**Executive Summary**

**TMDB Box Office Prediction:**

Group 3 — Krupa Chand Appikonda, Tagore Srinivas Pothuneedi, Rakesh Bommisetty, Sai Kumar Thallad, Swetha Gondi.

There are many factors that affect a movies performance at the box office, in this study we try to analyse the factors that effect a movies revenue. The data consists information about 7000 past films, data points provided include cast, crew, plot keywords, budget, posters, release dates, languages, production companies, and countries. IMDB is the source of this dataset.

**Exploratory Data Analysis:**

We try to focus which factors affect the revenue of a film.

* Film that have a homepage tend to perform better that those films that do not have a homepage.
* Films that belong to movie collection perform better than the rest, this may be because of the popularity of the movie collection.
* Movie genres that are popular are; Action, Comedy, Horror, Drama, Adventure, Thriller, Animation, Crime, Fantasy, Animation.
* Movies that belong to two or three genres perform better than movies those that belong to a single genre.
* Different production companies may be in involved for one film and films that have limited production companies (2 to 3) perform well, as the number of production companies grow the performance of the movies goes down.
* Movies that are made in English and Chinese perform well at the box office.
* Movies that are released on the 3rd day of the week perform very well compared to the movies that are released on other days of the week.
* Movies that have a tag line to the original title collect more revenue.

**Machine Learning Analysis:**

Clustering: With the help of clustering we try to find categorize them depending on the similarities, Based on the clustering analysis the movies can be categorized into 3 clusters, preferably movies that performed well, movies that did not perform well and movies that performed averagely in the box office.

Regression: With the help of regression analysis we try to find the most important factors that effect a movies revenue. Models such as Linear regression, Decision tree regressor, Random forest regressor are used for this analysis. Budget, orginal\_language, popularity, and release\_date are the most important factors that influence the revenue of a film.

Prediction: We try to find the best model to predict a movies revenue, Linear regression, KNN regressor, boosting regressor are used for this analysis. RMSE is used as the evaluation metric for this analysis. Boosting regressor gives the least error and should be considered for prediction.

**TMDB Box Office Predictions**

Krupa Chand Appikonda Tagore Srinivas Rakesh Bommisetty

Student id:801166358 Student id: 801164787 Student id: 801165740

kappikon[@uncc.edu](mailto:@uncc.edu) [tpothune@uncc.edu](mailto:tpothune@uncc.edu%20)  [rbommise@uncc.edu](mailto:rbommise@uncc.edu)

Swetha Gondi Sai Kumar Thallada

Student id: 801103431 Student id: 801154715

sgondi@uncc.edu sthallad@uncc.edu

**ABSTRACT:**

Movies are an important part of entertainment in today’s world, over the years the movie industry has grown into a multi-dollar industry. Many factors that affect a movie's success when the amount put for a movie to be produced is very high, the success of the movie is very important. Hence the moviemakers need to know the factors that affect the success of a movie. The data available on movies is massive.

In this study, we tried to learn the factors that affect a movie's performance. This study provides investors with the right information to avoid risk. We compare the movie success by its revenue collected, we implement various machine learning algorithms and find out the factors that affect a movie's revenue and provide an insight into the data using visualizations.

**Keywords:**

Data Visualization, Prediction, Clustering, Regression

**1.INTRODUCTION**

These days movies are not the only source of entertainment but also one of the major sources of income. Movies create a special craze among, especially young people. The popularity of movies is concerned not only with film distributors and box office officials but also with common people. Users in social media used to speak about these. Data analysts show a special interest in the data about the movies available through social media content. Apart from this actor’s popularity and directors, previous box office history will also come forefront. Data available from the various reviews given by the moviegoers in platforms like IMDB, Rotten Tomatoes, etc. and it’s not necessary that people from various regions should like the content. This data available online about reviews and movie ratings online are analyzed in this paper is used for prediction of the box office success statistics.

