Abstract :

This project will be based on classification problem to predict customer online shopping intention. we aim to help companies to be successful in a highly competitive eCommerce environment by predict whether the customer, visiting web pages of an online shop, will end up with a purchase or not. The dataset contains 12330 observation and 18 features divided into 80% training and 20% testing.

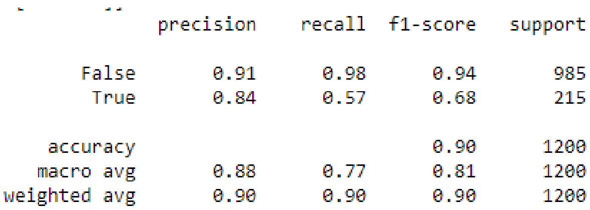
I have used three classification models which are logistic regression, Random Forest and decision tree .The results show The random forest outperform the other model in accuracy , precision, recall and F1 Score with 90%91%, 98%,94% respectively for not purchase . While for purchasing precision is is 84%, recall is 57% ,F1 score is 68%.

Design:

The dataset consists of 10 numerical and 8 categorical attributes. I used five attributes (Month , operating system, Browser , Visitor type , Weekend, Revenue) in Eda part and in model training. The 'Revenue' attribute is used as class label. The dataset is clean, there are no missing values, but the dataset is unbalanced. There is a risk of bias, so the analysis must take the unbalanced dataset into consideration.

Algorithms:

I have used Random Forest, Logistic Regression and decision, the results showing in the figure below:



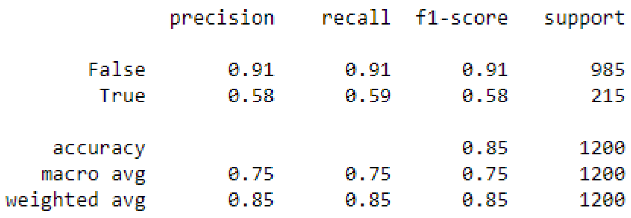
```
              precision    recall  f1-score   support

       False       0.91      0.98      0.94       985
        True       0.84      0.57      0.68       215

    accuracy                           0.90      1200
   macro avg       0.88      0.77      0.81      1200
weighted avg       0.90      0.90      0.90      1200
```

Figure 1 : Random forest performance

precision, recall and F1 Score with 91%, 98%,94% respectively for not purchase . While for purchasing the precision is 84%, recall is 57% ,F1 score is 68% .



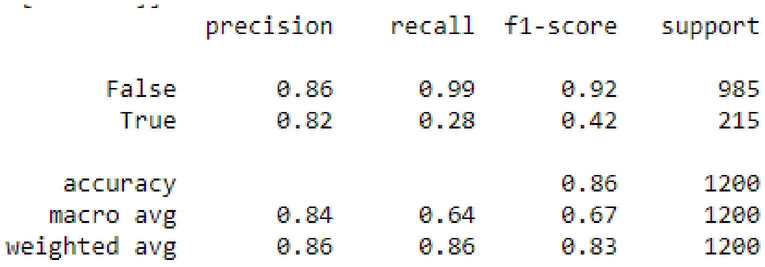
```
              precision    recall  f1-score   support

       False       0.91      0.91      0.91       985
        True       0.58      0.59      0.58       215

    accuracy                           0.85      1200
   macro avg       0.75      0.75      0.75      1200
weighted avg       0.85      0.85      0.85      1200
```

Figure 2 : Decision Tree performance

precision, recall and F1 Score with 91%,91%, 91%respectively for not purchase . While for purchasing precision is 58%, recall is 59% ,F1 score is 58% .



```
              precision    recall  f1-score   support

       False       0.86      0.99      0.92       985
        True       0.82      0.28      0.42       215

    accuracy                           0.86      1200
   macro avg       0.84      0.64      0.67      1200
weighted avg       0.86      0.86      0.83      1200
```

Figure 3 : Logistic Regression performance

Tool :

- Numpy and Pandas for data manipulation
- Scikit-learn for modeling, confusion matrix and feature extraction
- Matplotlib and Seaborn for plotting