Predicting Customer Responses to Phone Marketing Campaign

Rock Central DSA Case Study: **Neviya Prakash**

Business Objective

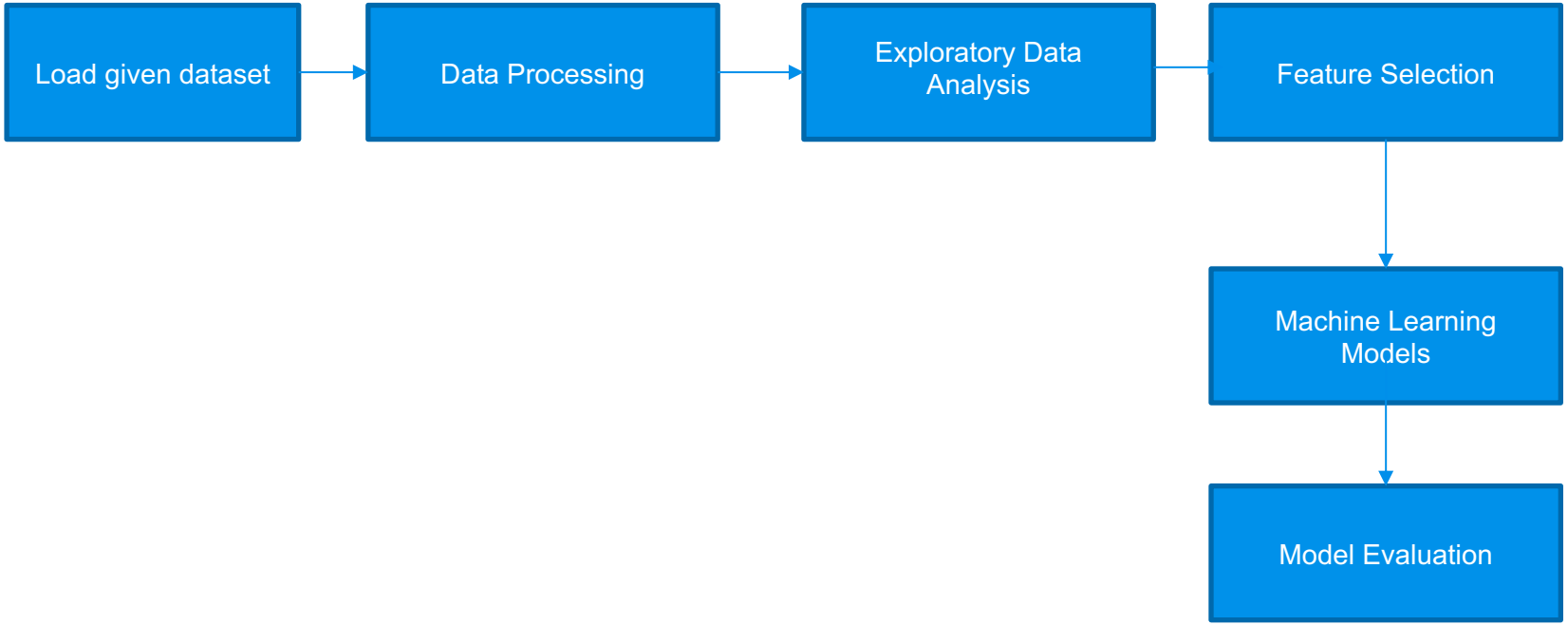
- This project aims to analyze the given data to identify data patterns & help the European banking institution to determine, in advance, clients who will be receptive to their phone marketing campaigns.

- We will try to identify factors affecting customer response & build a prediction model to determine if a customer will subscribe yes(1) or no(0) to a term deposit (y).

Approach

- **Load data :** Perform basic analysis verify number of rows, identify missing values, check the distribution of Dependent variables to the campaign outcome response.
- **Data Cleaning** : drop irrelevant columns, check for null/missing values, turn categorical columns into numeric using One hot encoding
- **Exploratory Data Analysis**: Identify patterns in the numeric and categorical columns with respect to response variable y
- **Feature Selection**: Identify methods to determine the relevant variables affecting the campaign.
- **Machine learning algorithm:** predict the marketing campaign outcome based on the relevant factors which affect the success of the campaign.