There is a very limited research study done on predicting the success rate using the movie's characteristics such as the history of the director's success, reviews, genre, region and other factors, etc. when compared to the research done based on the people's excitement to watch the movie. Being a very popular online review and rating platform, the data available is used to predict the movie ratings and success stats of the movie in IMDB through this research model.The success of the movie is relative depending upon research and study because some research studies might consider online popularity, the hype of the movie or based on the critic review and ratings as their success factor and classify the data accordingly. In this case study, our definition of success of a movie is considered as revenue collected by the movie. In this research, we considered three hypotheses that will let us find out on which factor does the movie success depends upon?

In this research study, we had used six methods which are K-Means clustering, Multi Linear Regression, Decision Tree Regressor, Random Forest Regressor, K-Nearest Neighbors and Ensemble bagging. We had used K-means clustering to find the similarities in the data and group them accordingly. All the regression modelling techniques like Multi Linear Regression, Decision Tree Regressor and Random forest regressor are used to know the feature importance which will help us know which features should be chosen to better implement in the prediction models. In these regressors for predicting the feature subset random forest regressor has the least error compared to other modelling techniques. The KNN and Ensemble bagging methods are powerful machine learning algorithms that are used for prediction. And the least error was found for ensemble bagging and work better than others.

In the next section, we mention the business use-cases. In section 3 we mentioned our chosen hypothesis for the research. In the following section 4, we had mentioned about the dataset. In section 5 we had briefly given the introduction to the prediction methods used in this paper. In section 6 we had predicted and provided an analysis of how the predictions are done. In section 7 we had given the recommendation based on the analysis done. Finally, we are concluding with conclusions, references and appendix

**2. BUSINESS USE-CASE**

This dataset has been taken from Kaggle. In this study, we try to find the factors affecting a movie's performance in the box office. Movies are one of the most enjoyable and revenue-generating platforms in the entertainment industry. So not only the big studios like Marvel, Warner Bros., etc but also the small studios will look for the statistics on their movie productions which will increase there movie success rate at the box office. Secondly, the Movie actors can also look for these statistics to know the current trends and entertain there audience accordingly.

**3. HYPOTHESIS QUESTIONS**

1. Identify if the movies can be grouped depending on any similarity.
2. Identify the most important factors that affect revenue.
3. To predict the revenue of the film based on factors such as budget, language, cast, crew, etc.

**4. DATASET**

The data set used in this project is TMDB Box Office Prediction taken from Kaggle. The data contains Train and Test data, attribute Revenue is not present in Test data, Apart from that all the attributes are the same.

The data set contains different details of a movie such as a name, budget, runtime, language, cast, crew, etc. They are a total of 7398 rows and 23 columns. Table 1 shows the description of all the features in the data set. All the columns except popularity and revenue give information about a movie before its release. This information can be very vital as we aim is to know the factors that affect the revenue and also to predict the revenue based on those important factors. Columns such as belongs\_to\_collection, homepage, poster\_path have a lot of null values. Most of the data is represented in JSON format for columns such as genres, cast, crew, etc. Columns such as Id, IMDB Id, title, tag\_line, over\_view are unique for every data point and such variables might not give any useful information during model building, some of these columns may provide a certain degree of information when they are converted into binary, this is discussed further in data pre-processing.

Table1: Feature Description

|  |  |  |
| --- | --- | --- |
| **Features** | **Types** | **Description** |
| id | Integer | Id of the movie in the data |
| Belongs to collection | JSON | the collection that the movie belongs to. |
| budget | Integer | amount spent on making the movie. |
| genres | JSON | genre the movie belongs to a movie that can have multiple genres. |
| homepage | Weblink | link to the movie's home page |
| imdb\_id | Integer | IMDB id of the film |
| Original language | Categorical | The original language in which the film was made. |
| Original title | Categorical | The original name was given to the movie. |
| overview | Categorical | overview of the film. |
| popularity | Integer | popularity of the film. |
| Poster path | JSON | link to the Movie Poster. |
| Production companies | JSON | names of the movie production company, there can be multiple production companies. |
| Production countries | JSON | names of the movies production countries, there can be multiple production countries. |
| release\_date | Date | movie’s release date. |
| runtime | Integer | movie run time information in minutes. |
| Spoken languages | JSON | languages spoken in the film, which can be more than one. |
| status | Categorical | release date of the movie |
| tagline | Categorical | Tagline of the movie. |
| title | Categorical | title of the movie. |
| Keywords | Categorical | keyword is given in the overview |
| cast | JSON | information of cast |
| crew | JSON | information of crew |
| revenue | Integer | information about the money collected. |

