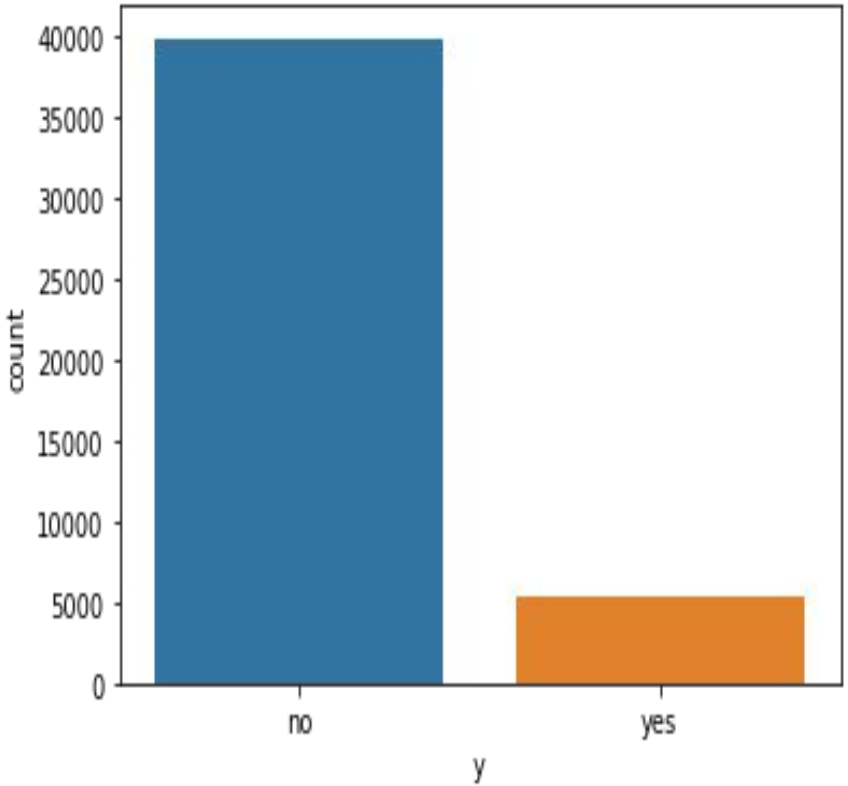
- y - has the client subscribed a term deposit? (binary: 'yes','no')

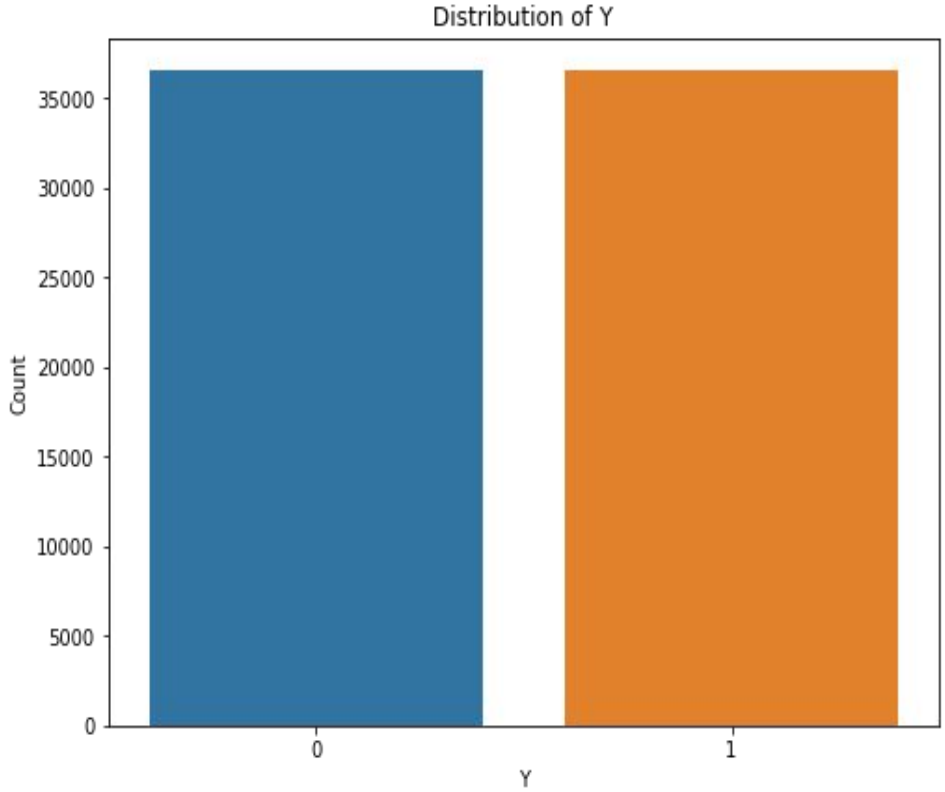
Numerical features

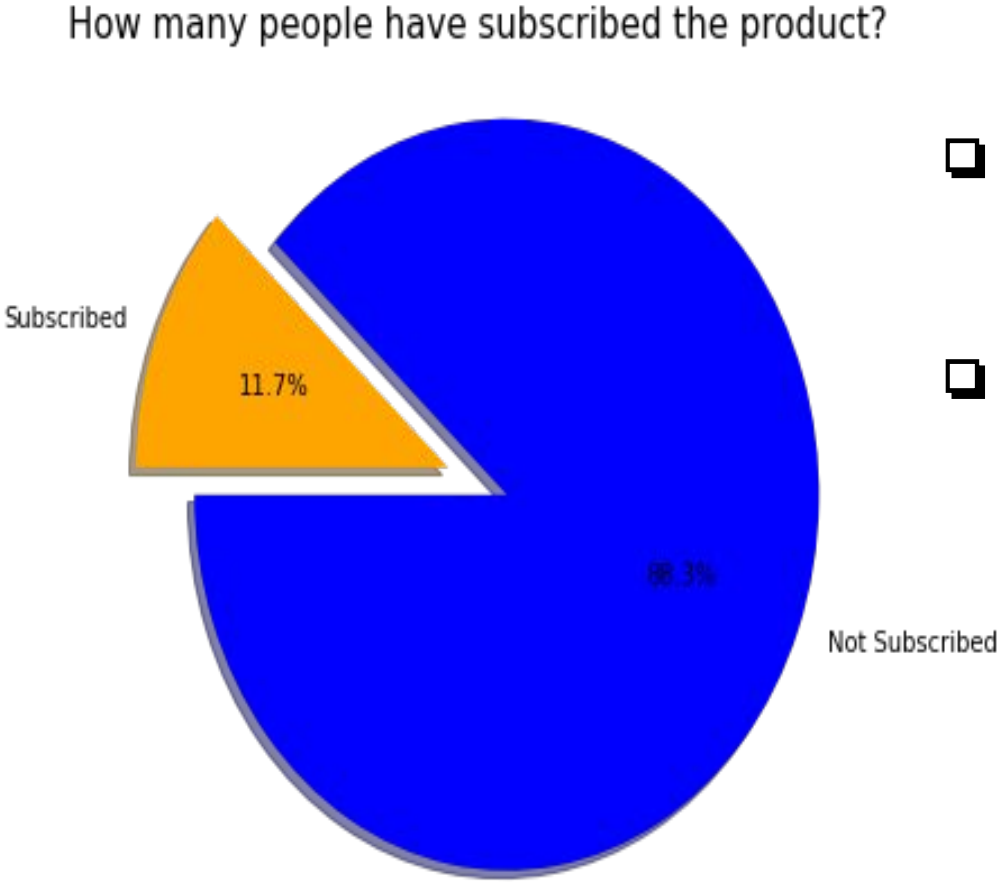
- Age
- Balance
- Day
- Duration
- Campaign
- Pdays
- Previous

