CMT307 Applied Machine Learning Coursework 1

Question 1:

The below steps are followed to compute the precision, recall, f-measure and accuracy for the given table in the assignment.

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known.

Confusion matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Predicted Class** | | |
| **Actual Class** |  | Class = True | Class = False |
| Class = True | True Positive  (TP) | False Negative  (FN) |
| Class = False | False Positive  (FP) | True Negative  (TN) |

\*\* Classifiers: derived based on confusion matrix

|  |  |  |  |
| --- | --- | --- | --- |
| **Id** | **Prediction** | **Target** | **Classifiers** |
| 1 | True | True | TP |
| 2 | True | True | TP |
| 3 | False | True | FN |
| 4 | True | True | TP |
| 5 | True | True | TP |
| 6 | False | True | FN |
| 7 | True | True | TP |
| 8 | True | True | TP |
| 9 | True | True | TP |
| 10 | False | False | TN |
| 11 | False | False | TN |
| 12 | False | False | TN |
| 13 | True | False | FP |
| 14 | False | False | TN |
| 15 | True | False | FP |
| 16 | False | False | TN |
| 17 | False | False | TN |
| 18 | True | False | FP |
| 19 | True | False | FP |
| 20 | False | False | TN |

|  |  |  |  |
| --- | --- | --- | --- |
| True Positive (TP) | The positive points that were correctly labelled by the classifier | Both Prediction and Target are True | **Number of TP = 7** |
| True Negative (TN) | The Negative points that were correctly labelled by the classifier | Both Prediction and Target are False | **Number of TN = 7** |
| False Positive (FP) | The Negative points that were incorrectly labelled as Positive by the classifier | Prediction is True but the Target is False | **Number of FP = 4** |
| False Negative (FN) | The Positive points that were incorrectly labelled as Negative by the classifier | Prediction is False but the Target is True | **Number of FN = 2** |

Performance Measures:

|  |  |
| --- | --- |
| **Precision**  (Measure of Exactness)  **[Appropriate for minimizing FP]** |  |
| **Recall**  (Measure of Completeness)  **[Appropriate for minimizing FN]** |  |
| **F – Measure**  (Combination of Precision and Recall)  **[Appropriate for balancing Precision and Recall]** |  |
| **Accuracy**  (Metric for performance evaluation) |  |

Conclusion:

We may infer from the above calculation that the accuracy and F1-score of the given model is close to 70 percent.

Report of Question 2:

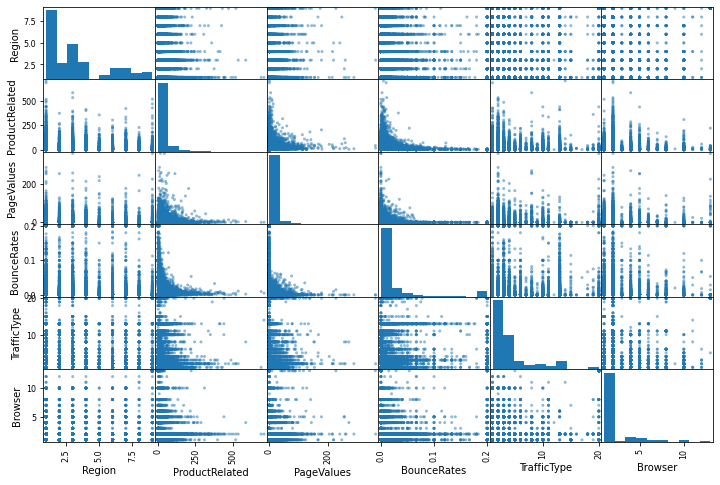
Following the given dataset Coursework\_1\_data.csv of the e-commerce visitors’, the objective is to develop the machine learning models to predict the purchasing intention using Python and to use most suitable performance metrics to assist the selection of the best model on characteristics of this dataset along with the selection of the classifiers.

Data Exploration:

The initial phase in Data Analysis is data exploration, where Data Analysts use data visualization and statistical techniques to explain data set characterizations with manual analysis and automated software solutions for data exploration that enable Data Analysts to gain greater insight into the essence of raw data.

