Foundational Business Analytics
Coursework 2022-2023

ID: 20439708

A Summarization

The dataset has 16 columns and 4000 rows. There are zero missing values and zero null values in the dataset. There are no duplicate rows.

A.1 Statistical Analysis

- **Numerical** - There are 7 numerical input features and their statistical summary is shown in Table 1. It is observed that there is a big difference in the mean and median values of "balance", "duration" and "pdays", which shows existence of outliers in their values.
- **Categorical** - There are 9 categorical values in the dataset including output variable and their mode values are are shown in table 2. "y" is the target variable.

Table 1: Statistical Summary for Numerical Data

	Mean	**Standard Deviation**	**Median**	**Min.**	**Max.**	**25%**	**75%**
age	40.78	10.9	39	18	90	32	48
balance	1281.21	2523.77	424	-1664	36252	82	1400.5
day	15.86	8.35	16	1	31	8	22
duration	289.08	292.09	199	4	3183	107	360.25
campaign	2.68	2.94	2	1	51	1	3
pdays	42.86	102.87	-1	-1	784	-1	-1
previous	0.61	1.88	0	0	24	0	0

Table 2: Mode Values for Categorical Data

Variable	**Mode Value**	**Frequency**	**Variable**	**Mode Value**	**Frequency**
job	management	878	**loan**	no	3385
marital	married	2302	**contact**	cellular	2689
education	secondary	1983	**poutcome**	unknown	3217
default	no	3930	**y**	no	3151
housing	yes	2094			

A.2 Relation between input feature and output variable

The following are the key analytical points observed in the dataset:

- **Relation between "duration" and "y"** - The more is the duration of call, the higher are the chances for the customer to buy the product. Duration time above 800 seconds suggests that customer mostly buys the product. Refer Figure 1.
- **Relation between "campaign" and "y"** - If a customer is contacted more than 10 times during the campaign in most cases the customer did not buy the product. Refer Figure 2.
- **Relation between "job" and "y"** - Students are more likely to buy the product because it has much more proportion of "yes" than "no" compared to other categories (Figure 3).
- **Relation between "poutcome" and "y"** - If outcome of previous campaign is "success", then there are high chances that the same customer will buy the product as it can be seen from Figure 4.

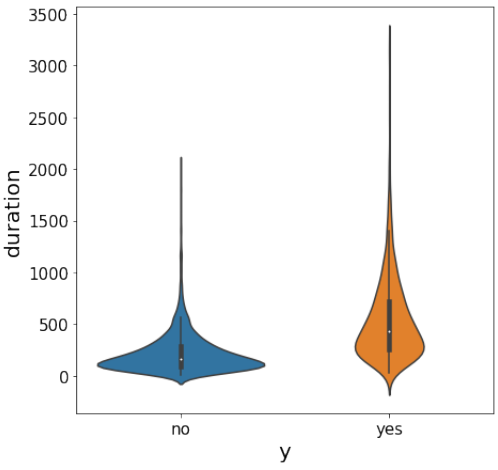

Figure 1: "duration" vs "y"

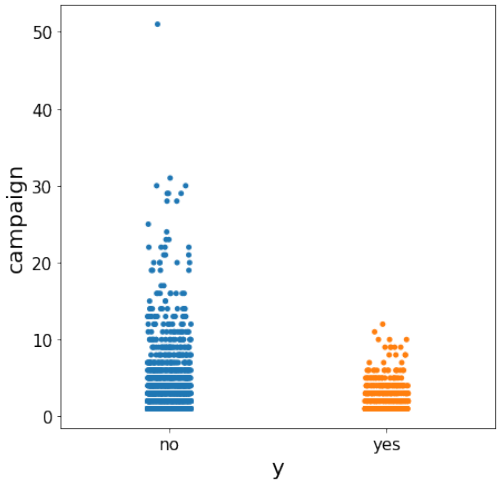

Figure 2: "campaign" vs "y"