Work Flow


Load given dataset
Data Processing
Exploratory Data Analysis
Feature Selection
Machine Learning Models
Model Evaluation

Data Cleaning

- Given dataset has **41188** records and **22** variables containing numeric & categorical data with y being the response
- Dataset has no Null values

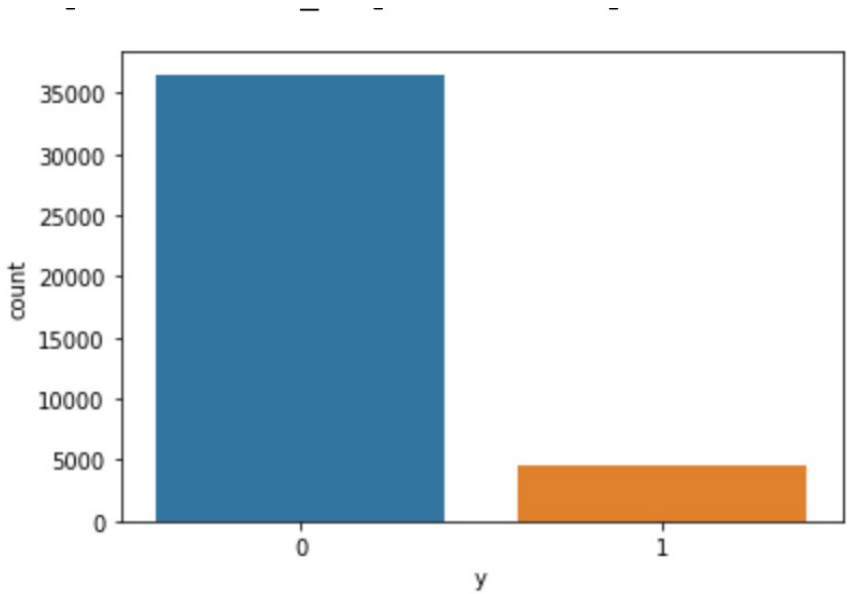
	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
count	41188.00000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000
mean	40.02406	258.285010	2.567593	962.475454	0.172963	0.081886	93.575664	-40.502600	3.621291	5167.035911
std	10.42125	259.279249	2.770014	186.910907	0.494901	1.570960	0.578840	4.628198	1.734447	72.251528
min	17.00000	0.000000	1.000000	0.000000	0.000000	-3.400000	92.201000	-50.800000	0.634000	4963.600000
25%	32.00000	102.000000	1.000000	999.000000	0.000000	-1.800000	93.075000	-42.700000	1.344000	5099.100000
50%	38.00000	180.000000	2.000000	999.000000	0.000000	1.100000	93.749000	-41.800000	4.8	click to expand output;
75%	47.00000	319.000000	3.000000	999.000000	0.000000	1.400000	93.994000	-36.400000	4.961000	5228.100000
max	98.00000	4918.000000	56.000000	999.000000	7.000000	1.400000	94.767000	-26.900000	5.045000	5228.100000

