Luke Bravo, Lai Jiang, Ali-Daanesh Sayyed

Python Data Analysis on Movie Dataset

Description of Project Goals

We are investigating a dataset on movies released between 1927 - 2016. The dataset includes information about the movie's production, reviews, genres, actors involved, directors involved, post production value as well as a variety of other variables.

We initially did exploratory analyses on the dataset to understand what relevant trends and insights we could extract. This allowed us to dive deeper into the content. We realized there was a significant amount of data on the production of each movie, which enabled deeper analysis on the aspects of production that led to review and revenue success. The bottom lines of our analyses are:

1. Whether we can predict a movie's ratings, revenue, and budget based on the variables we have in the data?
2. Which movies, actors, and directors are most notable in the features we are using to analyze?
3. Explore various predictive modeling methods to analyze our data and see which methods produce models that are reliable in comparing test sets and training.

Companies like Netflix, Hulu, and HBO Max want to understand what drives movies' success as they continue to transform the way media is consumed. We believe our analyses on budget, director and cast popularity, and genre provide insight for big media companies to create the most attractive content for their audience. The goal of these analyses is to help increase profits margins for media companies. The analysis also uncovered trends within the industry that provide economic insights to how movies generate revenue. The way media is consumed is rapidly changing. Understanding what factors contribute to movie success will help media companies select better casts and focus on popular genres in order to maximize gross profit.

Exploratory Analysis

Through our exploratory analyses, we got a better idea of which variables had high insight-potential. The first analysis done was to understand who the top 10 directors shown in figure 1. Our next analysis was to understand what the longest and shortest movies were which we found were 511 minutes long and 7 minutes long respectively.

Genre was another variable we wanted to explore. We discovered that the top 10 genres all had IMDb scores of 8.3/10 and above, however, that was suspiciously high. After reviewing the data, it appeared that the highest rated genres were those only mentioned in the dataset once. To get a better understanding of what the imdb scores for the genres were, we created a new series that only had single-category genres.

Once this new series was created, we were able to see the top three genres: comedy, action, and drama. After finding single category genres, we looked at their average imdb scores. We found that the highest scoring genres did not have the highest average IMDb scores. Comedy, action, and drama all had scores in the 6.1-6.7 range whereas the highest scoring genres were film-noir and history at 7.6 and 7.5 respectively, which makes sense as there was only 1 submission for those niche genres.

Another analysis done was to look at how many movies used the word “The” in their titles. Through a quick search, we found that the word “The” was used in 982 of the movies within the dataset.

From there, we moved on to regression analyses. First, we used IMDb score to predict gross revenue. There are two scatter plots (fig. 2) corresponding to this model. The second excludes outliers, and better illustrates the trend of the data. The second OLS predicted gross with budget (fig. 3); naturally there was a high correlation here, but there was also a low r^2, suggesting external variables play a role in the pattern we saw. In light of this, we would later examine gross and budget in time series analyses. Our last regression predicted IMDb score with budget (fig. 4), there was a weak relationship and low r^2, so we did not explore this relationship further.

One challenge with the data was that budget and gross were recorded in terms of dollars at the time of production. To get meaningful insights, we pulled data on CPI and inflation from 1913 to 2014 published on Data Hub. PANDAS was useful for calculating annual average indices, which we used to inflate each US-produced movie’s financials to 2020 USD. Then, we were able to build proper regression models examining movie gross and budget.

Solution and Insights

Our regression analyses suggest that we need additional data to identify key drivers of movie success. While our OLS models were able to illustrate general trends (e.g., a positive correlation between budget and gross), their r^2 values suggest that there are external variables influencing said trends. This is one of the routes for further research we uncovered with our analysis, and time series analysis helped to determine specific factors that may be driving the correlations OLS found.