EXPLORATORY DATA ANALYSIS

- **Target variable before and after**

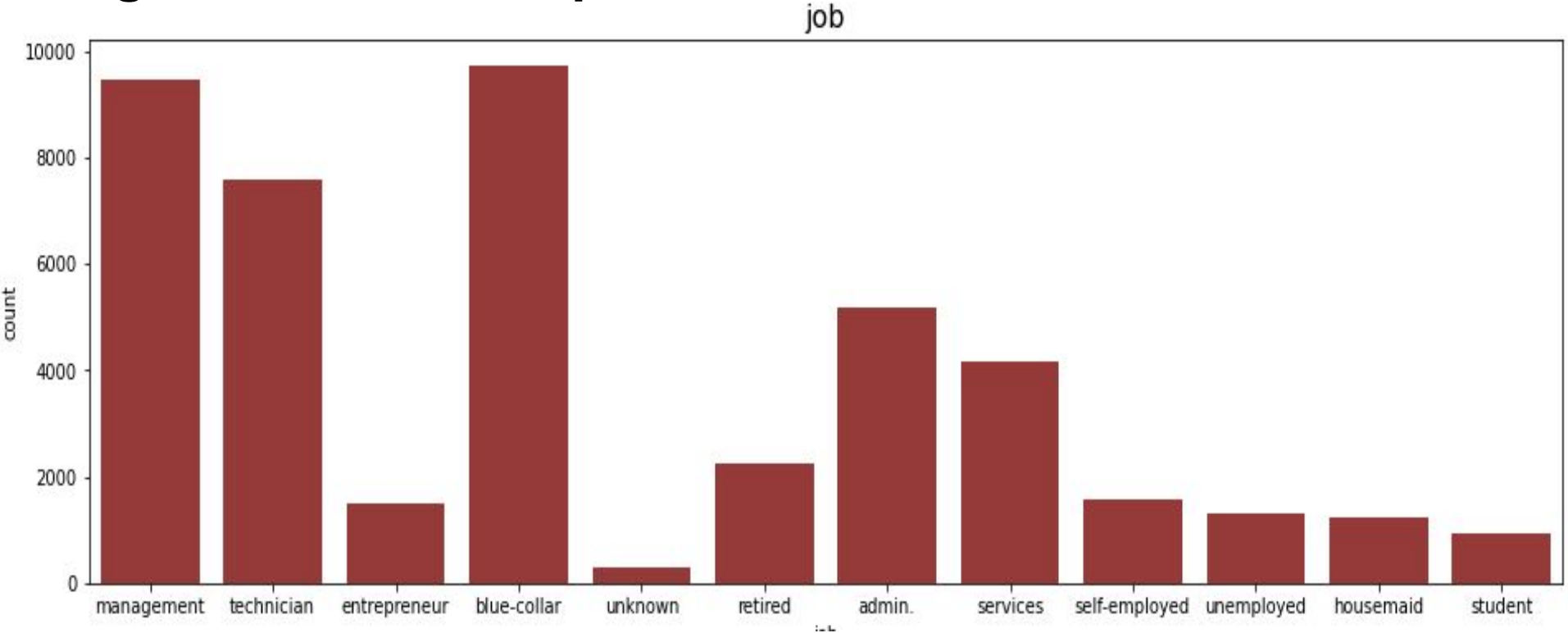






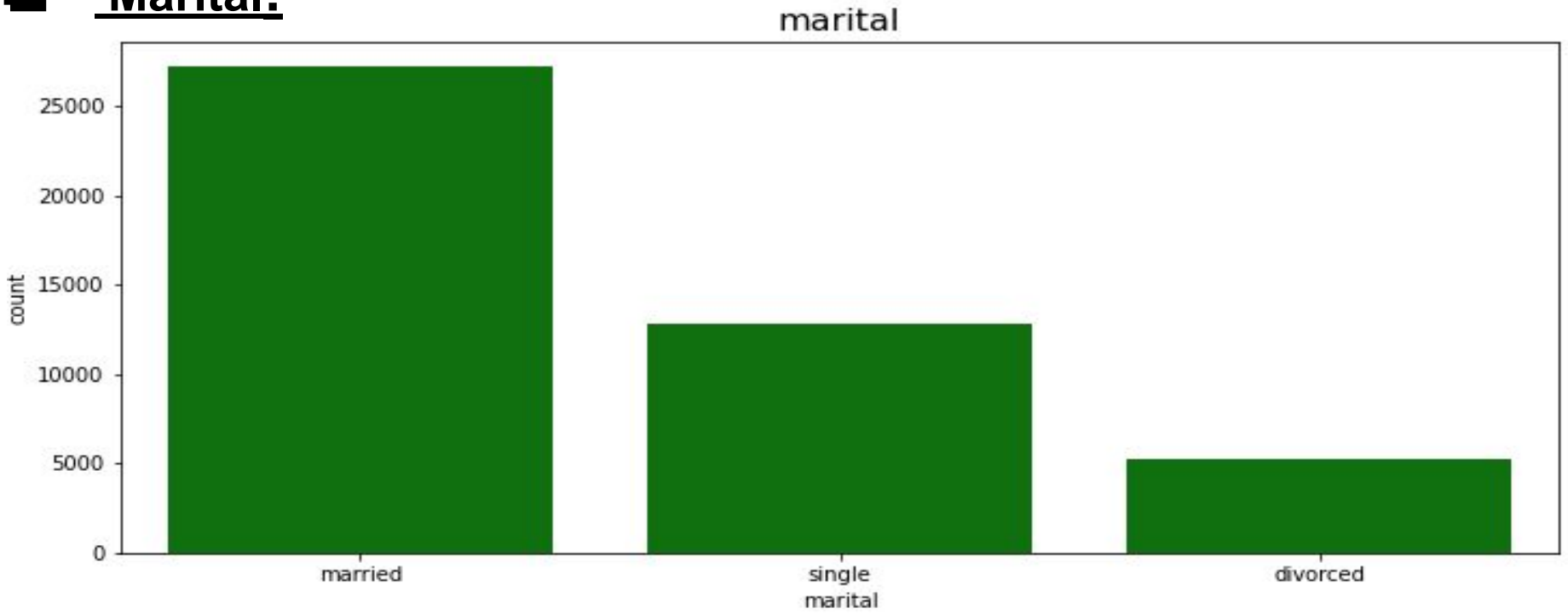
- From the given pi-chart we can see that 88% customers did not subscribe for the term deposit.
- Only 11.7% customers have subscribed for the term deposit.

