Capstone Project -3

Bank Marketing
Effectiveness Prediction

Suvendu Nayak

Contents

- Problem Statement
- Data Summary
- Exploratory Data Analysis
- Feature Engineering
- Encoding categorical features
- Sampling and feature scaling
- Model Training
- Model performance
- Challenges
- Conclusion

Problem Statement

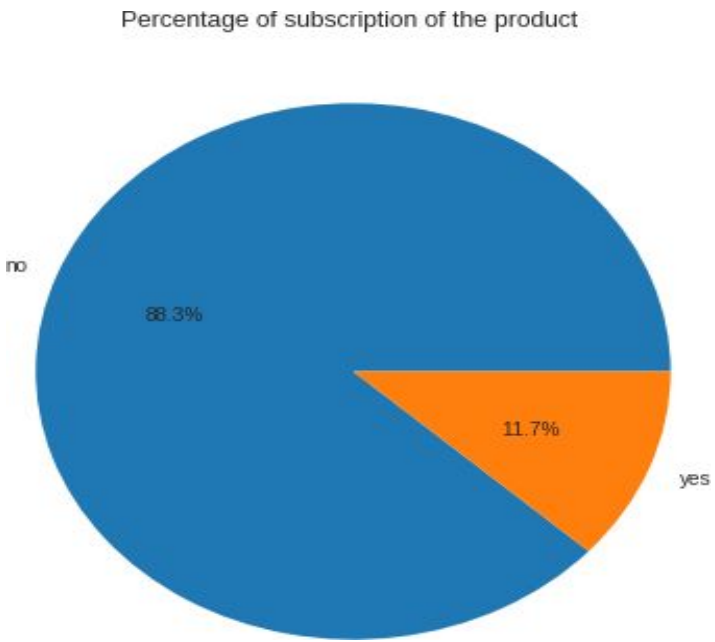
The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no')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

Data Summary

- 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 of previous campaign)
- Housing -(Yes/No)
- Loan -(Yes/No)
- Default -(Yes/No)
- Age
- Balance
- Day
- Duration
- Campaign (Number of contact performed during campaign)
- Pdays (Number of days passed by after the client was last contacted)
- Previous (Number of contact performed before this campaign)

Exploratory Data Analysis

Target variable distribution

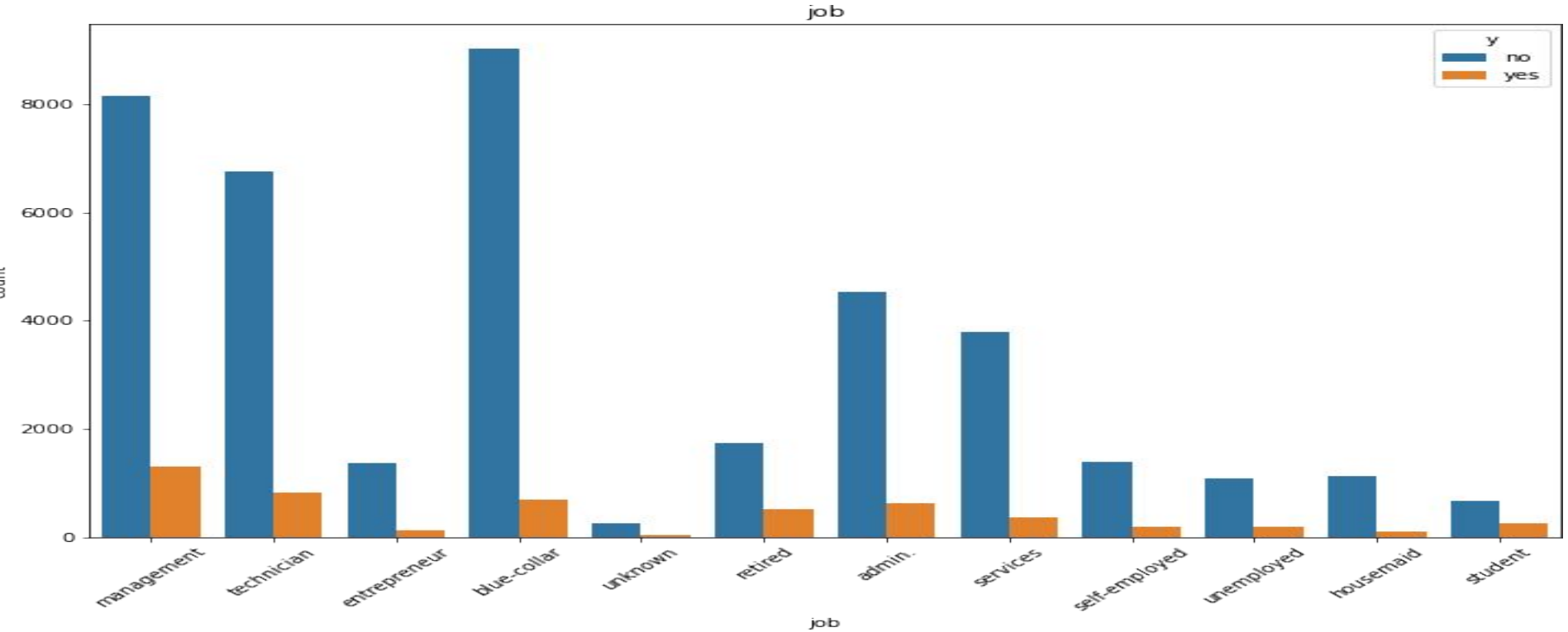


- The target variable ‘y’ tells us the outcome of the campaign whether they went ahead for the term deposit or not.
- Out of 45211 only 5289 people subscribed to the term deposit.

