Predicting IMDb Movie Ratings

By: Owen Doyle and Jusitin Self

Overview

- Introduction

- Design

- Models

- Results

- Conclusion

Background

- Data: Collected from Kaggle

  - Web Scraping: Movies with more than 100 votes on IMDb website

    - IMDb: popular online film database

  - Contained data on movie production, actors, and reviewer
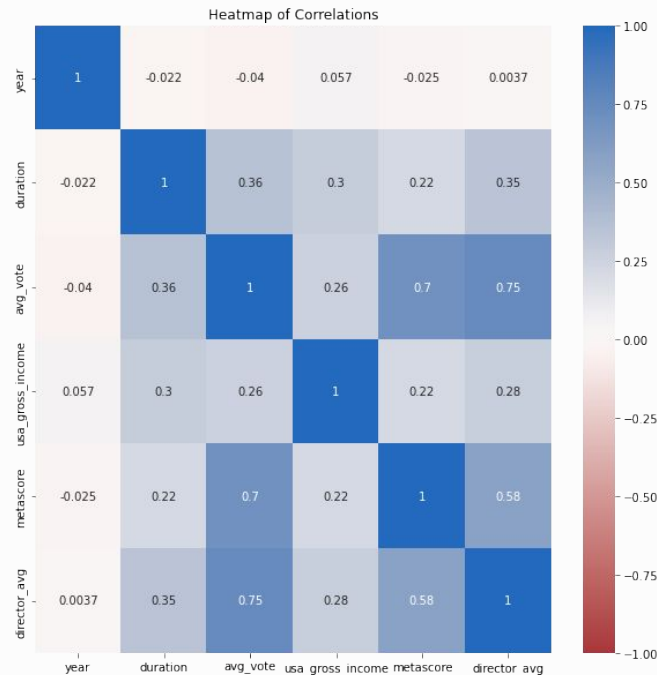
    demographics

Purpose and Question

- Our Question: What modelling method and combination of variables can best predict the average rating viewers will give a movie on IMDb?

- Why?

  - Want to understand what generally makes a well-liked movie

  - Interested in the history of films

  - Better movies can generate more revenue

Design: Exploratory Analysis

- Exploratory Analysis

  - Correlation coefficients (Pearson's Correlation)

  - Pairwise plots

  - Suspected duration, budget, USA gross income, metascore and director average to be most important

Heatmap of Correlations

Design: Modeling Techniques

- Modeling Techniques
  - Regression, both linear and multivariate
    - Extrapolation - estimating average vote
  - Lasso
  - Bagging

Models

- Linear Regression

  - RMSE: 0.52, R^2: 0.71

- Lasso Regression

  - RMSE: 0.64, R^2: 0.55

- Bagging

  - RMSE: 0.68, R^2: 0.51

- Decision Tree

  - RMSE: 0.63, R^2: 0.39

Results

- Linear regression with duration, budget, USA gross income, metascore, and director average was most successful model

- Metascore and director average are very highly correlated with average rating

- These variables can predict the quality of a movie reasonably well

Conclusions

- Multiple factors contribute to the success of a movie

- Model was more successful than we expected

- Possible flaws

  - Variables too closely related to average vote

  - Information is not available until after movie is released

- Learned that data might not contain exactly what we need

  - Had to create new variable

Future Research

- With out best model we are able to predict the average rating of a movies after they come out (gross income, metascore)

- Future investigation can include building a model with future that are know only prior to release

  - Budget, Director Average, Duration

  - Actors, Production Company

Demonstration