Applied Machine Learning

CMT307

Coursework-1

Student No: C21034767

Student Name: Safi Ammar Mohammed

Lecturer: Yuhua Li

Question – 2 (Report):

Dataset Information and Features:

- The dataset consists of 10 numerical and 8 categorical features.
- The 'Revenue' attribute is used as class label.
- The dataset is clean, there are no missing values but it is unbalanced.
- There is a risk of bias, so the analysis needs to consider the unbalanced dataset.

Features Correlation:

- According to the correlation matrix, the target is only correlated to a small number of variables in the dataset.
    - PagesValue (0,5)
    - ExitRates (-0,21)
    - BounceRates (-0,16)
    - ProductRelated_Duration (0,15)

Principal Components Analysis:

- There are 12 clusters in the 2D PCA that may correspond to the 12 months.
- Because the percentage of explained variance decays slowly, it is not possible to represent the dataset properly in 2 or 3 dimensions.

Dataset Balance:

- Comparing the target (Purchased) and VisitorTypes, we can see that the dataset is quite unbalanced. This means that we will need to downsample/undersample he data that leads to 'No Purchase'.

Classification Algorithm Exploration:

- To select the best classification algorithm, we measure the accuracy and F1-score of each algorithm.
- To have a better idea of the performance, we also plot the confusion matric of each prediction.

Hyperparameters Optimization:

- After a quick overview of each algorithm's performance on the dataset, we select the best three based on the Accuracy and F1-scores.
- We then try to improve each algorithm by optimizing the hyperparameters with Grid Search.

Final Model:

- The final model is a combination of the three best models which are:
    - Random Forest
    - Decision Tree
    - Stochastic Gradient Descent.
- The Accuracy, F1-score, Precision and Recall obtained by the ensemble model are 89.99%, 0.672, 0.712 and 0.637 respectively.
- After creating an ensemble model using the three best machine learning models, we cross validate the accuracy and get a final accuracy of 89.67%.
- This model can now be used to predict if new sessions lead to purchases/revenue or not.