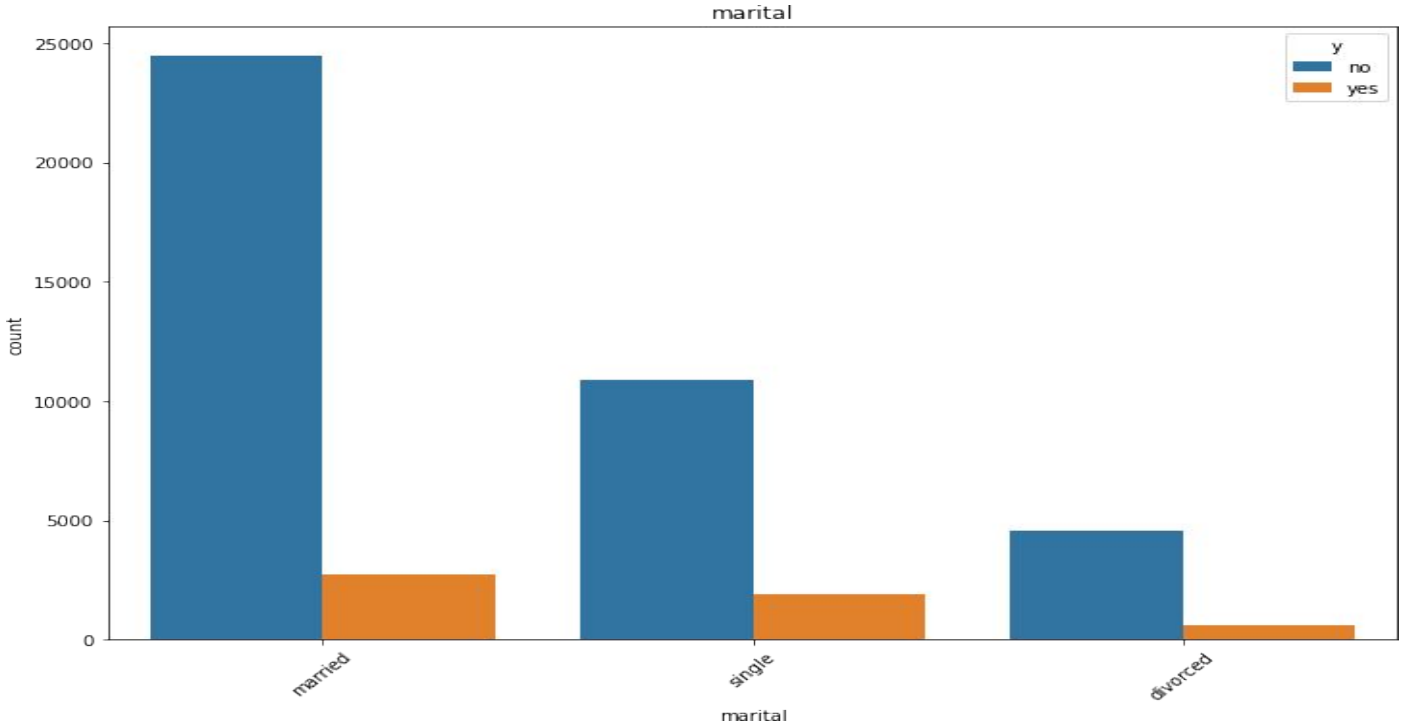
categorical features

Relationship between categorical feature and Target Variable

People with management jobs have the most number of term deposit

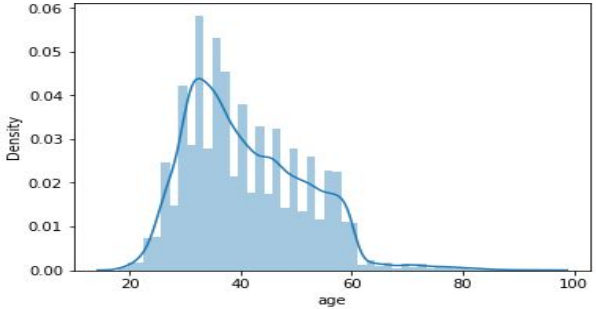


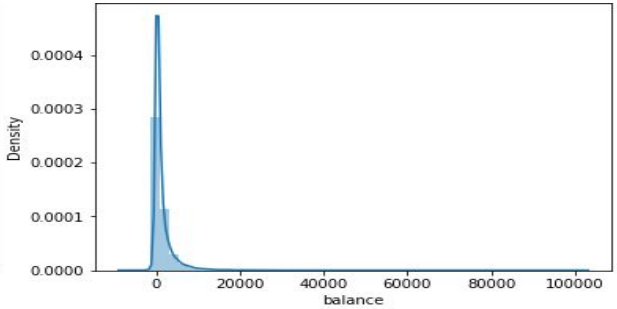
Marital status v/s target variable

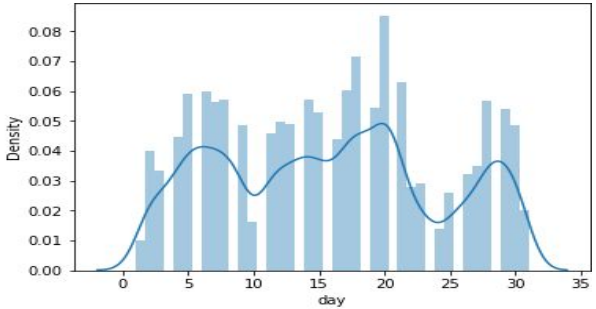


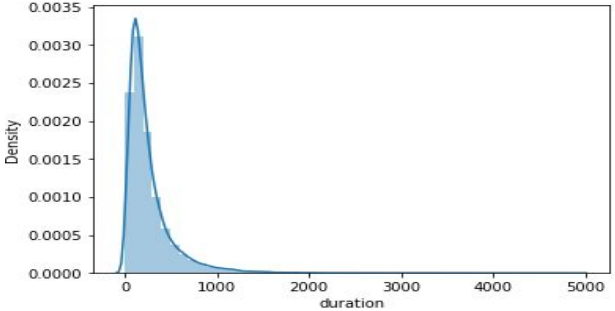