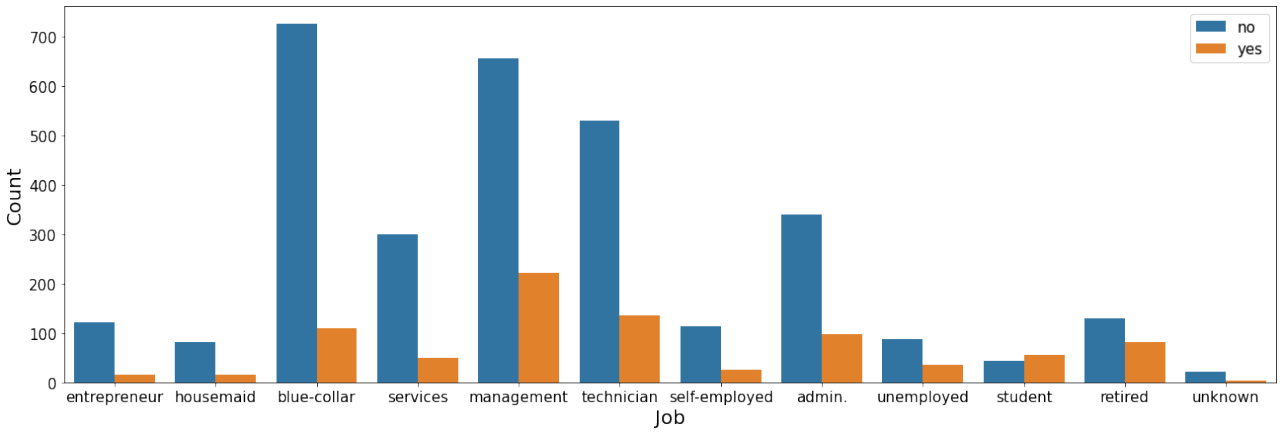

Figure 3: "job" distribution across output categories

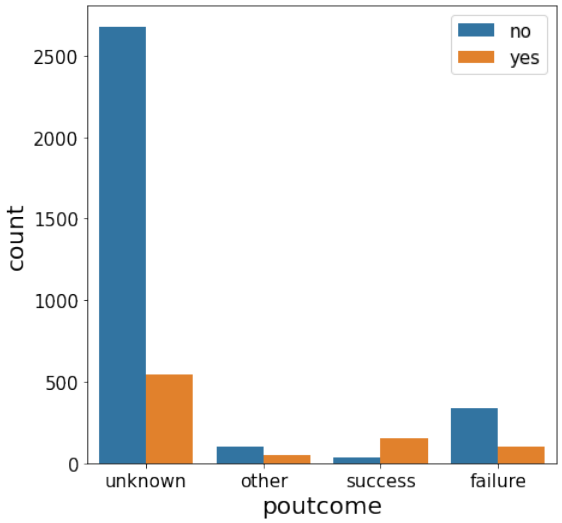

Figure 4: "poutcome" distribution across output categories

A.3 Relation between input features

- **Relation between "pdays" and "poutcome"** : Figure 5 shows that if pdays is less than 100, than there are negligible chances of success. It depicts that taking follow-up calls too soon or too late i.e before 100 days or after 250 days, usually does not lead to success in selling the product.

