Machine Learning and Movies

By: Owen Doyle and Justin Self

Overview of Project

- Movies are, to most, a great form of entertainment
- We want to learn more about what makes a movie “enjoyable” (IMDb ratings)
- There are many applications of this research
    - Generating Revenue
    - Oscar Contenders
    - Making Enjoyable Movies in General

Research Question

- What factors make a movie have a higher IMDb rating?
- We will be using an IMDb dataset found on Kaggle.
    - The dataset includes numerous categorical and quantitative variables.
    - There shouldn't be any issues with exploring what potential variables we will use.

Hypothesis

- We hypothesize that a combination of a movies budget and the presence of certain actors will produce the best prediction of IMDb rating.
    - The dataset already contains a variable for movie budget.
    - We will need to work on formalizing how to include actors as a predictor
        - One potential way is to create a new indicator variable that has levels 1 for a movie that contains one or more of the top 100 actors/actresses, and 0 for a movie that does not contain any of the top 100 actors/actresses.
    - Other variables that may have an effect on a movie's rating could include duration or Metascore rating.

Data and Methods

- As stated before our dataset comes from Kaggle.
- It has three different files that contain information about the movie, the actors/actresses, and the IMDb ratings.
- In total there are 85,855 movies in the dataset.
- The data has already been scraped to have at least 100 votes, but additional scraping will be needed to get a good concise set of working data.
    - Some movies have NAN for important variables like budget.

Data and Methods (cont.)

- We plan to use the methods learned in this class to answer our research question
  - That may include random forest and other machine learning techniques
- Pandas and Numpy will be used for most of the preliminary data analysis
  - Scraping our data
  - Generating preliminary reports (5-num summary, correlation, and covariance)
- Matplotlib will be used in creating visualizations of the data

Schedule

- September 27: Part 1
  - Proposal-Justin
  - Presentation-Owen
- October 21: Part 2
  - Report-Owen
  - Short Presentation-Justin
- November 16: Part 3
  - Paper Draft-Justin
  - Demonstration-Owen
- December 16: Part 4
  - Presentation-Both