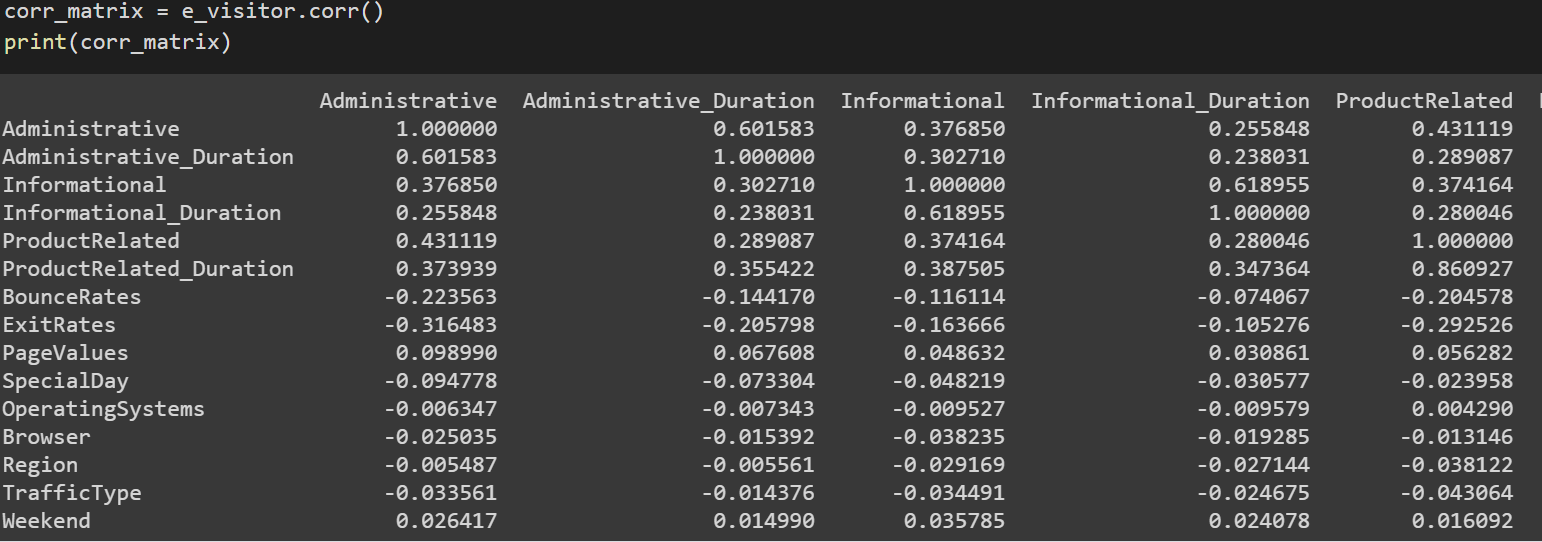
Steps and Inference of Data Exploration:

* We have imported our Coursework\_1\_data.csv file into Google Colab. Using the 'shape' property, first we tested the shape and dimension of the given data.
* To evaluate the measures of central tendency, we applied the 'describe' function. We found a single number representing the entire dataset by means of the specified dataset. The data being clustered around the Mean suggested a smaller magnitude of Standard Deviation and the greater magnitude of Standard Deviation indicated the data were more spread out. We roughly estimated from the quartiles and Inter Quartile Range (IQR) how far we must go on either side of the median before we can use half of the dataset values.
* To print data types, the function 'details' is added. It offered a brief description of the provided data frame.
* The function 'isna' was used to decide whether there is a missing value in our given dataset **(Fig. 1)**. This returned a Boolean object of the same size, indicating if the values were missing. Here we found that there was no missing value in our dataset because each missing value was mapped to zero (0), i.e. False, so we could skip the respective steps to deal with missing values.
* To see the bivariate relationship between all the combination of variables in our given dataset, the scatter matrix from the Pandas library was imported **(Fig. 2)**.A picture containing text

