A Report on

Bank Marketing Campaign

Predicting Term Deposit Subscription

Introduction

Marketing is the process of introducing and promoting the product or service to the market and encourages sales from the buying public. The study examines the fundamental role of marketing in the Banking Industry. Banks need effective marketing strategies to retain their existing customers and attract new customers. Due to increase in competition, it has become imperative for banks to use marketing tool to increase their marketing share by providing awareness of their products to their prospective customers. Good marketing has become increasingly vital ingredient for banks. It is embedded in everything we do Marketing practices are continually being refined and reformed in financial industries to increase the chances of success.

The aim of the study is to find the best strategies to improve for the next marketing campaign. How can the financial institution have a greater effectiveness for future marketing campaigns? To predict the term deposit subscription in Bank, whether the customer is going to subscribe the term deposit in Bank or not ? . This study is on classification problem for bank marketing team for their improvements in forth coming strategies for customers. How can the financial institution have a greater effectiveness for future marketing campaigns? In order to answer this, we have to analyze the last marketing campaign the bank performed and identify the patterns that will help us find conclusions in order to develop future strategies.

A Term deposit is a deposit that a bank or a financial institution offers with a fixed rate that is often better than just opening deposit account in which your money will be returned at a specific maturity time. Key takeaways are as follows, 1. A term deposit is a type of deposit account held at a financial institution where money is locked up for some set period. 2. Term deposits are usually short-term deposits with maturities ranging from one month to a few years. 3. Typically, term deposits offer higher interest rates than traditional liquid savings accounts, whereby customers can withdraw their money at any time.

Approach to successively come out with the best prediction, in order to optimize marketing campaigns with the help of the dataset, we will have to take the following steps: Import data from dataset and perform initial high-level analysis: look at the number of rows, look at the missing values, look at dataset columns and their values respective to the campaign outcome. Clean the data: remove irrelevant columns, deal with missing and incorrect values, turn categorical columns into dummy variables. Use machine learning techniques to predict the marketing campaign outcome and to find out factors, which affect the success of the campaign.

Thus, from the foregoing, the study is designed to examine to predict the term deposit subscription in Bank, whether the customer is going to subscribe the term deposit in Bank or not ? in general.

Benchmark

There is some work done previously for this particular task, have gone through some of the research papers and Kaggle notebooks and noticed their work.

As per one work done greatly on this dataset before, they have performed some of the informative analytics and modeling. To be specific, they have tried implementing 6-7 different machine learning models on this dataset to find the best fit algorithm. First and foremost, that work shows that KNN, Decision tree and logistic regression and the outcome they have got are 0.80, 0.78 and 0.82 with the best possible parameters. Furthermore, to generalize the accuracy or come out with best possible solution, they tried applying some method and then applied other ML algorithms such as Support Vector Machine(SVM), Random Forest, Naïve Bayes and neural classifier. Thus, after performing these many algorithms, that work has highest and best accuracy is 0.84 with SVM, Random Forest and Neural classifier. However, we have performed analytical methods and applied different ML models and tried to come out with best accuracy and another thing is we tried to implement four ML models at once and finding best one from that and ended up with best boosting algorithm which gives better results.

Methods

To start with all the processes and steps taken in order to successful completion of this project are explained in detail.

Data Sources

There are many data providing platforms available online where good amount of records and data are available, and they are open source so they can be used for practice. Here, we have found our interested topic and the dataset from Kaggle which largest data science community in the world and provides many datasets from different fields.

We have found one very reliable and good dataset for bank marketing analysis from Kaggle and worked on the following dataset. https://www.kaggle.com/janiobachmann/bank-marketing-dataset

Original source is [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

Exploratory Data Analysis

By observing the dataset, the necessary steps have been taken and have done some explanatory data analysis on this dataset.

