IDS ASSIGNMENT - 3

GROUP - 157 - Pothula Maruthi Chowdary , Ajit Mohan Pattanayak , Jayadhar Alla

Objective :

Predicting Top 5 Movies for a User based on his Movie Watching History

Dataset given from here > Data : https://www.kaggle.com/netflix-inc/netflix-prize-data

As per the question we have calculated Performance with the 2 Classifiers :

(Logistic Regression and Decision Tree) - file **: IDS_Assignment3_G157.pynb**

Note : I kept all the required input files and the pynb file in the same directory so you will see direct file path rather than absolute path in our Program.

combined_data_1.txt and **movie_titles.csv** are considered as inputs .

Input files has different type of data , Ordinal, Categorical etc..

Feature Engineering Techniques used : **Handling Outliers, Binning , Scaling**

> As part of Preprocessing and Data Preparation to handle the Empty Rows and missing values properly .

> Joint point for number of Ratings and histogram considering the Ratings feature are plotted and shown .

Based on the Requirement few of the features were removed and also few were modified .

> Rating is modified as Rating Category and used as the target Variable .

> Proper segregation is done for the Test and train splits .

LOGISTIC REGRESSSION :

Logistic Regression ensures that the values output as predictions can be interpreted as probabilities of class membership.

LOGISTIC REGRESSION is used and the Predict probabilities are calculated for the provided Data model .

Advantages of logistic regression:

- Highly interpretable (if you remember how).
- Model training and prediction are fast.
- No tuning is required (excluding regularisation).
- Features don't need scaling.
- Can perform well with a small number of observations.
- Outputs well-calibrated predicted probabilities.

Disadvantages of Logistic Regression :

Disadvantages of logistic regression:

- Presumes a linear relationship between the features and the log odds of the response.
- Performance is (generally) not competitive with the best supervised learning methods.
- Can't automatically learn feature interactions.

DECISION TREE :

The Decision Tree Algorithm is a simple classic supervised learning model that works surprisingly It is majorly used for Classification types.

Decision trees are a tree algorithm that split the data based on certain decisions. well. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

This classifier is used for the same data , tree model fit is performed on it and predicted for the Test set.

Note : To display the Decision tree in the pynb file, We have installed graphviz, condo-forge and pydotplus by following commands from terminal :

conda install graphviz , conda install -c condo-forge pydotplus

After Predicting and calculating the tree_model.score , by using the modules mentioned in the above note , we have shown the DECISION TREE .

After that Classification report is also generated to know the Precision Recall F1-score and Support and shown as output .

Second Approach :

Second Approach is the one ,which is for recommending the Movies using the Pearsons' R Correlation .

Data handling and Feature Selection Techniques are used to prepare the well structured data which is used in the model.

We have followed all the required flow to prepare the data and performed required steps to get the recommendation top 5 movies base on the input .

All the methods are executed and outputs are shown.

Best Regards

GROUP 157