Categorical Feature Exploration:



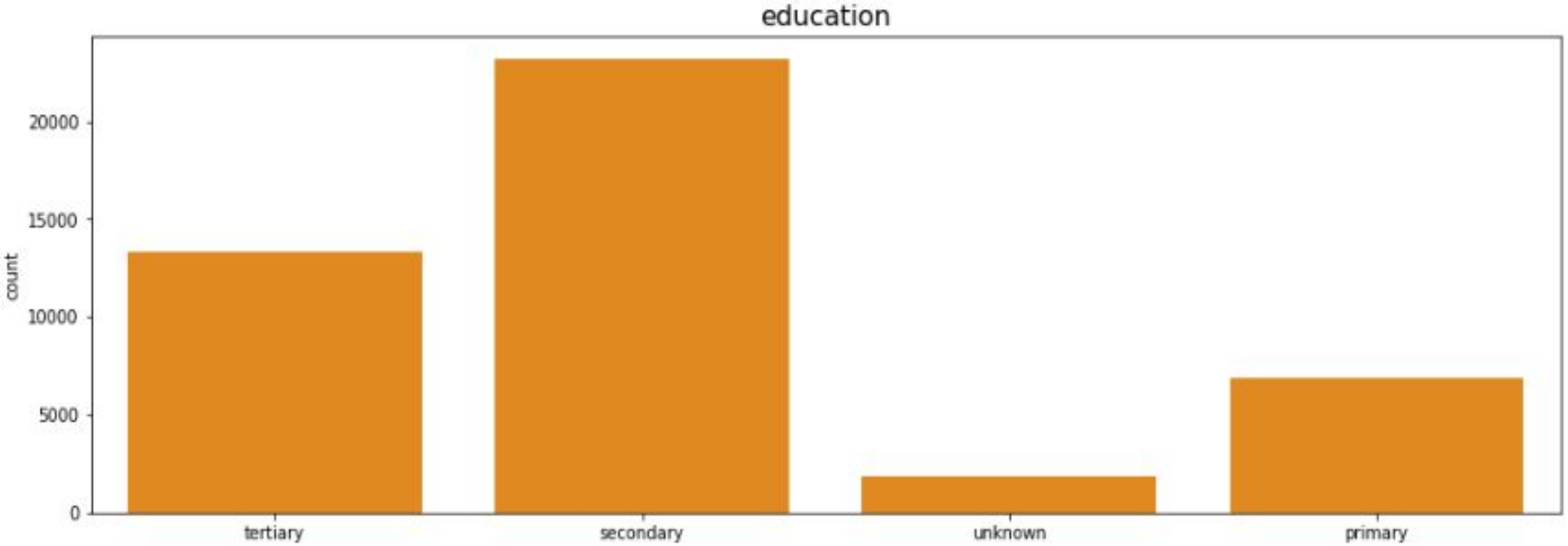
- **This bar graph shows 12 different job profile.**
- **We can say that people with blue-collar and management have subscribed more deposit than others.**
- **Management ,Technician and Blue-collar are top 3 subscriber.**

❑ Marital:



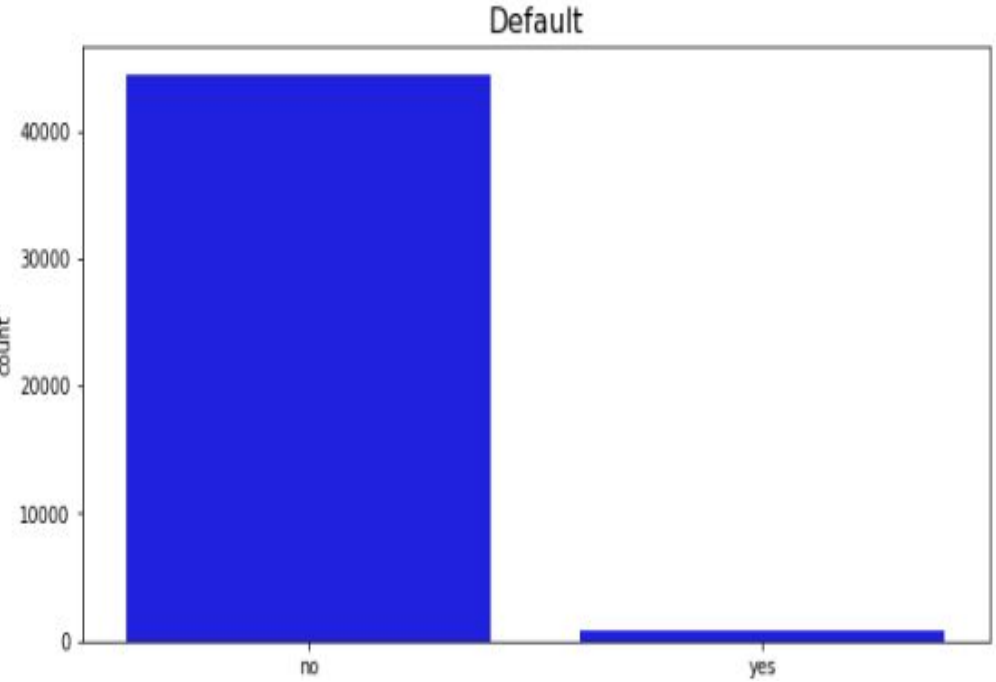
- **Approximately 60% client are married and 12%single and 12% are divorced.**
- **Client ,who are married have subscribed more than single and divorced.**

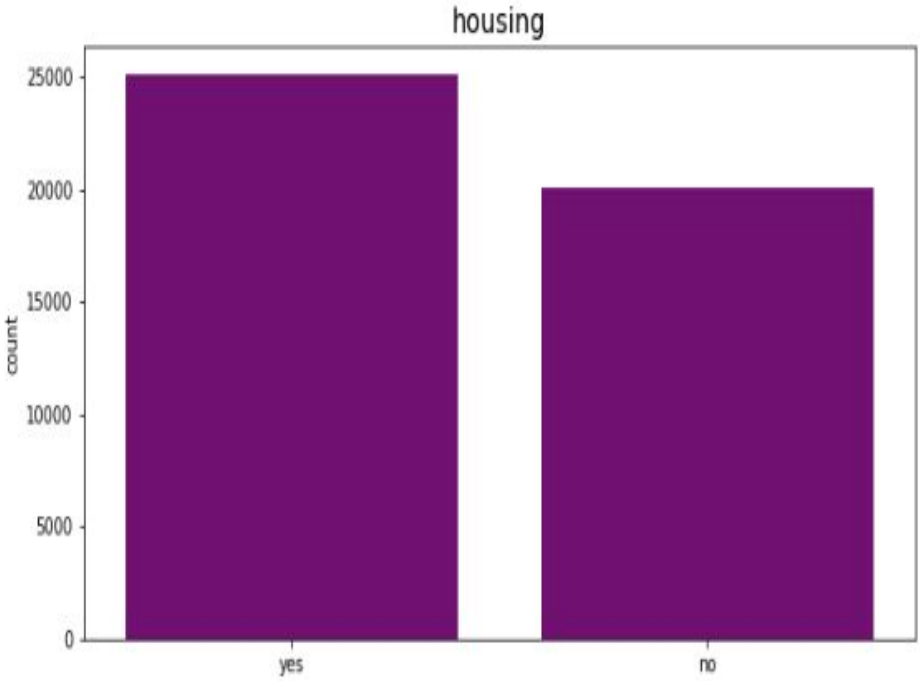
❑ Education:



- **People with secondary education qualification have subscribed more deposit than others.**

❑ Default & Housing:

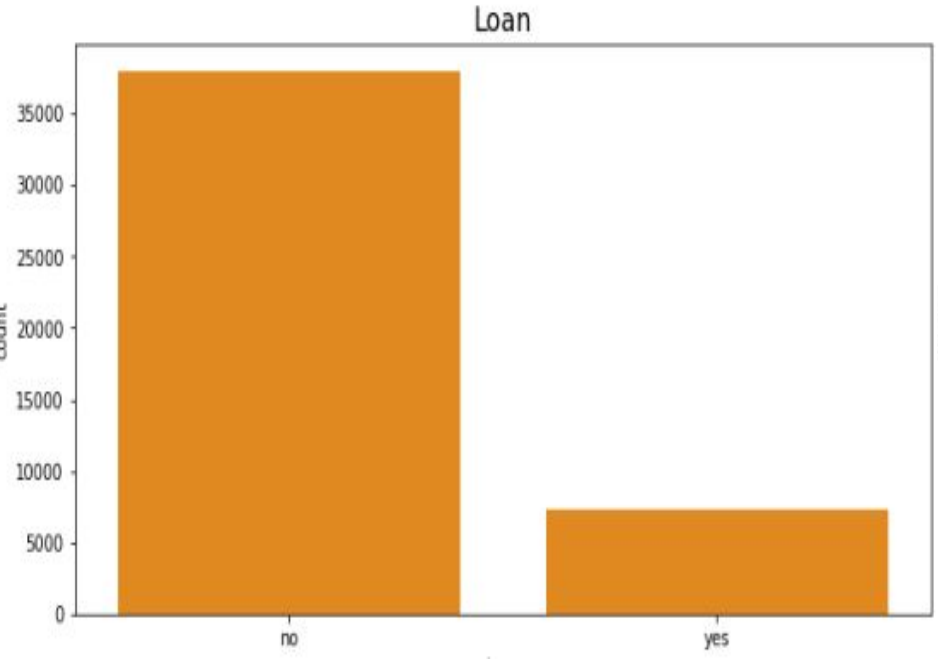


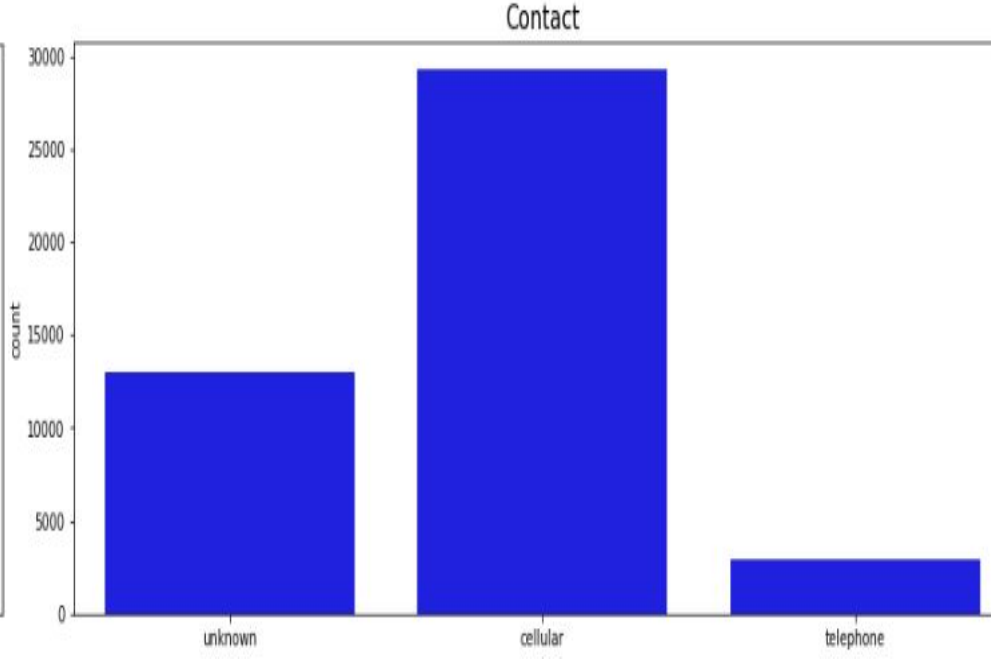


- **People with default status ‘no’ are the most one who have not subscribed for bank deposit.**

- **Most of client have taken housing loan.**
- **People with housing loan are most one contacted by bank .**

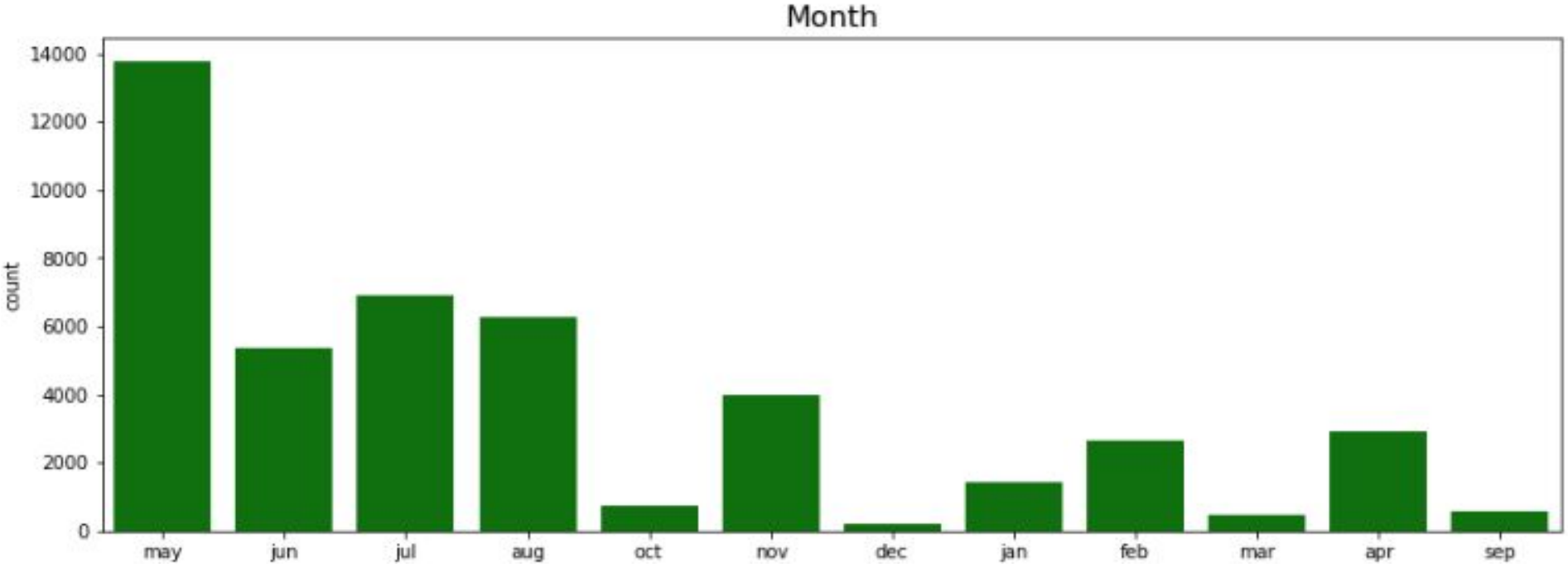
❑ Loan & Contact:





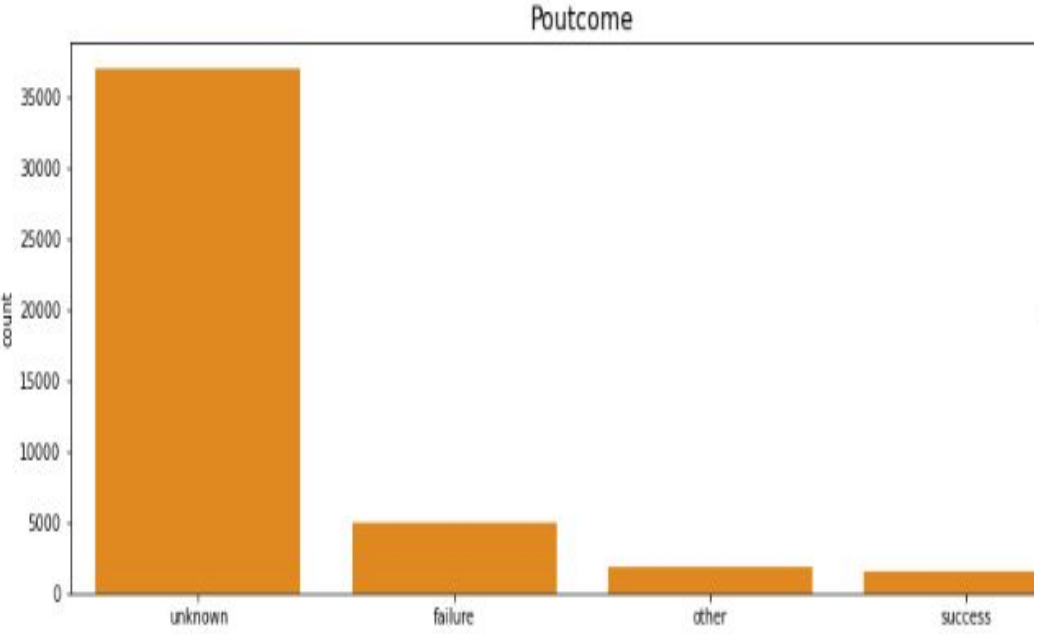
- **People with no personal loans are most one contacted by the bank for the deposit.**
- **Most people are contacted in cellular than telephone**

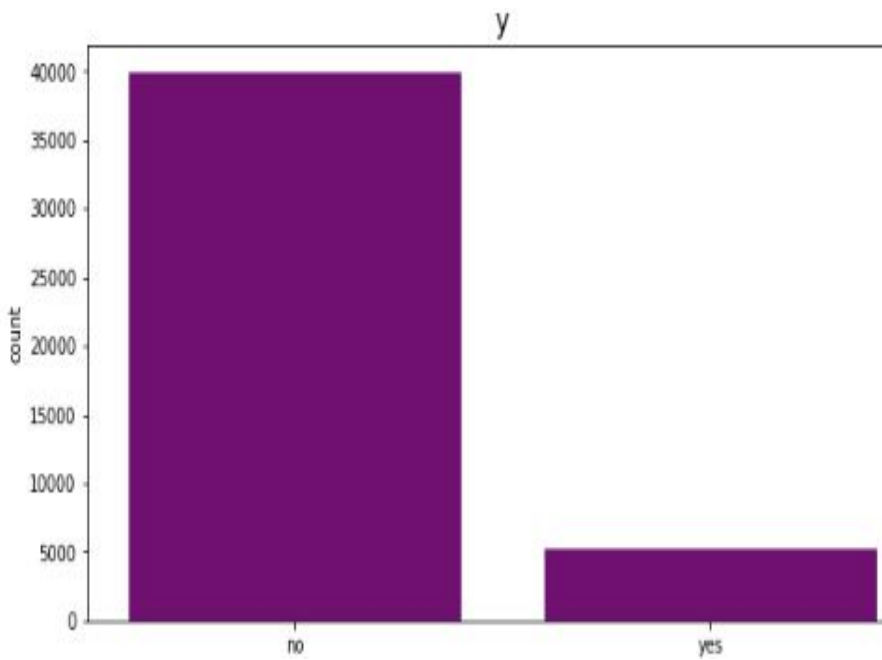
❑ Month:



- **From the graph we can say data in May is very high.**
- **Most of marketing activity was done in month of May.**

❑ Poutcome & y

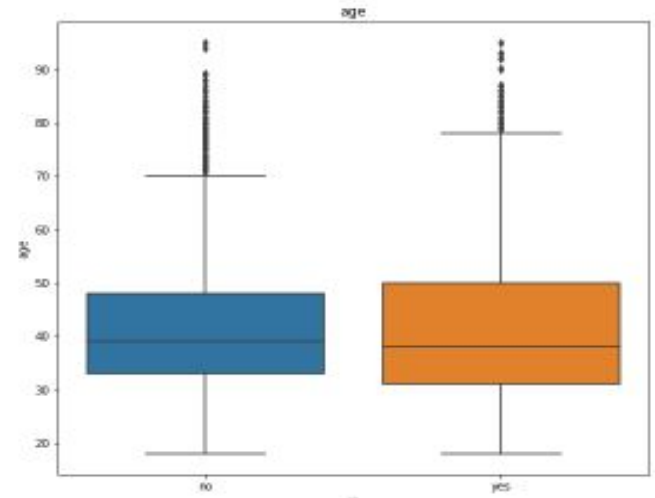


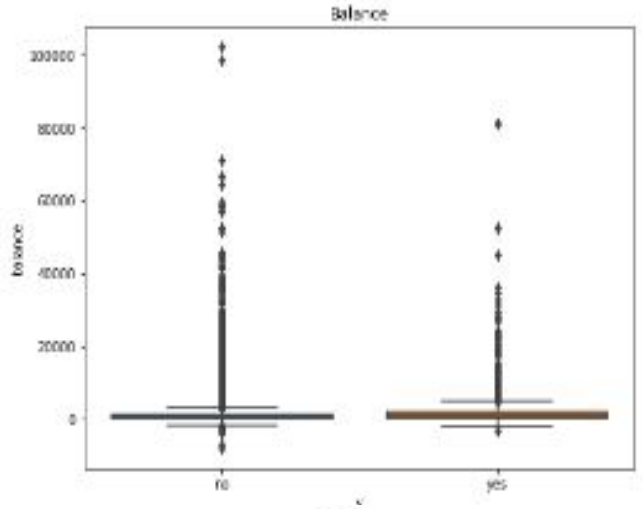


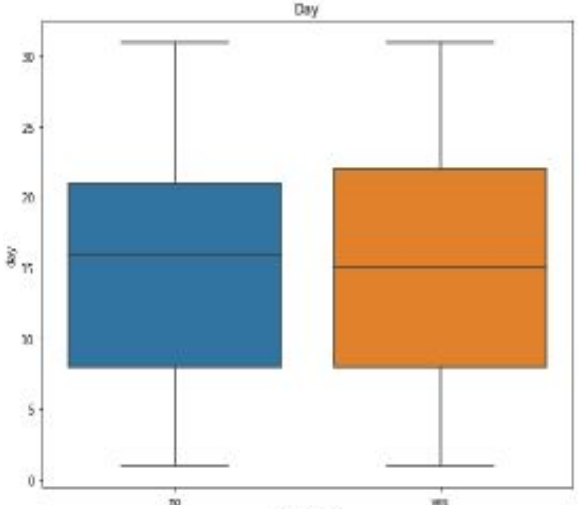
- **Majority of outcome of previous campaign is unknown .**
- **People whose previou outcome is not known have subscribed more.**

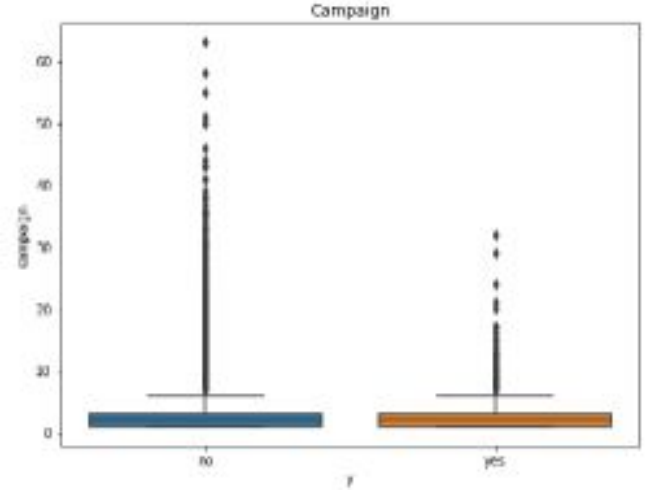
- **Most of clients have not subscribed for deposit.**

