Can Machines Predict Future? Let's find out!

- Project Background
- Exploratory Data Analysis
- Data Cleaning and Oversampling
- Machine Learning: Classification
- Stacking
- Survival of the Fittest
- Declaring the Winner
- Feature Importance(SHAP)
- Conclusion

What to achieve?

- **Main Objective**: Increase the effectiveness of the bank's telemarketing 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 Menu

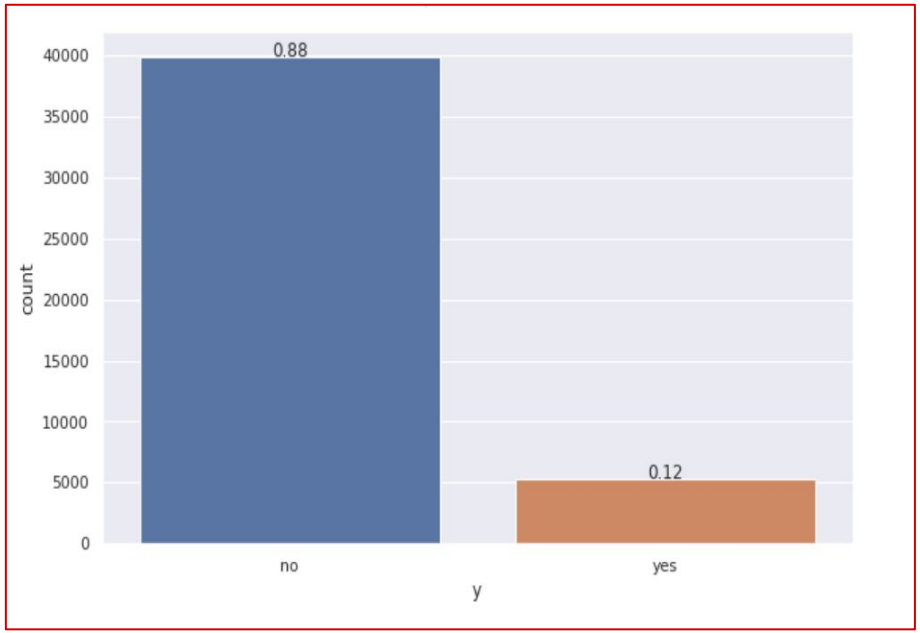
Categorical Features

- 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

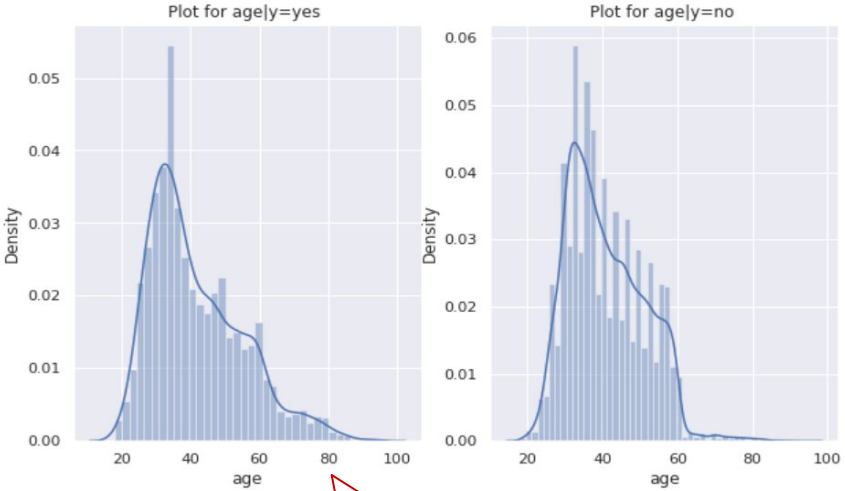
- Age
- Balance
- Day
- Duration
- Campaign
- Pdays
- Previous

Defining the Target



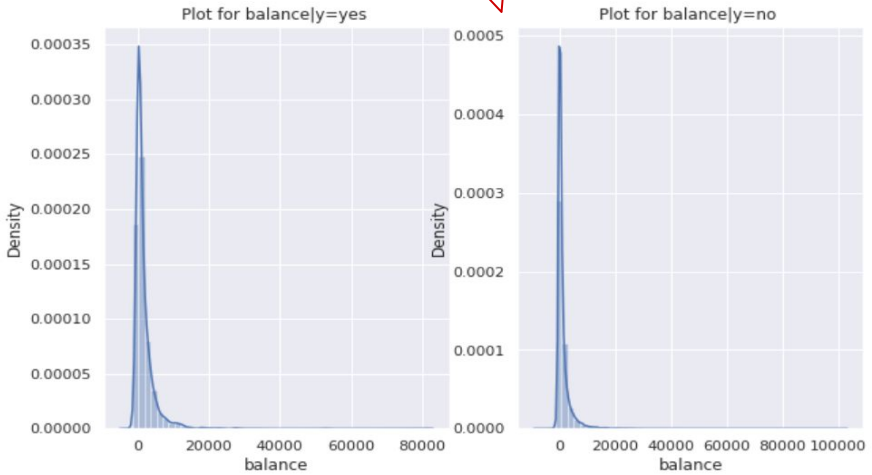
- The target variable tells us the outcome of the campaign whether they went ahead for the term deposit or not.
- Looking at the plot below we can say that the percentage of people subscribing for term is deposit is quite low, thus creating an imbalance in the data.