First and foremost, it is necessary to look at overview of the data and shape of data. So, for that we have first seen data overview and shape by pandas **.head()** and **.shape** methods respectively which can be seen in notebook_25. To move further, team members have taken out basic summary and statistical information by using **.info()** and **.describe()** methods in which we looked at mean, standard deviation, min-max values and datatypes

of all the features. Then, we saw unique values by using **.unique()** to see that what are the data and which feature should be considered for label and in return one variable we found which can be used for label y and that is deposit holding yes and no values.

Now, it is time to check the missing values as all the datasets do not come ready made, there are chances of missing values or unwanted features. So, to check missing values in all the features, we have used commonly used method that is **dataframe.isnull().sum()** which helped us to analyze the insights of records and this can be seen in notebook_25.

Next step is to see columns which are not important for this analysis which cannot make any difference by their presence. So, to see unwanted columns which are not practically required for analysis, we **group by** features with label feature that is deposit and by doing that we analyzed that couple of the columns are not required and thus, they are dropped from data frame with the use of **.drop()** pandas method.

After that all the features have no numeric values remained, in order to make data frame ready for model fitting, we must convert all the features into numeric where relates. So, here we have defined categorical and numerical columns for further analysis with the use of **.dtypes**.

Exploring categorical features by **unique values** and **visualizing** all the categorical features with the label feature using **seaborn** library and the output will be shown in results and then with help of **bar chart**, frequency of each category feature with label.

By performing previous visualizations, we were able to understand the categorical variables properly by looking deep into all the records of each column and how they relate to deposit that label feature and for looking at the distribution of categorical features with **matplotlib** .

After categorical features are understood, numerical columns are defined separately and will look into. For visualizing numeric features, we have used **seaborne library** with **bar graph.**

After this we are looking at continuous and discrete values in numerical data with help of number of unique values and in return, we have all the numeric features are continuous and there is no discrete feature.

To look into the distribution of continuous features, we have plotted one **histogram** with help of **seaborne library's distplot()**.

So, by looking at his graph we can see that there are some features such as duration, previous, balance, campaign and pdays are totally left skewed and there are some outliers as well. Whereas, Age and Day have distributed normally.

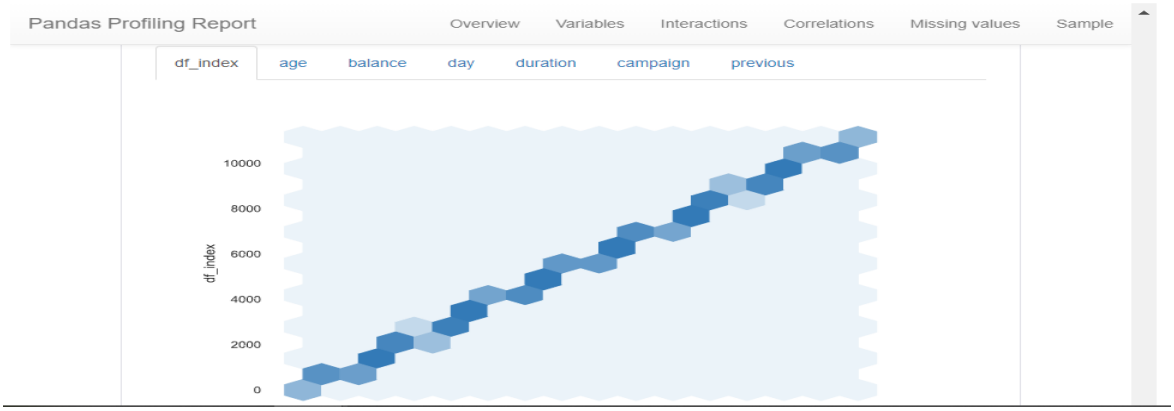
So, based on the previous graph we can see that there are some outliers in few features, and they need to be handled properly and for that we are using **box plot** here to find out outliers in numeric features.

There are some of outliers, but the problem is if we remove that one's then there could be effect in other features as well as most of the features are corelated. So, after that we are checking count for the features having outlier with deposit values(yes and no) and wherever there are more than 30 counts we are neglecting that other values.

