Abstract: The “Movie Box Office Prediction” includes different factors that influence the movie revenue at the Box Office. Some of the factors include Budget, Genres, Spoken languages, Cast, Crew. In this paper, various plots are made in order to understand and observe the relations between the variables and the amount of effect of factors on the Revenue. Linear Regression, Random Forest and XGBoost are the models used for Training and Testing the Data.

Keywords: Linear Regression, Random Forest, XGBoost.

1. INTRODUCTION
The Global box office revenue had hit a record of $42.5 billion in 2019. According to this, Movie Industries has become one of the major industries in the world which includes a lot of Budget, Crew, and Cast. So the prediction of movie revenue makes a great deal. Predicted revenues can be used to decide on planning the production and the distribution stages. For an instance, the projected revenue will be helpful to decide the remuneration of actors, members, and also the costs to sell the copyrights. Movie success or failure can depend mainly on the Stars that are acting in the film, release day, Budget, and also the Genre of the film. For an instance, the Adventure Genre films usually get more Gross revenue compared to other Genre. These all the factors have an impact on the Movie’s revenue. So it is not easy for humans to predict the Collections. Since this era became a Computer and Data Science era it is easier for the machines to perform mathematical computations in predicting the Movie’s revenue based on the historical Databases.

2. LITERATURE SURVEY
Matt Vitelli [1] has designed a model for the movie revenue prediction by observing the relationships by the combination between the features using graphs such as actor-actor relationship graphs, actor-movie relationship graphs, and movie-movie relationship graphs.
and concluded that by combining these features and observing the relationships using the actor-actor relationship graphs and actor movie relationship graphs, they were able to design more accurate model than using the features in the Dataset alone.

Prediction of movie box office success based on the Wikipedia Activity Big Data\cite{3} used a simple model based only on few variables, but the efficiency could be enhanced using statistical computations and also more on the content related parameters\cite{4}\cite{5}, for example, the controversial measure in the article. There can be a diverse number of changes that can be made to the defined model in this paper.

The model designed in this paper does not consider the movie genre and the actor's popularity \cite{6}\cite{7}\cite{8} involved in the movies which might lead to accurate predictions if these are considered in the paper \cite{9}\cite{10}. One biggest change that can be made to this model, is that increasing the Data in the Dataset to obtain more accurate prediction \cite{11}\cite{12}.

3. DESCRIPTION OF DATA

Training dataset has 3000 Movie records and 23 variables (including revenue), Testing dataset has 4398 Movie records with 22 variables.

4. EXPLORATORY DATA ANALYSIS

Exploratory data Analysis is a process of analysing the Datasets and understanding the relations between the features in Dataset using visualizations and also remove the unnecessary features that are not influential on the dependant feature.
From the above it can be understood that the Mean of the revenue is around 6 Million. And the Mean of the budget is around 2.5 Million.
Figure 4 Training Dataset

<table>
<thead>
<tr>
<th>entry</th>
<th>production_countries</th>
<th>release_date</th>
<th>runtime</th>
<th>spoken_languages</th>
<th>status</th>
<th>title</th>
<th>Keywords</th>
<th>cast</th>
<th>crew</th>
</tr>
</thead>
<tbody>
<tr>
<td>1288</td>
<td>['Japan']</td>
<td>7/14/07</td>
<td>90.0</td>
<td>['Japanese']</td>
<td>Released</td>
<td><em>Somewhere Between Time &amp; Space: A Legend to Remember</em></td>
<td><em>Pinnacle: The Path of Dollars</em></td>
<td><em>cast</em>: 3, <em>character</em>: <em>Tory</em>, <em>credit</em>: ...</td>
<td><em>credit_id</em>: 52c444e53a32bb46e0e03a5835, ...</td>
</tr>
<tr>
<td>3029</td>
<td>['United States']</td>
<td>5/19/06</td>
<td>85.0</td>
<td>['English']</td>
<td>Released</td>
<td><em>A Titanic Beauty Spreads a Massive Wave of its</em></td>
<td><em>Attack of the 50 Foot Woman</em></td>
<td><em>cast</em>: 3, <em>character</em>: <em>Maggie</em>, <em>credit</em>: ...</td>
<td><em>credit_id</em>: 52a774e5ca20595b0d5c0695, ...</td>
</tr>
<tr>
<td>194</td>
<td>['United States']</td>
<td>5/03/07</td>
<td>100.0</td>
<td>['English']</td>
<td>Released</td>
<td><em>A Comedy About Lost Loves and Last Laughs</em></td>
<td><em>Addicted to Love</em></td>
<td><em>cast</em>: 11, <em>character</em>: <em>Maggie</em>, <em>credit</em>: ...</td>
<td><em>credit_id</em>: 52919339bb046bb47bd0763c, ...</td>
</tr>
<tr>
<td>1012</td>
<td>['Canada']</td>
<td>9/4/10</td>
<td>130.0</td>
<td>['French']</td>
<td>Released</td>
<td><em>The Search Begins at the Opening of Their Mother's</em></td>
<td><em>Incendies</em></td>
<td><em>cast</em>: 6, <em>character</em>: <em>Naomi</em>, <em>credit</em>: ...</td>
<td><em>credit_id</em>: 5257ed32ea305b52f0f4b2c5, ...</td>
</tr>
<tr>
<td>1603</td>
<td>['United States']</td>
<td>3/11/08</td>
<td>90.0</td>
<td>['English']</td>
<td>Released</td>
<td><em>It was Time in 30 thousand color</em></td>
<td><em>Inside Deep Threat</em></td>
<td><em>cast</em>: 1, <em>character</em>: <em>Nanor</em>, <em>credit</em>: ...</td>
<td><em>credit_id</em>: 52a44e626514167b32bf156d, ...</td>
</tr>
</tbody>
</table>