EDA

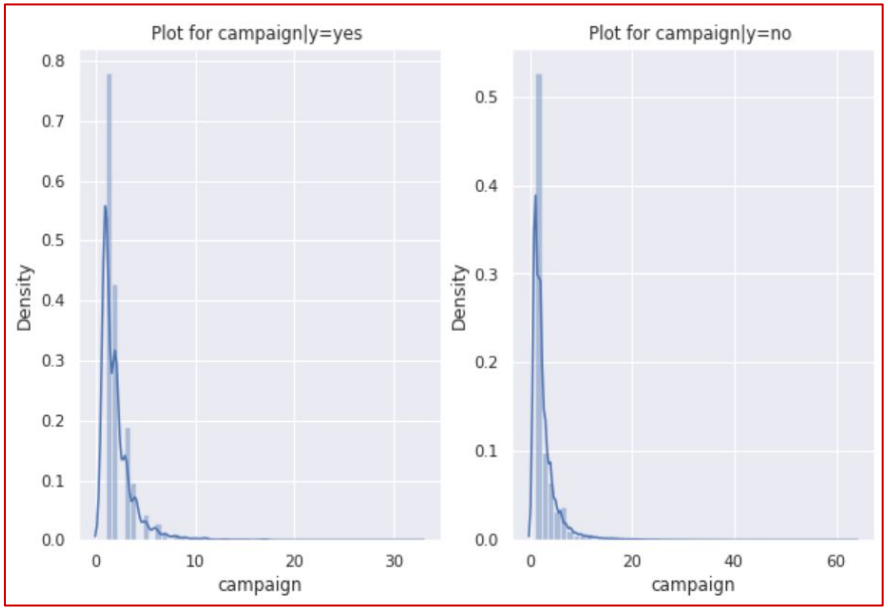


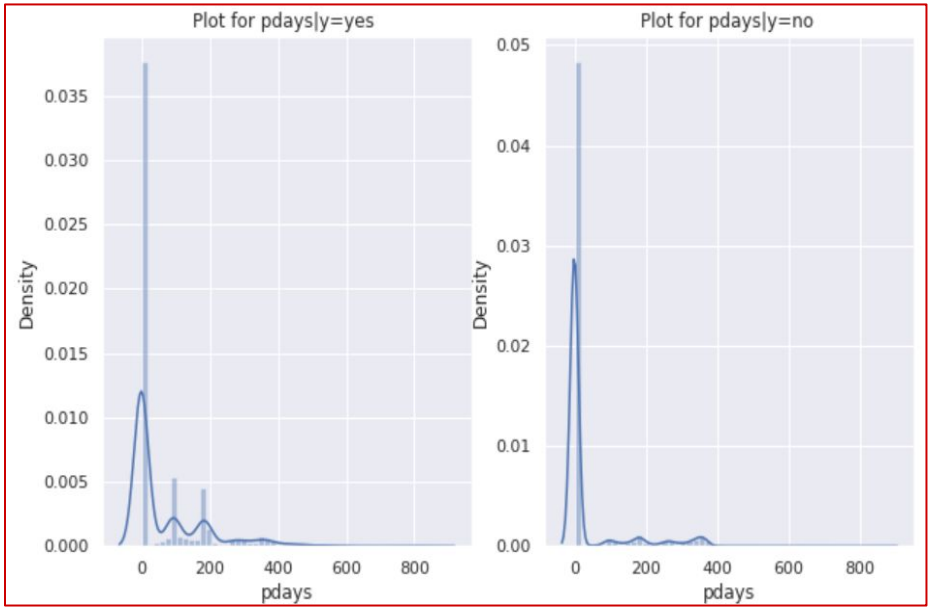
The distribution age of people deciding to say ‘yes’ or ‘no’ for term deposit.

The distribution account balance of people deciding to say ‘yes’ or ‘no’ for term deposit.

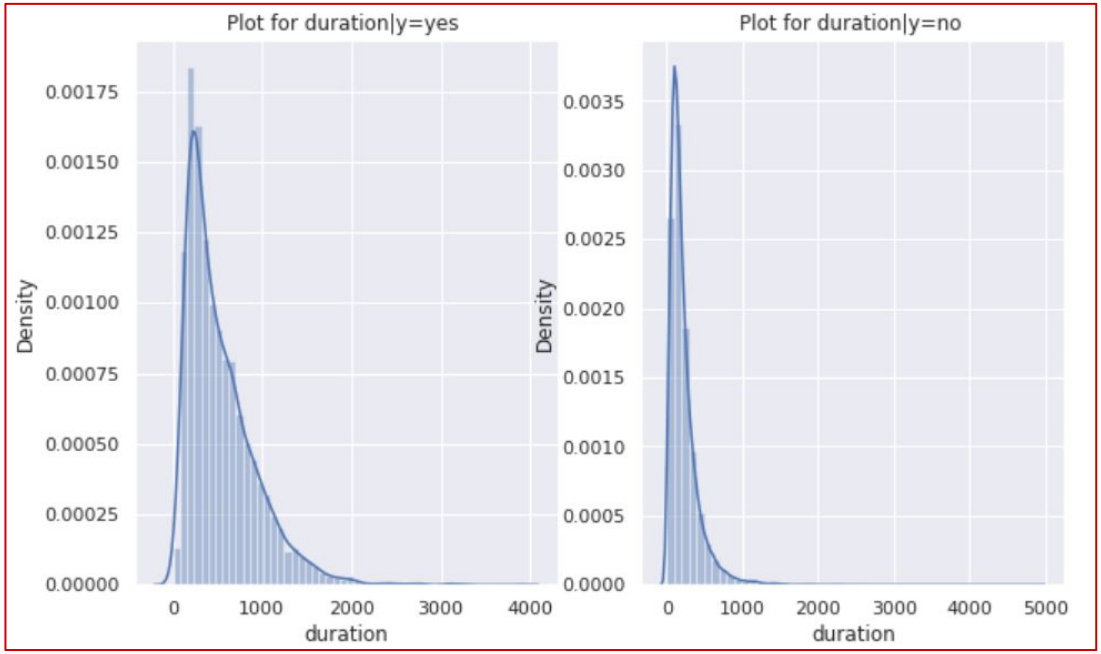


EDA(continued)



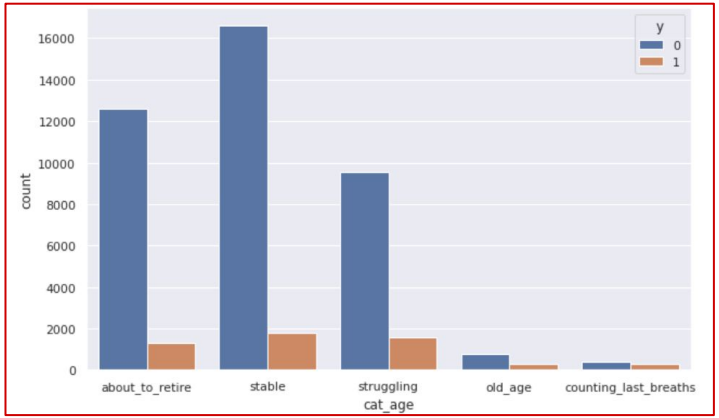


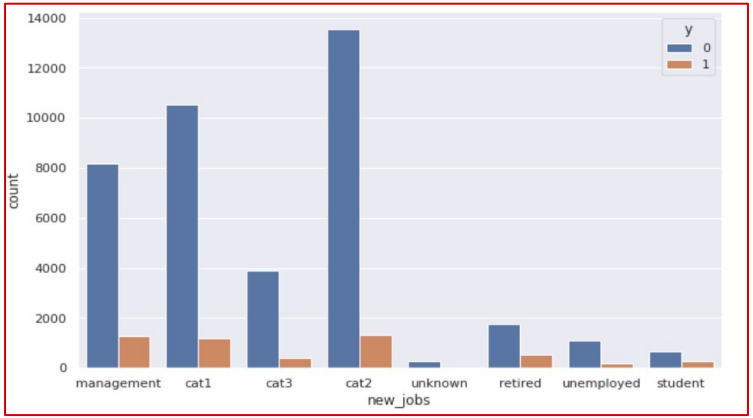
EDA(continued)

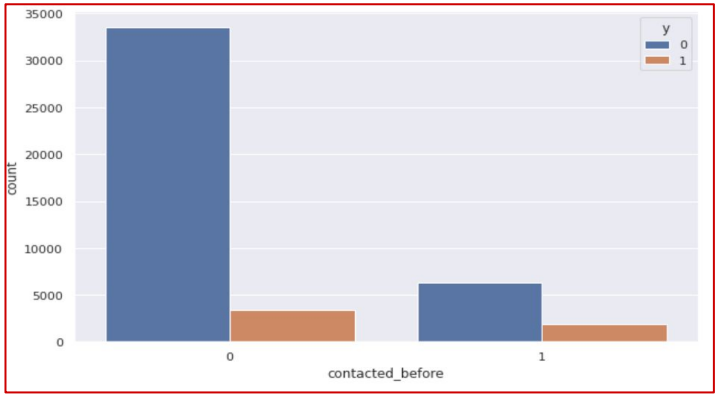


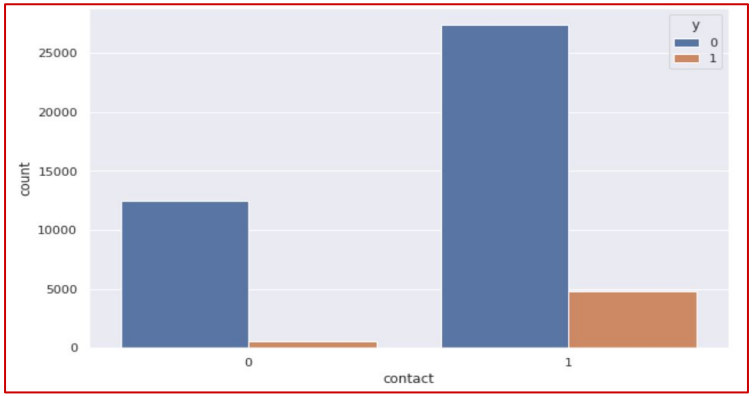
- This features on first glance looks of great relevance for the model, but it needs to be dropped as it is futuristic.
- It has nothing to do with prediction of successful subscriptions.