clients who has Married and single seems to be more interested on term deposit.

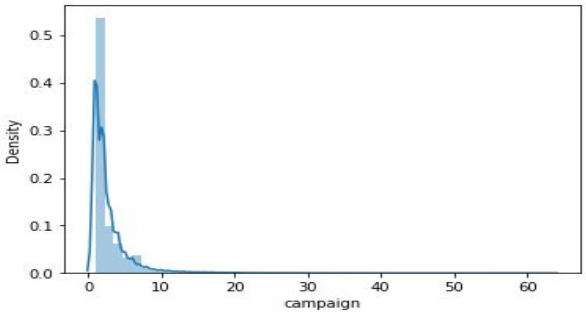
Numerical Features

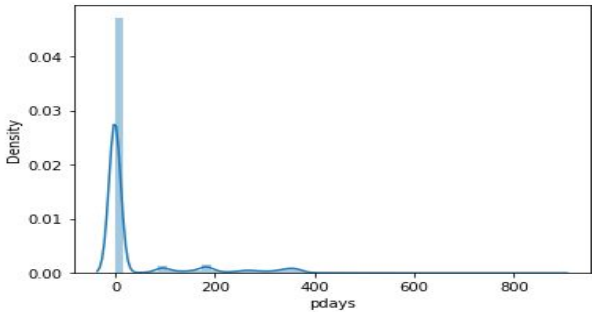


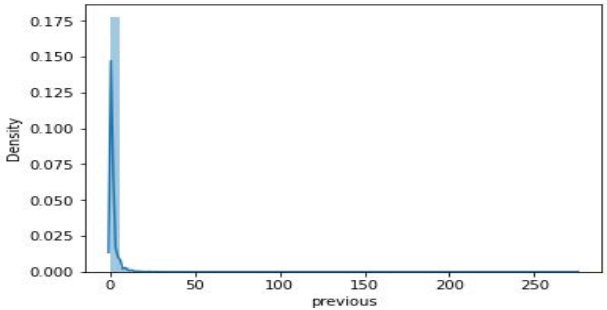










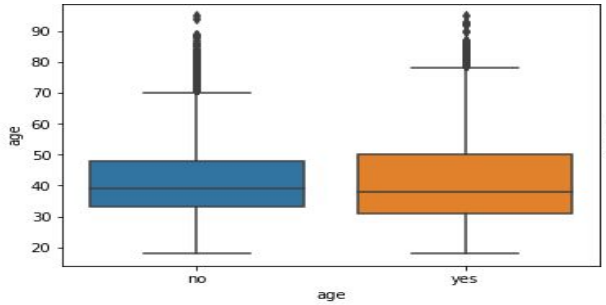


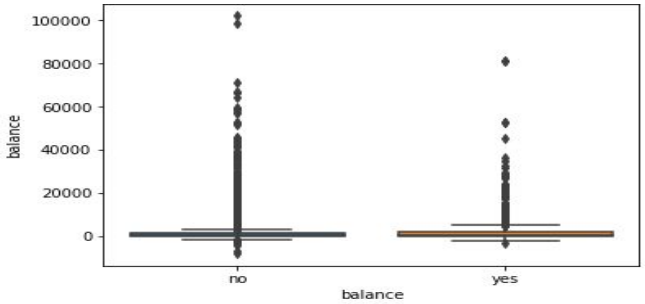
There are 7 Numerical Features.

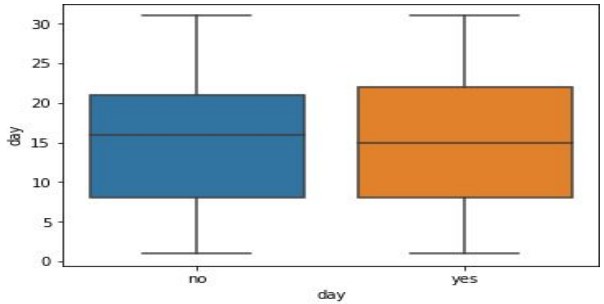
- It seems age and days are kind of normally distributed.

