Final Project: TMDB Movie revenue prediction

Sohan Thakur

Bellevue University, Nebraska

# Abstract

The objective of this paper is to discuss about different aspect of the final project that I have worked for. I selected TMDB movie revenue prediction as the topic of the project. I performed extensive exploratory data analytics and feature engineering to analyse and derive new features. At the end I used multiple Linear regression to develop a model that predicts the revenue with some accuracy with a lot of room for improvement.

*Keywords: model, EDA, Feature Engineering.*

• **Statistical/Hypothetical Question**

* **Are Revenue and the budget of the movie correlated?**
  + Yes
* **Is the correlation in above case statistically significant?**
  + Budget and revenue seems to be strongly correlated with a p-value less than 0.05.
  + Thus the correlation is statistically significant.
* **Do we have outliers in the budget column?**
  + Yes, based on histogram, there are quite a few value on the higher side with more than $250M as the budget of the movie, they are outliers.
* **What is the correlation between popularity of the movie and the revenue?**
  + Popularity of the movie is moderately correlated with revenue 0.4359925127357491
* **Is the correlation statistically significant?**
  + Yes, p-value 4.3153222778910924e-94
* **Does Genre of the movie plays an important role in revenue generation, get the revenue distribution by genres of the movie?**
  + Yes, based on Histogram it appears that movie with 3 or 4 different genres perform better than less than 3 or more than 4 genres on average.
* **Is long runtime impacts revenue generation positively or negatively?**
  + Movie Runtime and revenue doesn't seem to be correlated with a very small p-value less than 0.05.
  + Thus the correlation is statistically significant.

• **Outcome of your EDA**

* + 1.For Budget, 75% of the values are less than or equal to 40M , hoever max value is 380M. Thus the higher values seems to be outliers.
  + 2.For popularity score, 75% values are less than or equal to 11.89 but the max value is 295, seems there are outlier on the higher side of values.
  + Popularity score is a leaky predictor which will be available only after few weeks of movie release, hence I am not considering that in the analysis.
  + Day since release is also a leaky predictor.
  + Language of the movie is not impacting the revenue a lot. CDFs of all the movie’s revenue VS only movies with language as English does’t show any difference.
  + Average renenue of movies with 3 or 4 genres is : $123.30949550502511M
    - Average renenue of movies with other than 3 or 4 genres is :$93.86691896867231M
    - stat=3.280, p=0.001
    - Difference in mean is significant

• **What do you feel was missed during the analysis?**

* With many features available, there are possibility of creating dummy features based on the available ones. I couldn’t create many possible new featured.
* The model with few selected features seems to have moderate accuracy, I wanted to use more featured to train the model.

**• Were there any variables you felt could have helped in the analysis?**

* Yes, there are few variables which may play an important role in the revenue like is the name of the production company, name of the actors etc.

• **Were there any assumptions made you felt were incorrect?**

* I was assuming the popularity should have very strong correlation with revenue, however that was not the case.

• **What challenges did you face, what did you not fully understand?**

* While calculating the cdf, the variables that I choose were not normally distributed, hence I had to perform hit and trail to find a range of population which are normally distributed.
* I scaled down few variables like Budget and revenue only after it cause enough trouble without scaling.

• Submit a link to your repository to the assignment link during the final week of class.

References:

Think Stats , Exploratory Data Analysis in Python by Allen B. Downey