Report

Movie Gross Predictor

Viraj Bharadwaj

Siddharth Reddy

Raja Rakshit

Special thanks to **Viraj Bharadwaj**

Problem statement, dataset, approach taken

Problem statement :

To predict the gross collection of a movie given its budget,production,rating etc.

Dataset :

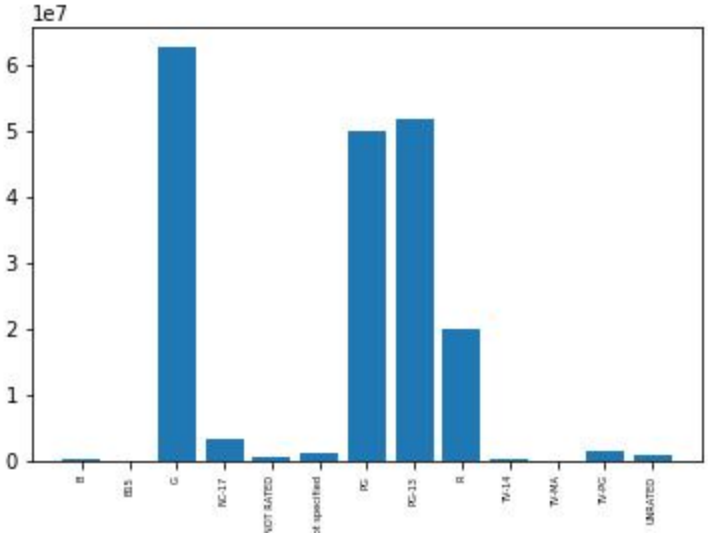
Three decades of movie data (1986-2016), taken from kaggle , which was scrapped from IMDB using Python.

Approach taken :

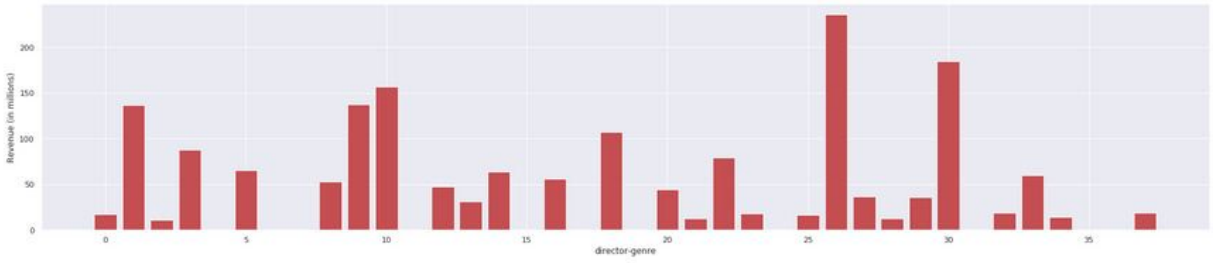
Analysis of dataset, visualisation, feature engineering

Analysis and Visualisation:

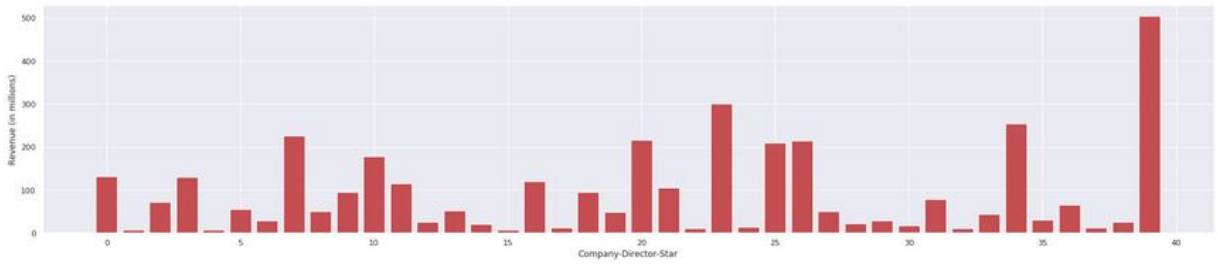
We saw correlations of all attributes to target variable and we found much more correlation when we club some subset of attributes.