Our time series analysis (fig. 5) uncovered unexpected trends, specifically in the change over time of movie budgets and gross. Controlling for the number of movies released each year by computing percent change each year, we found that movie spending has stabilized over the years. We also found that percent change in movie gross has stabilized concurrently with budget. Since we controlled for volume by taking annual averages, we can rule out the possibility that these trends are due to the increase in available data each year. Rather, we believe that this is a sign of movie studios recognizing the trend of decreasing marginal benefit of increasing their budget. The two time series plots appended to this report illustrate that while budget and gross are correlated and jointly increased steadily through time, they reach an asymptote in the 2000s. Based on conjecture, we believe this may be the result of streaming platforms. These services enabled a new type of content consumption which no longer translates directly to profit-per-view for movie studios.

For services like Netflix, Hulu, and HBO Max, this insight can be taken as confirmation that their services impacted media consumption; platform-exclusive movies and miniseries attract large audiences. We believe the rise of a new type of media consumption is likely (1) an external, moderator variable to the pairwise regressions we built, and (2) an key factor in movie studios' business decisions and the reason why we no longer see huge spikes in movie budgets in the 2000s onward.

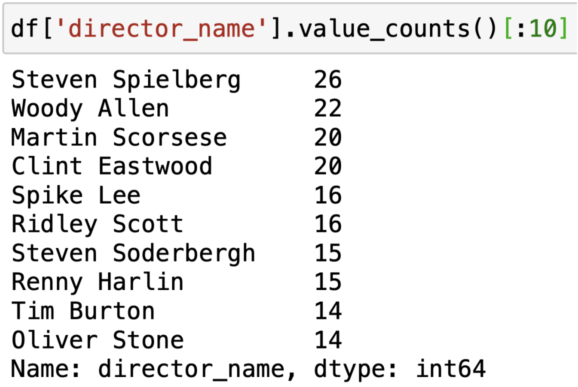
The central question we raised was whether we can predict the success of the movies with selected features in the data set. We excluded names of directors and actors since they are merely categorical variables in decision tree analysis which would be too complicated for data analysis. Thus I select color, number of critics for review, duration of the move, popularity of directors and actors, languages, content rating, budget adjusted for inflation, aspect ratio and etc. The classifier we used is imdb score greater than 6, which is a fair baseline to indicate relative success of the movie.

Initially when most variables are incorporated as features in the model, we got 100% accuracy on training data and only 68% on test data. It was clearly that the model was overfitting. Thus we decide to prune the tree by first deciding the most optimal depth of the decision tree. Set StratifiedKfolds allows us to compare 3 sets of randomized samples at each depth. We found the best and easily approachable depth is 3. Model predicted 74.5% success on the training set and 71.75% on the testing set.
The most important features are the number of voted users for the movies, duration of the movie, budget in 2020 in decreasing importance. Decision tree is shown in figure 6. Number of vote users are at the root. At the first depth, the sample is split on whether the movie's number of voted users equals 109532, the second depth on the left is whether the movie duration is greater or less than 110.5, On the right, the classifier is number of voted users less than 223,703. On 3rd depth on the left, the classifiers are based on different adjusted budgets in 2020. The branch on the left of every end branch are the good movies aka Imdb score greater than 6.

We also used color, duration, director, actor 1-3 facebook likes, aspect ratio to determine the budget of the movie adjusted for inflation. We purposely avoided categorical variables such as names of actors and directors since the data would become very difficult to analyze. We substitute this with facebook likes since more likes is correlated with popularity of the person. The R squared value is rather disappointing from our multivariate regression at 0.1646 and director facebook likes does not appear to contribute to budget estimation with p value >0.05 and Actor 1 facebook likes only has a T score of 2.233; these results are summarized in figure 6. The result indicates that while all consider equal actor 2 and 3 popularity, aspect ratio, color, duration of the movie is more important in budget estimation. However, there are other things not included in the data that contributes to the overall budgets.