Statistic distribution of Numeric variables

Exploratory Analysis

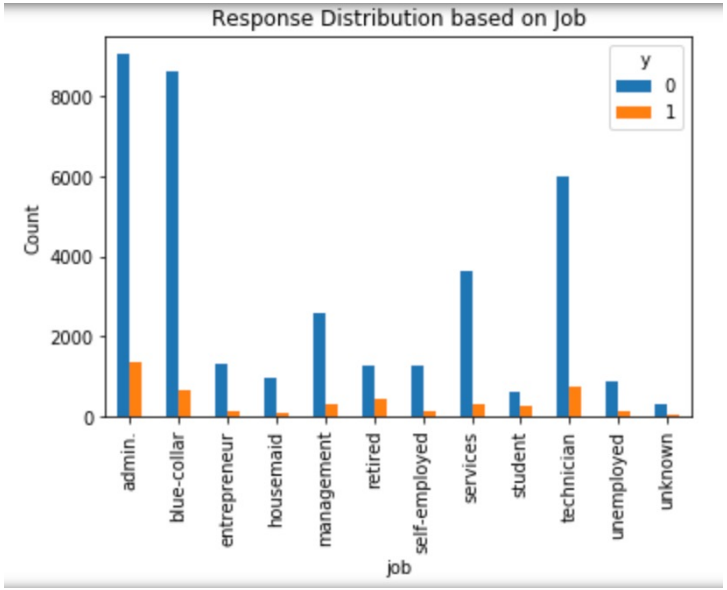
1. **Target Variable distribution**

- 88% of customers said no(0) and only 11% of customers said yes(1) to the campaign



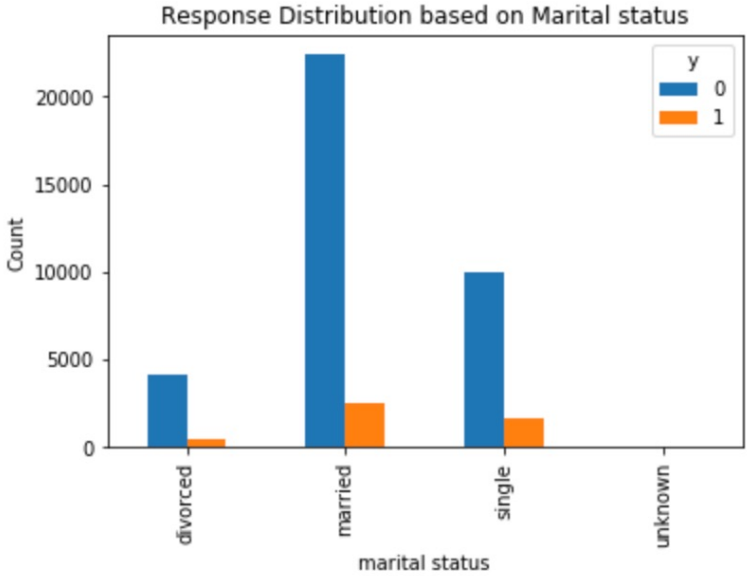
Exploratory Analysis: Categorical Variables

Jobs vs Response y



- **Admin** jobs have the highest **yes rate** of subscribing a term deposit, but they also have the highest no rate, since count of admin jobs are highest in our dataset.