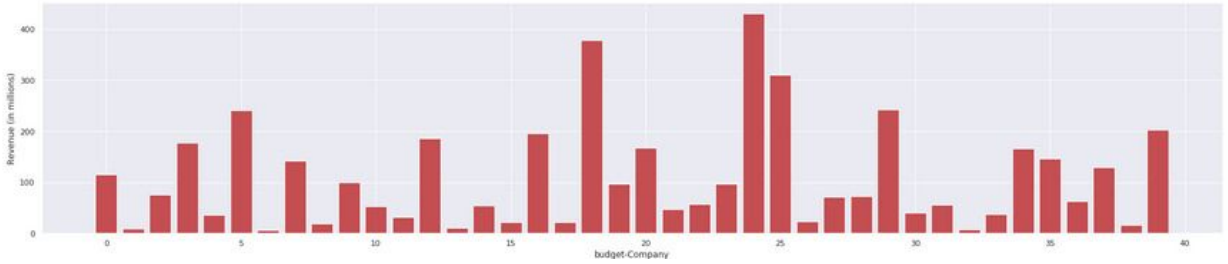
Rating vs gross



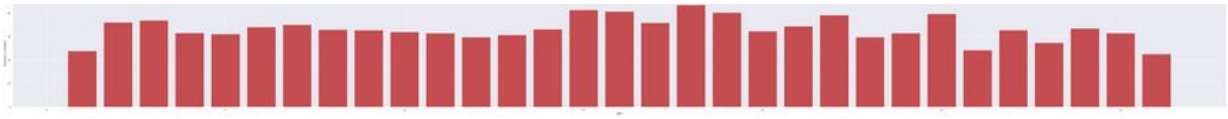
Director -genre vs gross



Director -company-star vs gross

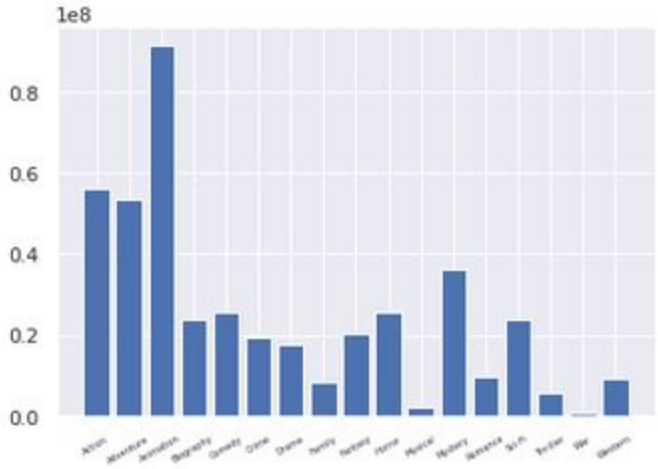


Budget -Company vs gross

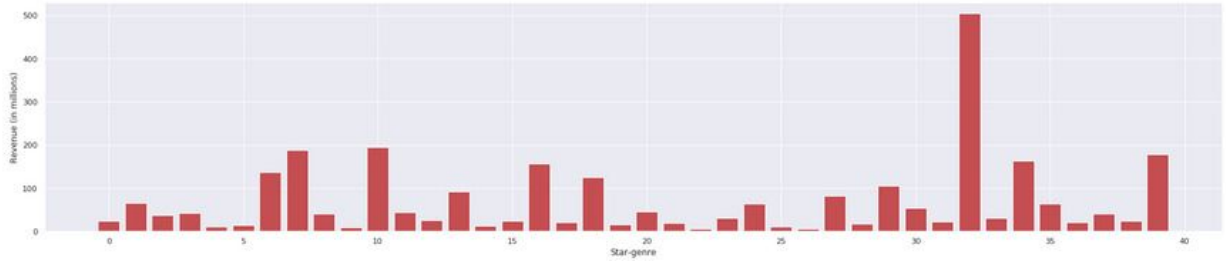


Date vs gross (less correlation)

