*Dhanushka Abeyratne*

*Daare21@gmail.com*

*Data Science Academy Capstone Project*

**1.0 Introduction**

The Capstone project for ML Course by Data Science Academy consists in building a classification model that capable to predict whether a customer will subscribe a term deposit or not given customer relationship data. The classification goal is to predict if the customers will subscribe a term deposit (target variable y).

This will benefit for bank marketing team to easily identify potential customers in existing database.

**2.0 Data**

The dataset was picked from [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/Bank+Marketing) which is an amazing source for publicly available datasets. There were four variants of the datasets out of which this project conducted with “ bank-full.csv” which consists of 45211 data points with 16 independent variables out of which 7 are numeric features and 10 are categorical features.

The list of features available to us are given below:

|  |  |  |
| --- | --- | --- |
|  | **Feature** | **Data Type** |
| 1 | age | Numeric |
| 2 | job | Categorical |
| 3 | marital | Categorical |
| 4 | education | Categorical |
| 5 | default | Categorical |
| 6 | balance | Numeric |
| 7 | housing | Categorical |
| 8 | loan | Categorical |
| 9 | contact | Categorical |
| 10 | day | Numeric |
| 11 | month | Categorical |
| 12 | duration | Numeric |
| 13 | campaign | Numeric |
| 14 | pdays | Numeric |
| 15 | previous | Numeric |
| 16 | poutcome | Categorical |
| 17 | y | Categorical |

**3.0 Methodology**

This case study can be identified as a binary classification problem. Two classes are “yes” denoting that the customer subscribed to a term deposit, and “no” denoting that the customer did not subscribe. Therefore, classification algorithms such as logistic Regression classifier, Decision Tree classifier and Random Forest classifier were tested on the dataset. Finally, Random Forest Classifier selected due to the higher performance.

Table

Description automatically generated

Comprehensive ML workflow is desgined to obtain best model with parameters to classify the customer base.

Graphical user interface, diagram, application, Teams

Description automatically generated

3.1 data analysis

Exploratory data analysis process conducted to identify feature distributions and potential importance. Highly imbalanced classification is identified in the dataset. Up-sampling and Down-sampling techniques used to handle imbalance data.

Treemap chart

Description automatically generated

Above chart shows the correlation of each feature in the data set, pday and previous shows the considerable correlation.

3.2 Data Preprocessing

One-Hot Encoding and feature transformation done for the potential features to improve the performance of the model.

3.3 Performance Metric Used

The performance metric used for this case study Precision and ROC score known as Receiver Operating Characteristics).

The reason to choose Precision and ROC over accuracy is because, according to the Exploratory data analysis, the dataset we are working with is an imbalanced dataset with the class “no” being the majority class. If we use accuracy as our metric, any random model can give us a very good accuracy. But at the end, it will be a random model. Precision and ROC gets over this problem by looking into both the True positive rate (TPR) and False positive rate (FPR). Only if both the TPR and FPR are well above the random line in the ROC curve will get a good AUC. Accuracy does not guarantee that.

**4.0 Results**

4.1 Model Performance

After applying all possible feature engineering and Hyper parameter tuning process following is the final table of all the train and test Precision and ROC scores.

Table

Description automatically generated with medium confidence

Highest performing model, which is ‘rf4’ model (RandomForestClassifier(max\_depth=20, n\_estimators=1000, n\_jobs=3, verbose=1) selected to be deployed.

4.2 Model Explanation

Considering the selected sample following is the model explanation on the prediction.



Chart

Description automatically generatedChart

Description automatically generated

Duration, pdays, balance and month features are more influential to identified as a potential subscribe customer for the term deposit.

**5.0 Conclusion**

Overall, the model deployed was highly efficient to classify any given customer with related features on 0.900 ROC, which means the model able to identify 90% of potential subscribing customer from any give data set.

Further, the top 4 key features that helped to predict class variable are duration, age, balance, month.

**Table

Description automatically generated**

**6.0 Discussion**

One of key challenging factors during this project was time constraint as this need to complete within two-week time frame. However, this project able to cover most of the best practices while developing the ideal ML workflow to the selected case study. It is evident by scoring ROC over 0.9. Further, it turns out that the most important features were the numerical features in the dataset. Additional engineering processes with different scaling techniques (Standard Scaler, Polynomial Features) might be able to improve the performance of the model.

Finally, it would be interesting if this model deploys in containerized application with a proper input and output UI for the users.