### Term\_DepositSubscription\_Project

## Project Description

\*\*\*

Banking is a personalized service-oriented industry that provides services according to the customers’ needs. In our project we are focusing on specific banking service – term deposits. Portuguese bank conducted a marketing campaign using 'telemarketing'. Contact of the client was required, to access if the bank term deposit would be subscribed or not subscribed. To make the upcoming marketing campaign more effective the Portuguese bank wants to identify the customers who are more likely to make term deposits and focus marketing efforts on those customers. This will not only allow the bank to secure deposits more effectively but also increase customer satisfaction by reducing undesirable advertisements for non-potential customers. The objective of our application is to identify the potential customers that subscribe for term deposits. Our application analyses customer features, such as demographics and transaction history, the bank will be able to predict customer saving behaviors and predict customers likely to make term deposits.

## Use case Description

The main task of the marketing team is to create an efficient, customer oriented, targeted marketing campaign dedicated to delivering the right message to the right people at the right time. Portuguese bank marketing team aims to conduct telemarketing campaign for the subscription of Term Deposits. The customer Optimization of this marketing campaign involves the application of machine learning model that produces list of prospective subscribers of term deposit. Machine learning algorithm used to predict whether a customer will subscribe to term deposit or not. Based on this predicted customer list the marketing team designs next telemarketing campaign.

## Useful features from your data & the descriptions of the features

\*\*\*

Description of Features:

1. age : customer age (numeric)
2. job : type of job (type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur","student",
   * 1. "blue-collar","self-employed","retired","technician","services")
3. marital : marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)
4. education : education level (categorical: "unknown","secondary","primary","tertiary")
5. default : has credit in default? (binary: "yes","no")
6. balance : balance level (numeric)
7. housing : has housing loan?(binary: "yes","no")
8. loan : has personal loan? (binary: "yes","no")
9. contact : contact communication type (categorical: "unknown","telephone","cellular")
10. day : last contact day of the week (numeric)
11. month : last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
12. duration : last contact duration, in seconds (numeric)
13. campaign : number of contacts performed during this campaign and for this client (numeric)
14. pdays : number of days that passed by after the client was last contacted from a previous campaign(numeric)
15. previous : number of contacts performed before this campaign and for this client (numeric)
16. poutcome : outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")
17. y : has the client subscribed a term deposit? (binary: "yes","no")

## Data Source

\*\*\*

The dataset consists of direct marketing campaigns data of a Portuguese banking institution. The dataset was available ics website and consists of 45211 records. Data used for the project can be retrieved from following link.

https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

##Model Description

Purpose: Predict the bank customers who might subscribe for term deposits.

The problem we are trying to solve is predicting likeliness of term deposit subscription. Hence, we are dealing with labeled data. This would imply that we are dealing with Supervised Learning. In addition, we are interested in predicting a labeled variable. So, we need predictive machine learning algorithm. Our labeled target variable is categorical (subscription: Yes or No). It requires the target variable to be continuous, numeric variable and the target variable has only two categories. As binary classifier would be suitable for this problem Logistic-Regression and Random Forests algorithms are used for model building.

The dataset does not contain any null or missing values. But there is a huge percentage difference between the potential clients who accepted for term deposit (11.69%) and declined for term deposit subscription (88.30%). Up sampling method was applied to on the data. The resulting balanced dataset contains 76847 records. As Logistic Regression and Random Forest algorithms requires that variables (feature and target) all need to be numeric, all the categorical variable of the dataset were converted into the numerical variables, by applying string indexing and then Hot encoder applied for converting the indexed categories into one-hot encoded variables. Vector assembler was used as feature transformer to merge the numerical columns in the dataset into a vector column of features. The numerical features of the data have been scaled by using Minmax Scaler. This results the features to scale in the range 0 and 1. Pipeline was used to chain multiple Transformers and Estimators together to specify our machine learning workflow. To train and test the model’s dataset was split 70 – 30% (record count: 53909 and 22938)

Logistic regression model and random forest model were built on training data. Test data was used for evaluation of the models. Prediction of subscription acceptance was made by both models using test data.

##Model Evaluation

Evaluation metrics were calculated for both models and AOC (Area under ROC) value was generated from BinaryClassificationEvaluator.

accuracy = (TrueNegative + TruePositive) / (TrueNegative + TruePositive + FalseNegative + FalsePositive)

precision = TruePositive / (TruePositive + FalsePositive)

recall = TruePositive / (TruePositive + FalseNegative)

F 1 score= 2 \* (precision\*recall) / (precision + recall)

Logistic regression model metrics:

Precision: 0.687= 0.69

Recall: 0.569

Accuracy: 0.672 =0.68

F1 score: 0.623

Area Under ROC : 0.7256643017596769

Random Forest Model Metrics

Precision: 0.774

Recall: 0.569

Accuracy: 0.716

F1 score: 0.656

Area Under ROC : 0.7898717840938961

##Conclusion

According to the metrics, performance of Random Forest model is better than that of Logistic Regression model. The best model for this use case is Random Forest model. The Random Forest model takes the client data from bank and generates list with predicted subscription column. Marketing team designs next campaign based on this prediction from our model.