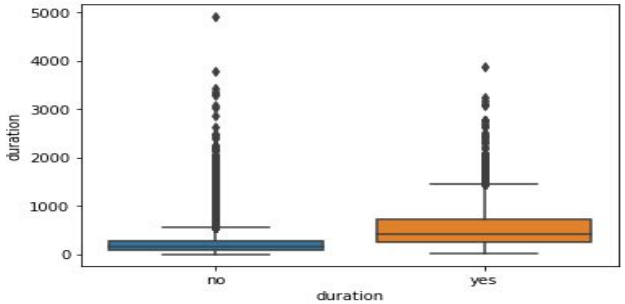
•Balance, Duration, Campaign, pdays and previous are positively skewed.

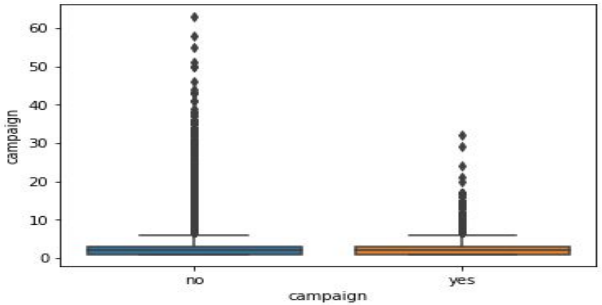
Relation between Numerical feature and Target Variable

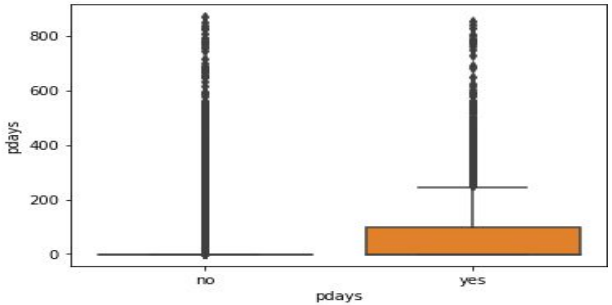


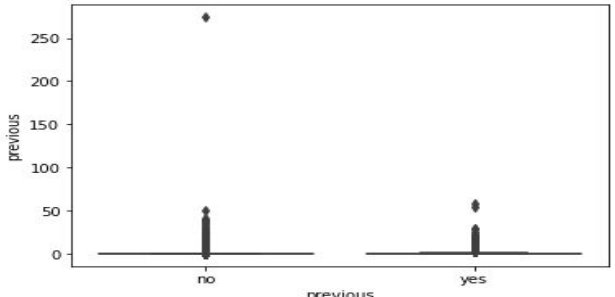








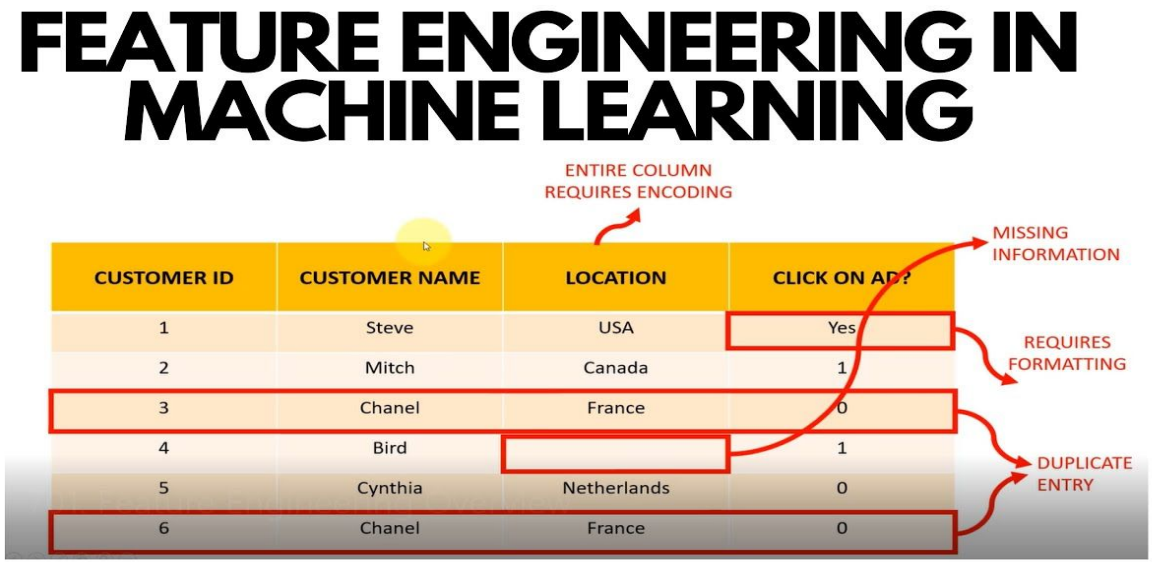




- Age distribution in dataset, shows that campaign is more centered to 30-50 age group.
- client shows interest on term deposit, who had discussion for longer duration.

Feature Engineering

Feature Engineering is a machine learning technique that leverages data to create new variable that aren't in the training set . We produce new features with the goal of simplifying and speeding up data transformation while also enhancing model accuracy.



