Capstone Project – 3

Bank Marketing Effectiveness Prediction

Content

- Problem Statement
- Data Summary
- Exploratory Data Analysis
- Data Preprocessing
- Sampling
- Model Implementation
- Conclusion

Problem Statement

- **Aim:- Predicting the effectiveness of bank marketing campaign**

- **Problem Statement:** The data is related with direct marketing campaigns 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 Summary

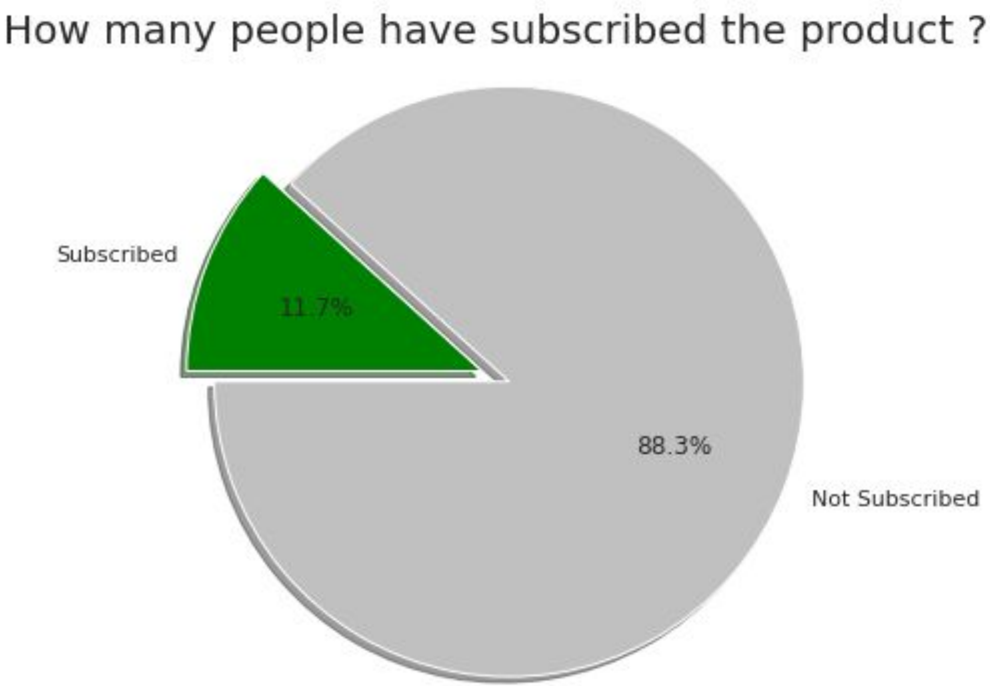
Categorical Features

- Marital - (Married , Single , Divorced)
- Job-(Management,BlueCollar,retired etc)
- Contact - (Telephone,Cellular,Unknown)
- Education (Primary,Secondary,Tertiary)
- Month-(Jan,Feb,Mar,Apr,May etc)
- Poutcome - (Success,Failure,Other,Unknown)
- Housing - (Yes/No)
- Loan - (Yes/No)
- Default - (Yes/No)

Numerical Features

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

Exploratory Data Analysis



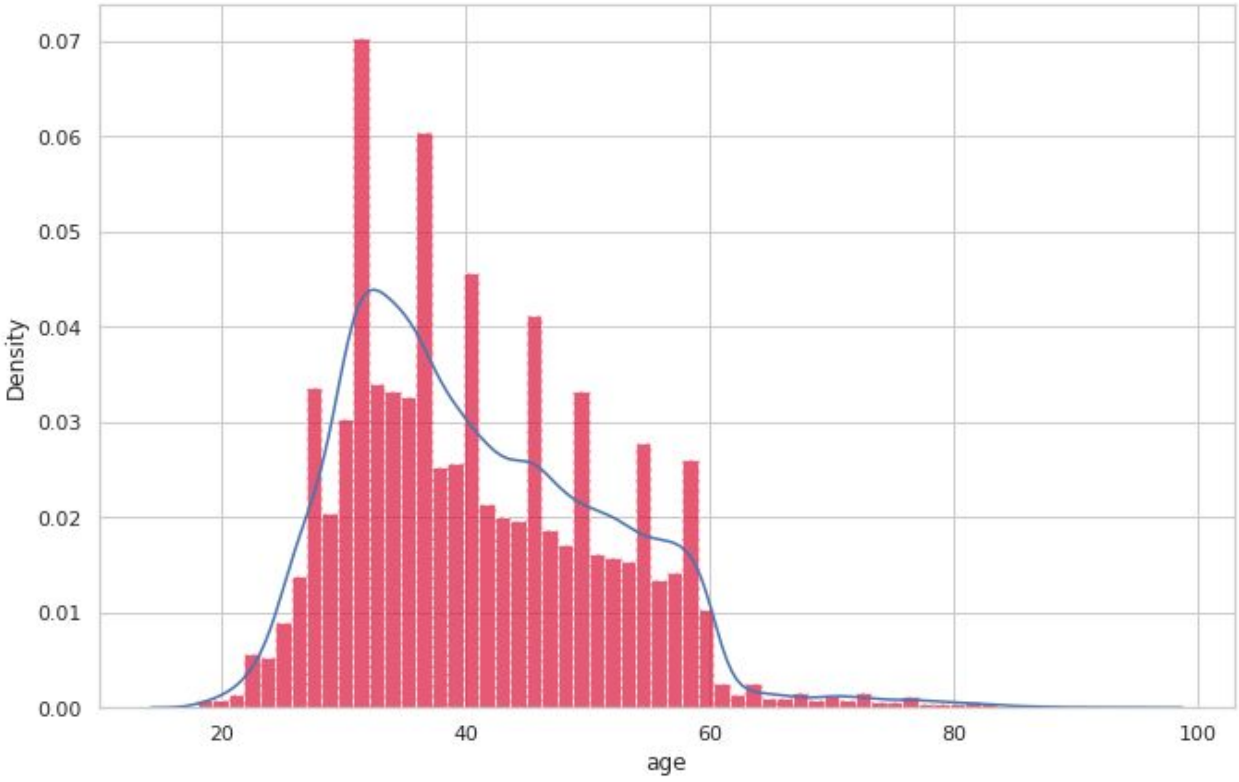
How many people have subscribed the product ?