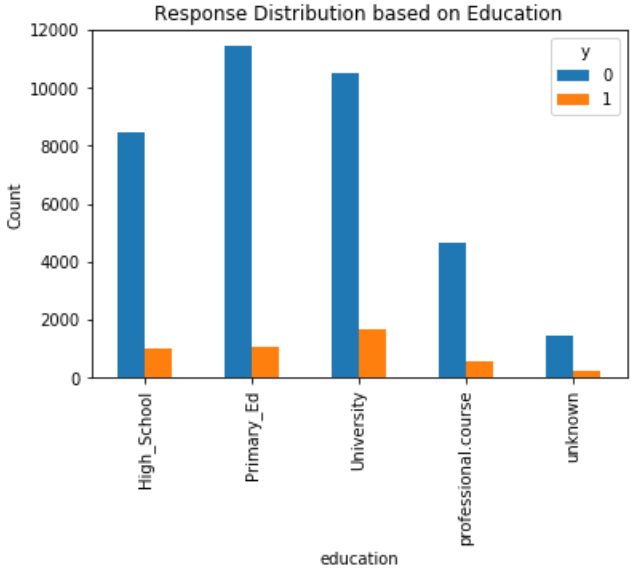
Marital Status vs Response y



- Married customers seem to be more responsive to the campaign

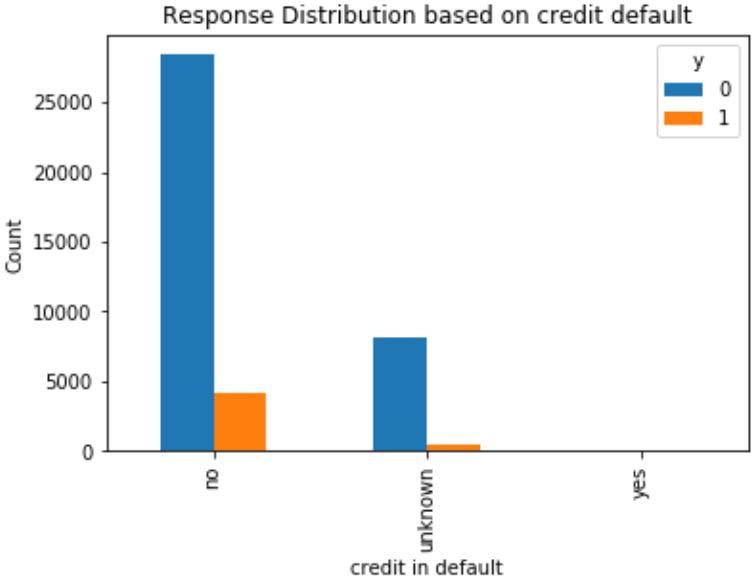
Exploratory Analysis: Categorical Variables

Education vs Response y



- **University** education has the highest yes rate of subscribing a term deposit.
- **Primary education** seems to have the highest **No rate**, it could possibly be due to the lack of banking knowledge in the younger age group

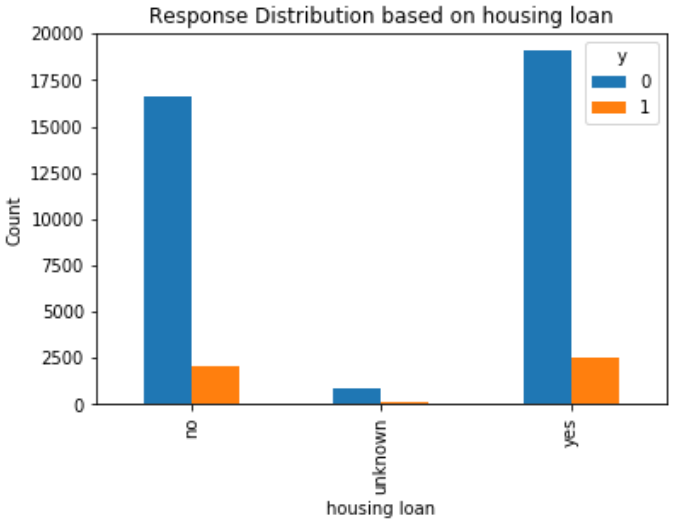
Credit Default vs Response y



- Customers with No default credit show a higher No rate to the campaign

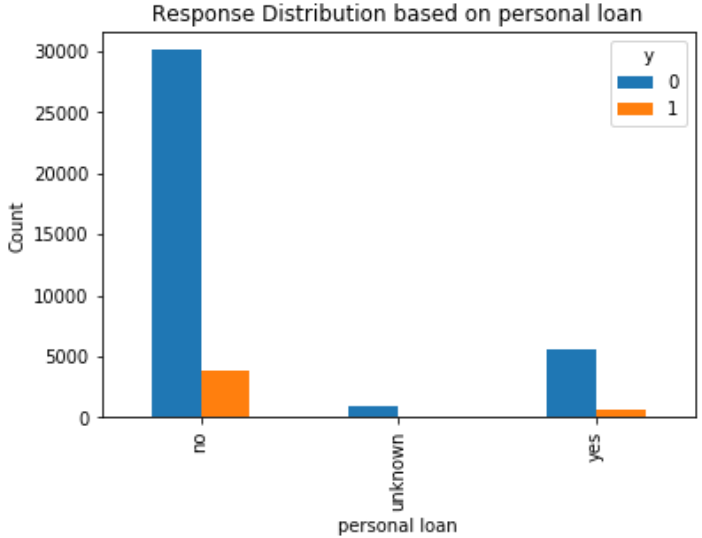
Exploratory Analysis: Categorical Variables