Plots, Images, & Data Tables

Top 10 directors in dataset

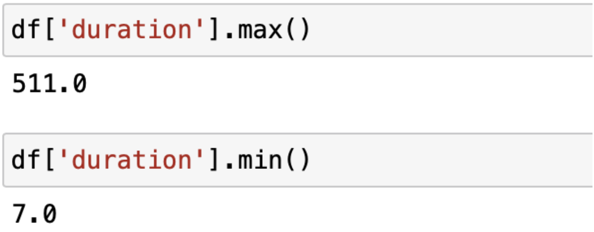


```
df['director_name'].value_counts()[:10]
```

```
Steven Spielberg      26
Woody Allen           22
Martin Scorsese       20
Clint Eastwood        20
Spike Lee             16
Ridley Scott          16
Steven Soderbergh     15
Renny Harlin          15
Tim Burton            14
Oliver Stone          14
Name: director_name, dtype: int64
```

Figure 1

Longest & Shortest Movie in Dataset



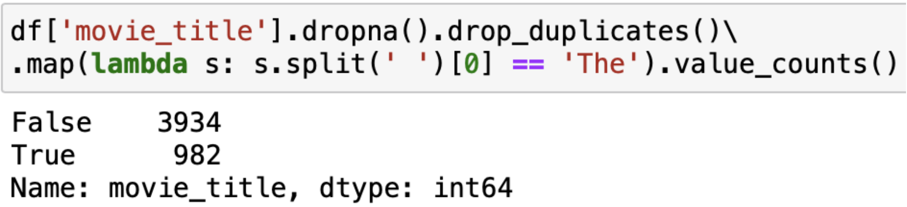
```
df['duration'].max()
```

```
511.0
```

```
df['duration'].min()
```

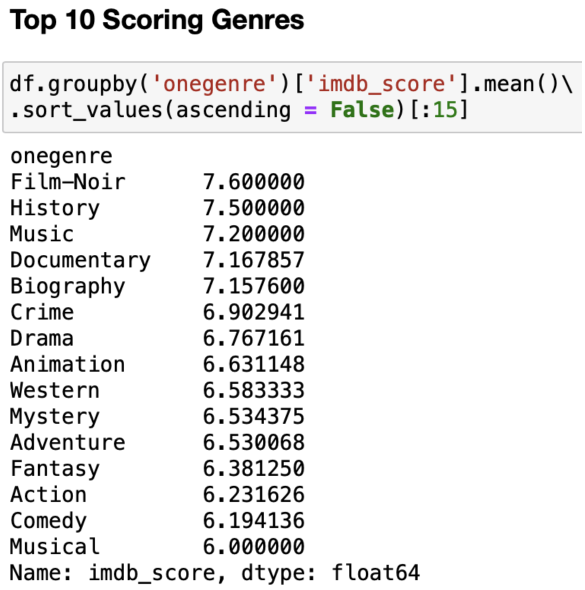
```
7.0
```

How many unique movie titles have The in their name



```
df['movie_title'].dropna().drop_duplicates()\
.map(lambda s: s.split(' ')[0] == 'The').value_counts()
```

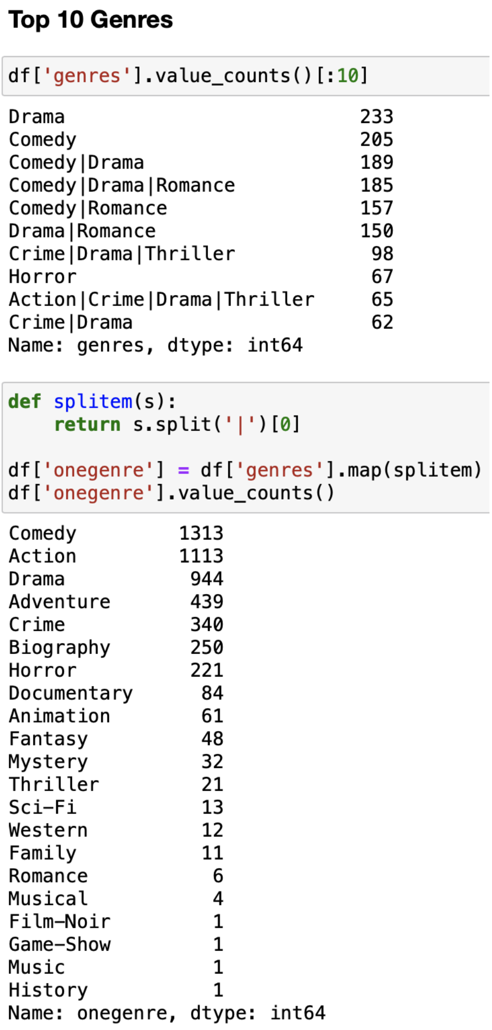
```
False    3934
True      982
Name: movie_title, dtype: int64
```