**4.1 Pre-Processing of Data:**

Information and raw data that is available and collected from different sources may be in different formats and structures. So the data should be cleaned and processed before the data is given to the model. Data preprocessing refers to a set of activities that are done to make raw data suitable for further processing in the context of web mining. As reported by Wei et. All data preprocessing is one of the first and critical steps to data mining and data analysis, the result of data preprocessing is directly inputted to the mining model and obtained the final result.

In the source dataset, some features are in JSON format like Genre, Cast, Crew, production companies, production countries, spoken language. These had been converted to there respective counts because converting these types of data into there counts is more intuitive and be understood easily rather than converting each category of a particular attribute into a separate column which will increase the number of dimensions in the data.

The categorical features like Tagline, Hompage contains a lot of null values which should not be present while data modelling. The naïve implementation is to drop such kind of data but this may lead to loss of important data to avoid this we had converted this variable to binary. So this will help the model and avoid loss of data.

The feature original language consists of a lot of categories like English, Chinese, Hindi, etc. But in the Exploratory data analysis, we observed that only English and Chinese languages have provided a major contribution to the revenue. So in our study, we had only converted all the English and Chinese original language attributes to 1 and all remaining as Zero. Thus making Original language to binary.

**5. DEMONSTRATING PROCEDURES**

In this study, we had used several regression models for the prediction modelling of our dataset. The models we used are K-Means clustering, Multi Linear Regression, Decision Tree Regressor, Random Forest Regressor, K-Nearest Neighbors and Ensemble bagging. Below is a brief description of the modelling techniques used in our study.

K-Means Clustering:

Clustering is one of the most used unsupervised machine learning algorithms. K-means Clustering is the unsupervised algorithm that doesn’t depend upon the outcome or the criterion value.

K-Means algorithm will groups the data depending on the centroids(k) initially taken. In other words, we choose k number of centroids and calculate nearest points to each centroid forming a cluster and parallelly keep the cluster as tight as possible [1].

Multi Linear Regression:

Multi Linear Regression is also called Multi Regression. This regression method is a fundamental technique used when the data is continuous. The only difference between Linear regression and Multilinear regression is that Multi regression has to examine the relationship between two or more variables [2]. When the relationship is directly proportional the dependent variable will increase with independent variables and vice versa.

Decision Tree Regressor:

A decision tree can be used for decision making and to visually and explicitly represent decisions. It uses a tree-like model to represent the decisions. A decision tree is drawn upside down with its root pointed at the top. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed [3].

Random Forest Regressor:

Random Forest is another famous machine learning algorithm and is a very powerful predictive modelling algorithm [4]. This method will provide a very reliable output because it takes the number of different weak learners and gives the best features from all the inputs. Some times this algorithm will identify the most complex dependencies which no other method can find but at a cost of more computation time.

K-Nearest Neighbors Regressor:

KNN is a supervised algorithm which is simple but effective in many cases. This algorithm takes a bunch of labelled points and uses them to learn how to label other points. If a new point arrives, it looks at the labelled points closest to that new point

So whichever label most of its neighbors have, becomes the label for the new point[5]. However, to apply KNN we need to choose an appropriate value of k, and the success of classification is very much dependent on this value. There are many ways of choosing the value of k, but a simple one is to run the algorithm many times with different k values and choose the one with the best performance[6].