  Description automatically generated 

**Fig. 1 Fig. 2**

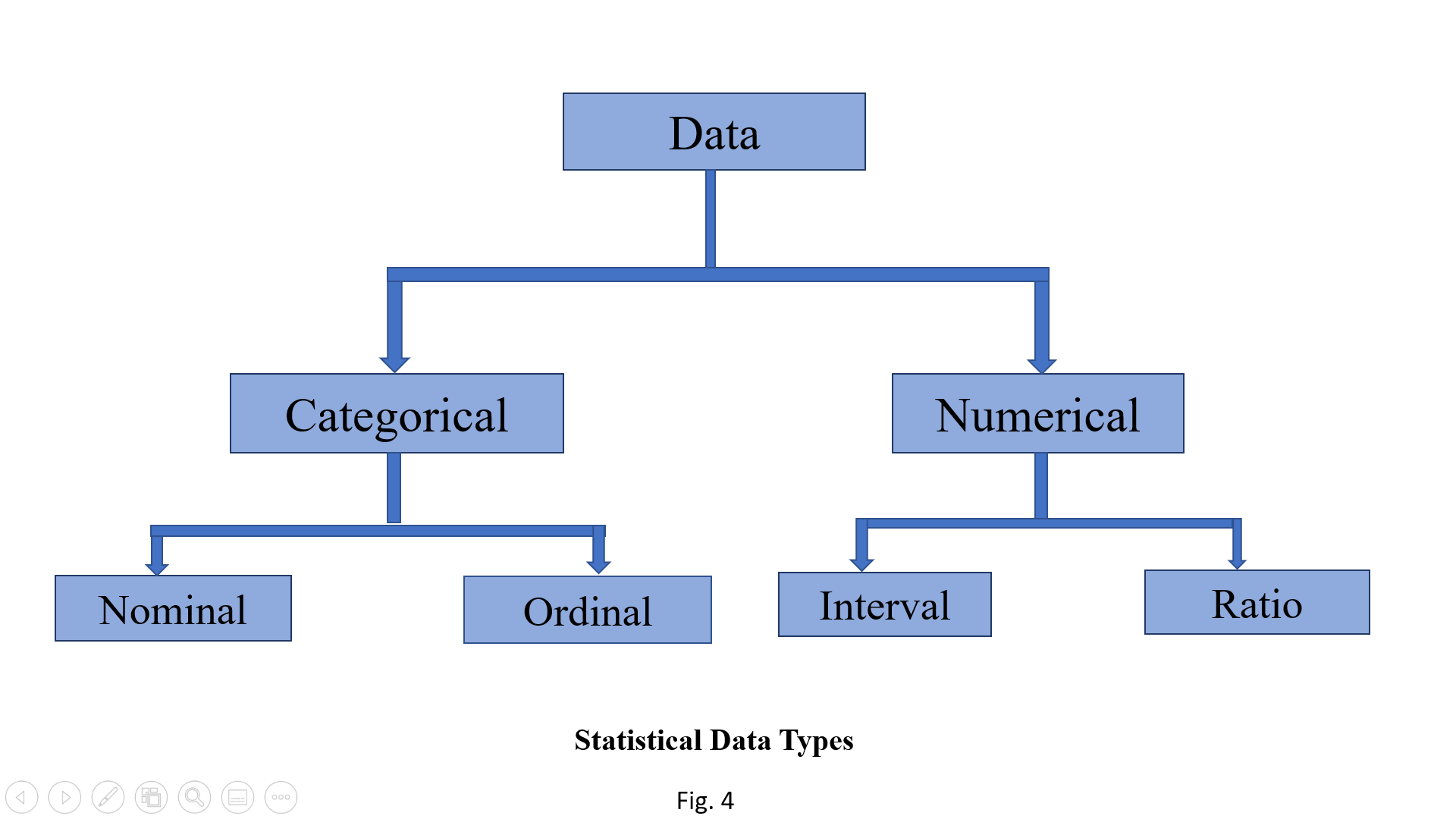
* Finally, to evaluate the regression form, we used the Correlation Matrix **(Fig. 3)**. The intensity of the defined linear relationship is given by:
* Exactly -1: Perfectly Negative.
* -0.70: A Strong Negative.
* -0.50: A Moderately Negative.
* -0.30: A Weak Negative.
* Exactly 0: No Linear Relationship.
* +0.30: A Weak Positive.
* +0.50: A Moderately Positive.
* +0.70: A Strong Positive.
* Exactly 1: Perfectly Positive.



**Fig. 3**

Data Pre-Processing:

Data Pre-processing is the step in which data is transformed or encoded to give a state that algorithms can easily interpret the characteristics of data. Typically, **Fig.4** displays the types of statistical data that we must deal with.



Steps and Inference of Data Pre-processing:

* The Categorical Data was selected from our dataset and we used One-hot encoding in arrays instead of matrix.
* Since ‘Weekend’ and ‘Revenue’ were in Boolean form in our given dataset, we separately used one-hot encoding and removed one column from each of them.
* To calculate the accuracy of our model, we split our data into 'train set' and 'test set' and saved our trained data into 'X-train'.
* Standardization is a scaling method determined by subtracting the mean value and dividing the standard deviation, converting the probability distribution for an input variable to standard Gaussian. But if the input variable has outliers, standardization may be distorted or biased. So, to overcome this issue, we used ‘MinMax Scaling’ on ‘X\_train’ (the trained data), which effectively encoded the outliers. Since our SVM model was taking more time after using ‘Robust Scaling’, so we did not use this.
* After encoding, we split our data and then used Robust Scaling to solve the data leakage problem.

Model Implementation:

The strength of Machine Learning includes the ability to manage high-dimensionality data and to map classes with very complex features. Our emphasis is therefore on the implementation of our model in reasonably mature techniques such as Support Vector Machine (SVM), Random Forest (RF), k-Nearest Neighbours (k-NN), etc.