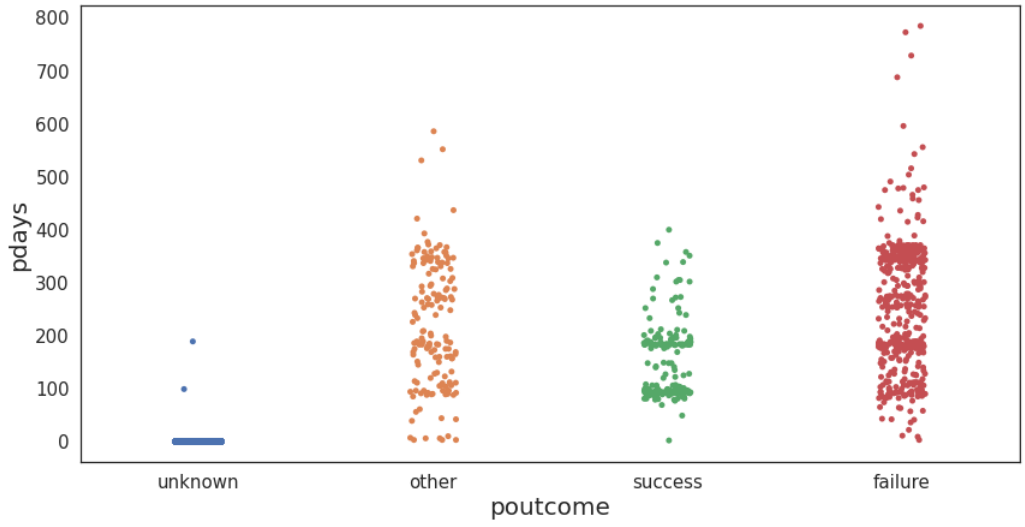

Figure 5: "outcome" vs "pdays"

- **Relation between numerical variables**: Figure 6 shows that variables "pdays" and "previous" are linearly correlated with moderate correlation.

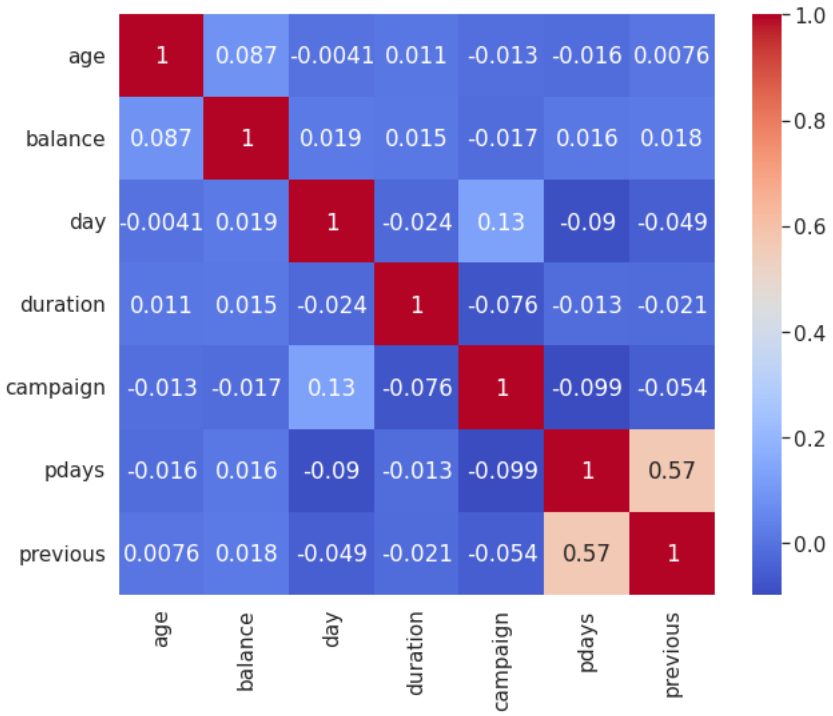

Figure 6: Pearson's Correlation Coefficient Matrix

B Exploration

To apply a decision tree in order to identify the important features in the data, we must first clean the data and prepare it, so that decision tree algorithm can process it properly.

B.1 Data Pre-processing

B.1.1 Numerical Data

The extreme values which are also called outliers must be removed from the dataset because they can affect the accuracy of predictions for some ML models.

- **Removing Outliers using Histogram Charts**: The features like "campaign", "pdays" and "previous", which do not follow a normal distribution and have highly undistributed values in their distribution must be analyzed visually for outliers using histograms. The extreme values (outliers) observed from the chart are removed.
- **Removing Outliers using Z-score**: The features like "age", "balance", "day" and "duration" which have a distribution similar to normal distribution can be analyzed using z-score. The data with extreme z-score values shows that they are outliers. The z-score is calculated using StandardScaler() module from sklearn package.
- **Standardization**: The numerical values are standardized after removing the outliers. StandardScaler() from sklearn package is used for that. Standardization is done to bring all the numerical values on the same scale with mean value equal to zero and standard deviation value equal to 1.

