Film Analytics

Data Dawgs

(Siddhant Sutar, Dalton Childers, Neil Matin, Abhimanyu Tomar)

PROGRESS REPORT 2

Achievements

- Previously, fetched movie data for 1.1 million movies from OMDb and wrote it into a csv file (approx 500 MB size).
- Implemented the multiple linear regression model using the movie data for building the movie ratings (as well as the box office gross) predictor.
- In the multiple linear regression analysis, the independent variables (x) are Rating (for example, PG-13, R, etc.), Genre(s), Director(s), Actor(s), Writer(s). Since all variables are categorical, not continuous, need to convert the categorical variables into a numerical form that "makes sense" to regression analysis.
- Using Pandas get_dummies() function, create feature vectors for the independent variables: rating, genre(s), director(s), writer(s), actors. For example, a typical feature vector for actors would be a dataframe that would have movie IDs as rows and list of actors (from the movies in the dataset) as columns (dummy variables). If a specific actor was present in a specific movie, their intersection would have value 1; 0 otherwise.
- The dependent variable (y) is the IMDb rating, the outcome prediction.
- Import the linear regression model using the scikit-learn module, which after instantiating "fits" the x and y dataframes corresponding to the feature columns and the prediction column; here the dataframe x would contain only those columns from the original dataset, and y would correspond to the dataframe that contains the dependant variable (IMDb rating).
- We have our new predictor model using multiple linear regression.

Observations

- Scatter points closely lying around the line of best-fit (blue line), indicating a strong positive correlation in the comparison between the actual and the predicted ratings.
- Root-mean-square error (RMSE): 0.8112

Future Plans

- The rating with certain predictors crosses the 10 point mark - needs to be fixed.
- To consider different kinds of machine learning models such as generalized additive model (GAM) and see which is more effective and gives a lesser RMSE error.
- Memory management: since feature vectors are being created for categorical data, such as multiple actors, currently only considering movies above a certain threshold due to limited processing capabilities on a regular machine. The training dataset includes only movies with more than 1000 votes. (Movie count: 10200, director count: 4528, actor count: 15129)
- Implement support for more features, more outcome predictors such as box office data: budget (feature), gross (outcome).
- Using box office data, analyze trends such as predicting the ideal time (month) for a movie release based on the budget and gross, showing a bar chart for the predicted box office collections depending upon the months.
- Incorporate sentiment analysis on plot synopsis using naïve Bayes classification and Python’s NLTK library; movie plots with a rating of 7.5 are considered positive. Add plot keywords as predictor features.