Feature Engineering



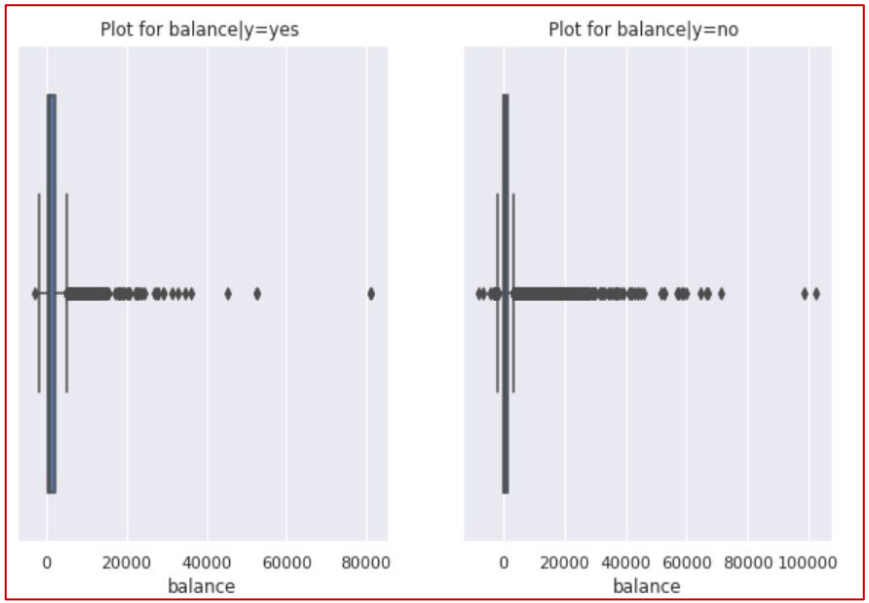




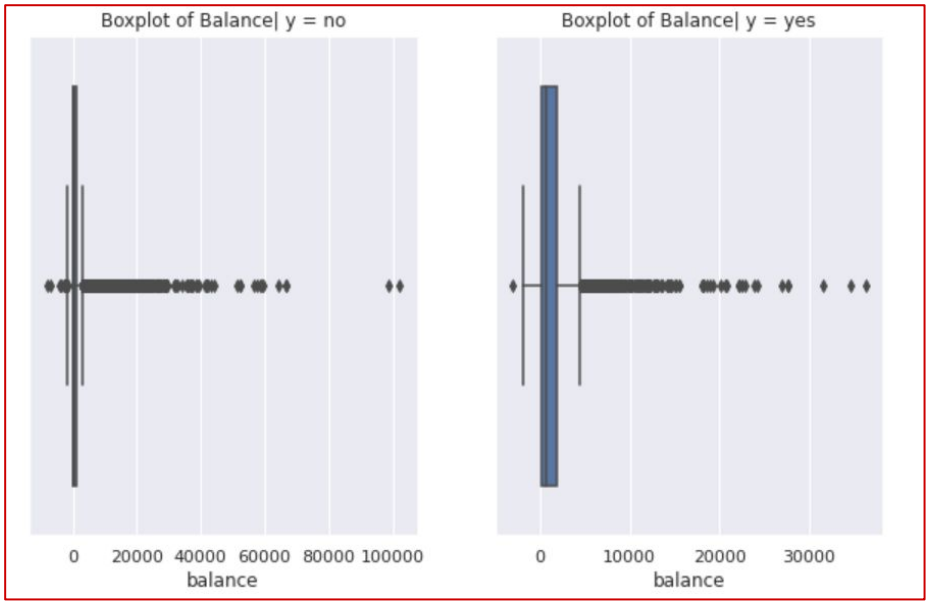


Eliminating Oddity(Isolation Forest)

Before

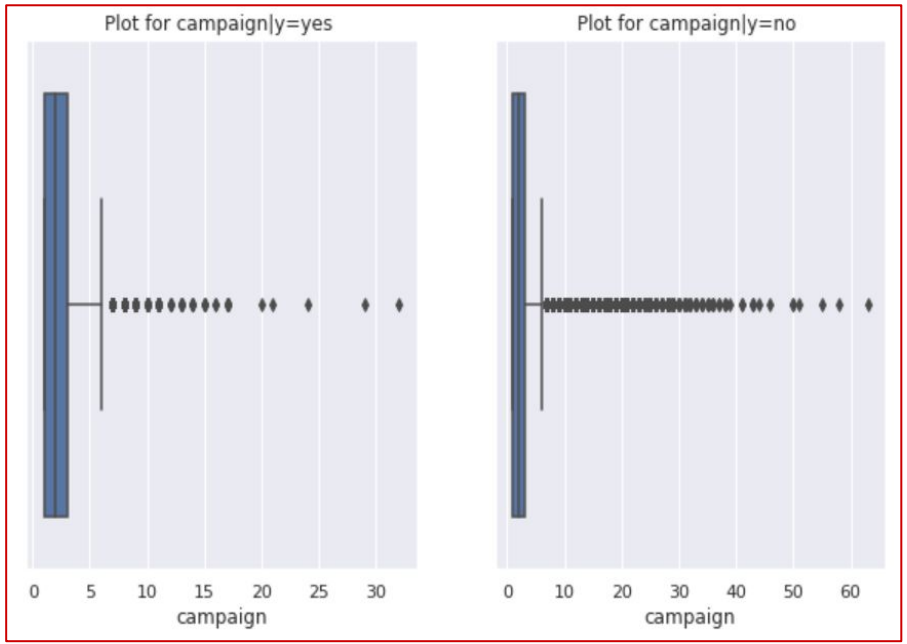


After

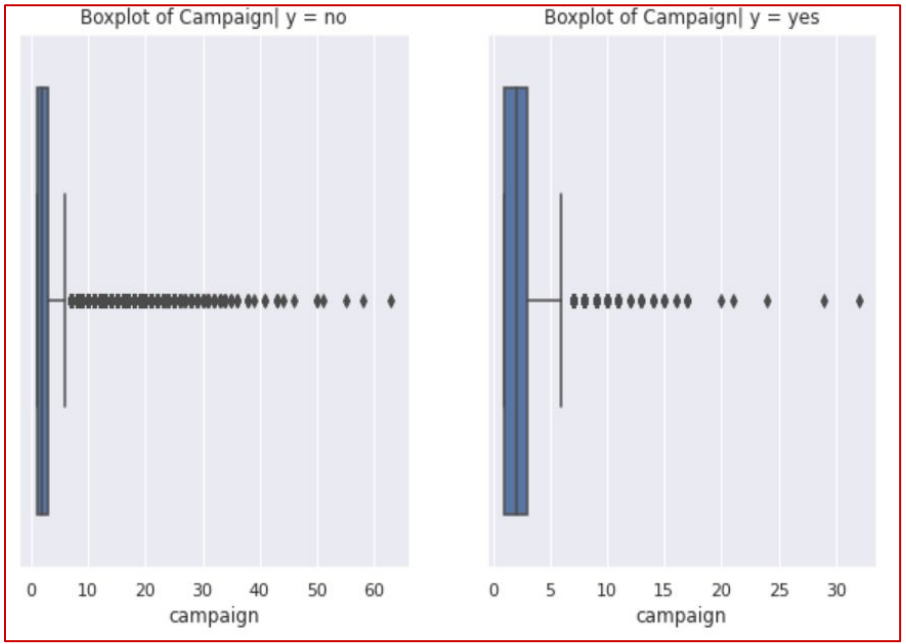


Eliminating Oddity(continued)