House Loan vs Response y



- Clients with Housing Loan have a higher No rate to the campaign

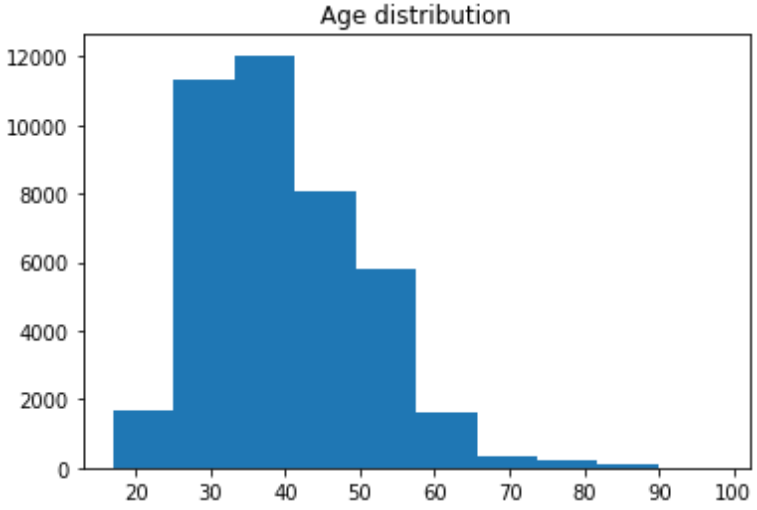
Personal Loan vs Response y



- Clients with personal Loan have a higher No rate to the campaign

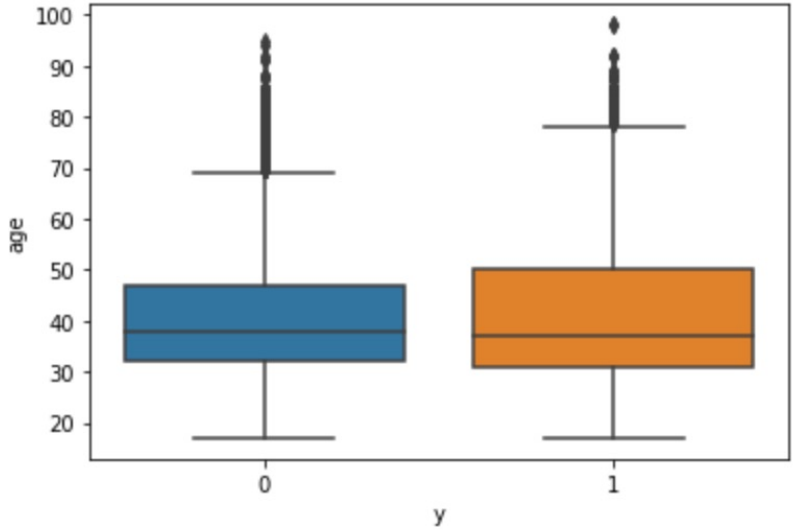
Exploratory Analysis: Numeric Variables

Age Distribution



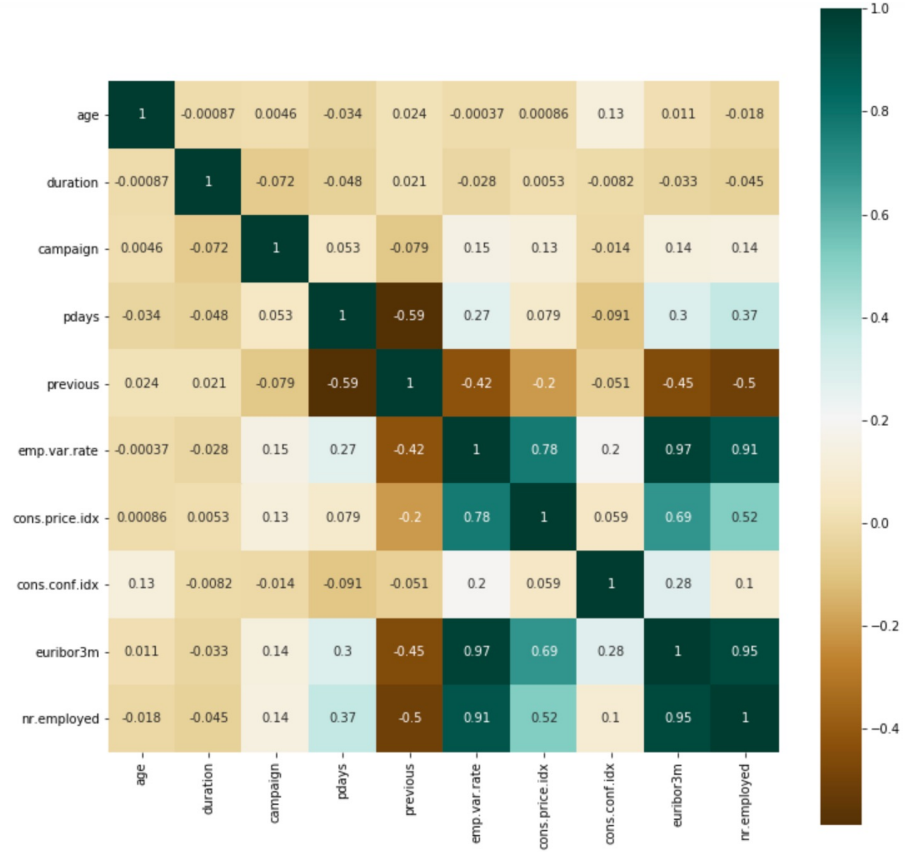
- Age category follows a normal distribution

Personal Loan vs Response y



