Movie Revenue Prediction

Submitted by:

Team Name: Kitna Kamayegi?

Members: 1. Sourabh Kondapaka (MT2018119)

2. Naman Bhatt (MT2018066)

3. Namanraj Varshney (MT2018067)

Abstract

Movie revenue prediction has been studied in a variety of contexts ranging from economics and business to statistics and forecasting.Fundamentally, revenue prediction is a regression task in which we seek to estimate a single number representing the gross revenue based on a variety of factors.

Primary Dataset is taken from Google dataset website.

URL:https://toolbox.google.com/datasetsearch/search?query=IMDB%205000%20Movie%20Dataset&docid=zDmMCvkxFert5jadAAAAAA%3D%3D

Since the primary features available in the dataset were not enough the rest of the data was scraped from the following websites:

www.facebook.com

www.twitter.com

Variables Used For Analysis

This **primary data set** contained the following features:

- popularity : Gives us a measure of how famous is the particular movie in respect to others
- vote_average: Provides a insight on how the movie was received by the crowd.
- Genres: Provides knowledge about under what genres does the movie falls()
- Budget: Amount spent on the making of the movie.
- Vote_count: Number of votes given to rate the movie on a scale 1 to 10 , 10 being the highest.
- Vote_avg: Average of all the votes given to particular movie.
- release_date: The date at which the movie was released(This dataset was given in the format of date which was converted to number of days with 18-11-1910 being the zeroth day).

Since the above features are not enough to predict the revenue generated by the movie, as the popularity of the lead character also provides some extra credebility to the number of people showing up for the movie. To know about the popularity of the lead character facebook likes and twitter followers were scraped. To maintain the authencity of twitter and facebook handles of the lead character it was scraped in follwing steps:

1. The name of lead cast was retrieved from the credits.csv file.
2. The google link to search the obtained actor was created.
3. The link to the twitter and facebook handle of actor was then scrapped.
4. Finally, the number of followers was scraped from the twitter and facebook handle links.

Derived Features with the help of web scraping:

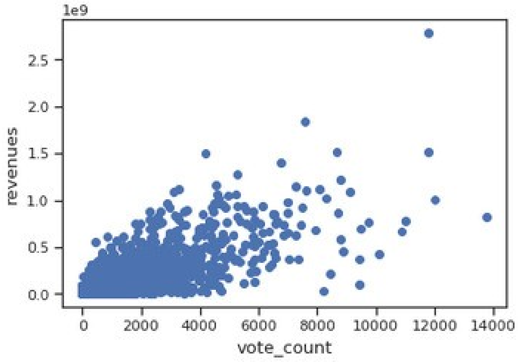
- Facebook likes of the lead character.
- Twitter followers of the lead character.

- Popularity_Index: With the help of facebook likes and twitter followers another column was created which would provided us a relative view of how popular the lead character of the cast is.

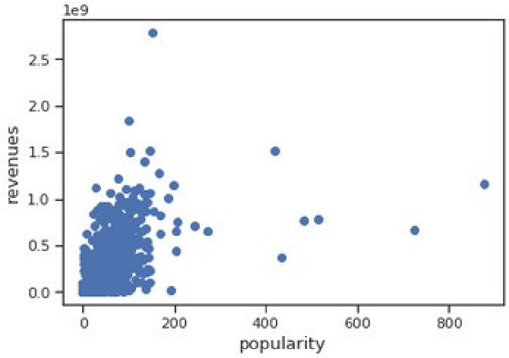
Target Variable:

- 'Revenue': A prediction of how much will the movie make in the box office.

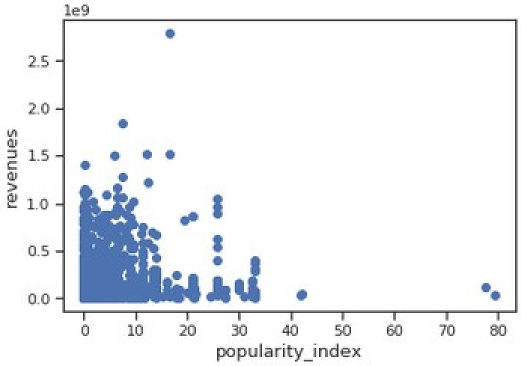
Exploratory Data Analysis (EDA)