Before



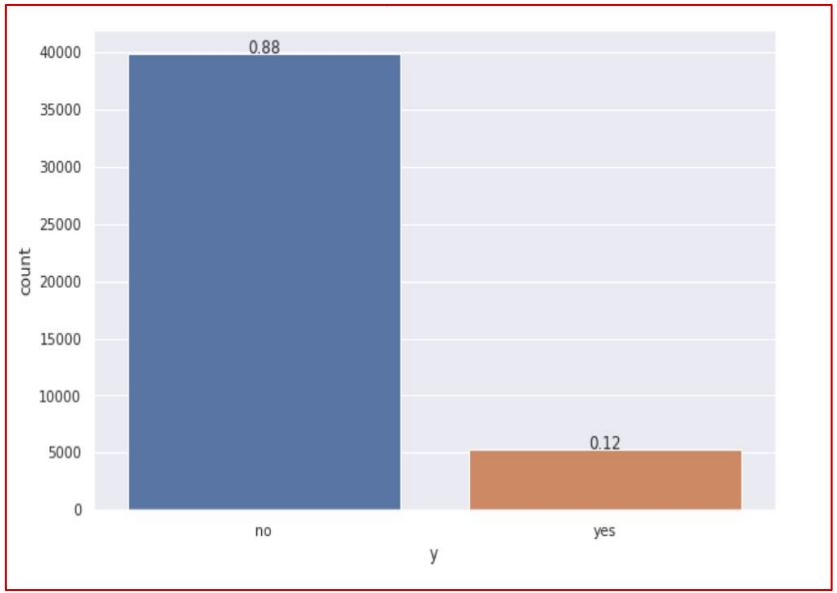
After



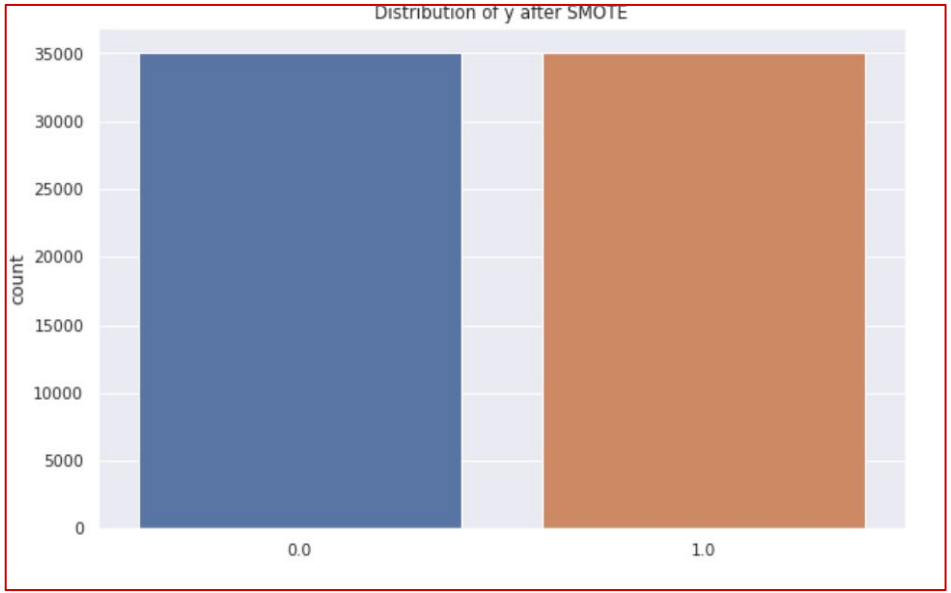
We scaled the data before using SMOTE

Oversampling(SMOTE)

Before



After

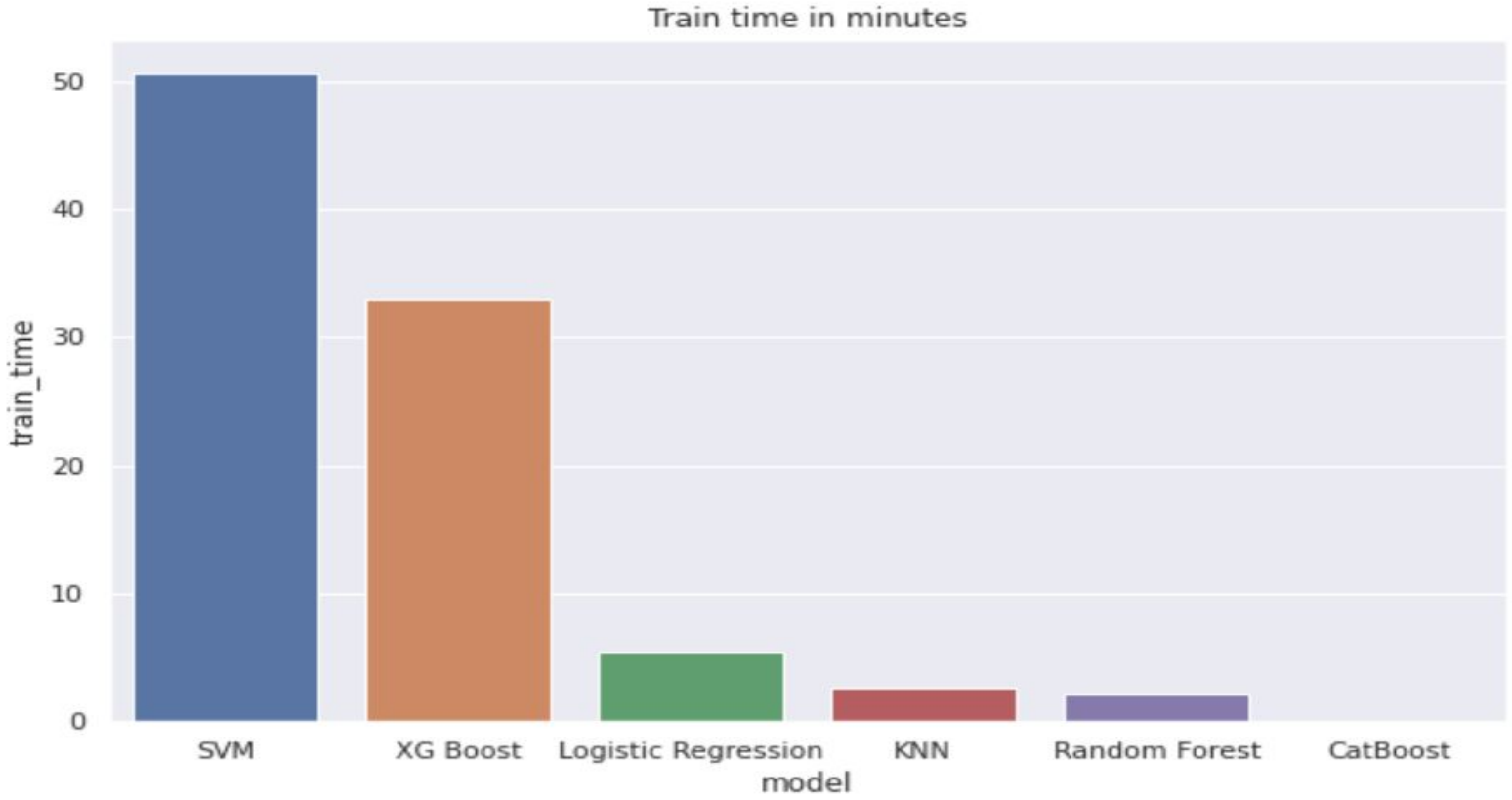


Oversampling(SMOTE) continued..

