**ASSIGNMENT REPORT**

1. **Problem statement:**

Dataset consists of two files one is customer behavioral dataset and other is final conversion dataset. So based on this dataset, various features and model has to be generated which will basically tell whether user will purchase the product or not, based on their browsing behavior. This result will be in the form of probability scores.

1. **Assumption:**

For developing model, a chunk of data of 5000 randomly selected userid from customer behavior dataset is used due to system constrain and high computation time.

1. **Approach:**

First raw data is loaded, and then various features are generated. Once features are available exploratory data analysis is performed to check whether features are useful or not. Next is model development, here various classification models are tried and among them best classifier is selected based on ROC-AUC and accuracy score. Once best classifier is selected it is tuned for hyperparameter for better results. When done with model development, this model is tested on unseen dataset i.e., not used while training and later probability scores for user conversion (user will buy or not) is computed. Steps involved in model development is explained briefly below.

* 1. **Feature engineering:**
* First raw data of behavior dataset is converted into 17 new features for each user. This feature is user count on each website section visited.
* Next a new ***'conversion'*** column is added in new data frame. This column has binary values, 1 is product purchased and 0 is product not purchased.
  1. **Exploratory data analysis:**

Before jumping directly into development of model, first we need to visualize the multi-dimensional data. To visualize the multi-dimensional features, here two methods are used

* T-distributed stochastic neighbor embedding (t\_SNE)
* Principal Component Analysis (PCA)

From the two plot it can be seen there is a difference between purchased user and not purchased user. So now we are ready to develop model.

* 1. **Model development:**

To solve this problem various classification model is developed on training dataset and tested on test dataset. Based on accuracy score and ROC-AUC sore best classifier is selected and tuned. Finally, from selected classifier probability scores are computed for conversion of user.

1. **Interpretation of Results:**

Model results are in the form of probability scores, based on the scores we can decide which users are interested in purchasing the product and target those users. This score can be used to segment into different classes based on there website section or product visited. If we segment into different classes, then this will also tell which user has spent time on browsing but not purchased any product, so seller should target those users by providing some offers. This score can be effectively used to build strategies for selling the product and results improving business.

1. **Conclusion:**

* The model developed has accuracy of 93% on test dataset for determining whether user will buy the product or not. This accuracy might change if tested on whole dataset.
* Accuracy can be improved by adding other features like total time spent on browsing etc.
* This model can be further used to target some specific users for selling product based on user product searched.

**Scope of improvement:**

* For generating features, timestamp of each dataset is not used. If we estimate total time spent by each user for browsing will tell whether he is really interested in buying product or not.
* Product purchased and total amount features is not used, this feature might be useful with, what kind of product user was searching in website section.
* Feature selection: Various univariate and wrapper-based feature selection technique can be applied. This will result in only important features to be selected.
* Other models like Dense Neural network or Deep learning method are not used. This model might give better results.