Ensemble Bagging:

Bagging is one of the powerful method in the ensemble and this method takes several weak learners and average there output and build a model based on the inputs given by the weak learners. This method is very complex and difficult to interpret. Generally, this method of bagging would give the best possible model if the data is huge.

**6. BUILDING MODELS**

In this section, various models are built to analyze the hypothesis questions.

**HYPOTHESIS 1:**

Identify if the movies can be grouped depending on any similarity

**Model 1: Clustering**

Clustering is a technique in which the data points are grouped into different clusters based on the similarity between the points, In this study, we use K-Means clustering which uses mean as a mode to check the similarity.

We run the K-Means Clustering on the whole data set by selecting the input variables as budget popularity, runtime, cast\_count, crew\_count, num\_prod\_countries, release\_day, num\_genres, spoken\_count, isTaglineNA, has\_homepage, original\_language, num\_prod\_companies and revenue.

We take K ( no of clusters) as 5 and 30 iterations for this model, Value of K is generally selected using elbow plot, for this case study we use 5 as the K value and then group the movies based on the distances between the clusters.

Below are the results for the K-Means clustering model.

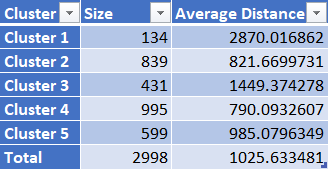


Figure 1. Average cluster distance(K=5)

Here we can see that a majority of records are grouped into cluster 2 and cluster 4 and also the average distance of the two clusters if very close.

Below are the inter-clusters distances

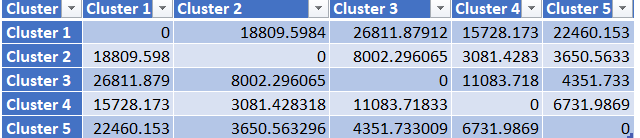
****

Figure 2. Intercluster distance(K=5)

From Fig. 2, we could derive that cluster 2 and cluster 4 are not far from each other and thus they can be similarities between the two clusters and after cluster 2 and 4; cluster 3 and cluster 5 appear to be close based on the distance.

We run K-Means Clustering again on the whole data, we take K as 3 and 50 iterations for this model.

Below are the results for K-Means clustering when K=3

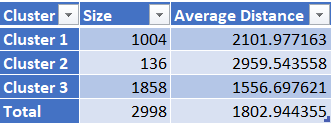


Figure 3. Average cluster distance(K=3)

From Fig. 3 we observe that cluster 3 is tighter than the remaining two and cluster 2 is very similar to cluster 1 when K was equal to 5.

Below are the inter-clusters distances

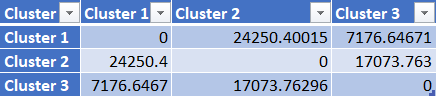


Figure 4. Intercluster distance(K=3)

Taking 3 clusters might be optimal for this data set and that may divide the movies into 3 categories; movies that performed well, movies that did not perform well and movies that performed averagely in the box office.

**Advantages with Clustering**

1. Simple to implement, even if the data set is large, it can easily scale up to the requirement.
2. Clustering guarantees Convergence and can also easily adapt to new examples as the model uses mean as the distance to categorize data.

**Dis-advantages with Clustering**

1. A lot depends on the K value and it is difficult to predict the value of K manually.
2. Clusters vary based on the initial centroid value taken, the fixed initial value may not be optimal every time.
3. If the no of dimensions are high (Variables) then dimensionality reduction is required and clustering is not optimal in such cases

**HYPOTHESIS 2:**

Identify the most important factors that affect revenue.