- Customers that subscribed or didn't subscribe to a term deposit, has a median age of around **38-40 years**

Correlation matrix of numerical features



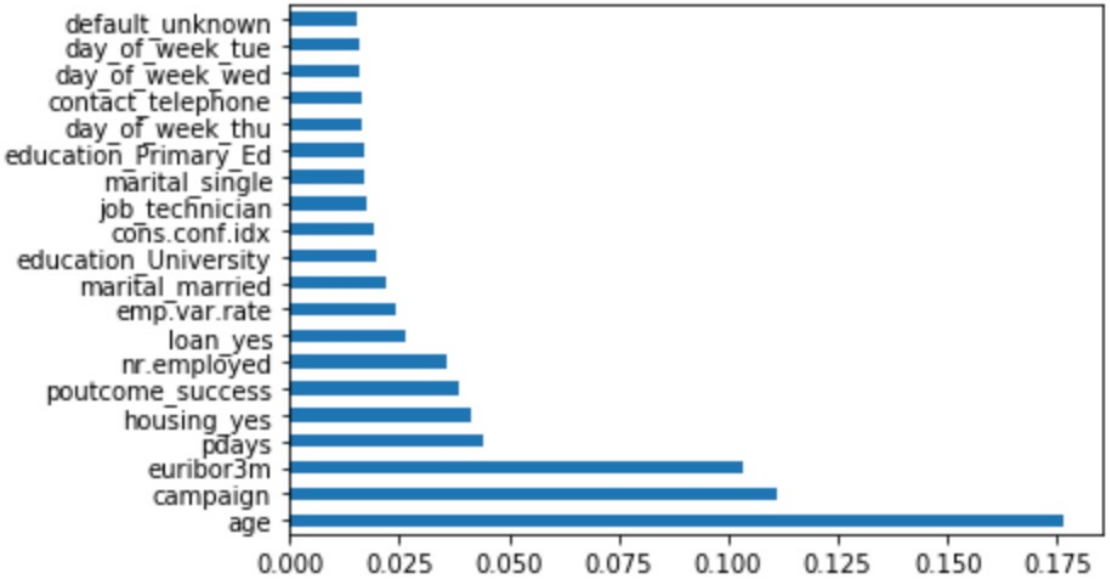
- From the heatmap we can see that there are some numerical features with a high correlation , e.g **nr.employed** and **euribor3m** these features share a correlation value of **0.95**
- **euribor3m** and **emp.var.rate** share a correlation of **0.97,** which is very high compared to the other features that we see in the heatmap.