Steps and Inference of Model Implementation:

* On our trained dataset, which is from the classification family and known as the supervised learning algorithm, we implemented the 'Random Forest Model.' First, this model created several random trees (or forests) and then the algorithm combined the trees during the prediction process using the estimated result and voting procedure. This implementation is worth using as this will increase the accuracy of the forecast for future data.
* After that, we introduced the 'Accuracy Score' function, which is used to measure the accuracy of the subset by the 'number of correctly classified data / total number of data' formula.
* We fitted the 'Random Forest Classifier' with estimator 100, maximum depth 500, random state 42 and jobs-1 into our qualified dataset and checked the accuracy, recall and f-1 score.
* Repeated Stratified K-fold cross validator was imported, which repeated Stratified K-fold with different randomization in each repetition.
* Hyperparameter optimization was implemented as this method seeks a tuple of hyperparameters that yield an optimized model that minimizes a predefined loss function on the trained dataset provided. The objective function takes a hyperparameter tuple and returns the corresponding loss. We rechecked the accuracy score of our model after this.
* We have now implemented Logistic Regression in our model as we saw that we were unable to apply Linear Regression in our dataset from the Data Exploration portion. This model's learning and prediction is based on calculating the likelihood of binary classification. The equation of Logistic Regression is given by: .
* In the multiclass case, the training algorithm uses cross-entropy loss scheme if the ‘multi-class’ is set to ‘multinomial’. Currently the ‘multinomial’ option is supported only by ‘lbfgs’, ‘sag’, ‘saga’ and ‘newton-cg’ solvers. In our model, to deal with elastic net penalty, we used ‘saga’ solver to get fast convergence with approximately same scale.
* Finally, we used Support Vector Machine (SVM) as the model implementation, which is also known as the supervised machine learning model, where the classes are predefined in the dataset. It classified the trained data by the pre-defined classes in the given dataset. One of many groups was allocated to optimize output accuracy.

Performance Evaluation:

The performance evaluation of the model is an important part of our project to deal with classification and regression problems as it seeks to estimate the generalization of a model's accuracy to generate flawless prediction.

Steps and Inference of Performance Evaluation:

* Performance Evaluation is commonly carried out in two categories:

1. **Holdout Evaluation** is the process of evaluating a model on different data than training to provide an unbiased estimation of the success of learning. There are three subsets of this approach, first is the **Training Set** to construct predictive models, second is the **Validation Set** to evaluate the performance of the model developed in the training process, and third is the **Test Set** to evaluate a model's probable future performance.
2. **Cross-Validation** is a technique of partitioning the original observation dataset into a training set, used to train the model and evaluate the analysis.

* A model assessment metric has now been used to measure model efficiency. We used accuracy, precision, and recall from the '**sklearn**' function for classification metrics.
* Since a Confusion Matrix provides a more detailed breakdown of correct and incorrect classifications for each class, we used Iris dataset to classify and compute the confusion matrix for the correct prediction.

Result Analysis and Discussion:

In this project, we used the system of models as the classification of our model for best prediction. We pre-processed our raw data and obtained qualified datasets on which SVM, RF and LR were implemented. We can infer the following after evaluating our results:

* We have no improvement in Accuracy Score after implementing the Random Forest Classifier, but a small increase in Precision and f-1 score.
* We obtained modifications in all dimensions, Accuracy Score, Precision, Recall and f-1 score after applying Logistic Regression to our trained dataset.
* After implementation of Support Vector Machine in our model, we got no change in Accuracy Score, Precision, and f-1 score, but only a slight decrease in Recall.
* After reviewing all the findings, we can conclude that the best fit for our given dataset is the Random Forest Classifier and Logistic Regression.

References:

* <https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74>
* [file:///C:/Users/c1893753/OneDrive%20-%20Cardiff%20University/Desktop/d8c51251e8c6126a1527e545bd78860a10f9.pdf](file:///C:\Users\c1893753\OneDrive%20-%20Cardiff%20University\Desktop\d8c51251e8c6126a1527e545bd78860a10f9.pdf)
* [file:///C:/Users/c1893753/OneDrive%20-%20Cardiff%20University/Desktop/yousefi.arzyabiamalkard12.pdf](file:///C:\Users\c1893753\OneDrive%20-%20Cardiff%20University\Desktop\yousefi.arzyabiamalkard12.pdf)
* <https://learning.oreilly.com/library/view/hands-on-machine-learning/9781492032632/>
* <https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/>
* <https://librarysearch.cardiff.ac.uk/permalink/f/1tfrs8a/44CAR_ALMA2167905940002420>