Encoding categorical variables

In simple words encoding means converting data into required format. Since ML models takes only numerical data to do computation we will convert all cat variable into numerical data.

We used two methods to encode data.

Label Encoding: Label Encoding refers to converting labels to numeric form.

One Hot Encoding: It is also the process of converting categorical data into numerical data but here we don’t give
labels to each category instead we create new columns for each category and gives binary values.

Type
AA
AB
CD
AA

Onehot encoding →

Type	AA_Onehot	AB_Onehot	CD_Onehot
AA	1	0	0
AB	0	1	0
CD	0	0	1
AA	0	0	0

Sampling

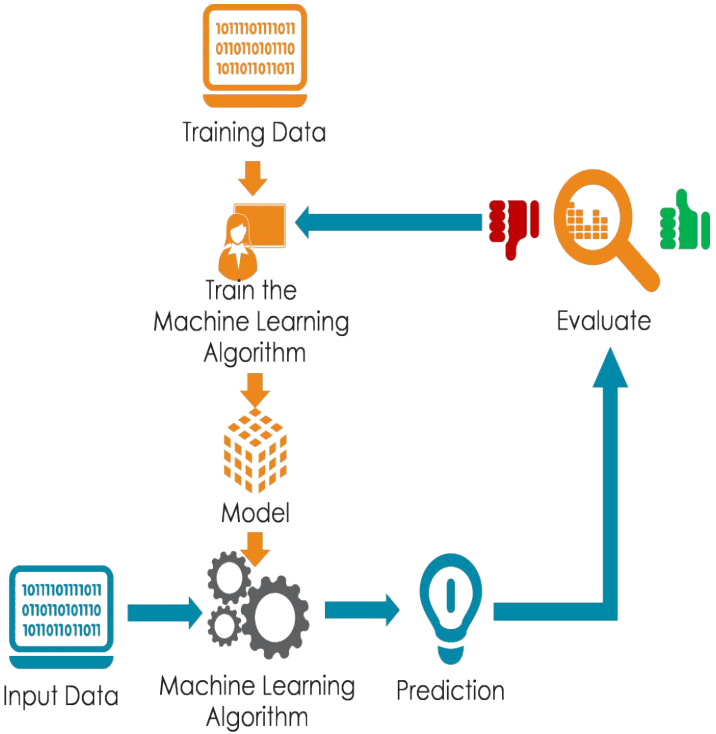
- The dataset was highly imbalance so to balance the dataset, we used technique called Random over sampler.
- Oversampling can be defined as adding more copies to the minority class.

Feature Scaling

- Feature Scaling is a technique to normalize/standardize the independent features present in the dataset in a fixed range.
- We used MinMaxScaler to scale feature ,it basically takes min and max value of column and scale feature according to that in range of 0 to 1

Model Training

Model training is the process of fitting a data into machine learning model from which model learns the patterns in data to predict the dependent variable. Model do it so by assigning a weight to each variable.
After our model is trained we test our model on test data to check how our model is performing.



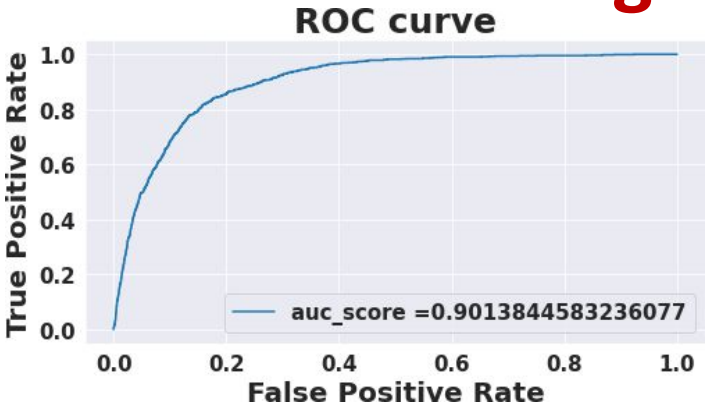
Model Evaluation

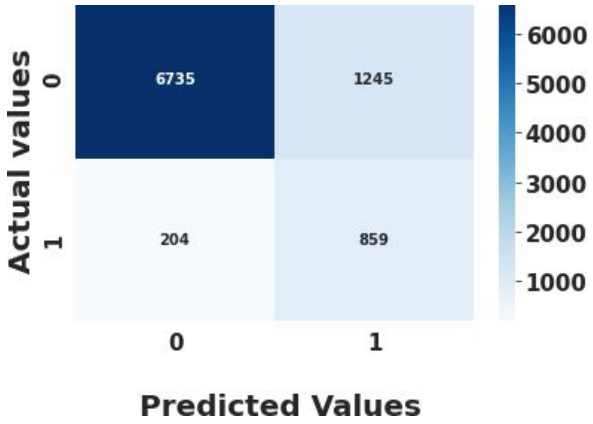
Choose the evaluation metric wisely.