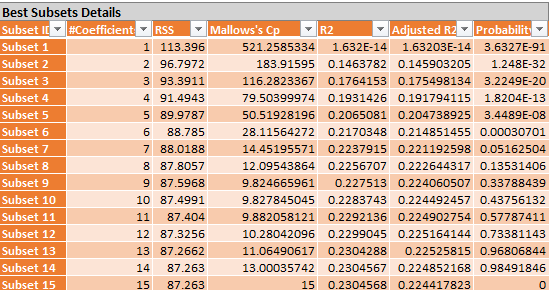
We use Multi Linear Regressor, Decision Tree Regressor and Ensemble-Random Forest Regressor to find out the factors that affect the revenue the most.

**Model 1: Multi Linear Regression**

Multi Linear Regression is used to explain the relationship between a dependent variable and many independent variables – generally two or more. Multi Linear Regression gives different coefficients to variables and we use those coefficients to extract the feature importance.

We apply the model on the data, which is portioned into Training and Validation, they contain 60% and 40% of the data respectively. Attribute Revenue is the target variable and remaining data is given as input except for the column Id.

In our study, we used the best subset feature selection for the model.

Figure 5. Feature selecting in MultiLinear Regression

The best subset is chosen depending on the Cp and Adjusted R2 value, In this case, Subset 15 is the best.

Below are the coefficients are given by the model.

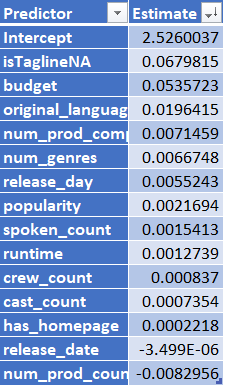


Figure 6. Multilinear Coefficients

As per the model variables budget, orginal\_language and is\_tagline\_na are affecting the revenue.

Below are the metrics for that subset.

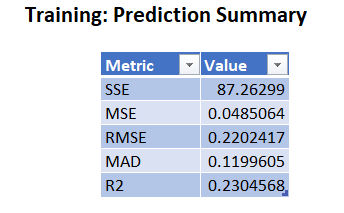
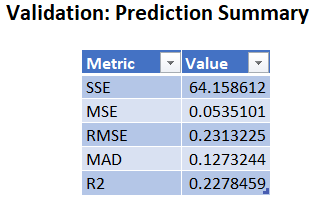
 '

Figure 7. Prediction Summary

**Advantages of Multi Linear Regression:**

1. It can determine the relative influence of different predictor variables to the criterion value.
2. It can easily identify the outliers present in the data.

**Limitations of Regression Tree:**

1. In multilinear regression, there is a linear relationship between both dependent and independent variables.

2. It assumes no major correlation between the independent variables.

3. It is normally distributed in all nodes which makes regression complex in terms of relating to another independent variable.

**Model 2 – Decision Tree Regressor**

Decision Tree Regressor uses Standard Deviation as a metric to find the best attribute which is homogeneous and further follows the same process to divide all the attributes.

In our study, we have seen a fully grown tree with a limit of 150 values at each node split.

Below is the decision tree.