This time we have come through one informative method for this data frame insights and information and that is **profile report**, **pandas-profiling** is a library that provides profile report method to print out all the and it makes data profiling and EDA process a breeze.

```
from pandas_profiling import ProfileReport
PR = ProfileReport(bank_df1)
PR
```

```
Summarize dataset: 100%  30/30 [00:12<00:00, 2.39it/s, Completed]
Generate report structure: 100%  1/1 [38:27<00:00, 2307.88s/it]
Render HTML: 100%  1/1 [07:57<00:00, 477.75s/it]
```



Now we want to look at the correlation matrix hit map to show which variable is having a high or low **correlation** in respect to another variable.

```
correlation=bank_df1.corr()
fig = plt.figure(figsize=(10,10))
sns.heatmap(correlation,annot=True)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2f290eb2b48>
```

So now, data is read to go but for cross check we are visualizing all the features with the label feature that is deposit with the help of seaborne.

Data Pre-processing

Now, in this part first thing is we are converting all the yes and no values into Boolean values that is 0 and 1 and all the values are type numeric and having no missing values.

Data Modeling

Now that the whole data frame is ready, it is time to define x and y values and then we are splitting the data frame into 80/20 ratio which for train and test respectively by importing **train_test_split module**. And we are using **stratify** for label feature for equally distribution of data.

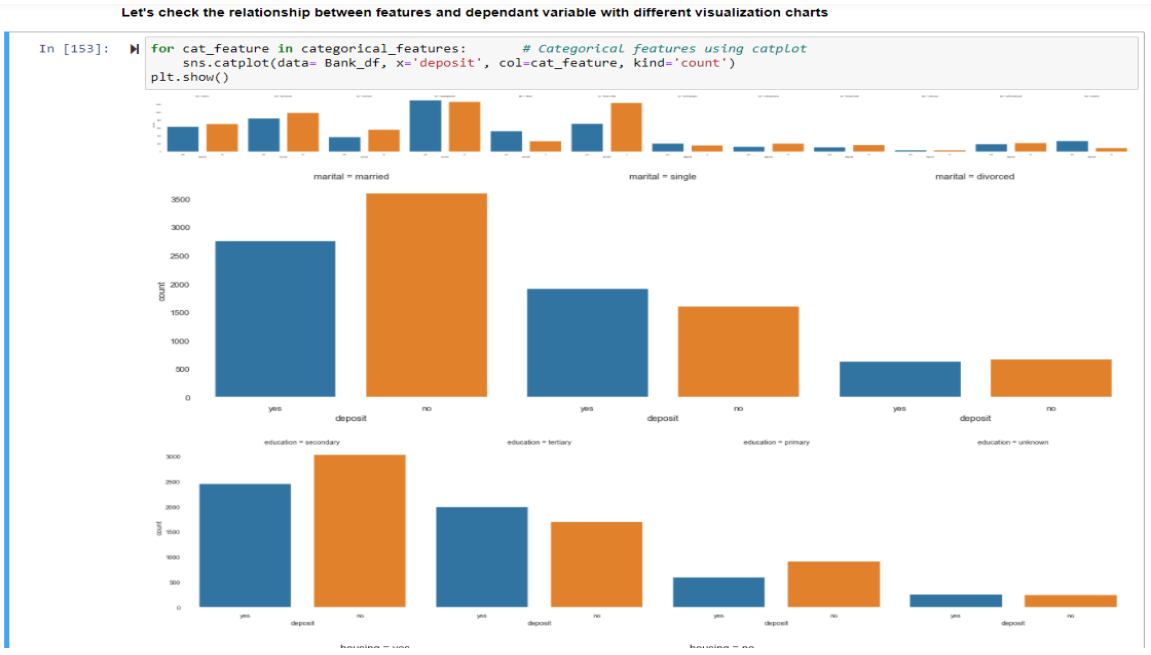
After splitting the data frame, we have implemented some Machine learning algorithms such as **KNN, Logistic regression, SVM(Support Vector Machine), Random Forest Classifier, Decision tree, catboost and xgboost algorithms** on train and test data and we took out training and testing accuracy.