B.1.2 Categorical Data

The categorical data must be converted to numerical data, so that machine learning models can understand and use the data. The ordinal, nominal and binary categorical data are as follows :

- **Ordinal** - To convert ordinal categorical data to numerical data, we use the label encoding technique. We have only 1 ordinal categorical variable i.e **"education"** - having categories: primary, secondary, and tertiary which is converted to 0, 1 and 2 respectively
- **Nominal** - To convert nominal categorical data to numerical data, we use the one-hot encoding technique. We have 4 nominal categorical variables which are: **"job", "marital", "contact" and "poutcome".** New columns are created for each unique value present in the original category columns.
- **Binary** - We can use either of the mentioned techniques for binary categorical data, since both give the same result. We have 4 binary categorical variables which are: **"default", "housing", "loan" and "y"**. Their values are converted to 0 and 1.

B.2 Feature Importance and Selection

- **Decision Tree Visualisation**: For conceptual understanding, Figure 7 shows how decision tree algorithm makes decisions. It can be observed that with tree depth of 2, two important features can be identified here.

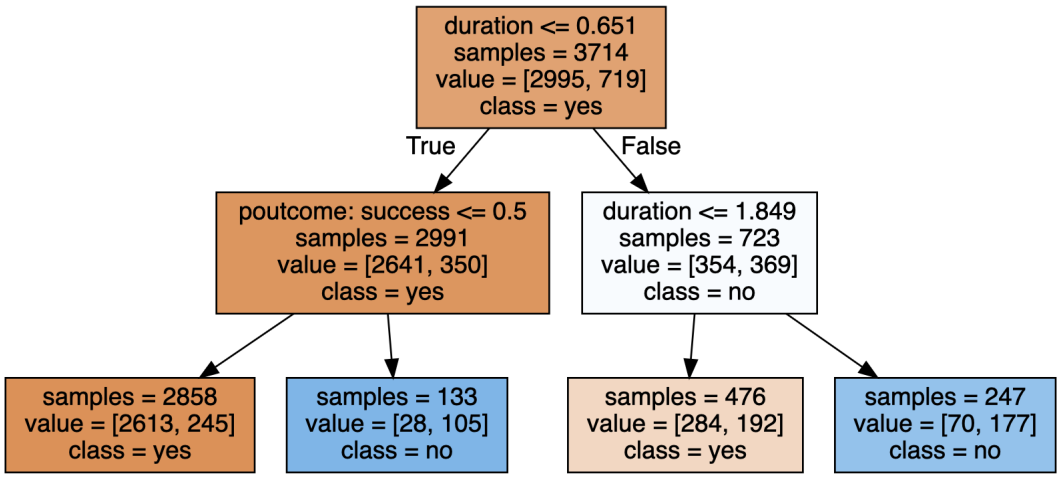

Figure 7: Visualization of Decision Tree with tree depth = 2

- **Important Variables**: The importance of features obtained by applying Decision Tree Classifier is as shown in Figure 8. To get the importance of features, the tree nodes are expanded until all leaves are pure. The feature importance tells us how important each feature is in predicting the output feature. It is observed from Figure 8 that few variables appear to be important in predicting the output variable which are **"duration", "poutcome:success", "age", "day" & "balance"** , hence not all input variables are useful. Therefore, around 70% of input features having high importance are selected for training machine learning model. However, it is important to note that we cannot use "duration" feature in training the predictive machine learning model, since its value will not be known for the test data because duration will not be known for future calls.

