NAME: K SUNIL JOSHI
SOCIAL COMPUTING
HOMEWORK -3 REPORT

"INFLUENCERS IN SOCIAL MEDIA USING TWITTER DATA"

Problem Statement: Our aim is to predict the influencer between two twitter users.

Dataset:

https://www.kaggle.com/c/predict-who-is-more-influential-in-a-social-network/data

The dataset contains Twitter user information like follower count, mentions, retweets, posts etc., which has train data of 5500 rows and 23 columns, test data with 22 columns and also a sample predictions dataset which has only choice labels to validate the outcomes.

Ground Truth Establishment: The Choice labels are marked 0/1 based on Peerindex application who provided the data for the challenge , there is no further information from the company saying how they did it.

1) Feature Engineering :

- Columns like follower_count, mentions_recieved,retweets_recieved are mainly used as train data as they are highly correlated with choice labels.
- The data is converted into log scale as input columns have different scales, so by converting them to log scale and then used for marking choice labels.

2) Classifier:

- The data set has already separated as train , test split with test data doesn't contain choice lables as we have to predict them.
- This is a binary classification task , so I used Logistic Regression, SVM, Random Forest Classifier to train the model on x_train and y_train.

- The Predcitions (y_pred) are made on test data (x_test) and used to calculate model performance metrics.

3)Model Evaluation:

- Metrics like accuracy_score , classification_report and confusion_matrix are used to evaluate the model performance.
- By training the model and tested on pedcitions using the above models we finally came up with a classification task with highest accuracy and prediction and performance metrics like Accuracy , Precision and Recall.

4)Implementation:

- Data is converted to log scale and converted into arrays to fit in the model.
- The metrics of the models are calculated using accuracy_score, confusion_matrix, classification_report , roc_score using python inbuilt sklern.metrics module, in all of these the input arguments passed are y_test and y_pred as the prediction is done on test data and compared with sample predcitions data set already labelled.

5) Results:

- The model accuracy and classification report are evaluated for each model and among all Logistic Regression gives the highest accuracy , with 90.1% compared to other models.
- By using python seaborn module I plotted heatmap for confusion_matrix from which it is evident to know how many of them are correct (true) predictions and which are incorrect (False) predictions.

```
Logistic Report              precision    recall  f1-score   support

           0       0.93      0.90      0.92      2954
           1       0.91      0.94      0.92      2998

   micro avg       0.92      0.92      0.92      5952
   macro avg       0.92      0.92      0.92      5952
weighted avg       0.92      0.92      0.92      5952

Logistic Regression confucion matrix [[2662  292]
 [ 191 2807]]
```