After applying ML algorithms, we are printing feature importance for all the features with the f1-score by using **plot_importance()** method to look what are the features can be considered for bank marketing campaign.

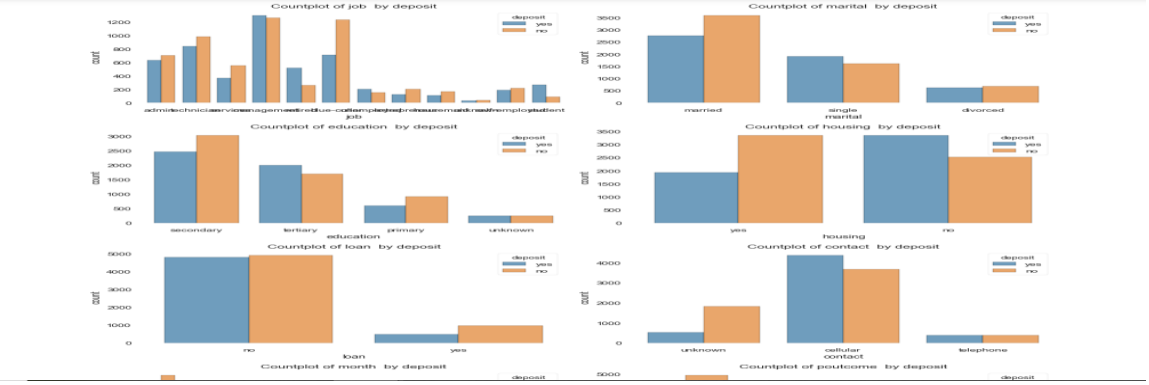
Results

All the above discussed methods have some outcomes in form of different charts and graphs.

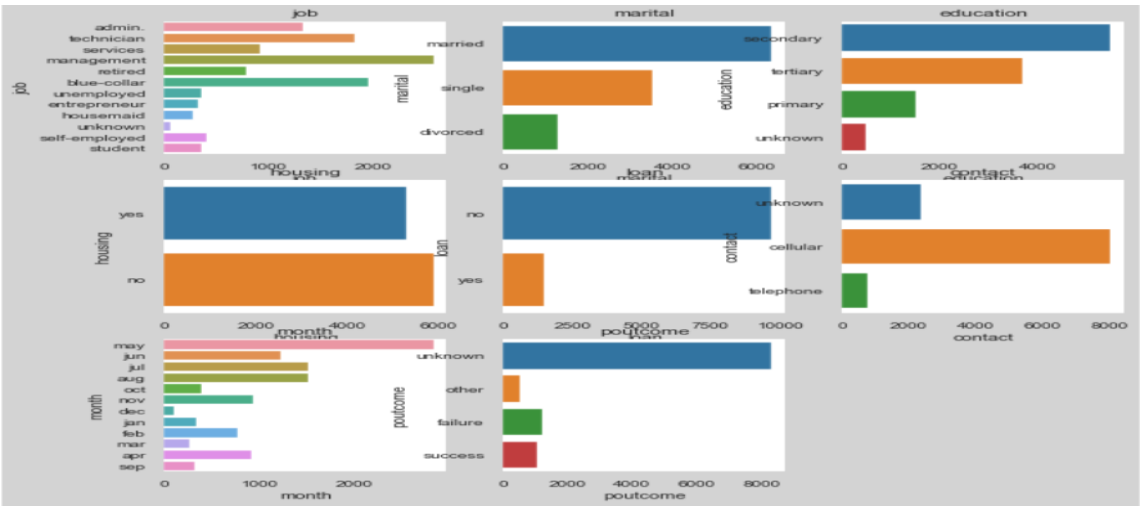
To starts with, we first took out categorical features visualization and the output is in the following:



The above chart tells us the relation of label feature with each feature that how they corelate.

