Data Science: A Programming Approach
Mahyar S Vaghefi
University of Texas Arlington

Group Project - Fall 2020

You need to work as a team for this project. Your job is to develop a predictive model that can predict whether or not a movie is a *Comedy*. In order to do so you need to use the textual features of the movie stories and creat your predictive models. There are totally three different files in this project.

1. **movie_story_student_file.csv**: This file contains the movie stories that should be used by students for model development.
2. **movie_story_evaluation_file.csv**: This file should NOT be used for model development purposes. Students should only use this file after developing their predictive models and selecting their best final model. They then need to use their best predictive model and predict whether or not the movies in **movie story evaluation file.csv** are Comedy.
3. **movies.csv**: This file contain the movie genres.

Additional Note: This note provides additional guideline for the project.

Step 1: Explore all files to become familiar with the dataset
Step 2: There are 20,000 movies in `movie_story_student_file.csv` file. Use the `movies.csv` file to determine whether a movie is Comedy or not. The final output of this step should be development of a dataframe file that contains three columns: 1) movieId, 2) story, 3) ComedyGenre. The value for the thirrd column should be in a binary format. If the movie is comedy the value should be 1, otherwise 0. Name the new file as `main_dataset` .
Step 3: If you want to just have one train and one test set, then split the `main_dataset` to train and test sets. Otherwise you can use cross validation methods.
Step 4: Use your supervised learning and text analysis knowledge to develope different predictive models (i.e. logestic regression, random forest, ...). Test the performance of your models in terms of accuracy, precision, recall, and F1 scores.
Step 5: Choose one of your best models (only one). Then use the full data available in `main dataset` to train that model.
Step 6: Prepare data in `movie_story_evaluation_file.csv` for prediction. You need to perform the exact same steps that you have done in **Step 2** to prepare this new dataset. Name the new dataset as `evaluation_dataset` .
Step 7: Use your selected model in Step 5 to predict whether or not movies in `evaluation_dataset` are Comedy movies.
Step 8: Report the final accuracy, precision, recall and F1 score.

Keep in mind that the above steps are only provide guidelines. There are more detail works in this project that should be captured by students.

Extra Credit The group with the best prediction model is going to get extra points.

Output

- Make sure to put descriptive comments on your code
- Use the markdown cell format in Jupiter to add your own interpretation to the result in each section.
- Make sure to keep the output of your runs when you want to save the final version of the file.
- The final submited file should be very well structured and should have a consistent flow of analysis.
- You may want to use an additional word document to report the result of your analysis

Due Date: Nov 30 2020 at 11:59 PM

Comprehensiveness	**30%**
Correctness	**30%**
Complete Report	**20%**
Clear Code	**20%**
Total	**100%**