- The target variable ‘y’ tells us the outcome of the campaign whether they went ahead for the term deposit or not.

- Out of 45211 only 5289 people subscribed to the term deposit.

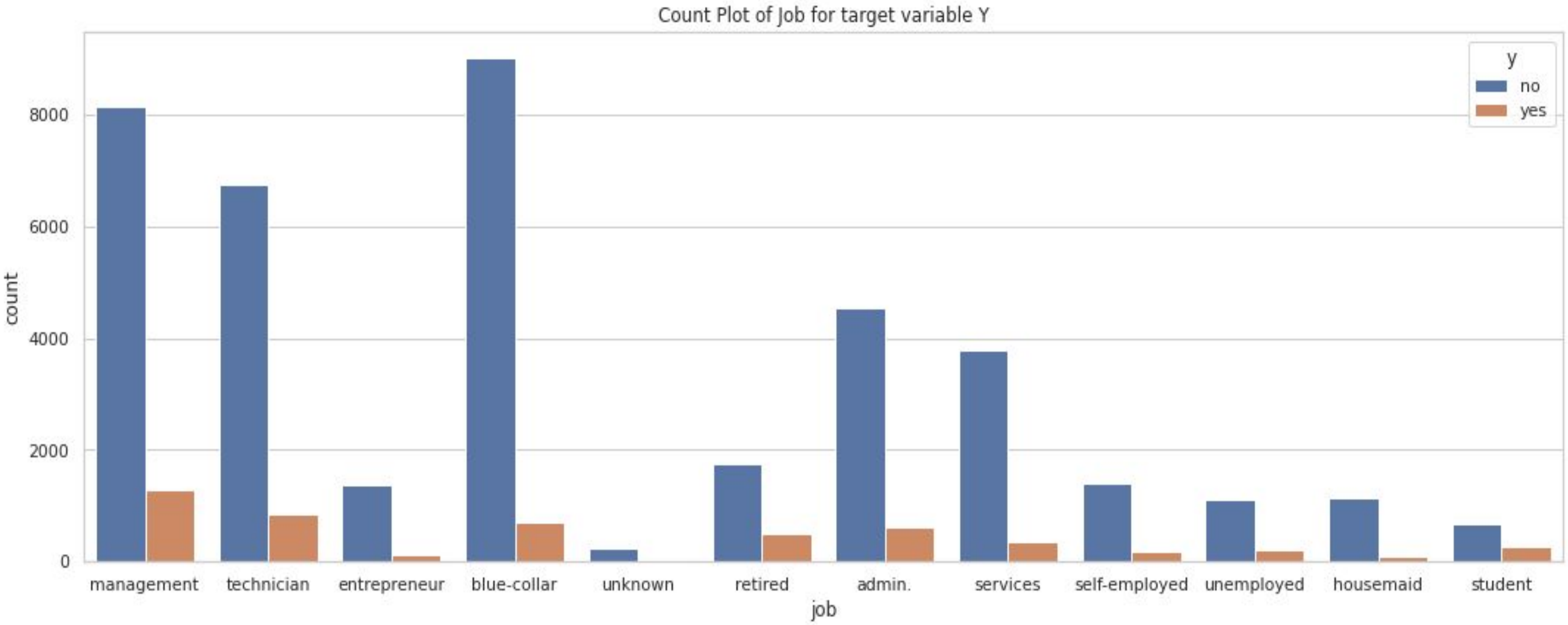
EDA(continued)

Age distribution in dataset. This shows that campaign is more centered to 30-50 age group.