Accuracy is not the best metric to use when evaluating imbalanced datasets as it can be misleading. Metrics that can provide better insight are:

- **Confusion Matrix**: a table showing correct predictions and types of incorrect predictions.
- **Precision**: the number of true positives divided by all positive predictions. Precision is also called Positive Predictive Value. It is a measure of a classifier's exactness. Low precision indicates a high number of false positives.
- **Recall**: the number of true positives divided by the number of positive values in the test data. The recall is also called Sensitivity or the True Positive Rate. It is a measure of a classifier's completeness. Low recall indicates a high number of false negatives.
- **F1 Score**: the weighted average of precision and recall.
- **Area Under ROC Curve (AUC-ROC)**: AUC-ROC represents the likelihood of your model distinguishing observations from two classes. In other words, if you randomly select one observation from each class, what's the probability that your model will be able to "rank" them correctly?

Logistic Regression



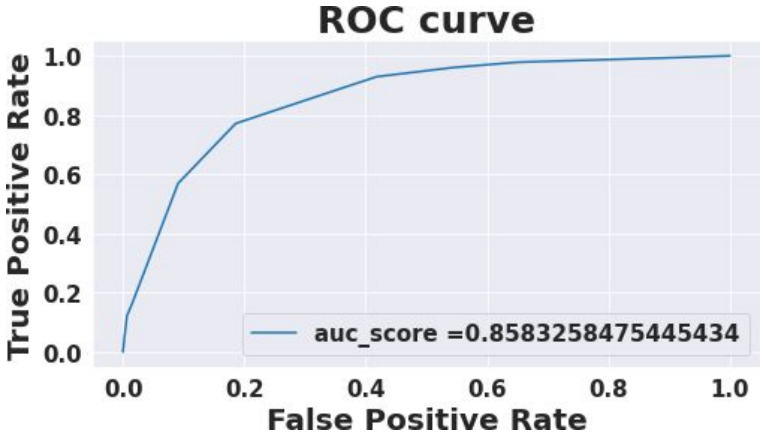


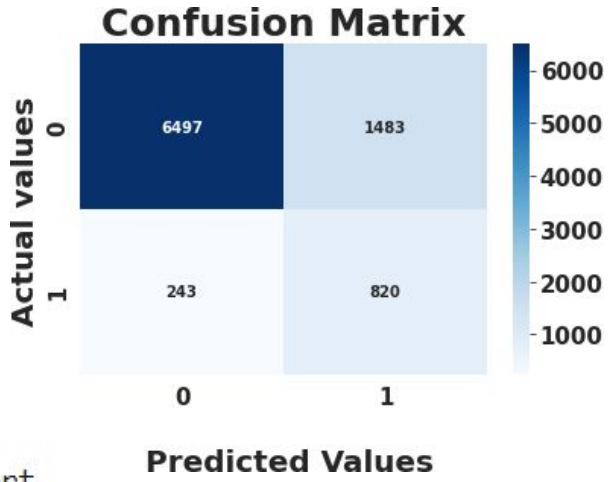
Classification report for Testing

	precision	recall	f1-score	support
0	0.97	0.84	0.90	7980
1	0.41	0.81	0.54	1063
accuracy			0.84	9043
macro avg	0.69	0.83	0.72	9043
weighted avg	0.90	0.84	0.86	9043

Test score :0.839765564525047

Decision Tree Classifier



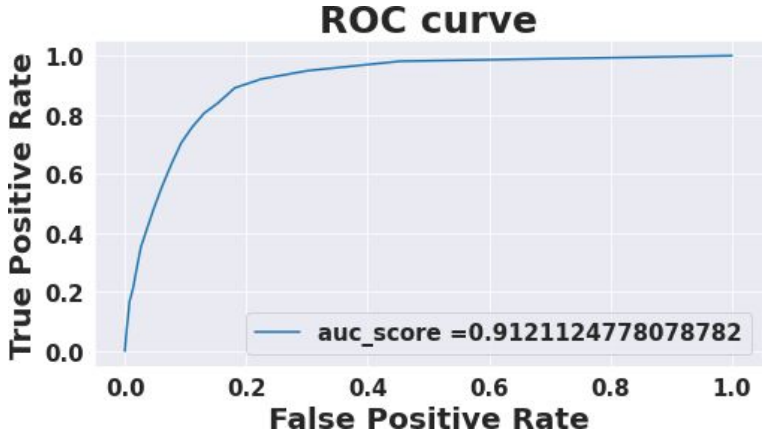


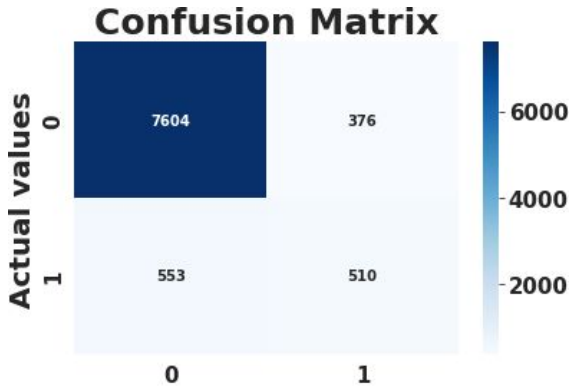
Classification report for Testing

	precision	recall	f1-score	support
0	0.96	0.81	0.88	7980
1	0.36	0.77	0.49	1063
accuracy			0.81	9043
macro avg	0.66	0.79	0.68	9043
weighted avg	0.89	0.81	0.84	9043