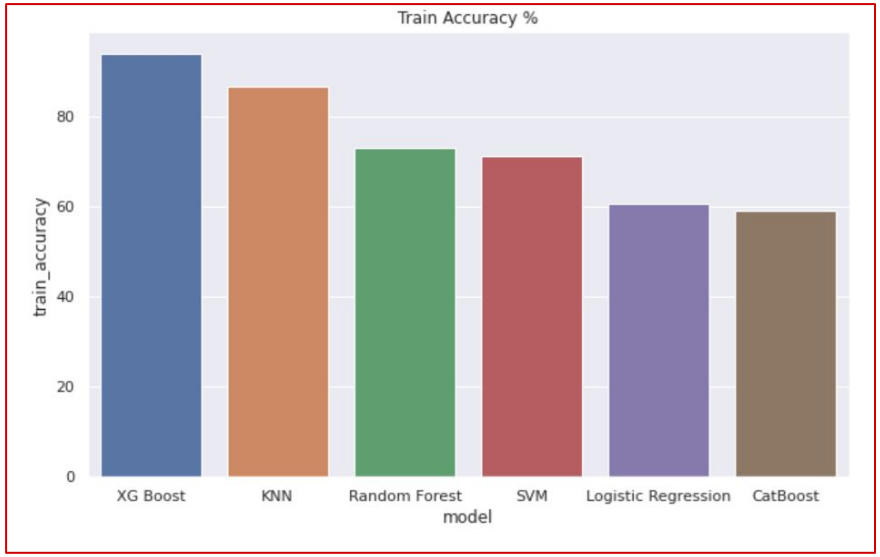
Before

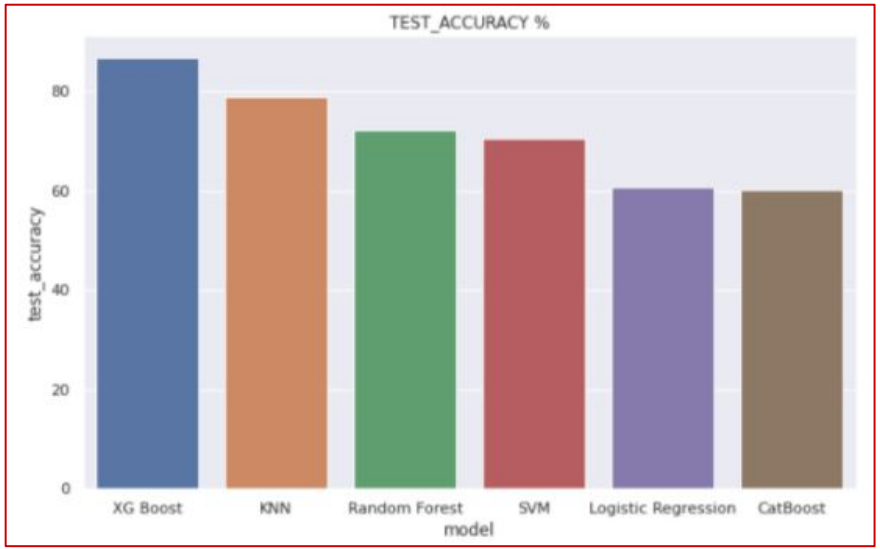
After

Applying ML Models

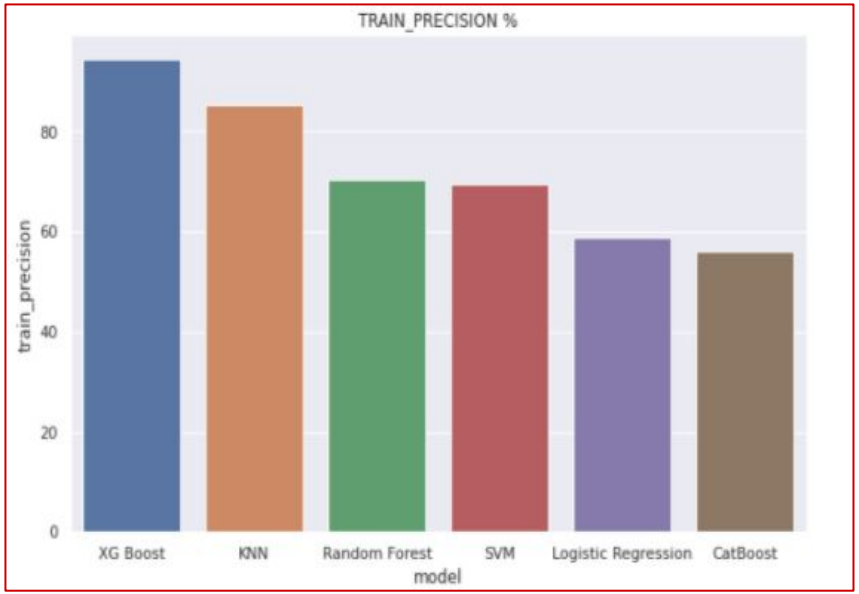


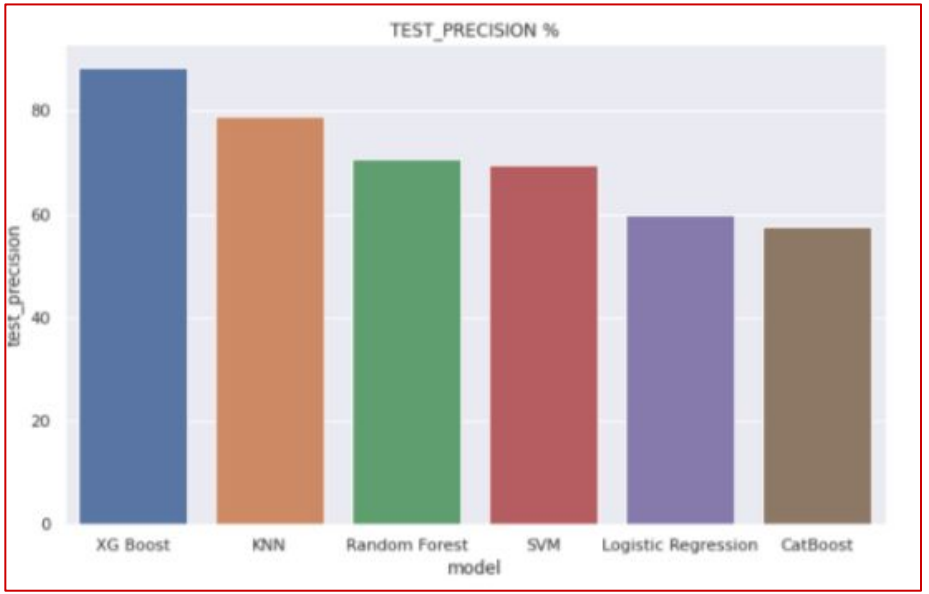
Survival of the Fittest(Accuracy)