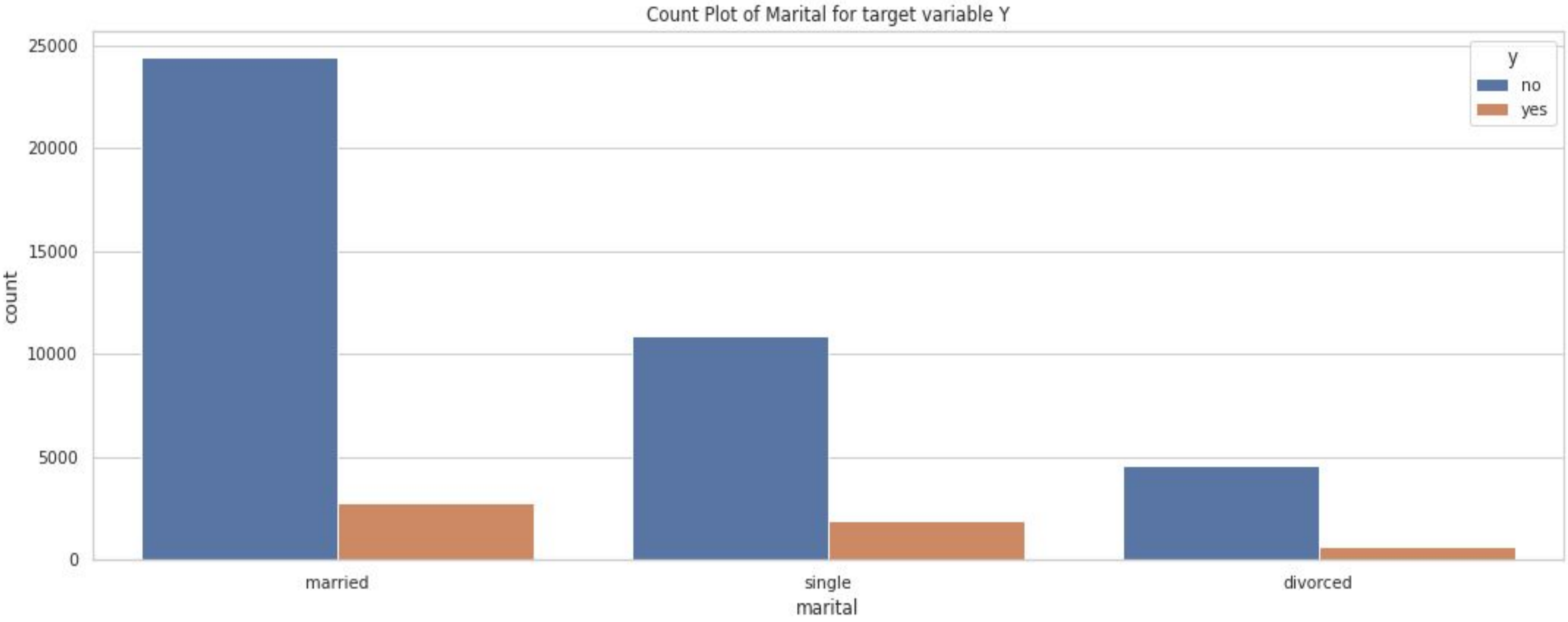
EDA(continued)

Job types v/s target variable. People with management jobs have the most number of term deposit and blue collar people rejected the most. Students have >28% success rate.



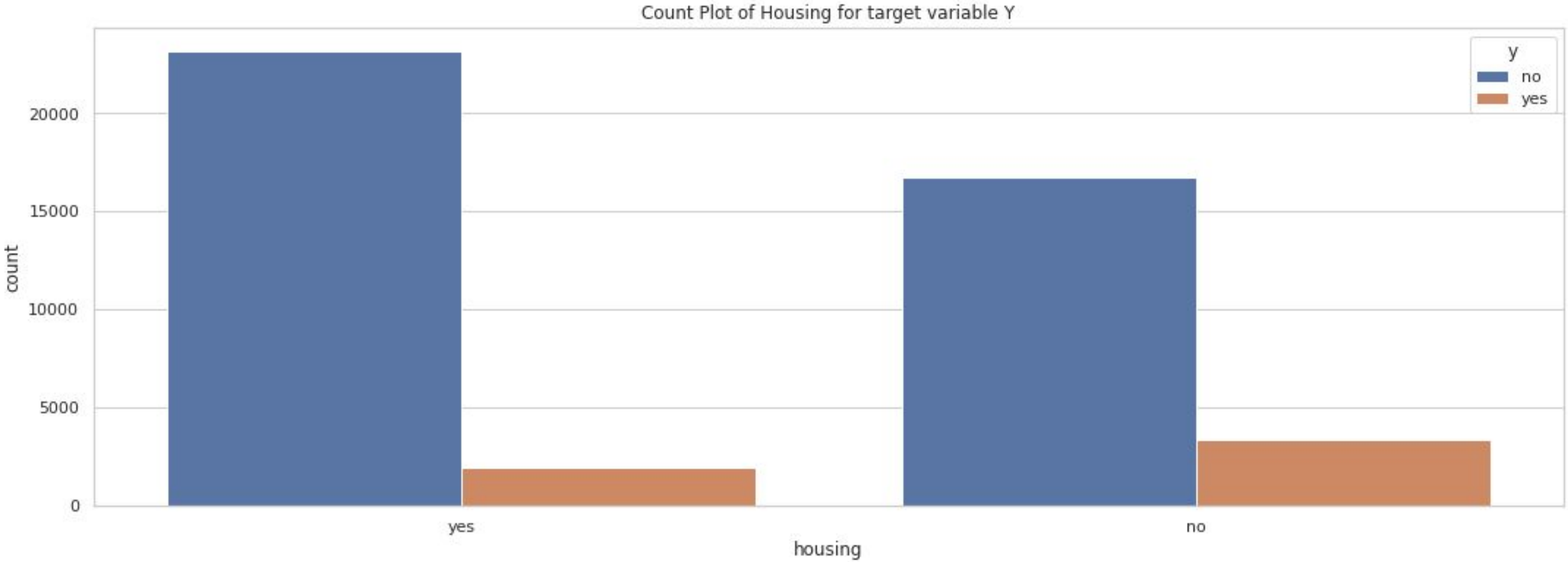
EDA(continued)

How marital status of a person effects target variable ‘y’ (Subscribed to term deposit yes/no)?