Train Score :0.8142754989273874
Test score :0.8091341369014707

Random Forest Classifier



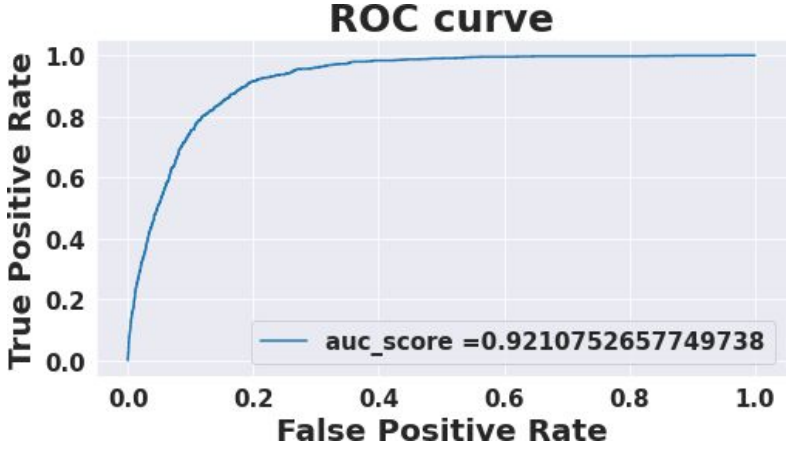


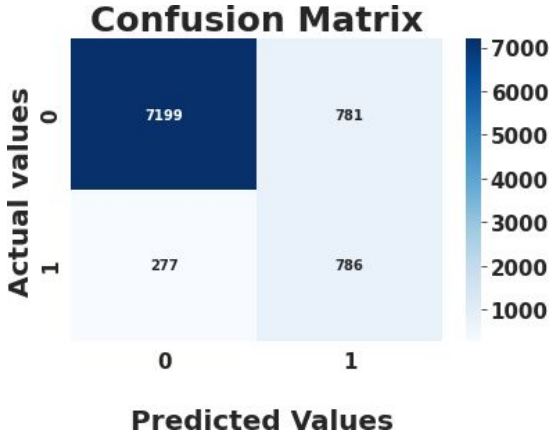
Classification report for Testing

	precision	recall	f1-score	support
0	0.93	0.95	0.94	7980
1	0.58	0.48	0.52	1063
accuracy			0.90	9043
macro avg	0.75	0.72	0.73	9043
weighted avg	0.89	0.90	0.89	9043