Feature selection: RFE (Recursive Feature Elimination)

	Rank	Feature
24	1	poutcome_nonexistent
26	1	contact_telephone
25	1	poutcome_success
47	1	default_unknown
37	1	day_of_week_wed
38	1	month_aug
39	1	month_dec
40	1	month_jul
41	1	month_jun
33	1	education_unknown
16	1	job_student

The **RFE** has helped us select the following features:
'poutcome_nonexistent','contact_telephone','poutcome_success','default_unknown','day_of_week_wed','month_aug','month_dec', 'month_jul', 'month_jun','education_unknown','job_student', 'job_services','month_mar','day_of_week_mon','job_retired','euribor3m','previous','job_blue-collar'

Feature selection: ExtraTreesClassifier



ExtraTree Classifier has given us the following variables:
['age','campaign','euribor3m','pdays','housing_yes','poutcome_success','nr.employed','loan_yes','emp.var.rate',
'marital_married','education_University','cons.conf.idx','job_technician','marital_single',
'education_Primary_Ed','day_of_week_thu','contact_telephone'
,'day_of_week_wed','day_of_week_tue','default_unknown'}

ML Model : Logistic Regression

1. Model Fitting with all the variables

Results:

- **Test Accuracy Score** 0.8976 Train Accuracy Score: 0.900
- **F1 Score**: 0.29
- The LR algorithm achieves a test accuracy of 89.79% with all the features, suggesting high level of strength to classify the people who are more likely to say no

2. Model Fitting with Selected Important variables

Results:

- **Test Accuracy Score** 0.8981 Train Accuracy Score: 0.901
- **F1 Score**: 0.271
- The LR algorithm achieves a test accuracy of 89.79% with all the features, & 89.8% with Selected Important variables, suggesting high level of strength to classify the people who are more likely to say no

ML Model : Random Forest

1. Model Fitting with all the variables

Results:

- **Test Accuracy Score** 0.8942 Train Accuracy Score: 0.899.
- **F1 Score**: 0.276
- The LR algorithm achieves a test accuracy of 89.79% with all the features, suggesting high level of strength to classify the people who are more likely to say no

2. Model Fitting with Selected Important variables

Results:

- **Test Accuracy Score** 0.8999 Train Accuracy Score: 0.903
- **F1 Score**: 0.271
- The RF algorithm achieves a test accuracy of 89.4% with all the features & 89.9% with Selected Important variables

Result

	Model Name	Test Accuracy	Train Accuracy
0	Logistic Regression with all	0.8976	0.9000
1	Logistic Regression with Selected	0.8981	0.9000
2	Random Forest with All	0.8942	0.8994
3	Random Forest with Selected	0.8990	0.9020

- Looking at the above table, we can see that **Random forest** with important selected features performed the **best** and gave better accuracy compared to all.
- We should go with Random Forest with important features as our final model as it has a **better F1 score, Accuracy & Confusion Matrix**

Conclusion

- We see that from the following parameters related to a target customer we were able to predict customers' response to the term deposit campaign.
'age','campaign','euribor3m','pdays','housing_yes','poutcome_success','nr.employed','loan_yes','emp.var.rate','marital_married','education_University','cons.conf.idx','job_technician','marital_single'

- We see that customer's **age** affects campaign outcome as well. Future campaigns should concentrate on customers between **30 - 48 years old**.

- Number of **calls** with the customer during the campaign is also important. Bank should avoid calling clients more than **6 times** as it could lead to customer dissatisfaction.

Thank You!

Any questions?

Contact me: neviya19@terpmail.umd.edu

LinkedIn: https://www.linkedin.com/in/neviyaprakash/

Github Notebook