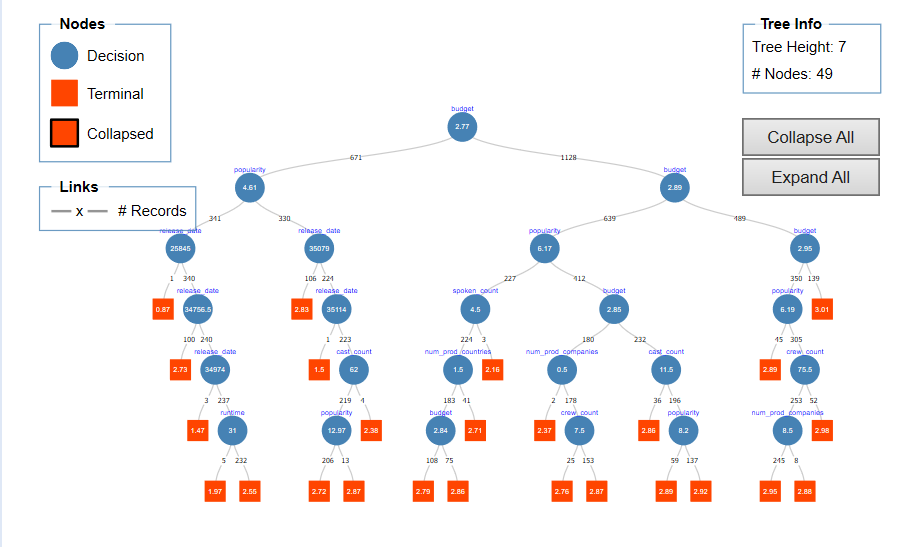


Figure 8. Decision Tree

The most important variable as per the model is budget, followed orginal\_language, popularity and release\_date.

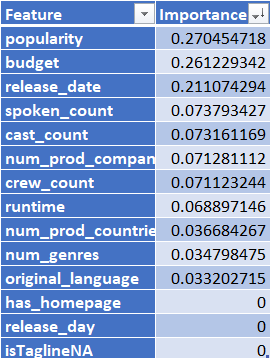


Figure 9. Feature Importance

Below are the metrics for the decision tree

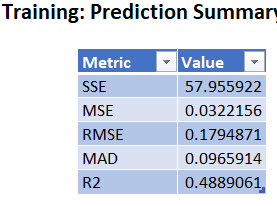
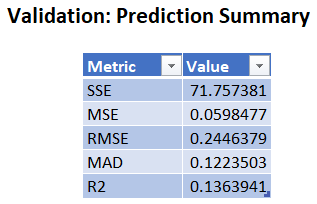
 

Figure 10. Prediction Summary for training and validation

The Model gives an R2 value of 0.488 on training data and an R2 value of 0.136 on validation data.

**Advantages of Decision Tree:**

1. Less data pre-processing steps required when compared to other algorithms
2. There is no requirement to Normalize or scale the data.
3. Missing values do not affect the data and also the trees are very intuitive.

**Dis-Advantages of Decision Tree:**

1. A small change in data can change the whole tree structure.
2. In some cases, the calculations may be complex.
3. Not generally preferred while doing regression.

**Model 3 – Random Forest Regressor:**

Random Forest Regressor is an ensemble method that uses weak learners such as decision trees to form a strong learner. It is a very powerful algorithm as different decision trees are used to get the required output.

We apply the model to the portioned data and observe that budget, orginal\_language, popularity and release\_date are the most important variables that affect the movie revenue as per the model

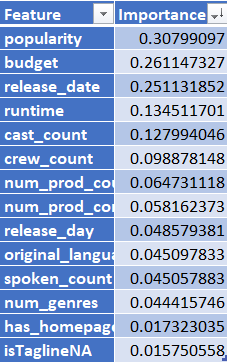


Figure 11. Feature Importance

Below are the metrics for the Random Forest Regressor.

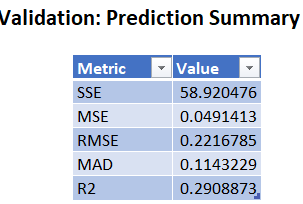
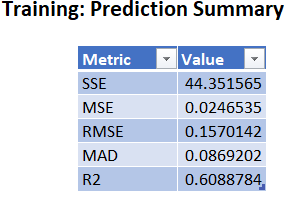


Figure 12. Prediction Summary for training and validation

**Advantages of Random Forest:**

1. Random Forest has one of the best predictive performance when compared to the best-supervised algorithms.
2. The random forest provides us a reliable feature importance estimate

**Dis-advantages of Random Forest:**

1. They are generally difficult to interpret when compared to the decision tree.
2. High computational costs and are much slower compared to other algorithms.
3. Each entity in a random forest splits out as class predictions and the class with the most correlation becomes our model prediction.

**Hypothesis 2 - Best Model:**

Multi Linear regressor, Decision tree regressor and Random Forest regressor are the 3 models used to find the factors that influence the revenue. Among all the three models Random Forest Regressor has the least error and so we consider Random Forest as the best model to identify the factors that influence the revenue of a film, as per Random Forest regressor budget, orginal\_language, popularity and release\_date are the most important factors that influence the revenue of a film.