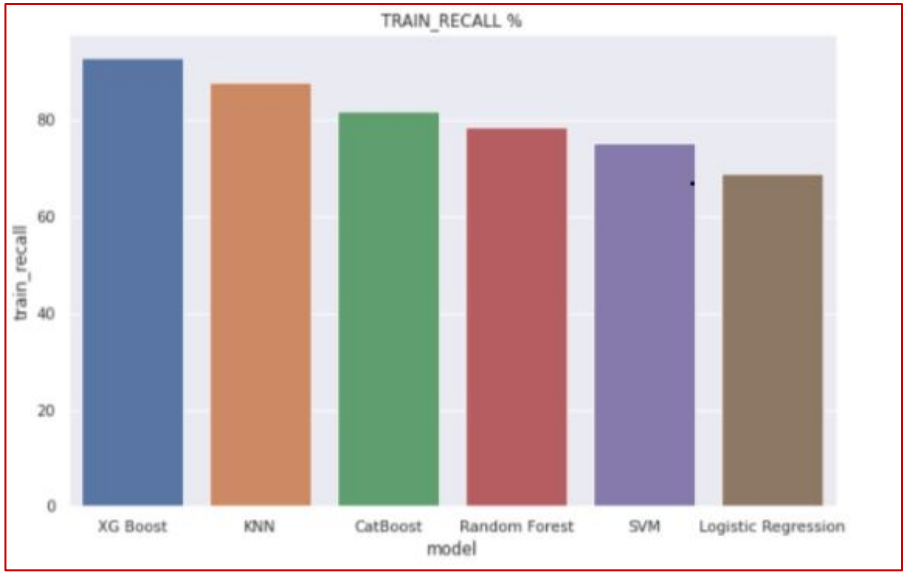


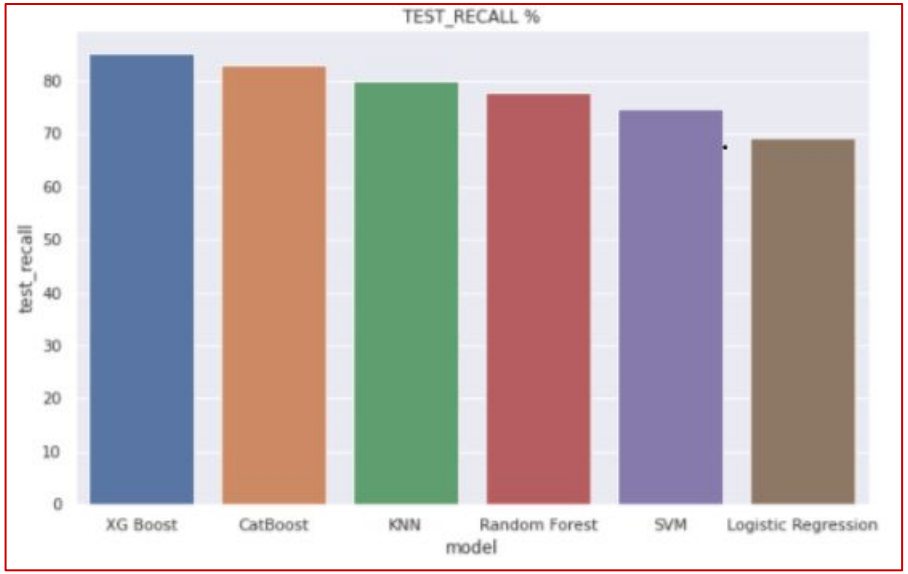
Survival of the Fittest(Precision)



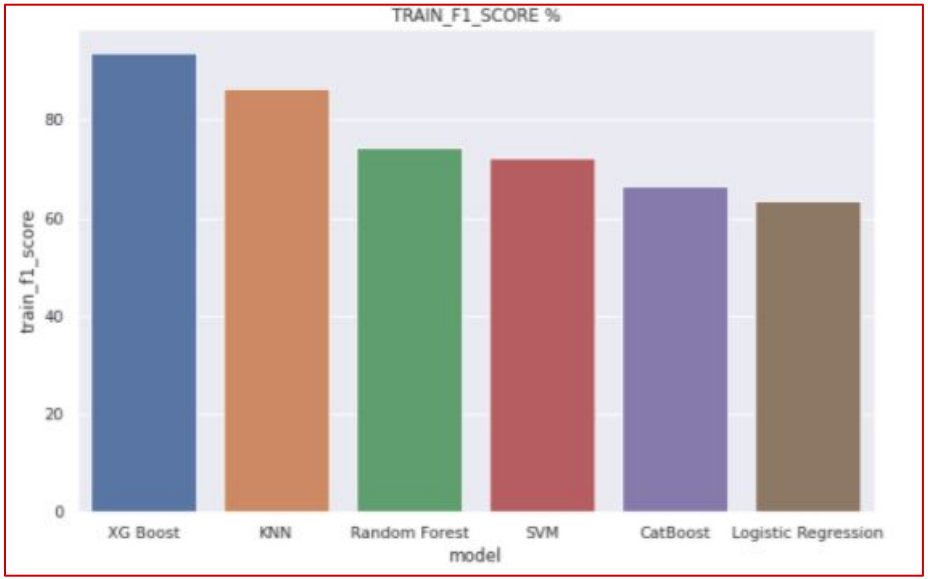


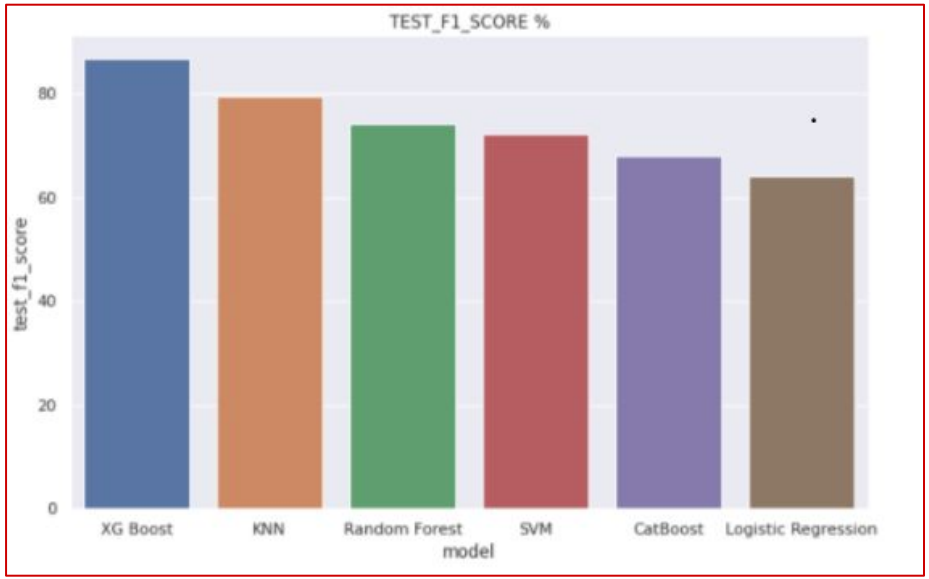
Survival of the Fittest(Recall)