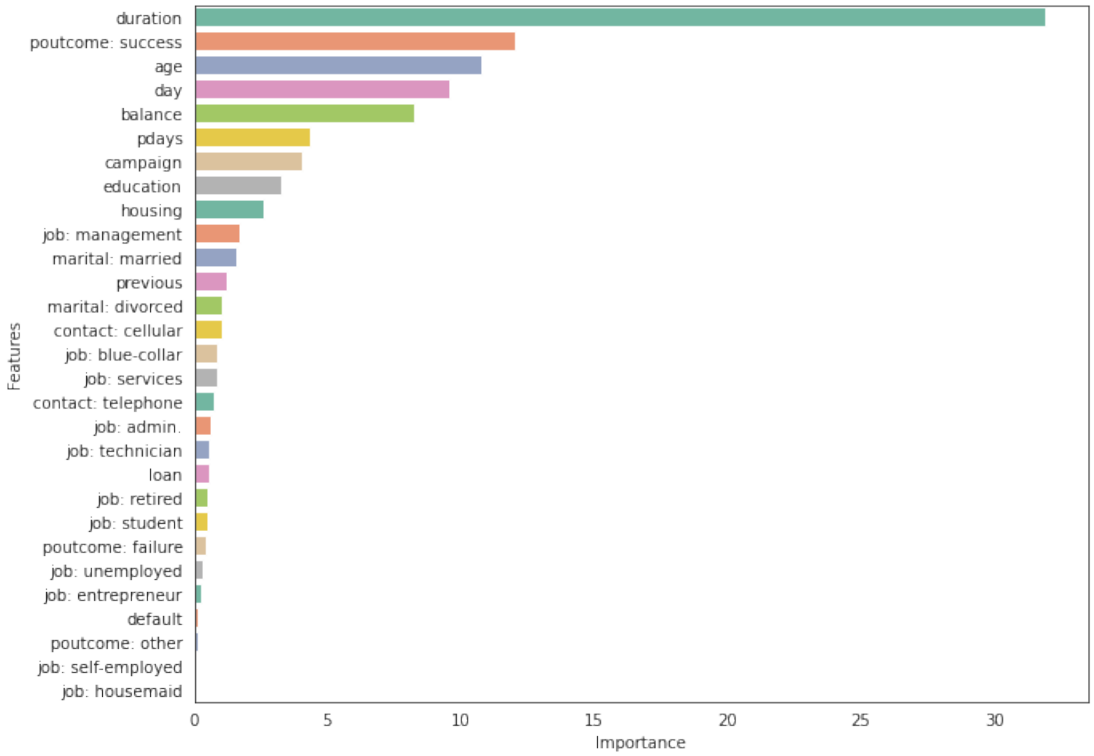

Figure 8: Importance of Features (in percentage) obtained by applying Decision Tree Classifier

- **Analysis**: As discussed in the section A.2, the variables "duration", "outcome:success" and "campaign" had a clear relationship or influence on the output variable "y". Additionally, it is confirmed by the decision tree that these variables indeed have a higher influence in predicting the output variable.

C Model Evaluation

C.1 Model Selection

The various classification models and the reason for selecting them are as follows:

1. **Logistic Regression**- The input features are independent of each other and the output feature is binary. In such scenarios, logistic regression is very effective in predicting. Moreover, the non-meaningful features has been removed which will help to increase the efficiency of the logistic regression model. The parameter selected is as follows:

    - Regularization parameter (C): To avoid overfitting of the model, value of C must not be high. Value of C selected for the model is 1.

2. **Random Forest** - The training data is non-linear as evident from correlation coefficient matrix. Random Forest can handle both non-linear data and outliers if any. It is also good at handling unbalanced dataset as in our case. There are less chances of overfitting the model as is the case with decision tree model. The parameters selected are as follows:

    - n_estimator: It is the number of trees selected for the model and is taken as 100.

- Max. depth: Value of max. depth selected for the model is 3.
- Min. samples leaf: Value of min. samples leaf selected for the model is 3.

The first parameter helps to yield a more robust ensemble and the last two parameters help in pruning trees, thus avoiding over-fitting of data.