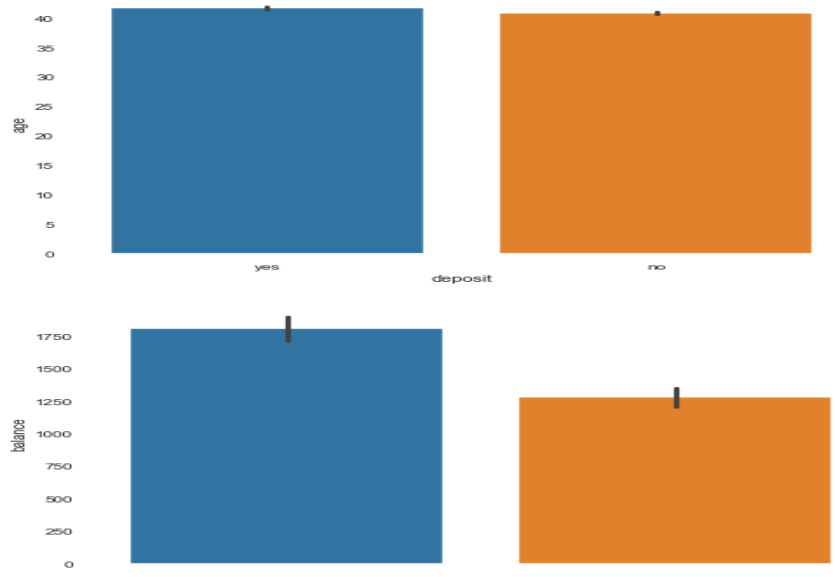


Now, let's see the distribution of categorical features with **matplotlib**.

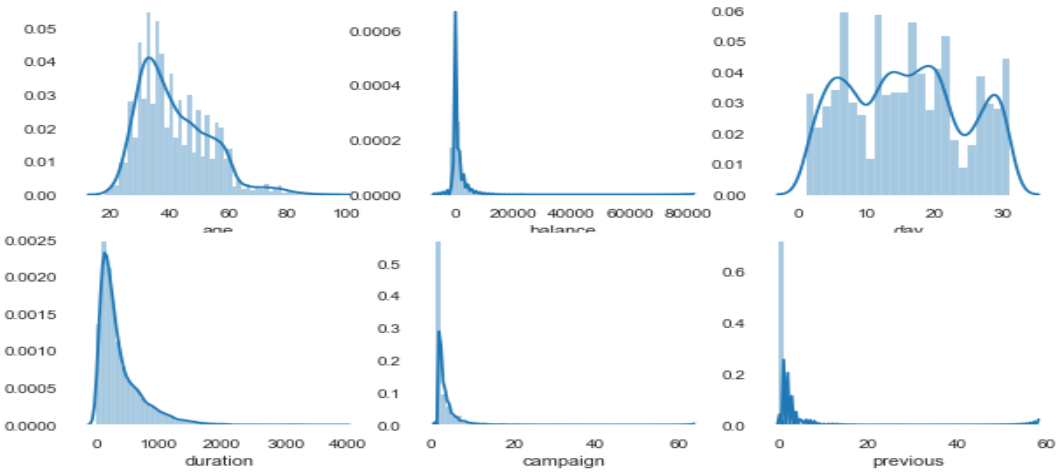


Same as categorical features, the numeric features are also visualized as discussed in methods.

```
In [159]:  # Exploring numeric data with deposit label using bar chart
           for i in num_features:
               sns.barplot(Bank_df.deposit, Bank_df[i])
               plt.show()
```