Top 10 Scoring Genres

```
df.groupby('onegenre')['imdb_score'].mean()\
.sort_values(ascending = False)[:15]
```

```
onegenre
Film-Noir      7.600000
History        7.500000
Music          7.200000
Documentary    7.167857
Biography      7.157600
Crime          6.902941
Drama          6.767161
Animation      6.631148
Western        6.583333
Mystery        6.534375
Adventure      6.530068
Fantasy        6.381250
Action         6.231626
Comedy         6.194136
Musical        6.000000
Name: imdb_score, dtype: float64
```



Top 10 Genres

```
df['genres'].value_counts()[:10]
```

```
Drama                             233
Comedy                            205
Comedy|Drama                      189
Comedy|Drama|Romance              185
Comedy|Romance                    157
Drama|Romance                     150
Crime|Drama|Thriller               98
Horror                             67
Action|Crime|Drama|Thriller        65
Crime|Drama                        62
Name: genres, dtype: int64
```

```
def splitem(s):
    return s.split('|')[0]

df['onegenre'] = df['genres'].map(splitem)
df['onegenre'].value_counts()
```

```
Comedy         1313
Action         1113
Drama           944
Adventure       439
Crime           340
Biography       250
Horror          221
Documentary      84
Animation        61
Fantasy          48
Mystery          32
Thriller         21
Sci-Fi           13
Western          12
Family           11
Romance           6
Musical           4
Film-Noir         1
Game-Show         1
Music             1
History           1
Name: onegenre, dtype: int64
```