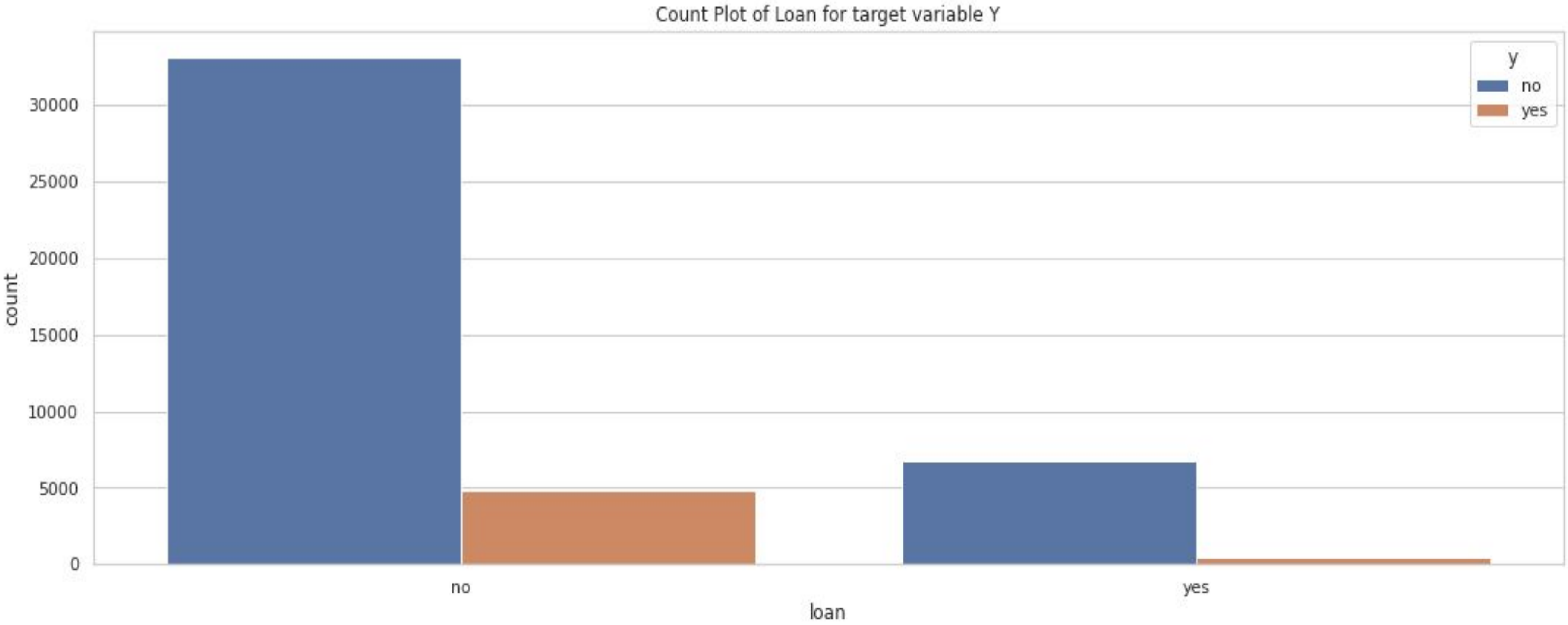
EDA(continued)

Effect of housing loan on target variable ‘y’ (Subscribed to term deposit yes/no)?
People having no housing loans are tend to subscribe more.



