- Machine Learning model to predict if a client will subscribe to the product, given his/her demographic and marketing campaign related information.

- **Data Exploration:**

1. There are 45211 observations of 17 variables in original dataset (7-Numerical Variables and 10-Categorical Variables).
2. No explicit missing values but there are many 'unknowns' values for some Categorical Variables that I treated as missing values
3. From the distribution of Target variable: "is_success" it is found that data is imbalanced because there is approx. 88% is 'no' and 12% is 'yes'.

- **Exploratory Data Analysis:**

1. **For Numerical Variables**
   a. Analysis of each Numerical variable by plotting Boxplot with respect to target variable.
   b. Some Independent numerical variable ('balance', 'duration', 'campaign', 'pdays', 'previous') contains many outliers.
   c. I applied capping analysis to remove all the data points which were out of range from lower and upper quartiles

2. **For Categorical Variables**

   a. Analysis of each Categorical variable by plotting Crosstab with respect to target variable.

   b. If any Categorical variable has more than 50% 'unknown' values('poutcome') or seems highly unbalanced ('default') or seems having negligible impact on target variable ('contact'), I dropped that variable from dataset.

3. Variables having less than 50% 'unknown' values are imputed by Mode of respective variable.

- **Feature Engineering**

1. Created new dummy variables to convert Categorical into Numerical.
2. Total variables after creating dummies becomes 39.

- **Feature Selection**

1. Feature selection by Principal Component Analysis. I have selected first 32 components out of 39.

- **Model Training**

1. Implemented Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbor, Decision Tree, Naive Bayes, Support Vector Machine along with Cross Validation.

- **Model Selection**

1. "Support Vector Machine" has highest Accuracy (89.35%) but it is taking more time compare to other algorithms.

2. "Logistic Regression" also has nearly same accuracy (89.17%) and it is very faster than SVM.

3. So I have considered Logistic Regression as Best model for prediction.

- **Prediction**

1. Prediction on Validation Dataset by Logistic Regression with following result:

Accuracy – 0.88

F1-score – 0.87