Train Score :0.9997183186754718
Test score :0.8972686055512551

XGBoost Classifier



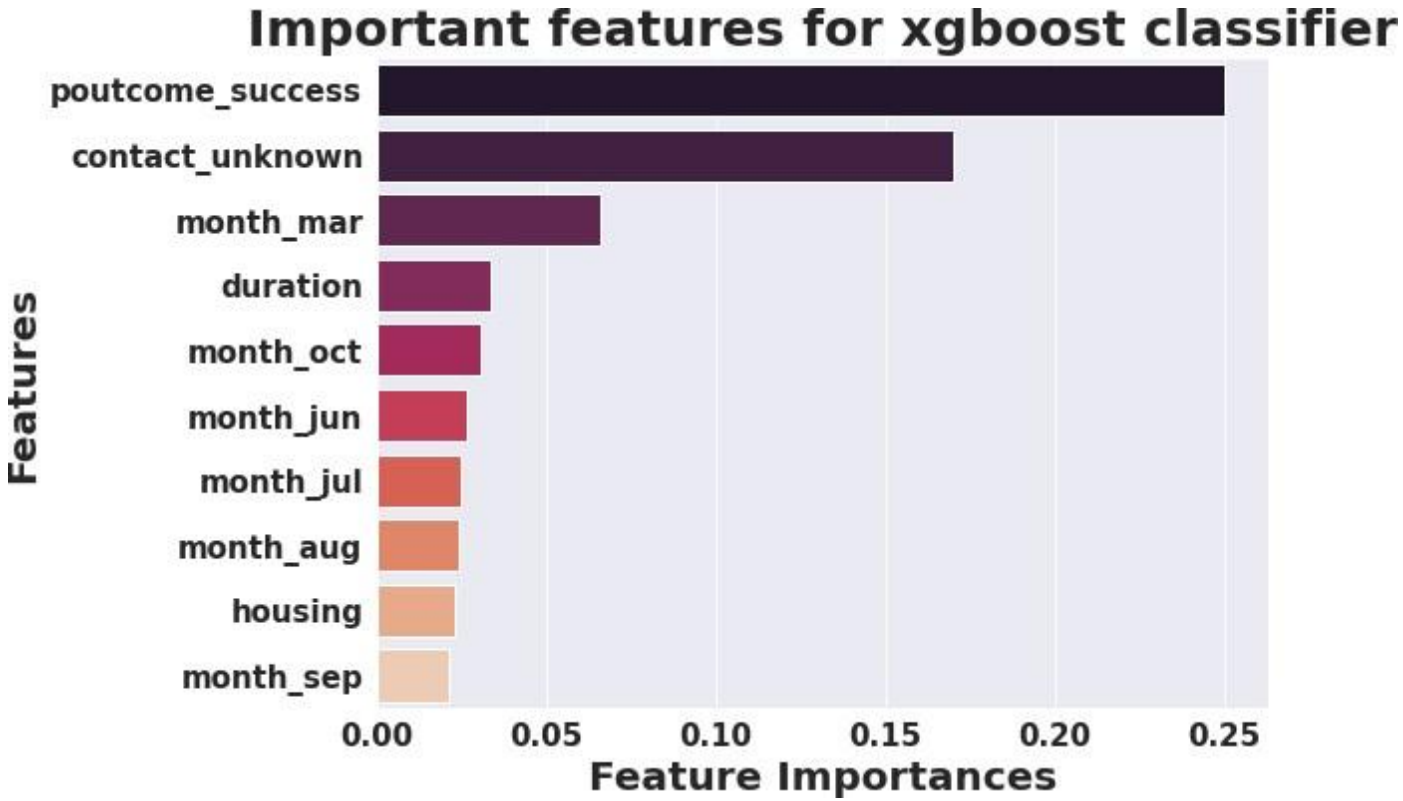


Classification report for Testing

	precision	recall	f1-score	support
0	0.96	0.90	0.93	7980
1	0.50	0.74	0.60	1063
accuracy			0.88	9043
macro avg	0.73	0.82	0.76	9043
weighted avg	0.91	0.88	0.89	9043

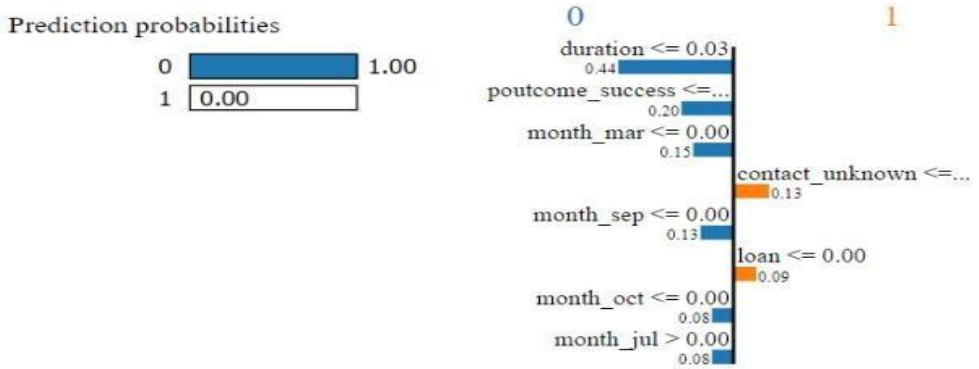
Train Score :0.9566108231891045
Test score :0.8830034280659074

Feature Importance



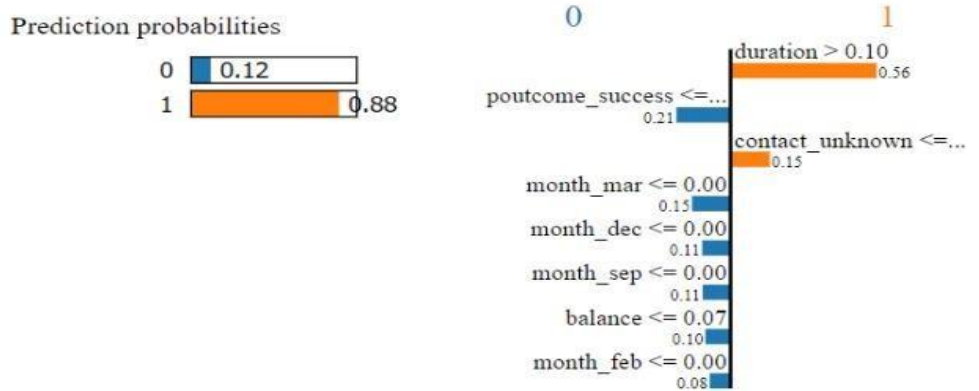
LIME (Local Interpretable Model:Agnostic Explanation)

```
# Get the explanation for observation 1
exp = explainer.explain_instance(X_test[0], predict_fn_xgb, num_features=8)
exp.show_in_notebook(show_all=False)
```



Feature	Value
duration	0.02
poutcome_success	0.00
month_mar	0.00
contact_unknown	0.00
month_sep	0.00
loan	0.00
month_oct	0.00
month_jul	1.00

```
# Get the explanation for observation 2
exp = explainer.explain_instance(X_test[1], predict_fn_xgb, num_features=8)
exp.show_in_notebook(show_all=False)
```



Feature	Value
duration	0.22
poutcome_success	0.00
contact_unknown	0.00
month_mar	0.00
month_dec	0.00
month_sep	0.00
balance	0.07
month_feb	0.00

Challenges

- Feature Engineering.
- Handling Class Imbalance.
- Selecting feature to train model.
- Model training, hyperparameter tuning and improving corrected of prediction for both category.

Conclusion

- First we trained our model before handling class imbalance our model performed very good on 0 category and very poor for category 1.
- After solving class imbalance we trained and compared performances of logistic regression, Decision Tree classifier, Random forest classifier and Xgboost classifier.
- After tuning hyperparameter Xgboost model gives best performance TP = 7199, FP=781, TN = 786 and FN=277) and we got auc score of 0.9210.

Thank You