**Assignment Report**

Jaunel Ansari <jaunelansari@gmail.com>

**Abstract**

This task is based upon determining whether a particular client will take term deposit or not. 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 relevant information about the client.

**Dataset**

For this Assignment, we are using the dataset on bank-marketing-uci from [www.kaggle.com/competitions](http://www.kaggle.com/competitions) on Kaggle. The dataset consists of several independent variables include:

* age
* job
* marital
* education
* default
* balance
* housing
* loan
* contact
* day
* month
* duration
* campaign
* pdays
* previous
* poutcome
* y

**Exploratory Data Analysis (EDA)**

Portuguese banking institution runs different marketing campaign based upon their business plans. During this process they collect client information like age, job, marital etc. Since the data is not in right format which machine learning algorithm can understand so a thorough data processing is required. In this process we analyse the data by plotting several graphs, we try to find out correlation between those features, we look for any kind of missing values in the dataset so that intelligent changes can be made in further processes.

Here we try to find out more insight about the dataset using several EDA techniques like checking for duplicated rows, checking balance of the dataset with respect to the Label etc. Based upon all these details we plan our further course of action. Mostly, Seaborn & Matplotlib library is used during exploratory data analysis (EDA).

**Feature Engineering**

After EDA we have a concreate idea about the dataset so in this step we utilize all those information. Here we convert all categorical feature into numeric one using Label Encoding or One Hot Encoding. Here we keep only required features in the dataset and drop all less relevant feature. Issues like missing values, duplicated rows are also addressed in this step.

**Data Splitting**

In this step we divide the dataset into train and test data. We standardize the input feature in single scale so that machine learning algorithm can process the data easily. We have used StandardScaler module of Scikit-Learn library for standardization. While splitting the data we have to ensure a fair share of all the categorical data gets divided into train & test sample. This can be achieved by using stratified sampling of the data.

**Model Training**

Here we train our model with several machine learning algorithms. For our problem statement, we have used Logistic Regression, Random Forest & Gradient Boosting algorithm for model training. we have calculated Accuracy, Precision, Recall, F1 score & ROC\_AUC \_score of all the model using metric module of Scikit-Learn library.

|  |  |  |  |
| --- | --- | --- | --- |
| Performance Metrics | Logistic Regression | Random  Forest | Gradient Boost |
| Accuracy | * 0.9104 | * 0.8839 | * 0.8983 |
| Precision | * 0.7090 | * 0.4920 | * 0.6001 |
| Recall | * 0.375 | * 0.2980 | * 0.3461 |
| F1 – score | * 0.4905 | * 0.3712 | * 0.4390 |
| Cross validation Accuracy | * 0.9021 | * 0.8901 | * 0.9018 |
| ROC – AUC score | * 0.8816 | * 0.8746 | * 0.8702 |

By comparing all three model we can conclude that :

* Logistic Regression model has slightly better cross validation accuracy as compared with other
* Logistic Regression also has slightly more AUC with respect to other

So finally, we have selected Logistic Regression as the most suitable model for our problem

**Model Optimization**

Some outliers are there in numerical columns like duration & previous so model performance could be improved if these outliers are handled intelligently.