Figure 5 Test Dataset

Revenue vs. Budget

The scatter plot for Revenue vs Budget recapitulates that if the Budget increases then the Revenue also increases. If the Budget of the movie is high it means it might have star actors or the production costs.

In [11]: train.plot.scatter('budget','revenue')
Out[11]: <matplotlib.axes._subplots.AxesSubplot at 0x1a35d56650>

Figure 6 Revenue vs. Budget

Revenue vs. has_homepage

If the movie has a homepage then it might be a high Budget film as per the catplot the movies having homepage had more revenues than the movies without having a home page and in other cases, the home page doesn’t matter.
Since the use of homepage is an ambiguous attribute, we drop the attribute homepage else it will yield false results.

Revenue vs Collection

Collection for an instance is a movie belonging to a particular series, and since this is also an ambiguous attribute we drop collection attribute else it may yield false predictions.

Revenue vs Language

According to this Box plot, ‘en’ that is English yields high revenues. Since English is the most spoken language in the world. English is the most Profitable language for the Movie.
Genre is a major factor to be considered to predict the movie’s revenue. As the below Box plot shows the Adventure Genre has the highest revenues compared to other Genres, Secondly Animated movies. The Genre will have high correlation with the Budget and as well as the Movie’s Revenue.

Revenue vs. different Number of Genres in the Film
It can be seen from the below plot that the movies having Four Genres has more revenue than the others. Because if the movie only contains one Genre people might not like it which may result in the less revenue and even if the movie has more number of movies it may result in less revenue because people could not absorb all the genres or it may reduce the quality of the film by just focusing on the Increasing the genres.

Revenue vs. Number of Production Companies

This Number of production in some cases represents that the movie is of high budget, so the more number of companies are involved to produce the film, So the number of production companies is indirectly linked to the Budget. As per the plot obtained, we can see the three production companies are having very high revenue compared to others.

Revenue vs. Number of Production Countries

The number of production countries is an ambiguous factor and correctly could not have impact on the movie’s revenue. So we drop the production countries column from our DataFrame.

Revenue vs. Overview

Mapping overview present to 1 and nulls to 0. This variable is unnecessary variable and can be dropped, because this does not affect the movies revenue much. So we drop the Overview column from our DataFrame – Train.

Revenue vs. Cast Members

Cast members is one of the most important Factors for the Revenue prediction. The cast members include the actors in the movie. As the actors will have their own individual popularity if the number of cast members or various actors from the various language will yield much revenue. So cast members become one of the major factor. For a movie the more crew members are required than the cast members.

Revenue vs. Crew Members

Crew members is one of the most important Factors for the Revenue prediction. The crew constitutes the people who are hired by a production company, for the purpose of producing a movie which includes every person who is not visible on the screen but works for the film. In case of High Budget movies the will require more number of people as a crew. So crew members become one of the major factor. For a movie the more crew members are required than the cast members. So, the revenue gets high when the crew numbers increased respectively.

Correlation between Variables
The following shows the relation between the left variables after dropping in our DataFrame. Higher the intensity of the colour, higher the correlation between the variables.