3. **K-nearest neighbours** - The number of features has been reduced in our training dataset and KNN performs better with low dimensionality. It is also a non-parametric model with few hyperparameters to tune. Moreover, it supports non-linear classification. The parameters selected are as follows:

   - No. of neighbours (K value): It influences the prediction accuracy of the model. Value of K selected for the model is 4.
   - Distance function: It will affect the distance measuring method of the algorithm. Euclidean distance is selected for the model, as it works well in most scenarios.

C.2 Evaluation Strategy for Comparing Models

Defining Positive class: Our target class for this classification problem is "yes" because we need to market to potential clients or those who are likely to buy the product. Therefore, while evaluating various performance measures for the ML models, "yes" is identified as positive and "no" is identified as negative.

Selection of Performance Metrics:The dataset is imbalanced because one class (i.e "no") is more frequent than the other class (i.e "yes"). In such cases, accuracy alone is not adequate measure for quantifying predictive performance. Additionally, the company wishes to prevent making pointless calls to uninterested clients and reduce wasteful expenses. This means we need to minimize False Positives in our ML model. **The performance measure that should be compared to minimize False Positives is "Precision".** The performance measures are calculated using confusion matrix given below. Here TN is True Negative, FP is False Positive, FN is False Negative and TP is True Positive.

Confusion Matrix	Point Predictor	Logistic Regression	Random Forest	KNN
$\begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix}$	$\begin{bmatrix} 2995 & 0 \\ 719 & 0 \end{bmatrix}$	$\begin{bmatrix} 2956 & 39 \\ 573 & 146 \end{bmatrix}$	$\begin{bmatrix} 2968 & 27 \\ 590 & 129 \end{bmatrix}$	$\begin{bmatrix} 2900 & 95 \\ 594 & 125 \end{bmatrix}$

Table 3: ML Model Evaluation using different Performance Measures

	Accuracy	**Precision**	**Sensitivity**	**Specificity**
Point Predictor	80.6%	Undefined	0%	100%
Logistic Regression	83.5%	78.9%	20.3%	98.6%
Random Forest	83.4%	82.7%	17.9%	99.1%
K-Nearest Neighbours	81.4%	56.8%	17.4%	96.8%

D Final Assessment

Considering the business case for measuring success, we must consider "precision" of models along with its "accuracy". The winning classifier model is the **"Random Forest"** model because of the following reasons:

- It has the highest precision value among all models
- Its accuracy is also high, exceeding the baseline accuracy provided by the point predictor model.

E Model Implementation

The submission contains 4 files which are:

- **Part 1.ipynb**: It contains code related to data summarization and contains various charts that show the relationship between variables of the dataset.
- **Part 2.ipynb**: It contains code related to data cleaning, data pre-processing, feature exploration and selection, model evaluation and implementation.
- **Predict.ipynb**: It contains code to test a new dataset which will then be pre-processed and used by the ML model to make predictions. Follow the steps below to process the new test set and to make predictions.
  - **Model_final.pkl**: It is the trained ML model that will be loaded in the "Predict.ipynb" file to make predictions.
  - Note: "Model_final.pkl" must be in the same location as of the "Predict.ipynb" file or else it must be ensured that correct file path is provided for "Model_final.pkl" file in Step 5 of the "Predict.ipynb" file.
  - The user must provide the file name and path of the test data inside the read.csv() function under Step 2 of the "Predict.ipynb" file.
  - The entire file must be run by going to Runtime> Run all in order to obtain the predicted output for the test data.

F Business Case Recommendation

Summary:

The data about previous marketing campaign of a similar financial product is utilized to implement targeted marketing in order to improve product sales and avoid unnecessary business costs. The highlights of the business case are :

- The data was used to identify the major factors that influence the decision about this product purchase. Specific insights about the buying pattern were also explored.
- The prediction is based on previous data i.e on the previous buying trend. If the buying trend changes in future, the prediction accuracy of the model may be reduced if the same data is used.

Future Recommendation:

- The prediction accuracy of the model can be further increased by collecting more data which is accurate, recent and relevant.
- Additionally, potential analysis can be done using the data to do customer segmentation using an unsupervised clustering technique. The segmentation can help to identify customers with similar trends which can then be targeted for marketing.