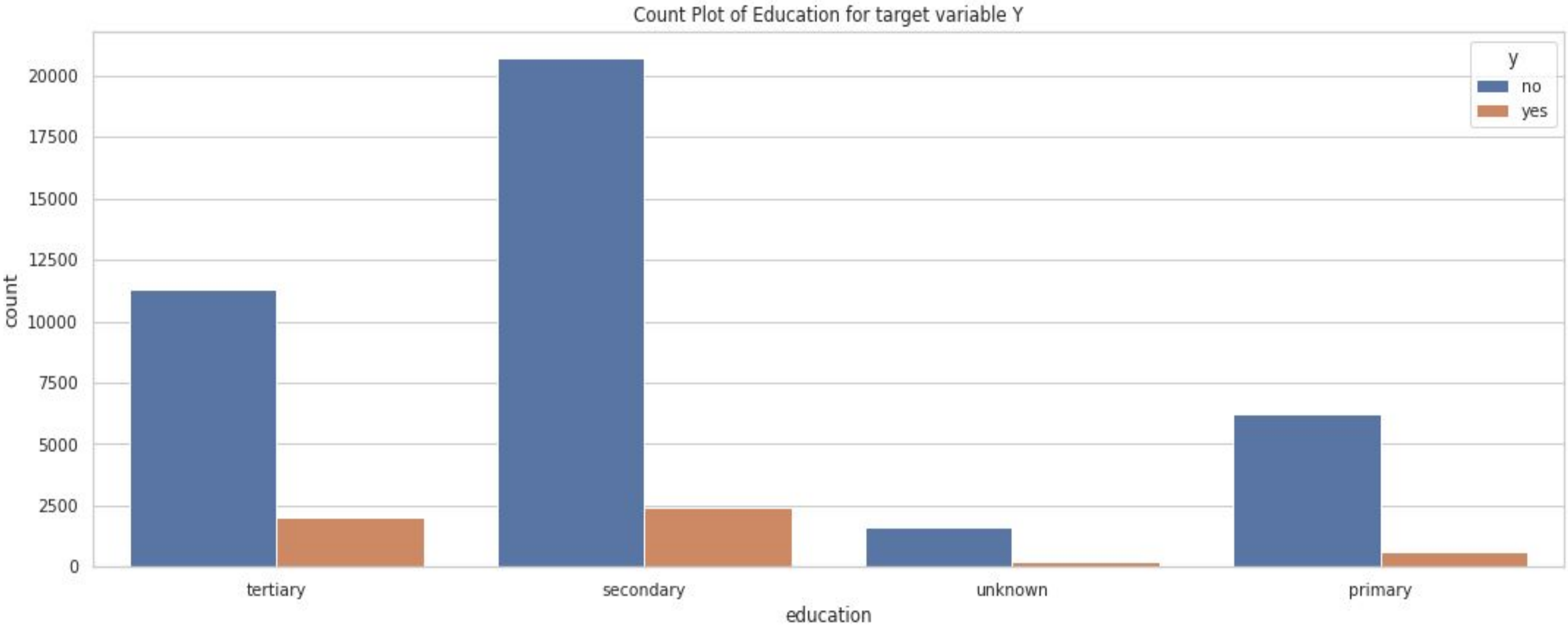
EDA(continued)

Effect of personal loan on target variable ‘y’ (Subscribed to term deposit yes/no)?
People having no personal loans are opting more for term deposit.



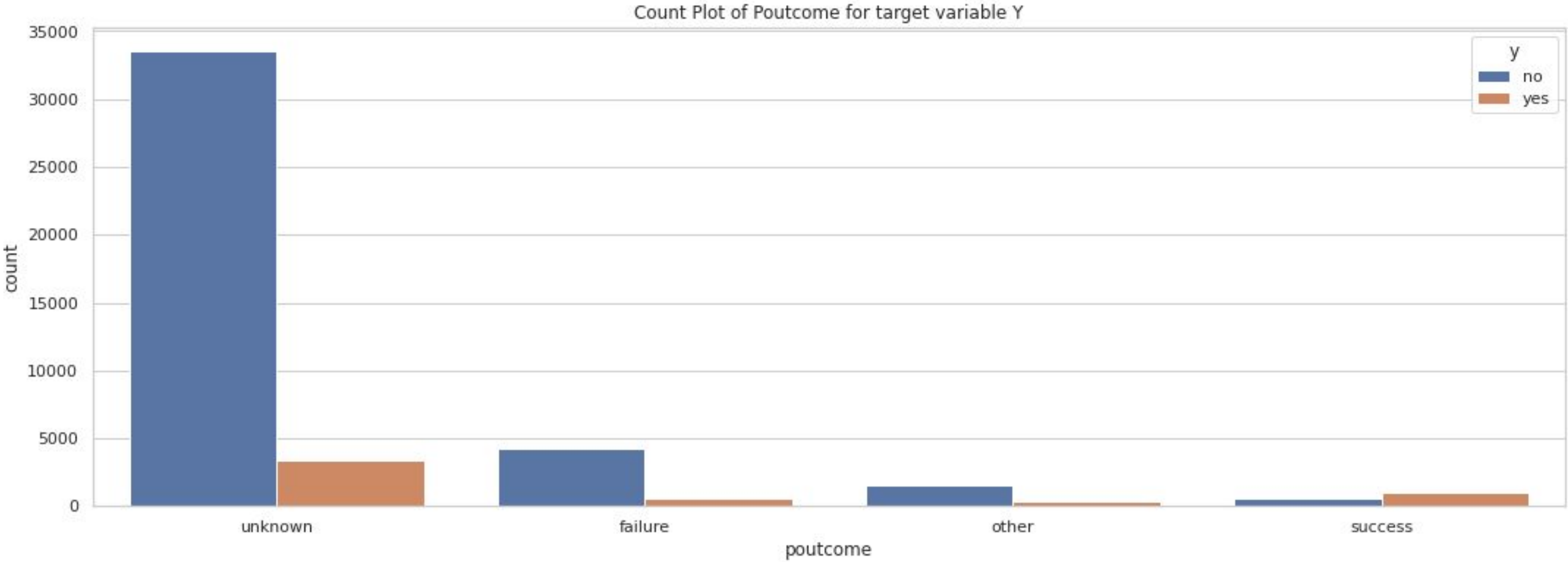
EDA(continued)

Education v/s target variable ‘y’. People with tertiary education are good prospect as they are more inclined towards term deposit.