From the below correlation plot, the Budget and Revenue has the Highest correlation and it shows that the value of the Revenue mainly depends on the Budget of the Movie. Secondly, the popularity has the second correlation between Revenue and the Popularity. And, then the runtime is the least correlation in our given variables.

**Revenue vs. Release_date**

From the below bar plot, 0 represents Sunday and 6 represents Saturday respectively. We observe that the movies released on Tuesday had got high revenue and the revenue has less as we got towards the weekend.
Revenue vs. Tagline
Since Tagline is a column given, which might be an ambiguous factor and may lead to the false results. So we drop the tagline column in order so that no false factors are remained.

5. THE FINAL DATASET
The below are the final variables left in our DataFrame after Dropping all the unnecessary variables.

![Figure 12 Train set](image12)

![Figure 13 Test set](image13)

6. TRAINING THE MODEL
Linear Regression

![Figure 14 Linear Regression](image14)

Root Mean Square value of the Linear Regression obtained for our DataSet is 2.4236

Random Forest Regression

![Figure 15 Random Forest Regression](image15)
Root Mean Square value of the Random Forest Regression obtained for our DataSet is 2.2127.

**XGBoost**

Model3: XGBoost

```python
from xgboost import XGBRegressor
xgb = XGBRegressor()
xgb.fit(X_train, y_train)
rmse = np.sqrt(mean_squared_error(y_test, xgb.predict(X_test)))
print(rmse)
```

Figure 16 XGBoost

Root Mean Square value of the Random Forest Regression obtained for our Dataset is 2.3558.

**7. TESTING**

**Linear Regression**

```python
X_test = test[col not in ['id']]
cols = test.columns
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)
pd.DataFrame({'id': test.id, 'revenue': y_pred}).to_csv('submission_linearreg.csv', index=False)
```

Figure 17 Linear Regression Prediction

From the above we obtained a CSV file containing the MovieId and the revenue. So we also wanted to include in which range the Movie Falls in. This will have more accuracy compared to that of discrete values. So we did the following code execution and obtained the ranges for each and every Movie. The ranges are as follows:

- Less than 1 Million
- 1 to 5 Million
- 5 to 10 Million
- 10 to 50 Million
- 50 to 100 Million
- Greater than 100 Million

```python
linearreg = pd.read_csv('submission_linearreg.csv')
for i in linearreg_df['id']:
    if (linearreg_df['id'][:, 'revenue'] < 100000):
        linearreg_df['id'][:, 'finalrange'] = 'Less than 1 Million'
    elif (linearreg_df['id'][:, 'revenue'] > 100000 and (linearreg_df['id'][:, 'revenue'] < 50000000)):
        linearreg_df['id'][:, 'finalrange'] = '1 to 5 Million'
    elif (linearreg_df['id'][:, 'revenue'] > 500000 and (linearreg_df['id'][:, 'revenue'] < 100000000)):
        linearreg_df['id'][:, 'finalrange'] = '5 to 10 Million'
    elif (linearreg_df['id'][:, 'revenue'] > 100000000 and (linearreg_df['id'][:, 'revenue'] < 5000000000)):
        linearreg_df['id'][:, 'finalrange'] = '10 to 50 Million'
    elif (linearreg_df['id'][:, 'revenue'] > 500000000 and (linearreg_df['id'][:, 'revenue'] < 1000000000)):
        linearreg_df['id'][:, 'finalrange'] = '50 to 100 Million'
    else:
        linearreg_df['id'][:, 'finalrange'] = 'Greater than 100 Million'
```

Figure 18 Revenue in Millions
8. RESULTS

ID 3010 - Toy Story 2
Revenue Collected - 497.4 Million

<table>
<thead>
<tr>
<th>Model</th>
<th>Predicted Revenue</th>
<th>Predicted Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Model</td>
<td>357817471.1728</td>
<td>Greater than 100 million</td>
</tr>
</tbody>
</table>
9. CONCLUSION

The film industry is an unpredictable business either the producer gain higher profits or they might get into huge loss, so it is difficult for a human to predict that the movie box-office prior to the release. Although it is very important for production studios to be able to predict the movie box office revenues before they are released, the prediction of box office revenue is still classified as an art rather than a science because most experts predict revenue based on their own rules of thumb, hunches, and their experience. This project helps the production studios to predict box office revenues that can be used to decide for planning the production and the movie distribution stages.

This process gave outputs with less Root mean square Error for Random Forest is less when compared to others and Random forest gave better results than the Linear Regression and the XGBoost Algorithms.

10. REFERENCES