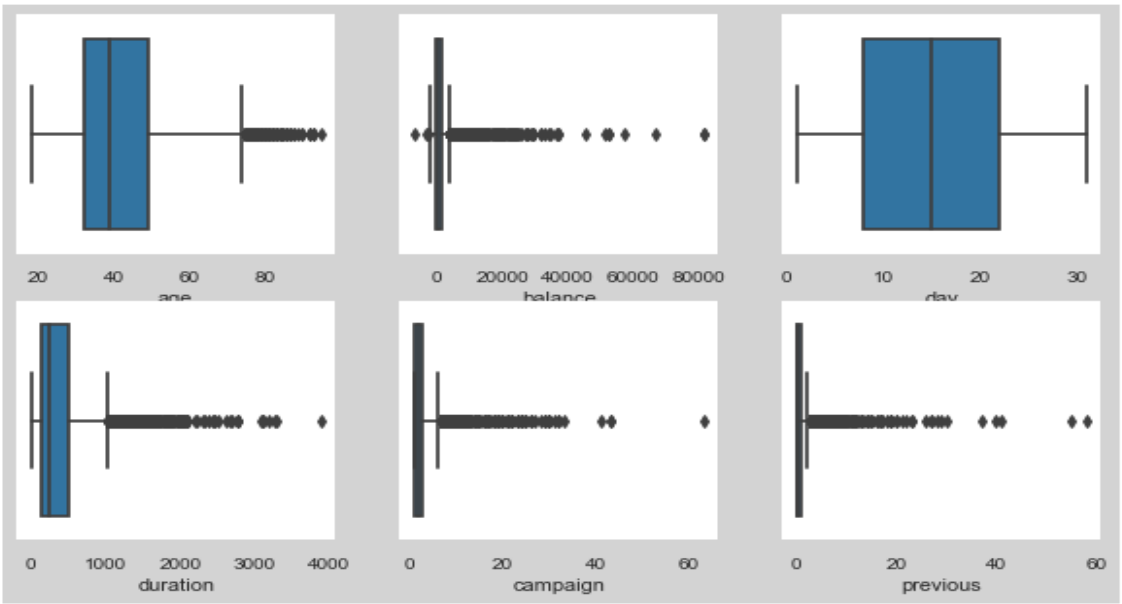
Foe numeric continuous features, we are visualizing it with histogram.



There are some features such as duration, previous, balance, campaign and pdays are totally left skewed and there are some outliers as well.

To look at the outliers in each feature, boxplot graph is used to visualize it.

```
plt.figure(figsize=(10,40), facecolor='lightGray')
pltnum =1
for num_feature in num_features:
    ax = plt.subplot(12,3,pltnum)
    sns.boxplot(Bank_df[num_feature])
    plt.xlabel(num_feature)
    pltnum+=1
plt.show()
```

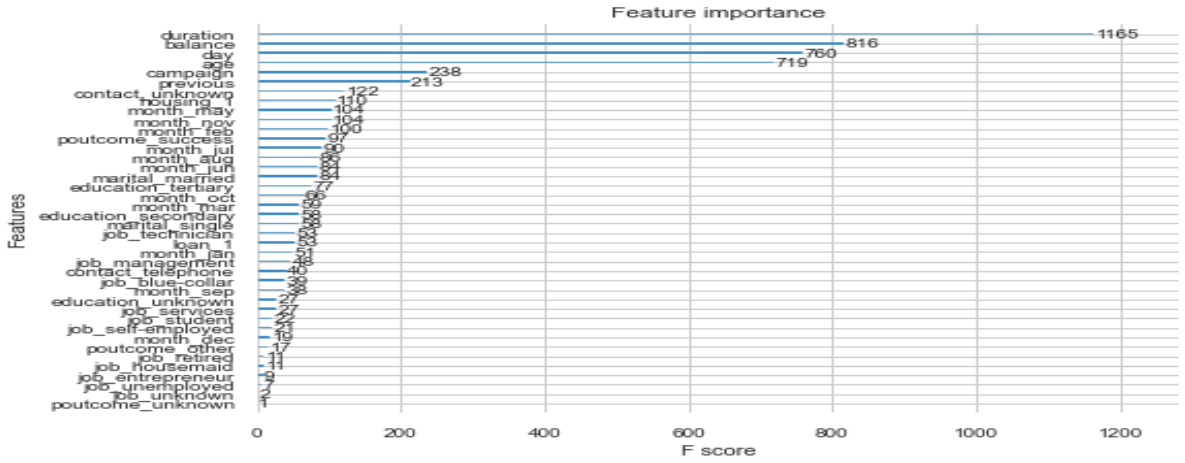


There are some of outliers, but the problem is if we remove that ones then there could be effect in other features as well as most of the features are corelated. So, we are removing more than 30 values as outliers in campaign.