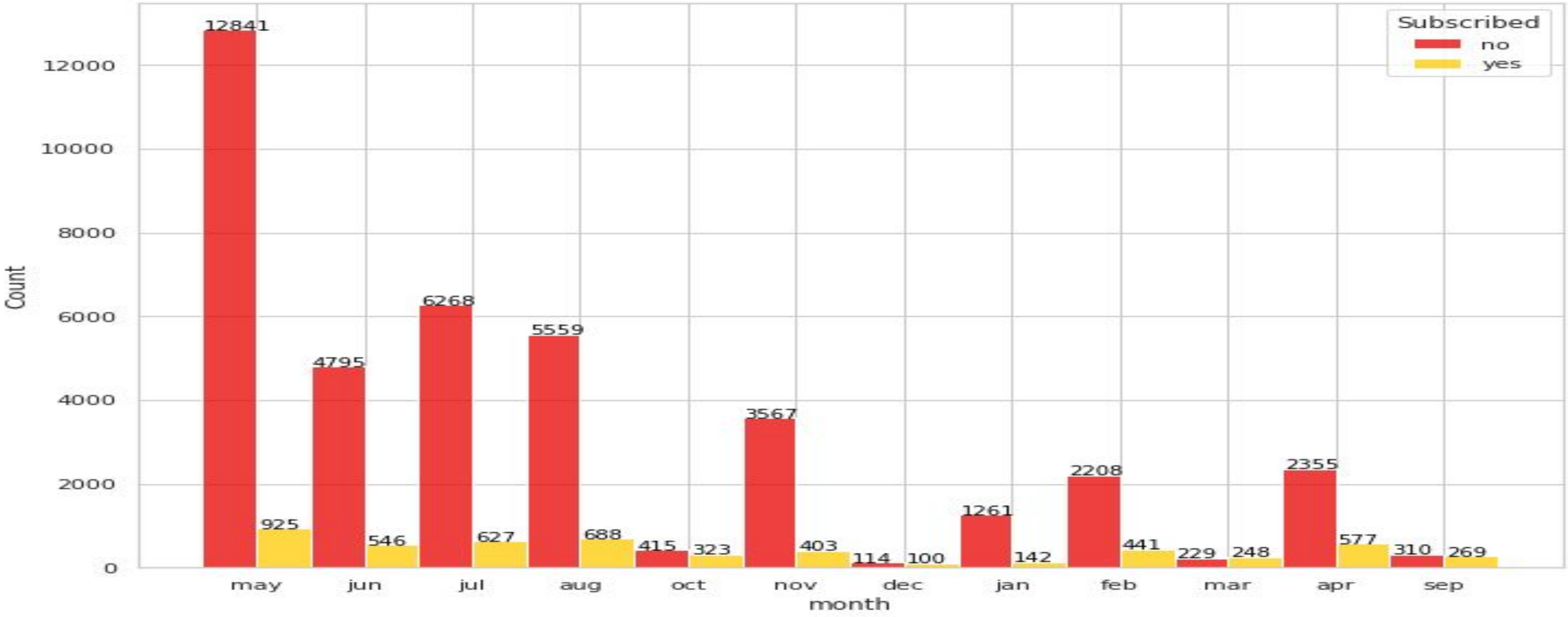
EDA(continued)

Outcome of the previous marketing campaign shows that people who already had success with the bank for earlier product are more prone to opt for term deposit.



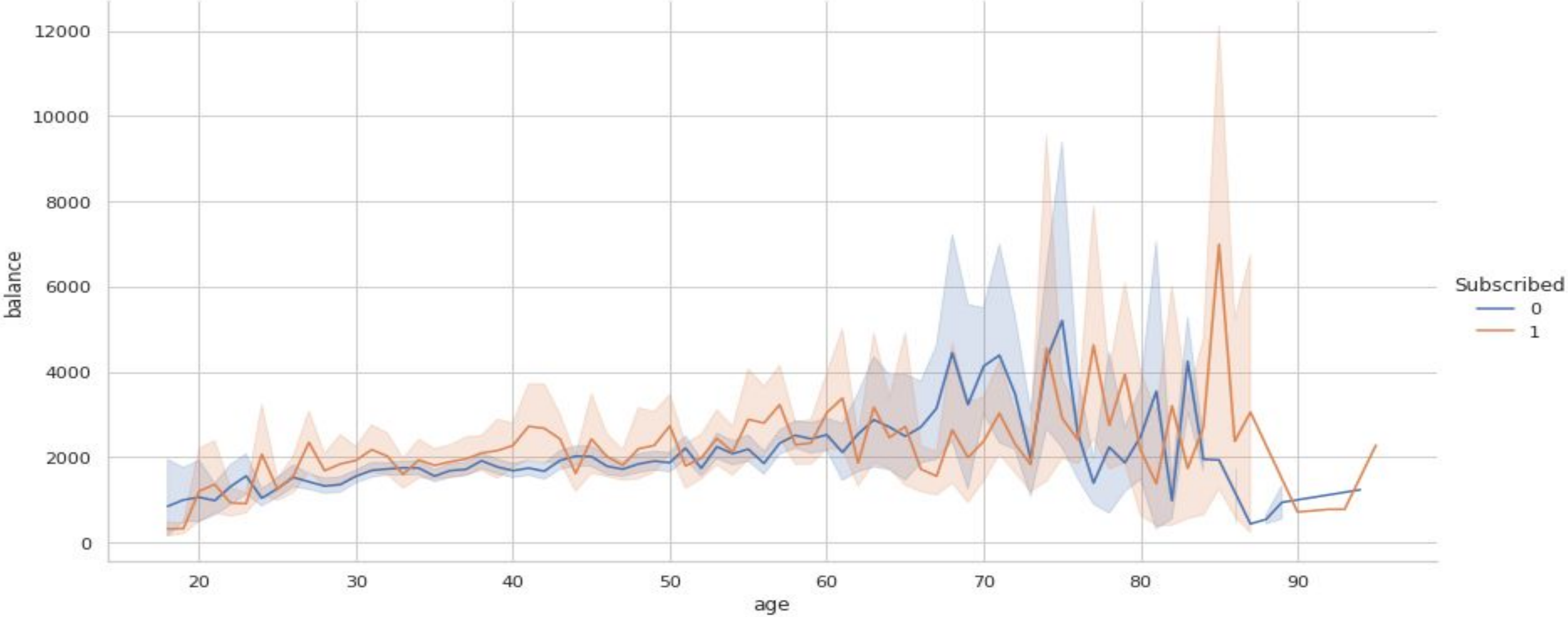
EDA(continued)