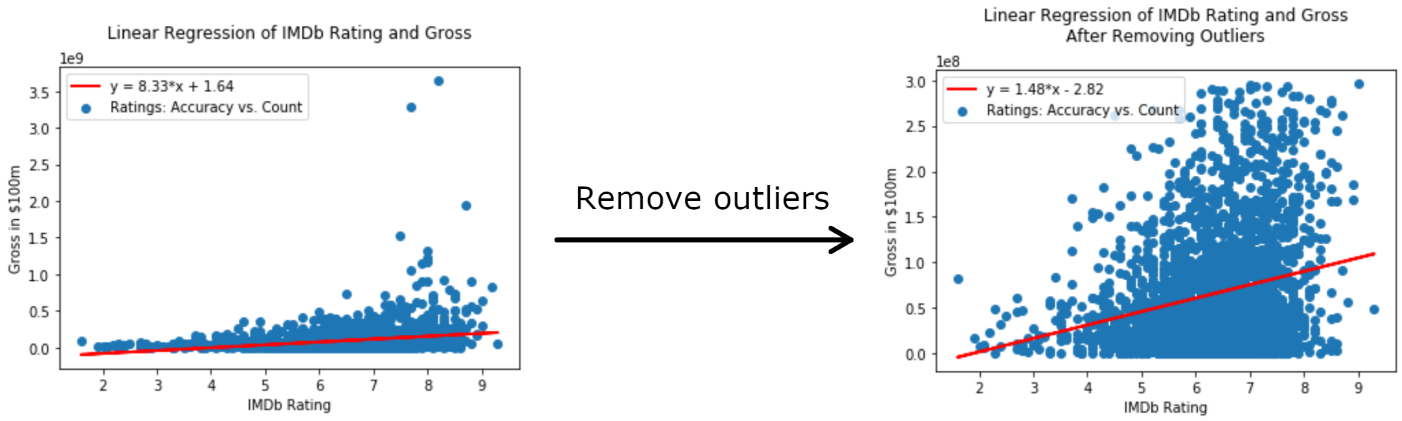


Figure 2, Predicting Gross with IMDb Score

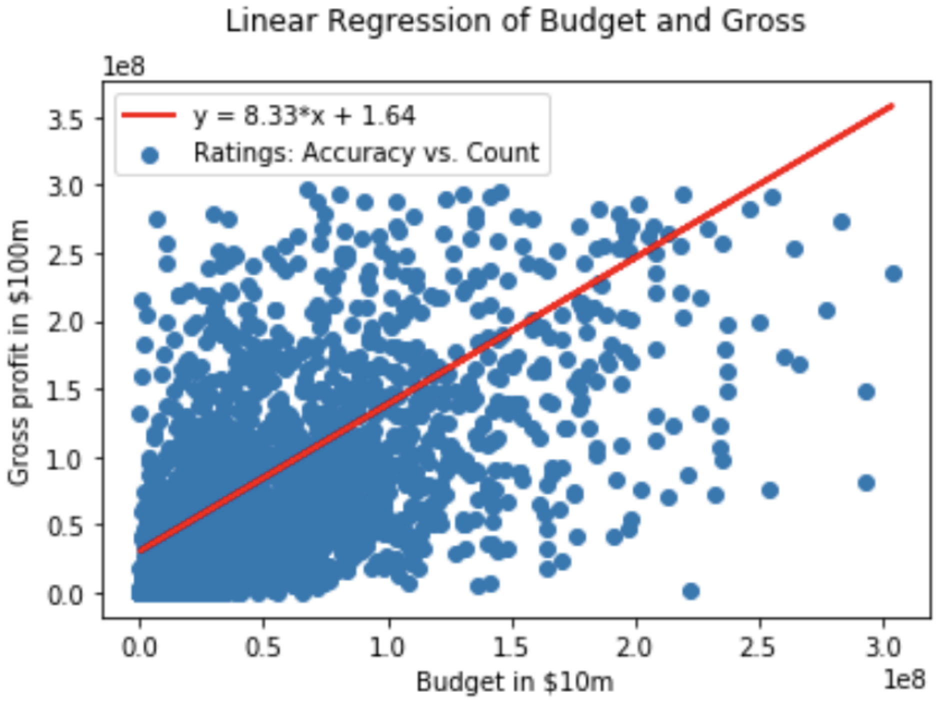


Figure 3, Predicting Gross with Budget

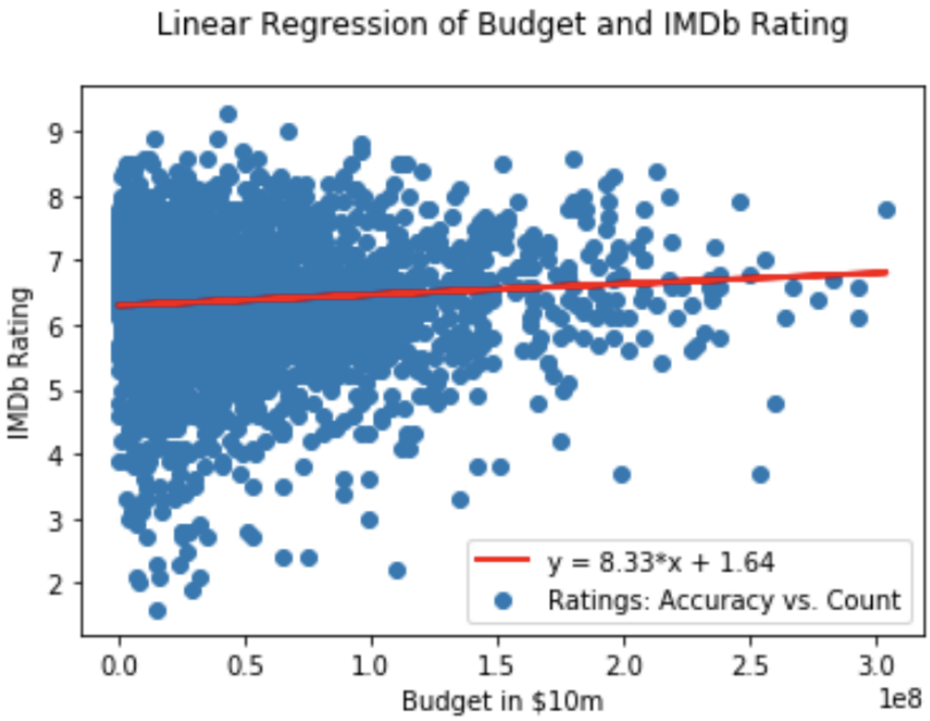


Figure 4, Predicting IMDb Score with Budget

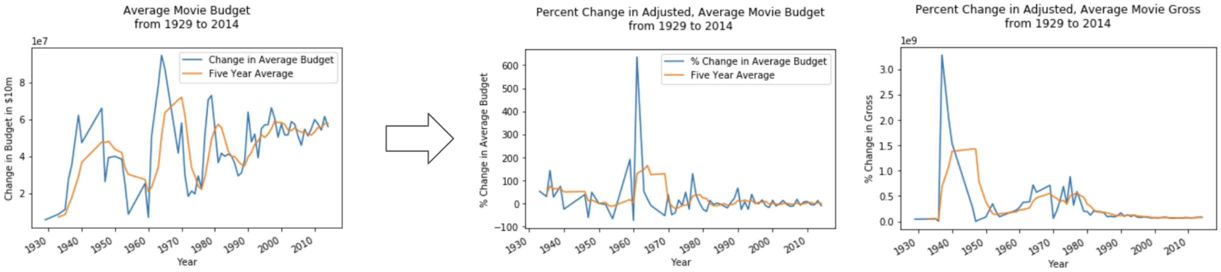


Figure 5, Time Series Analyses

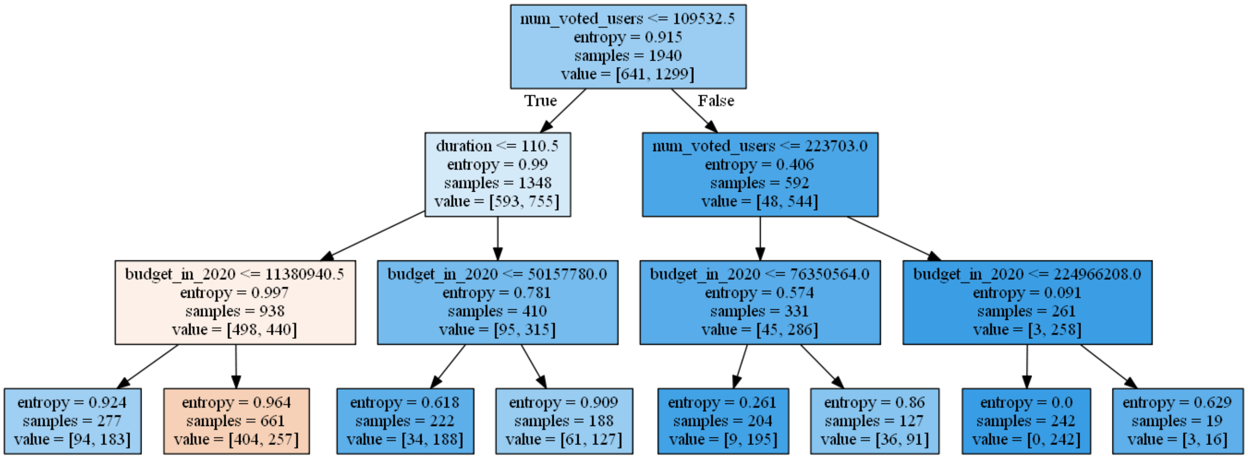
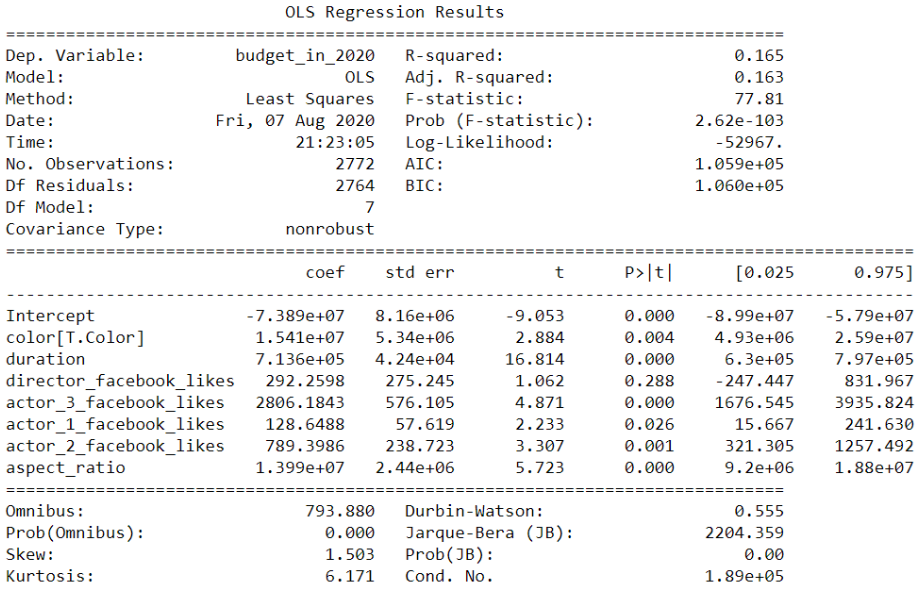


Figure 6, Decision Tree Model



```
                            OLS Regression Results
==============================================================================
Dep. Variable:         budget_in_2020   R-squared:                       0.165
Model:                            OLS   Adj. R-squared:                  0.163
Method:                 Least Squares   F-statistic:                     77.81