ML Models	Accuracy
KNN	0.76
Logistic regression	0.83
SVM(Support Vector Machine)	0.82
Random Forest Classifier	0.83
Decision tree	0.77
catboost	0.8613
xgboost	0.8649

We have performed above mentioned models and printed out test accuracy of all the ML models and the best accuracy is coming from boosting algorithms. Thus, we can say that xgboost is fitting best on this data frame which giving highest accuracy of 0.8649

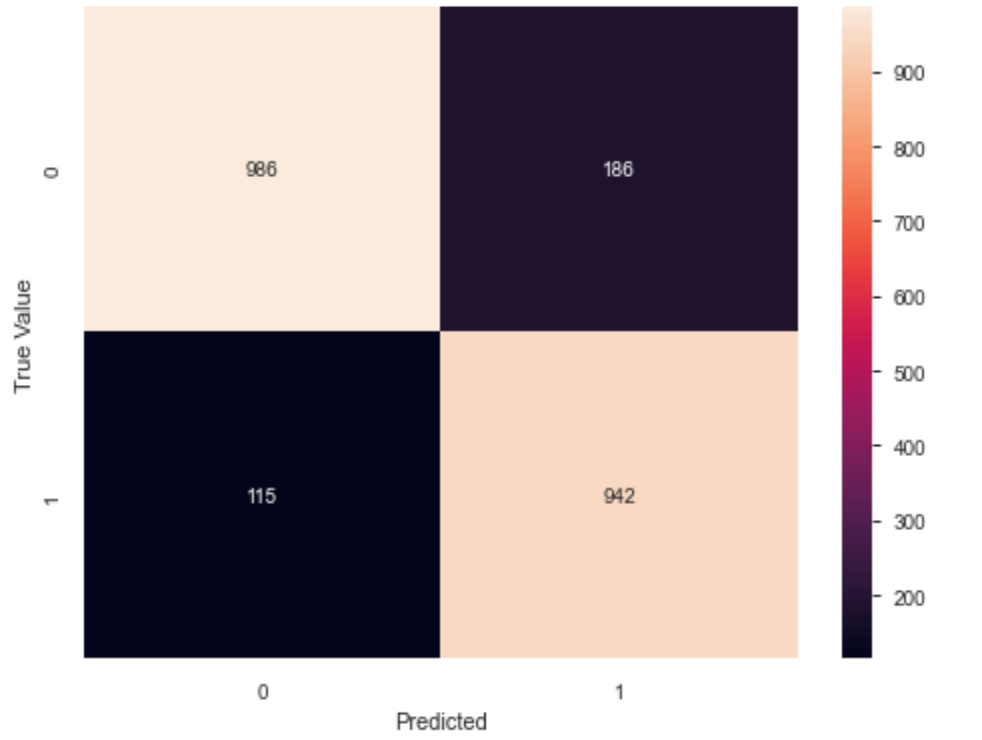
For looking at the features from this data frame that which are the features important and can be considered for bank marketing campaign, we are using plot_importance() method.



Confusion Metrix

```
#plot the graph

from matplotlib import pyplot as plt
import seaborn as sn
sn.heatmap(cmat, annot=True,fmt='g')
plt.xlabel('Predicted')
plt.ylabel('True Value')
plt.show()
```



Discussion

As discussed above that there are many projects done on this topic and they have done really great job doing this task. If we talk about visualization, then there are a lot of different aspects of methods to look into data and understand it and so, ours as we have also done some different and understandable methods for EDA. If we talk about modeling, there are many ML models performed by others and they are getting nearly same accuracy in comparison to our accuracy with many models. But the main significance is we have used boosting algorithms and we would like to add here is previous works have also used boosting algorithms and they have ended up getting accuracy near 84%. Whereas, we applied boosting algorithms such as catboost and xgboost in which we are getting 86% which would be the accuracy we expected to achieve. Talking about challenges, we had challenge of cleaning this dataset and making ready this dataset for better fitting and we cleaned it very precisely and learned some things which we do not know, and we learned it and performed.

