Bollywood Movies regression Model

Muqaddas

January 2017

1 Data Exploration:

We have data-set of 49521 movies from 2001 - 2015. It includes data of 175 countries.

Among this data-set, there were only 4274 objects of Indian movies.

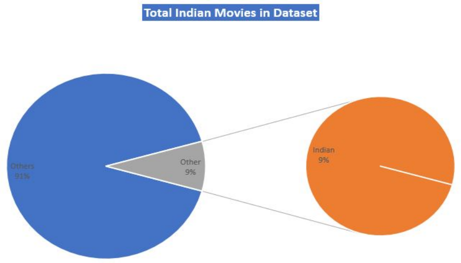

Figure 1: Total Indian Movies in Dataset

There were a lots of missing values in Indian dataset of movies. Which can be seen in figure 2.

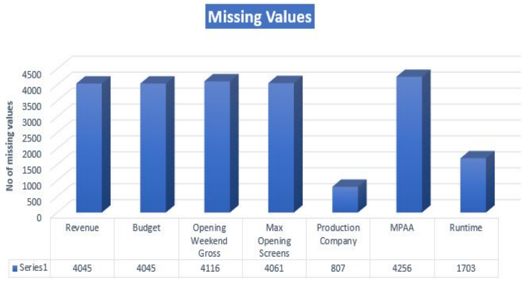

Figure 2: Missing Values in Indian Movies Dataset

As we can see that there are a lots of missing values. That's why we are not able to replace them with mean or median. So, ultimately we remove all the objects and the remaining dataset left with 192 objects.

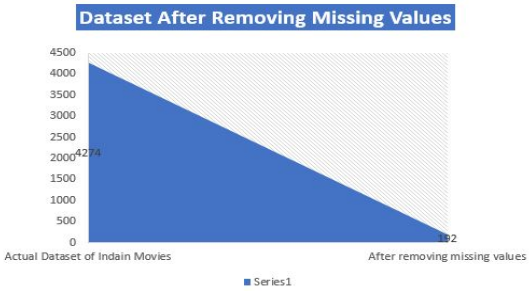

Figure 3: Dataset afer removing Missing Values

There were some outliers in which budget and revenue was less than $100. After removing them only 162 objects left. Among these 162 movies, total successful and unsuccessful movies can be seen in figure 4.

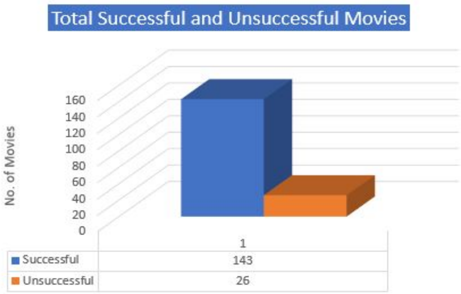

Figure 4: Total successful and unsuccessful movies

2 Linear Regression Model:

For linear regression model we are using 90% of data as training dataset and 10% as testing dataset. Random movies are chosen as testing data.

2.1 Independent Variable:

We use revenue as an independent variable in our regression model.

2.2 Dependent Variables:

Budget, star power,IMDb ratting, number of votes(on IMDb), genre, runtime and success(which is 1 if it is successfull and 0 if not) are used as dependent variables in our regression model.

2.3 Result:

The reult of linear regression model can be seen in figure 5.

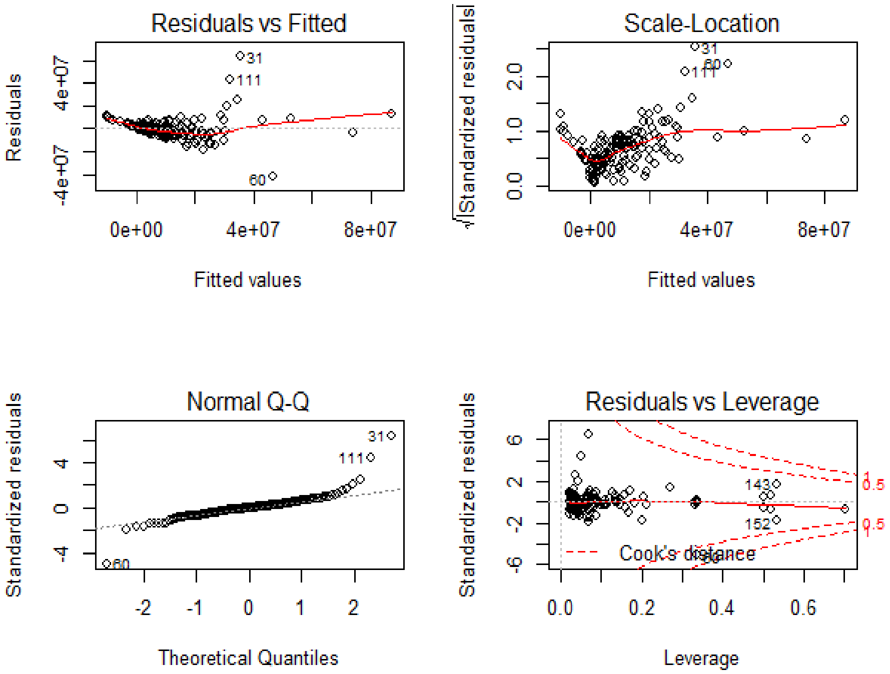

Figure 5: Linear Regression Model for Bollywood Movies

Residual error of model is as follow:

Residual standard error	Multiple R squared	Adjusted R squared
10230000	0.6705	0.6259

Table 1: Error of Linear Regression Model

And the intercept, coefficients are as follow

Coefficients	Estimate
Intercept	-7.275e+06
Production Budget	1.503e+00
Star Power	3.955e+05
Ratting	1.461e+05
Number of Votes	2.719e+02
Runtime	1.597e+04
Success	1.130e+07
Genre(Adventure)	-1.817e+07
Genre(Animation)	-5.060e+06
Genre(Biography)	-6.530e+06
Genre(Comedy)	-7.122e+06
Genre(Crime)	-7.357e+06
Genre(Drama)	-4.900e+06
Genre(Family)	-4.115e+06
Genre(Horror)	-3.218e+06
Genre(Musical)	-6.086e+06
Genre(Mystery)	-7.412e+06
Genre(Romance) -6.237e+06	
Genre(Thriller)	-2.041e+06

Table 2: Regression table of model 1

2.4 Problem with this Model:

- There are a lots of industries in India e.g. Bollywood, Tollywood etc, can be seen in figure 6.

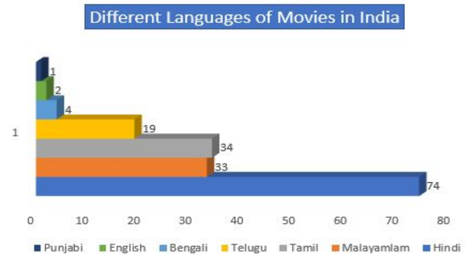

Figure 6: Different languages of movies in India

- Dataset is really small and there are a lots of missing values too.