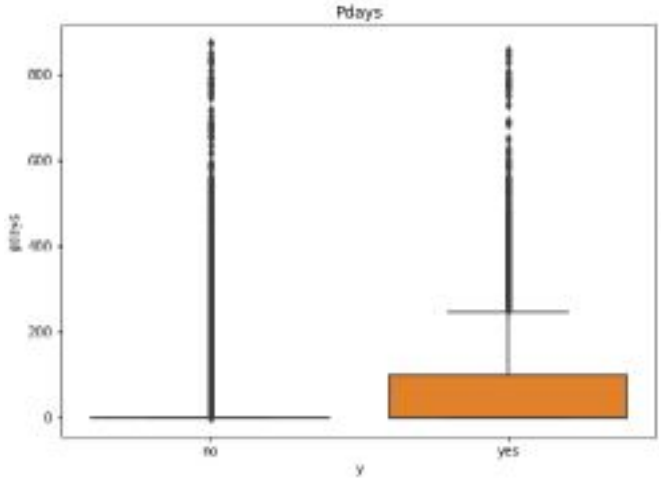
❑ Target distribution with respect numerical features:

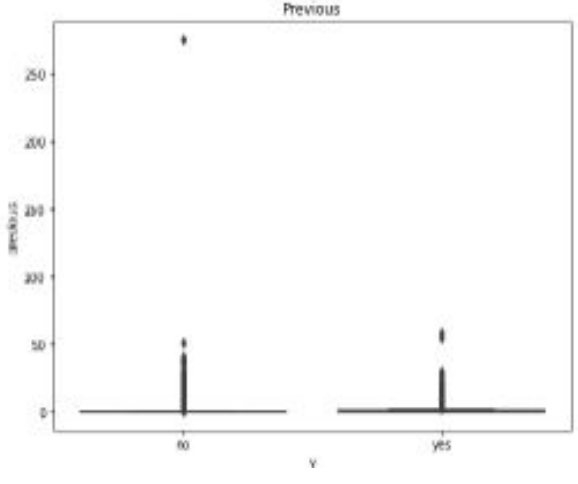




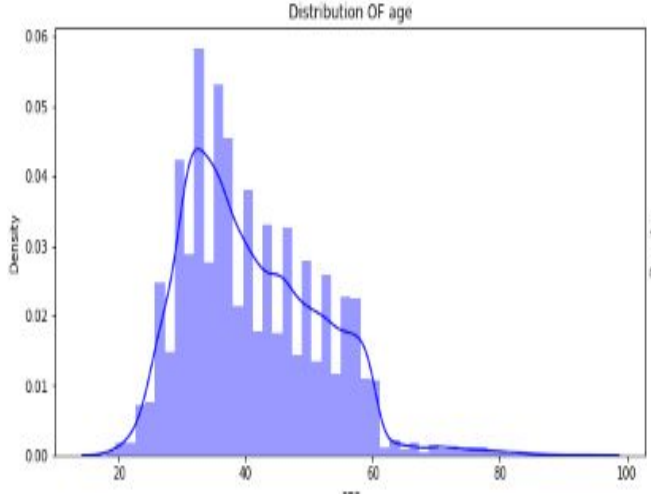


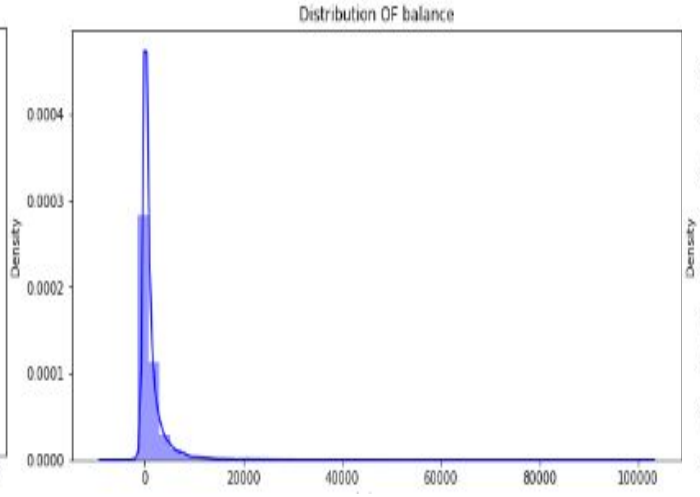


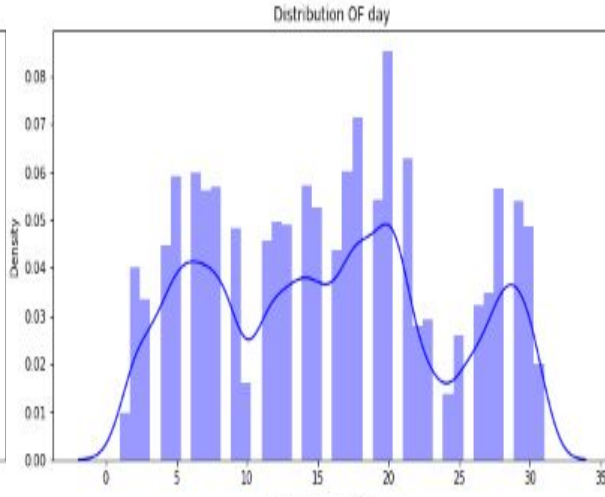


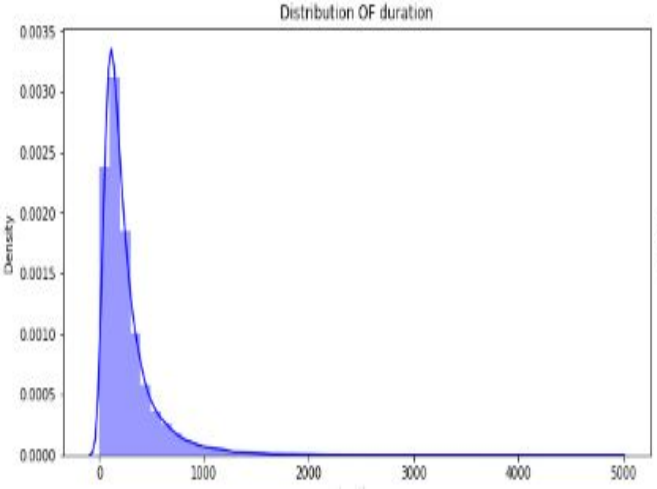


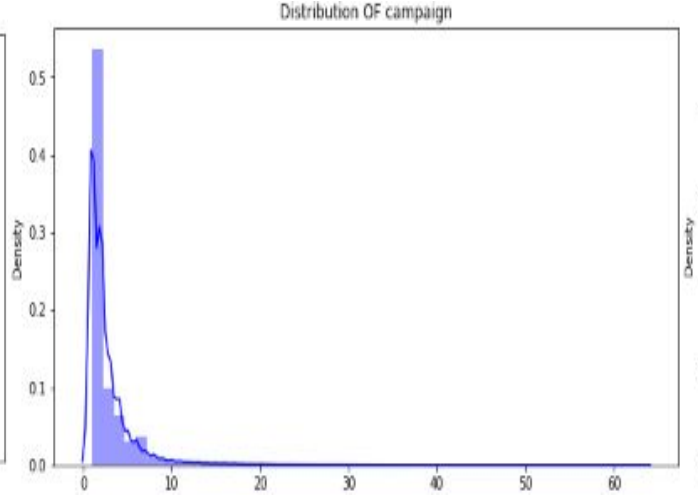
❑ Distribution Plot Of Numerical Features:

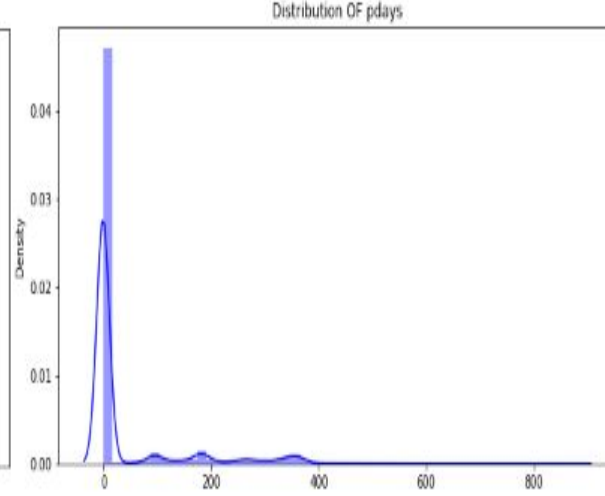




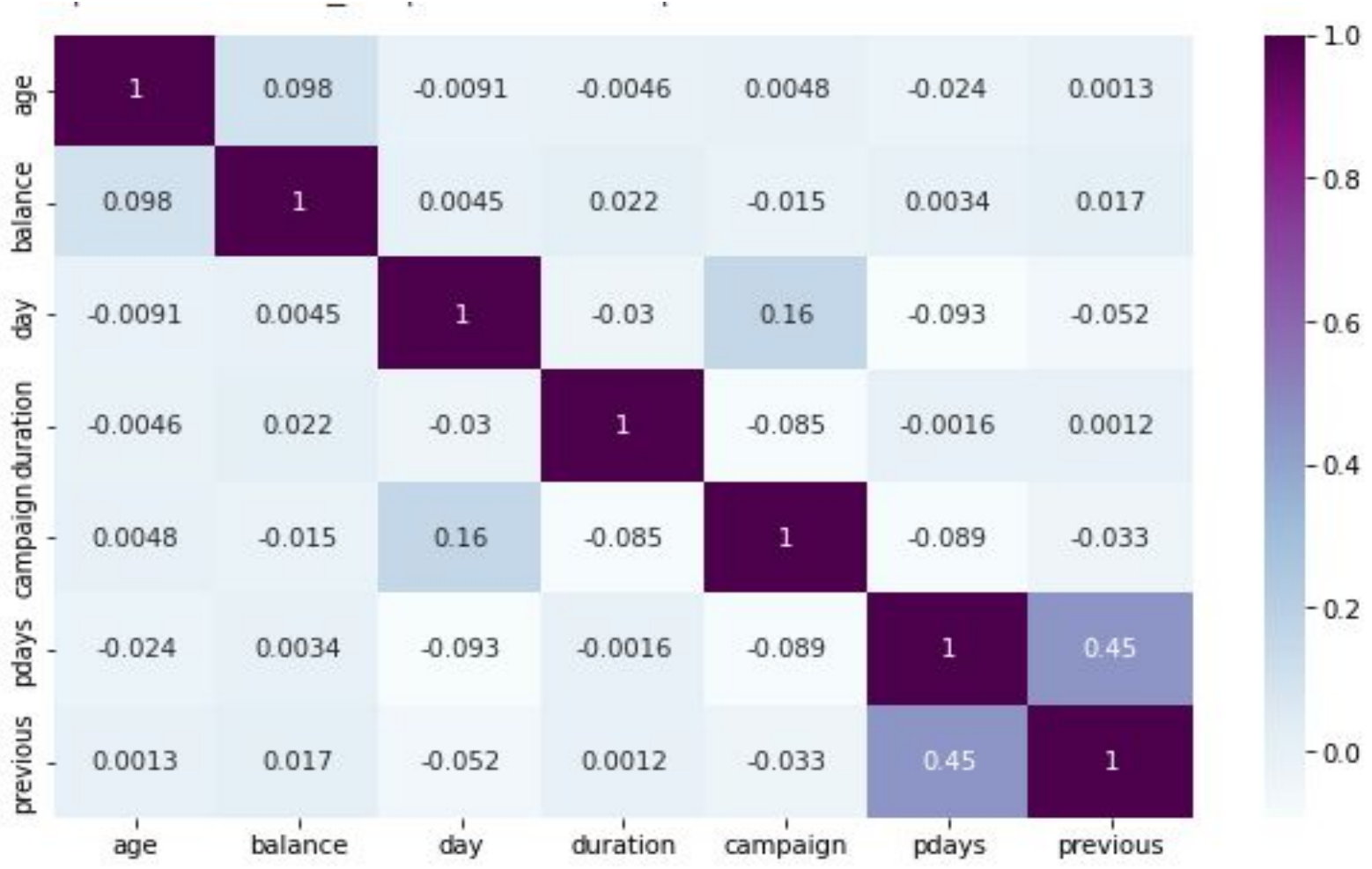








CORRELATION



MODEL IMPLEMENTATION

❑ Logistic Regression:

- **Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.**
- **Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.**
- **ROC-AUC score:**

