Predictability of TV Show Ratings

Sandy Weng

Abstract

The goal of this project was to use Linear Regression to see the likelihood of predicting tv show ratings in order to produce shows with high ratings. I worked with data scraped from IMDB and used a few Linear Regression models to get the best R-squared value.

Design

This project originates from the producers at Netflix who want to predict what kind of show will receive high ratings. This is a preliminary model exploration to see whether a good model can be built using Linear Regression given the public information. The data is scraped from IMDB and includes the following features – year released, number of reviews, genre, certificate, number of episodes, actors, network, runtime, and awards. Being able to predict ratings accurately would save money and increase viewership for the company.

Data

The dataset contains 4499 tv shows with 5 numeric features and 4 categorical features. The data was scraped from IMDB with BeautifulSoup. The numeric features were used to form a baseline model and then the categorical features were added for the other models.

Algorithms

Exploratory Data Analysis

- Cleaned and imputed missing data
- Created dummy variables for categorical features

Models

Simple Linear Regression, Ridge, and LassoCV were used before settling on LassoCV because LassoCV had the most stable R-squared values across training, validation, and test datasets.

Model Evaluation and Selection

The entire dataset of 4499 titles were split into 80% training data and 20% test data. Then, within the training data, 20% were split into validation data.

The final LassoCV R-Squared value is 0.084.

Tools

- Numpy and Pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib and Seaborn for plotting
- Statsmodels for modeling

- BeautifulSoup for web scraping

Communication

- Presentation of slides and visuals
- Project writeup summarizing my work