Project Name: Predicting a Movie's Success on Total Gross Revenue
Hannah Kim

Abstract:

The goal of this project is to use linear regression to predict a movie's total revenue based on certain features calculated. The topic brought upon interest is due to our client who is a movie investor. Analyzing the question of whether a movie's revenue can be predicted is to ultimately help movie investors who want to work on new movie projects and those who would like to find a way to recognize movie profitability. Due to the fact there can be risks when determining which movie to invest in, this project aims to help alleviate those worries and bring upon a model that can establish results needed.

Design:

This project utilizes the boxofficemojo website and to supplement will use the Kaggle dataset to bring upon more features like director and star of the movie. For the model, the target will be total revenue which is based on the sum of grosses domestically and internationally. The individual sample/ one row represents the impact certain features have in association to total revenue. The features I will work with will be: movie title, gross worldwide, director, distributor, genre, mpaa rating , star, year, runtime, and budget.

Data:

Data collected was from 2008 - 2019 and after cleaning the data I had a total of 1263 data points with 37 features for the final model before cross validation tests. Later on I eliminated certain features based on p-values and noise from VIF values I found in the model.

Algorithms:

Feature Engineering:

To add more complexity to the model I used dummy variables after analyzing the baseline model of just numerical features. Prior to implementing dummy variables I categorized within the feature of interest as 'other' for features that had a lot of value counts.

Models:

Used K-fold, CV on linear regression, lasso, ridge, polynomial regression, poly + ridge and poly + lasso to get the best model on test values. I ended up using the log transformed model to find out which regression model helps with better interpretation. I first did train_test_split on the log model to calculate train and validation scores of R2. Later I used test values to find out the best model. Ridge was used in the final model due to the highest R2 score. The value of R2 was 0.343501. Using Ridge, I was able to calculate the predicted and actual value for total gross on the movie, Jojo Rabbit. I got ~$9.12 million for predicted and ~$90 million for actual. This shows there is still much to implement and feature engineer as the actual total gross is ten times more than the predicted value. Due to lack of features and residual error on the target variable, the R2 could no longer be further improved.

Tools:

Will use the following:

- BeautifulSoup
- Pandas
- Numpy
- Sklearn
- Statsmodel

Communication:

Will deploy on my github after revision.