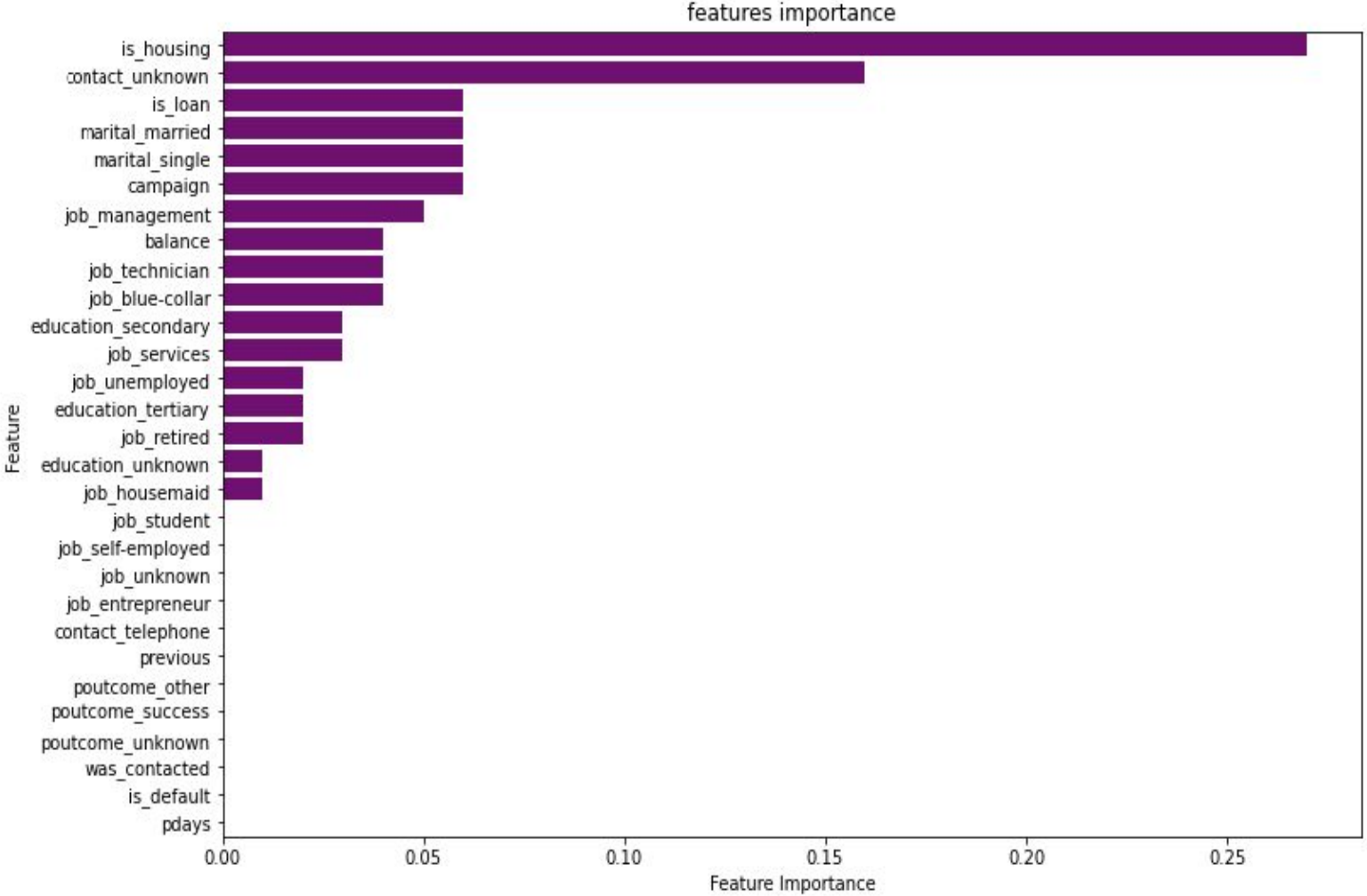
Train data	0.94
Test data	0.93

❑ Decision Tree:

- **Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems.**
- **It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.**
- **ROC-AUC score:**

Train data	0.92
Test data	0.90

Decision Tree feature importance:

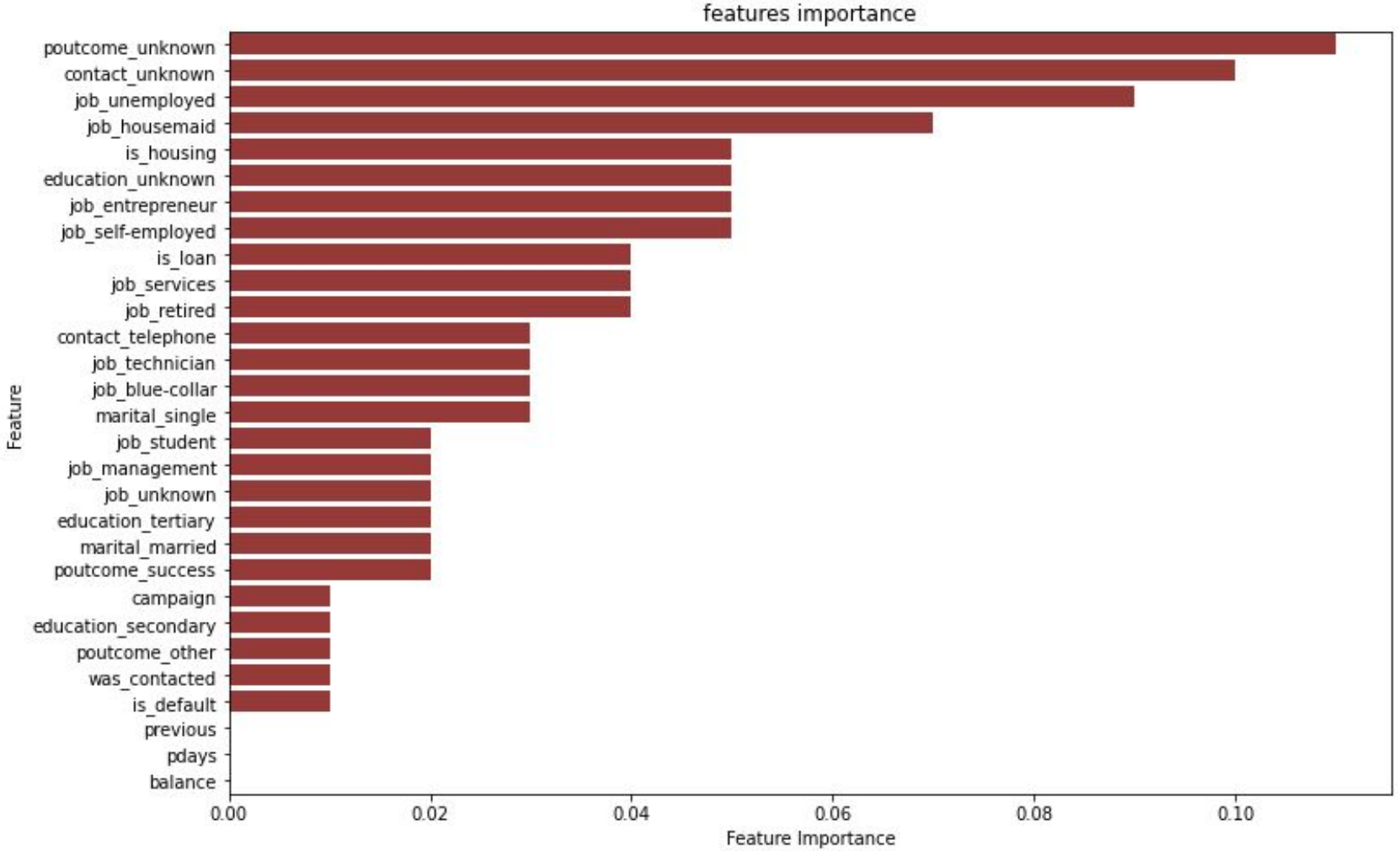


❑ XG boost classifier:

- **XGBoost is an efficient implementation of gradient boosting that can be used for regression predictive modeling.**
- **In this algorithm decision trees are created in sequential form.**
- **ROC-AUC score:**

Train data	0.95
Test data	0.91

XGboost feature importance:



❑ K-nearest neighbor:

- **The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems.**
- **The KNN algorithm can compete with the most accurate models because it makes highly accurate predictions.**
- **ROC-AUC score:**

Train data	0.95
Test data	0.93

Hyperparameter Tuning Evaluation:

Model	Test AUC	Test Accuracy	F1 score	Precision
Logistic Regression	0.93	0.86	0.87	0.89
Decision Tree	0.90	0.87	0.84	0.85
XG boost	0.91	0.86	0.91	0.92
K-NN	0.93	0.88	0.88	0.90

CHALLENGES

- Optimising the model
- Feature Engineering
- Handling Imbalanced Dataset

CONCLUSION

- For age , most of the customers are in the age range of 30-40.
- For balance , above 1000$ is like to subscribe a term deposit .
- The model can help to classify the customers on the basis on which they deposit or not.
- The model helps to target the right customer rather than wasting time on wrong customer.
- Comparing to all algorithms XGboost algorithm has best accuracy score and ROC-AUC score . So it is concluded as optimal model.

THANK YOU