Conclusion

To conclude, after performing ML models on the data frame we have got different accuracy with different models, as least accuracy we got is 77% in Decision tree and there are models having accuracy near 84% but highest accuracy we got from catboost and xgboost. Thus, Gradient Boosting classifier is the best model to predict whether a potential client will subscribe to a term deposit or not. 86% accuracy. When performed feature importance on xgboost model, we concluded that the most important features are customer's account balance, customer's age, number of contacts performed during this campaign and contact duration, number of contacts performed before this campaign. So the main outcomes of the modelling are customers of greater age are more likely to subscribe for the term deposit, customers with greater account balance are more likely to subscribe for the term deposit and number of contacts with the customers really matters. Too many contacts with the customer could make him or her decline the offer.

Contribution

So far, the distribution of work is as per below but may change according to the knowledge of the individual and have planned to do every work of this project as a team an d will always work together.

Name	**Distribution**
Anant Patel	Research work , EDA, Data Preprocessing, Model Building, Report Building
Darshankumar Patel	Research work , Data Cleaning, Data Exploration, Model Building, Report Building
Rushil Patel	Research work , Feature Engineering, EDA, Data Preprocessing, Report Building
Parth Sutariya	Research work , Data Cleaning, Data Exploration, Model Building, Report Building
Ruchita Tamboli	Research work , Data Cleaning, Feature Engineering, Model Building, Report Building

References

Dataset
https://www.kaggle.com/janiobachmann/bank-marketing-dataset

How important is AI/ML in financial services –
https://www.youtube.com/watch?v=DHp16TdV8bg

Will Machine Learning Transform Finance? –
https://www.youtube.com/watch?v=chgG0TgEPGo

AI ML IN FINANCIAL MARKET–
https://www.youtube.com/watch?v=OhOvr8_o9GI

Bank Marketing Project | Machine Learning Approach–
https://www.youtube.com/watch?v=dNdleMni4J8

Research Paper -Bank direct marketing analysis of asymmetric information based on machine learning
https://ieeexplore.ieee.org/abstract/document/7219777

Research paper -Classification of Bank Direct Marketing Data Using Subsets of Training Data
https://link.springer.com/chapter/10.1007/978-81-322-2757-1_16

https://www.investopedia.com/terms/t/termdeposit.asp

https://stackoverflow.com/questions/45201514/edit-seaborn-legend

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.429.563&rep=rep1&type=pdf

https://www.researchgate.net/profile/Alaa_Abu_srhan/publication/334943218_Visualization_and_Analysis_in_Bank_Direct_Marketing_Prediction/links/5d67e007458515a1c05f5a4c/Visualization-and-Analysis-in-Bank-Direct-Marketing-Prediction.pdf

https://github.com/Bharat-Reddy/Bank-Marketing-Analysis

https://github.com/topics/bank-marketing-analysis

https://medium.com/swlh/exploratory-data-analysis-on-the-bank-marketing-data-set-with-pandas-and-seaborn-72e5c05e0076

https://towardsdatascience.com/machine-learning-case-study-a-data-driven-approach-to-predict-the-success-of-bank-telemarketing-20e37d46c31c

Appendices

The whole code used for the analysis of this chosen dataset that is bank marketing campaign is in the **notebook_25.ipynb.**

Group Members

Name	Student ID	Section
Anant Patel	W0756333	1
Rushil Patel	W0755980	1
Darshankumar Patel	W0753849	2
Parth Sutariya	W0756042	5
Ruchita Tamboli	W0756325	5