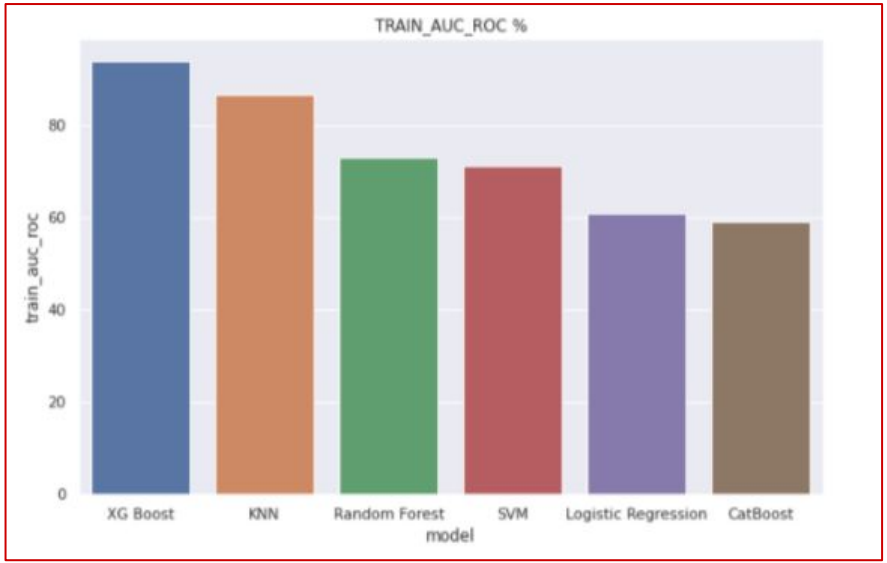


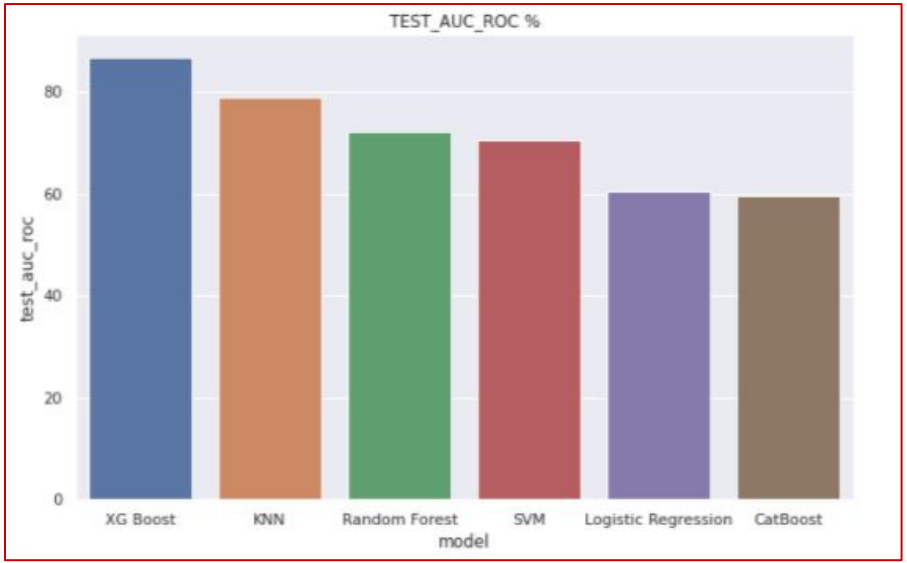
Survival of the Fittest(F1-Score)





Survival of the Fittest(AUC-ROC)





XGBoost

Parameters for GridSearchCV

{'learning_rate': [0.1,0.3,0.5], 'max_depth': [5,7,10], 'n_estimators': [50,100], 'reg_lambda': [0.1,1,10], 'seed': [123]}

Best Parameters

{'learning_rate': 0.3, 'max_depth': 10, 'n_estimators': 100, 'reg_lambda': 1, 'seed': 123}

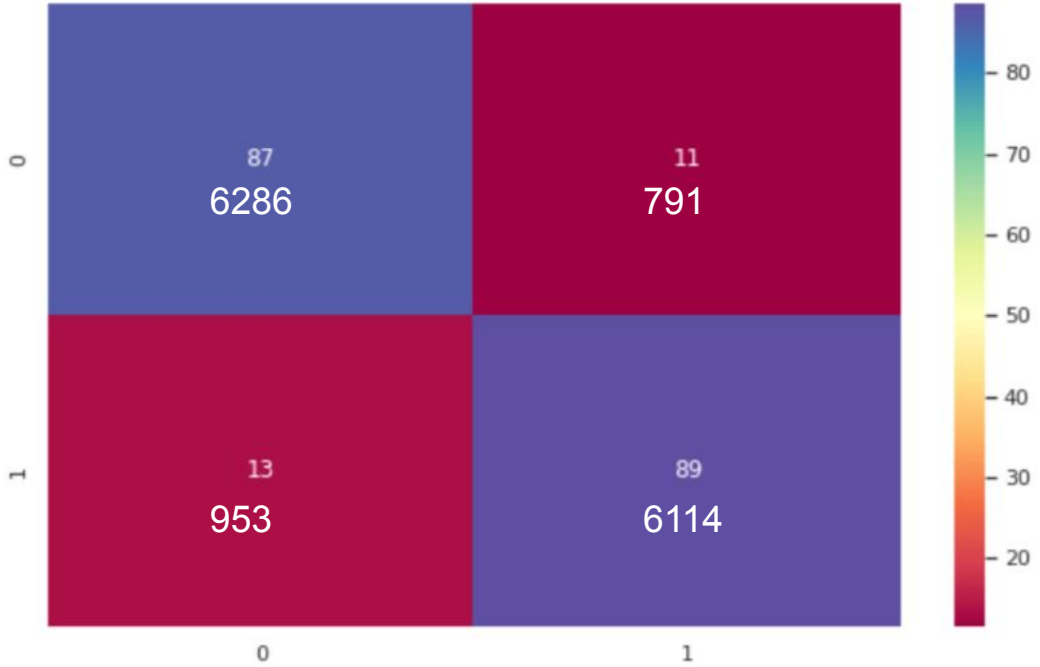
Declaring the Winner

- As seen from all the metric above , the best metric to measure the goodness of the model is AUC-ROC and Accuracy(as we have a balanced data set after using SMOTE).

- Thus,we would declare XG Boost as the winner among all the models having the best accuracy and AUC-ROC, taking a decent amount of time for training as shown above.