The campaign was more aggressive during 2nd quarter of the year especially in May. March is the best month as it is the only month having higher acceptance rate than rejection.



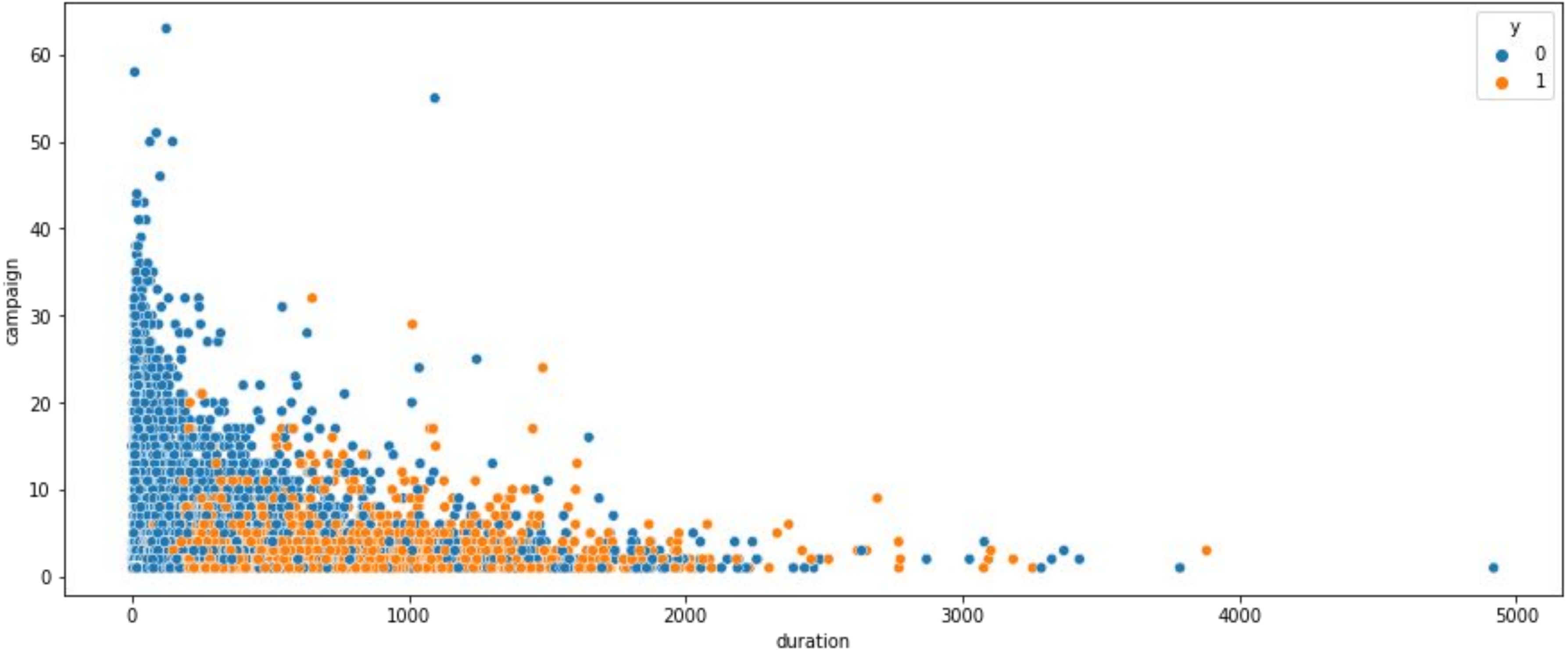
EDA(continued)

How balance & age relationship effects target variable ‘y’?



EDA(continued)

Increasing in duration leads to more term deposit but as it is highly misleading so we discarded this attribute for realistic predictive models



Data Preprocessing

- ‘duration’ attribute was highly affecting the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed.