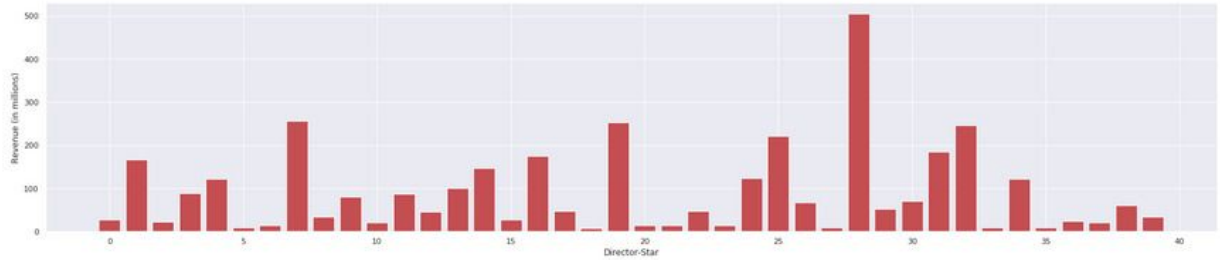
Month vs gross



Genre vs gross

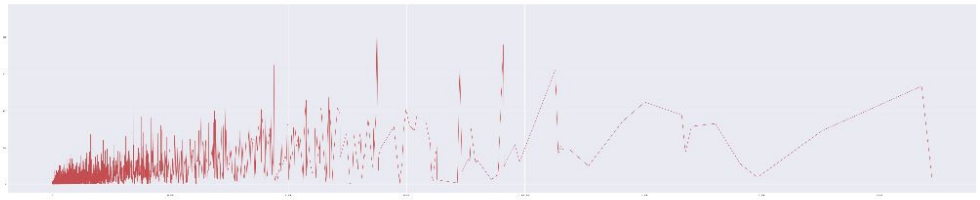


Star- genre vs gross

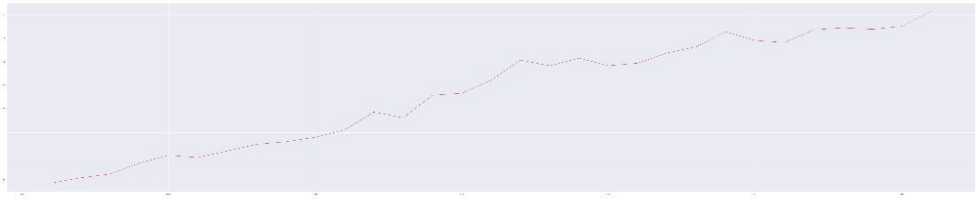


Director -star vs gross

Score vs gross



Votes vs gross



Year vs gross

These things has more correlation than standalone attributes.

Feature engineering:

The gross collection of a movie depends on numerous number of factors , even top experts of the movie industry sometimes fail to correctly predict the outcome of the movie.so , it's not an easy task to predict the gross over a few features.

Even though it's tough to predict the outcome there are some features on which the gross depends to some extent.

- Budget:

  having a bigger budget in most cases means bigger gross.this may not be true in all cases but there sure is some average increase in gross as budget increases.

- score:

  not all good rated movies may have collected more but all movies with good gross have an above average score.

- Rating:

  Movies with R rating tend to get less gross when compared to movies which are available to everyone, since they missout teenage audience.

- Genre:

  Popular genres like action,romance,etc seem to get higher gross than less popular genres like documentaries,slice of life.

- Votes:

  Two movies with same score may not get same gross , because both movies may not be equally popular, the number of votes received during a score can be taken as a measure for popularity of a movie.

- Country:

  Movies from countries like USA,UK where movies are more popular and also have developed movie industry seems to get more gross than movies from other countries.

- Release date:

  The time around which day the movie was also played a role in the gross. A movie is generally released around the weekends and movies released during holiday season always had more gross.

- Release month:

  The time around which month the movie was also played a role in the gross.

- Release year:

  The time around which year the movie was also played a role in the gross.

- Director-genre:

  A movie director may not good in making movies in every genre so, when a director makes a movie on one genre may collect more than other

- Budget-company:

  High budget movies are generally made by big production companies, which are popular with the audience also a big prod-company does many things better like publicity, etc which will help the movie to get higher gross

- Star-genre:

  When a famous actor acts in a movie it ends to get higher gross than the movies with less popular actors.but more peculiar pattern is seen when an actor acts in a certain genre.for example when an actor makes a action flick it gets more gross than when he makes a classic.

- Director-star:

  A director-actor combo who had success before tends to attain higher gross when they pair-up again.

- Company-director-star:

  Similar to director-actor combo when a production company-director-actor team up again like for a sequel seems to get a higher gross.

Model building, training, validation, test metrics

Model building:

We used one-hot encoding for all attributes except score and votes where we used label encoding.

Company-director-star was dropped and we got a accuracy of 71.6%. So,this feature must be included. Similarly when we did for Director-star, Star-genre, Budget-company, Director-genre, we got are 72.1% ,71.6%, 72.1%, 71.9%.

So, all these features are important and we didn't remove it.

Model we used is **Random Forest Regressor**

Training:

As it would be more practical , we trained the model on movies released from 1986-2015 and we tested it to movies which were released in 2016

Test metrics:

Accuracy we got from the model is **72%**

RMSE is 2.465