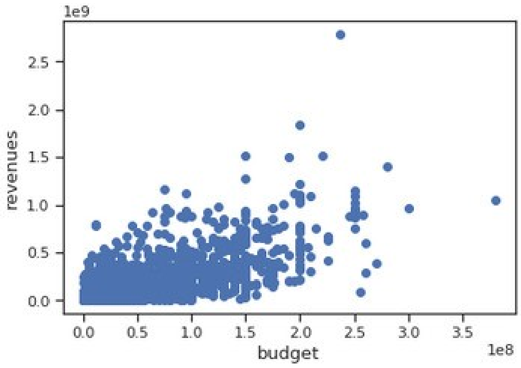
Corelation between revenues and vote_count



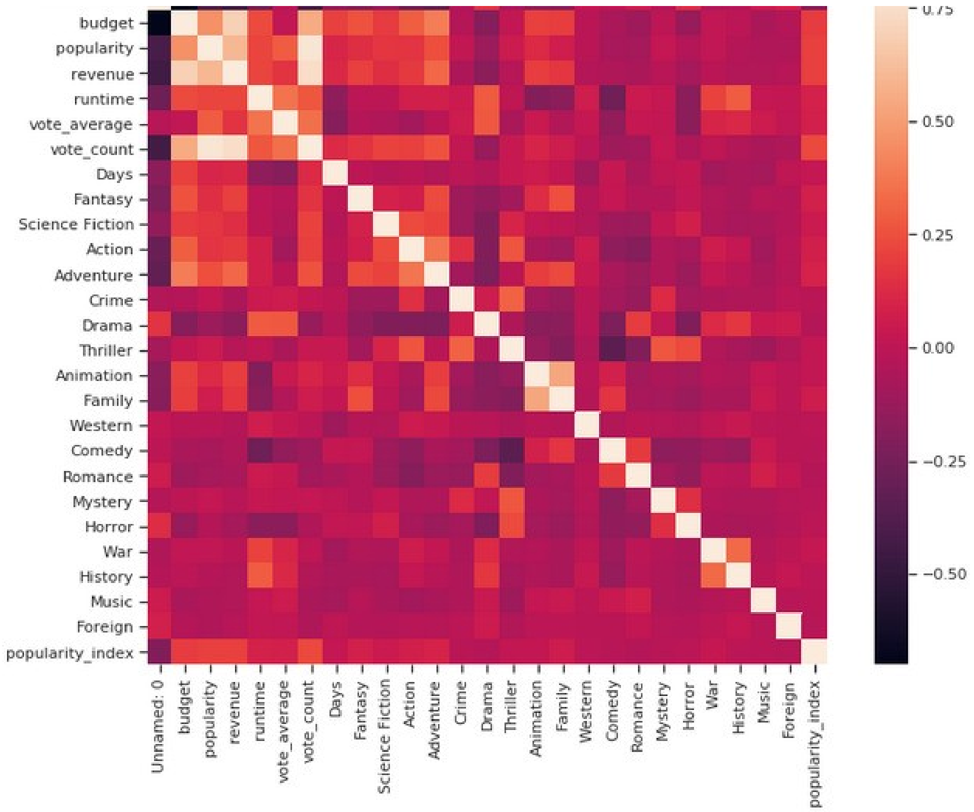
Corelation between revenues and popularity



Corelation between revenues and popularity_index



Corelation between revenues and budget



Correlation between various features and MSRP

Approach

- First we found missing values and visualized that features to observe it's pattern and accordingly filled the missing values.
- We scrapped www.facebook.com , www.twitter.com for likes and twitter followers respectively of the lead characters in the movies.
- We preprocessed certain columns such as release_date of the movie and filled out missing values appropriately.
- Based on likes on facebook and followers on twitter, we created derived attribute called popularity_index which infers the popularity of the actor/actress on social networking sites.
- We visualized relationship between features by finding correlation matric.
- We observe outliers using correlation metric and Scatterplot and drop some of the columns or tried to remove outliers.
- The models used to predict the revenue generated by the movies were linear regression, lasso and ridge.

Conclusion

1. With linear regression the prediction of the revenue of movies have an average accuracy of 73.15 .
2. With ridge and lasso the marked price prediction of the revenue of movies have the accuracy of 73.114 and 73.095 respectively.