Movie recommendations based on rating history, movie metadata, credited cast

Ryan McGuinness

Capstone Project Proposal

**Objectives**

The objective of this project is to build movie recommendation systems using several techniques of predictive analytics to suggest viewers new movies by finding similar movies to those the viewer rated highly. Additionally, this project will test whether such a model performs better than simpler models, which might for example consider only the ratings the target viewer and other viewers gave movies. Ideally, this project will test a variety of models to find similar movies.

This project aims to synthesize many techniques taught all throughout the Business Analytics curriculum. Specifically, this project will combine attributes of both predictive analytics and descriptive analytics, in suggesting movies the model predicts the viewer would like and in determining how similar two movies are respectively.

There will be a secondary emphasis on data manipulation and cleaning in this project. Even though this emphasis is unlikely to show in the final product, this is an essential part of all analytics projects and I specifically found a dataset that I believe would test my abilities there.

**The data**

The dataset I intend to use consists of metadata for 45,000 movies from the MovieLens dataset and ratings from 270,000 viewers. There are over 26 million records in the files of this dataset. It is published on Kaggle under the name “The Movies Dataset” by Rounak Banik.

The structure of the data itself is akin to a small database, and will require a healthy amount of cleaning and descriptive analytics to begin to understand. The movie metadata, cast, keywords, and ratings all come in separate .csv files under this dataset. Additionally, several of the .csv files contain data which is not “flat”, and uses convoluted JSON-like objects to store information. Of course, this also means these files are not truly comma delimited.

This particular dataset is linked to hundreds of notebooks on Kaggle alone. While this high usage might suggest that many analytics projects have already been done here and that this dataset might be less desirable for a new capstone project, there are a number of things which make this dataset quite a good candidate for projects. First, cleaning the data here is no trivial task, so my project is very likely to diverge from similar projects at the very onset. Second, the data is real and comes from a reputable source. And third, while the data isn’t remotely tidy, there are few surprises to be found in it.