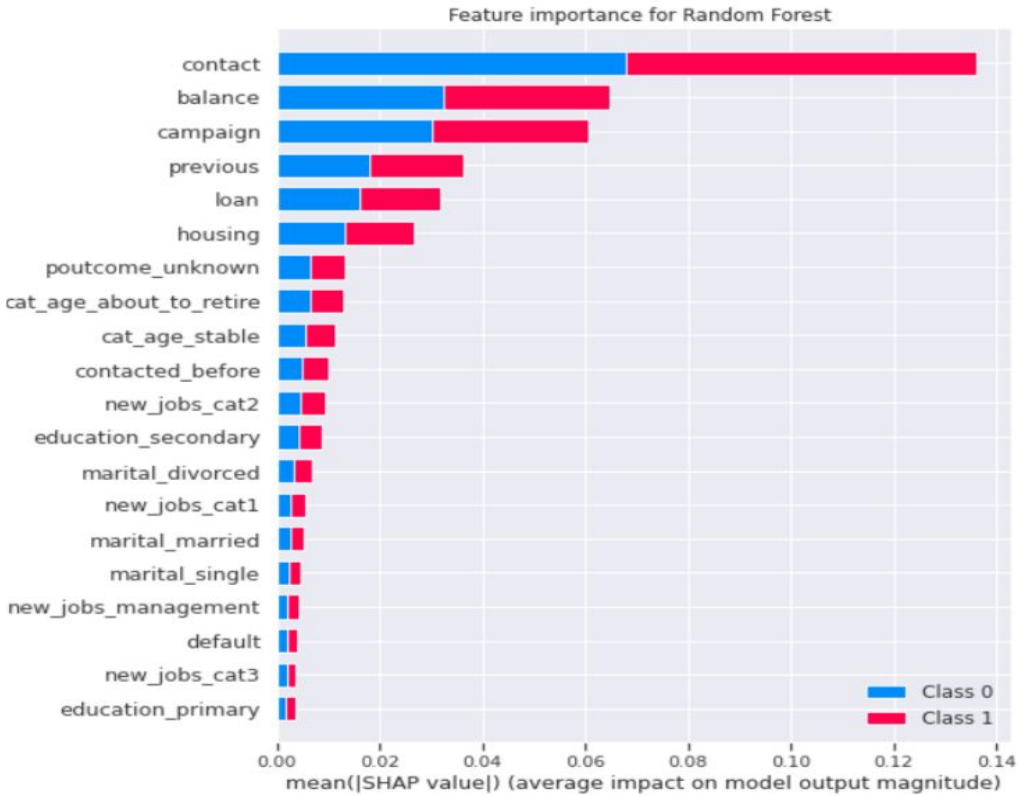
Stacking

- Used logistic regression as the meta model.

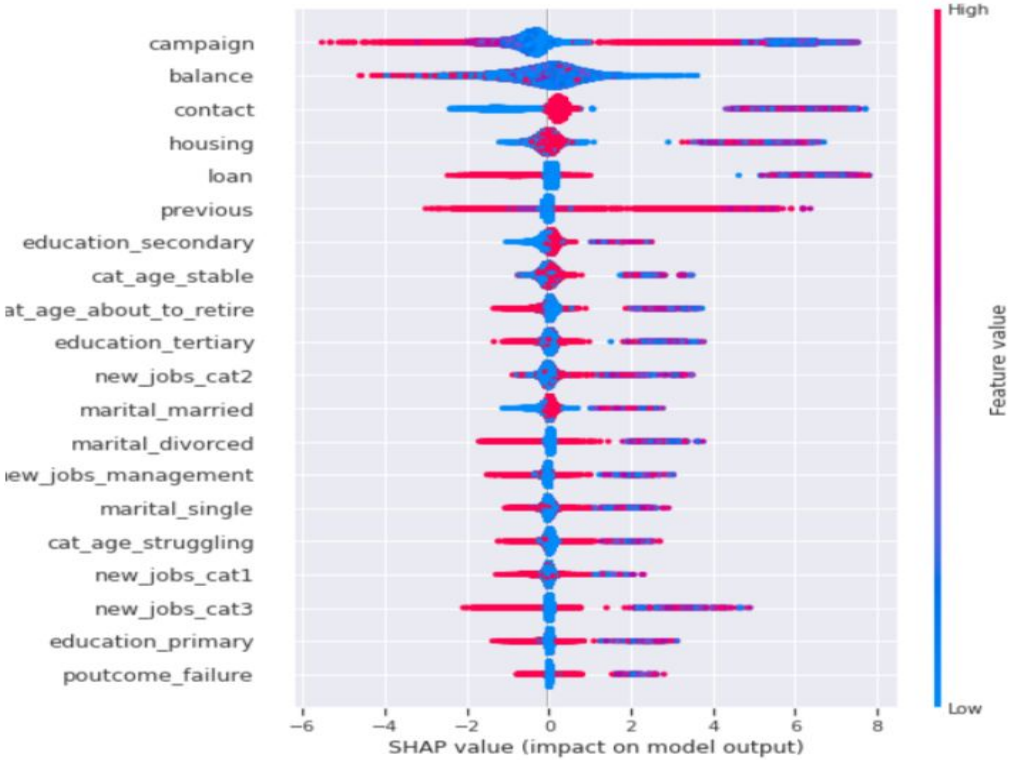


Stacking Accuracy : 0.8760
Stacking Precision: 0.8854
Stacking Recall : 0.8651
Stacking AUC : 0.8761

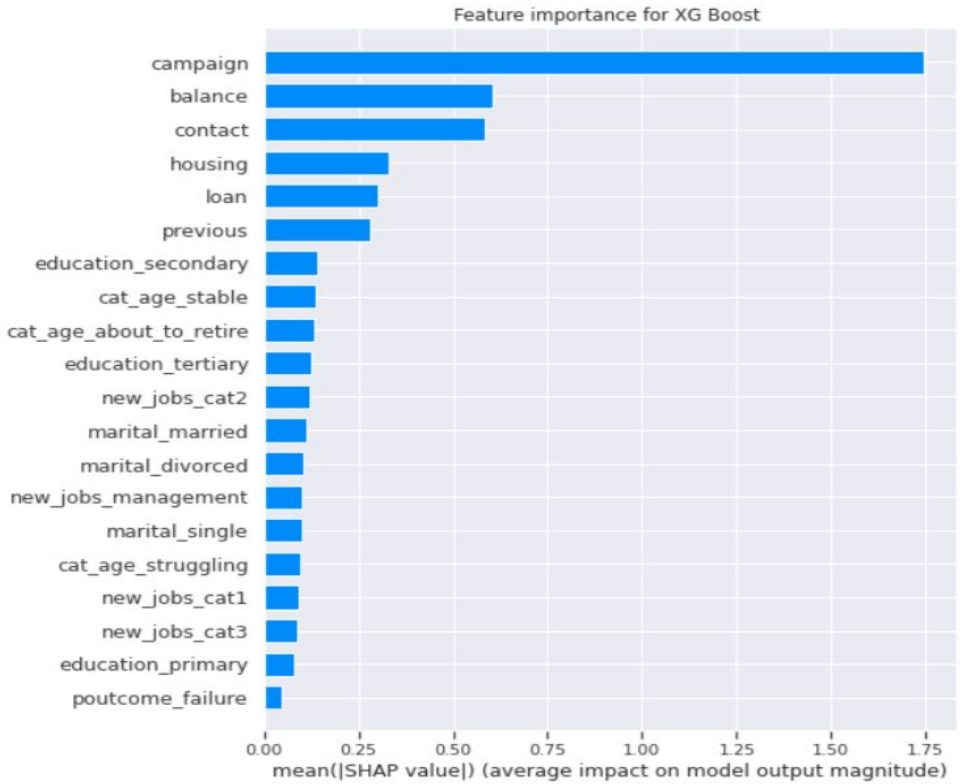
Feature Importance(Using Shapley Value)



Feature Importance(Using Shapley Value)



Feature Importance(Using Shapley Value)



Conclusion

- XGBoost has shown the best performance, but in the end it was able to identify slightly more that a half of positive outcomes, which tells me there must be ways to improve it.
- The customer's account balance has a huge influence on the campaign's outcome. People with account balance above 1490$ are more likely to subscribe for term deposit, so future address those customers.
- The customer's age affects campaign outcome as well. Future campaigns should concentrate on customers from age categories below 30 years old and above 50 years old.
- Number of contacts with the customer during the campaign is also crucial. The number of contacts with the customer shouldn't exceed 4.

Q & A