- So dropping ‘duration’ column for realistic predictive model.

- Converting categorical values into numerical values using label encoding.

Sampling

- The given dataset was highly imbalanced ,so to balance this we used the technique called sampling.
- Oversampling and undersampling of classes to fix imbalanced dataset.
- Standard scaling of balanced dataset.

Model implementation

On raw data set:-

Logistic regression, Decision Tree classifier.

Undersampling

1) Logistic regression
2) Random Forest classifier

Oversampling

1) Random Forest classifier
2) KNN
3) XGBoost classifier

Model Evaluation

Model	Test AUC	Test Accuracy	F1-score	Precision
Logistic Regression (Under sampling)	0.89	0.85	0.83	0.94
Random Forest (Under sampling)	0.95	0.89	0.88	0.92
Random Forest (Over sampling)	0.93	0.88	0.81	0.82
KNN (Over sampling)	0.87	0.82	0.69	0.78
XGBoost (Over sampling)	0.91	0.86	0.77	0.82

Random Forest classifier (undersampling)

We selected Random Forest Classifier as the best model.

Having-

AUC score = 0.94

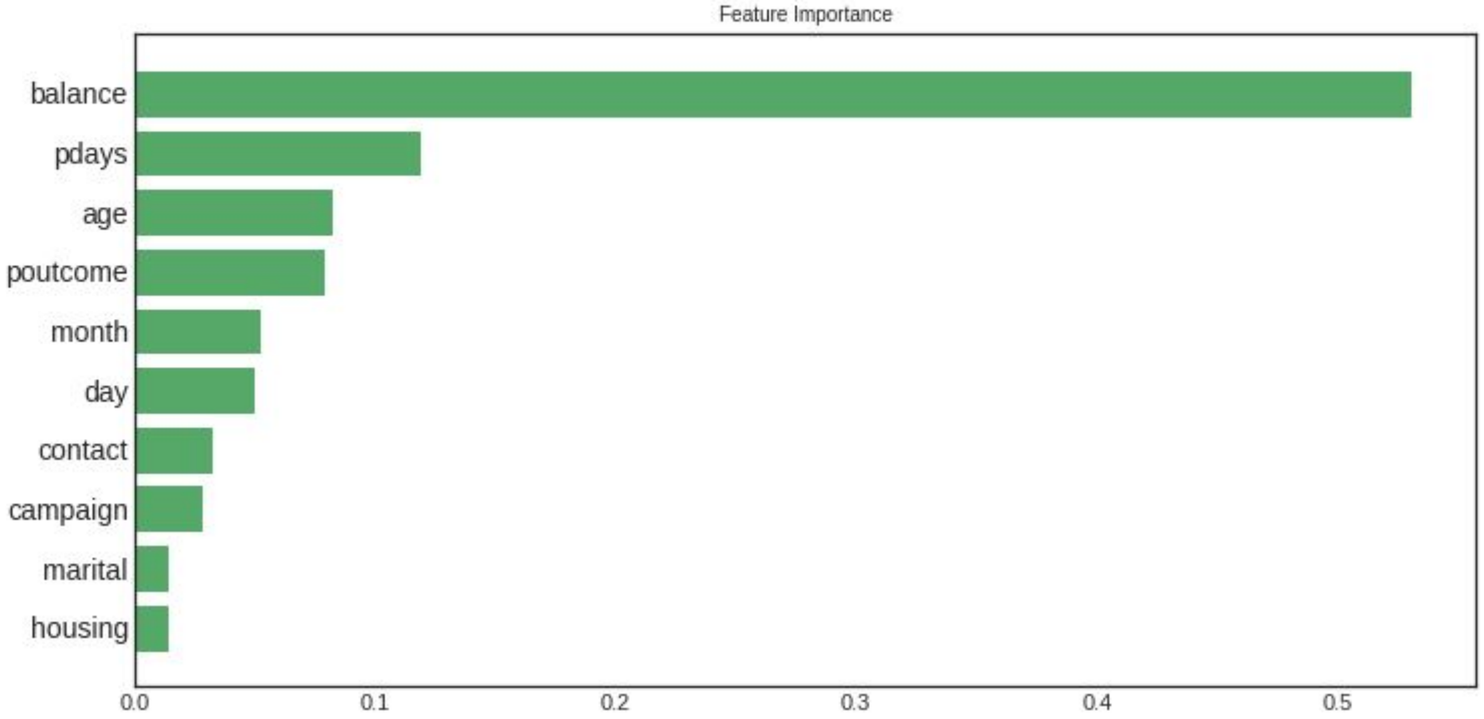
F1 score = 0.88

Accuracy = 0.88

Precision = 0.92

Here Precision is important as more number of false positive (type 1 error) can lead to poor marketing campaign because if a prospect is labelled as false positive we are completely losing him/her.

Feature Importance



Conclusion

- Random Forest and XGBoost have shown the best performance.

- The customer's account balance has a huge influence on the campaign's outcome. So we can address those customers having good account balance .

- The customer's age affects campaign outcome as well.

- Number of contacts with the customer during the campaign is also crucial.

- Outcome of previous marketing campaign also plays an important role. So we can focus on previous customers more in order to increase success of the campaign.

Thank You