Date:                Fri, 07 Aug 2020   Prob (F-statistic):          2.62e-103
Time:                        21:23:05   Log-Likelihood:                -52967.
No. Observations:                2772   AIC:                         1.059e+05
Df Residuals:                    2764   BIC:                         1.060e+05
Df Model:                           7
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
Intercept              -7.389e+07   8.16e+06     -9.053      0.000   -8.99e+07   -5.79e+07
color[T.Color]          1.541e+07   5.34e+06      2.884      0.004    4.93e+06    2.59e+07
duration                7.136e+05   4.24e+04     16.814      0.000     6.3e+05    7.97e+05
director_facebook_likes  292.2598    275.245      1.062      0.288    -247.447     831.967
actor_3_facebook_likes  2806.1843    576.105      4.871      0.000    1676.545    3935.824
actor_1_facebook_likes   128.6488     57.619      2.233      0.026      15.667     241.630
actor_2_facebook_likes   789.3986    238.723      3.307      0.001     321.305    1257.492
aspect_ratio            1.399e+07   2.44e+06      5.723      0.000     9.2e+06    1.88e+07
==============================================================================
Omnibus:                      793.880   Durbin-Watson:                   0.555
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2204.359
Skew:                           1.503   Prob(JB):                         0.00
Kurtosis:                       6.171   Cond. No.                     1.89e+05
```

Figure 7, Summary of Decision Trees

References

https://www.kaggle.com/prmohanty/pandas-movie-dataset

https://datahub.io/core/cpi-us