Final Project: Movie Gross Prediction

[Note: Links and other resources below are provided primarily as hints and guidance only and you are encouraged to explore other places for relevant data and code.]

Working in teams of 2, you are asked to predict gross revenue of movies that will be released in the week of Thanksgiving. So the movies of interest in this prediction will be those that will release in the week of November 26—December 2.

Here is one list of movies that will be on the screen in that week:
https://www.imdb.com/movies-coming-soon/2021-11/

Data Collection

Your team is expected to gather data from various sources, to clean the collected data, and to merge the data into a proper format for machine learning. Example data sources are as follows.

General movie information

IMDB (www.imdb.com), RottenTomatoes (http://www.rottentomatoes.com/), Metacritic (http://www.metacritic.com/): These are online databases of information related to films, television programs and video games, including cast, production crew, fictional characters, biographies, plot summaries, trivia and reviews.

Financial data

The site The Numbers (http://www.the-numbers.com/) provides budgets (http://www.the-numbers.com/movie/budgets/all) and box-office records (http://www.the-numbers.com/movie/records/).

Open source API for movie data collection

Movie db (https://www.themoviedb.org/?language=en) provides APIs that help movie-related information.

A Python wrapper for this API is here: https://github.com/celiao/tmdbsimple Or you may find another or write your own.

Data Analysis

- Dependent variable (class attribute): Do this as both a classification and regression problem.
  1. Regression: predict the gross revenue of movies that will be released in the week after Thanksgiving
  2. Classification: convert the gross revenue attribute (or another appropriate attribute of your choice) into a multi-valued nominal attribute as we have done multiple times in class
- Possible independent variables (input attributes):
  - Budget
  - Number of theater where the movie is released
  - Ratings from reviewers
  - Time/month of year
  - Genre

Possible procedures

1. Correlation analysis
2. Linear Regression, Regression Trees, KNN Regression, or Support Vector Regression
3. Classification models covered in class and researched/explored on your own
4. Fine-tuning of model parameters, including use of feature selection and ensemble techniques
5. Model comparison between models with different independent variables

Languages

You may choose between Python or R for your project work.

General notes

You are data scientists! This assignment is deliberately broad, because you are to fill in the details yourselves from things we have learned in class and things that you research on your own. Your directions are to build regression and classification models to predict movie success. How you do that is up to you. That means experimenting with data collection, cleaning, feature selection, regression and classification models, and the settings of those models – all to achieve success using the metrics we have learned. Don't restrict yourselves to only what we've done in class! Experimentation and learning on your own will count for a lot in the final assessment.

Train your models on the datasets that you are able to build, employing all the tricks you can think of to fine tune them to high accuracy. Then use the post-Thanksgiving week movie data as test data. Final presentations will be during lab on Monday, December 6, so your team will have a few days to do this testing on the real data and prepare a report of your results. You should document your experiments and steps along the way so that this report is easy to create, along with your final prediction results.

You may think of this kind of like a Kaggle competition – very few rules, only results. But of course I will be looking for appropriate application of sophisticated machine learning techniques.

Logistics

Training data, test data, and code files for each team will be submitted via Moodle. Details to be determined.

References

A couple of other similar project that you may use for inspiration:
http://www3.cs.stonybrook.edu/~skiena/591/final_projects/movie_gross/
http://jse.amstat.org/v17n1/datasets.mclaren.html