**HYPOTHESIS 3:**

To predict the revenue of the film based on factors such as budget, language, cast, crew, etc.

To predict the revenue of the film we use regression models such as Decision Tree Regressor, Multi Linear Regressor, K-nearest neighbors and Ensemble-Bagging.

**Model 1 -** **Decision Regression Tree**

Decision Tree Regressor uses Standard Deviation as a metric to find the best attribute which is homogeneous and further follows the same process to divide all the attributes.

We divide the data into Training, Test and validation, they contain 60%,10% and 30% respectively. Attribute revenue is considered as the target variable and variables such as is\_tag\_line\_na and has\_homepage is neglected as could see that they do not provide any vital information to the model.

The model considers budget, popularity, release\_date, and run\_time as important features.

Below is the decision tree of the model built.

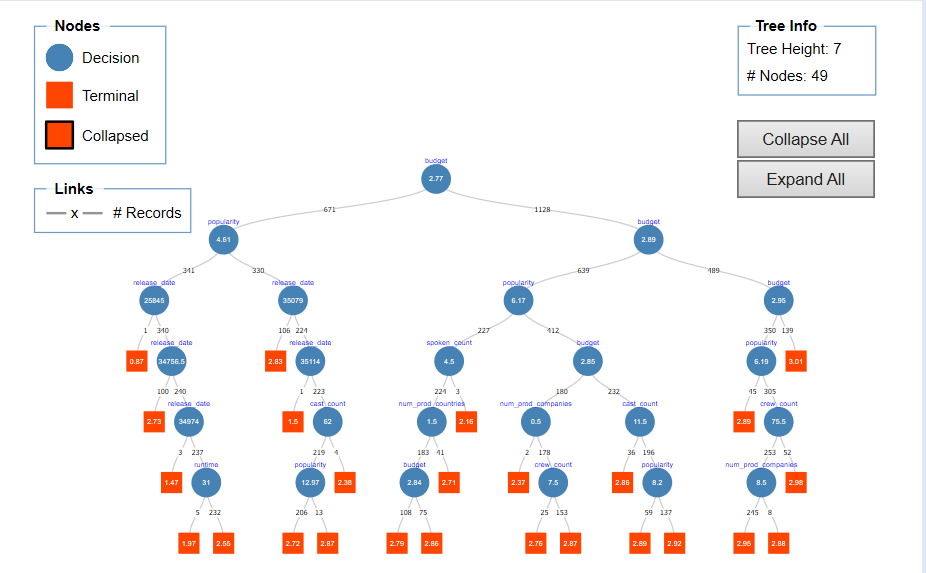
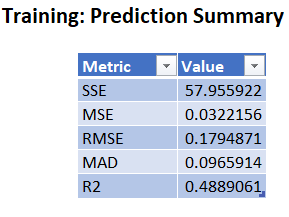


Figure13. Decision Tree

**Metrics for Regression Tree:**

****

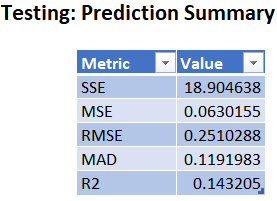
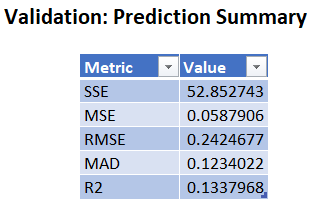
****

Figure 14. Prediction Summary for training, validation and testing.

The model is performing well as it is giving a low error value and also the RMSE value for validation and test is close. Also, we see that the model does not overfit or underfit the data.

**Model 2 - Multiple Linear Regression:**

Multi Linear Regression is used to explain the relationship between a dependent variable and many independent variables – generally two or more

The model is applied to the portioned data and we use the best subset for feature selection.

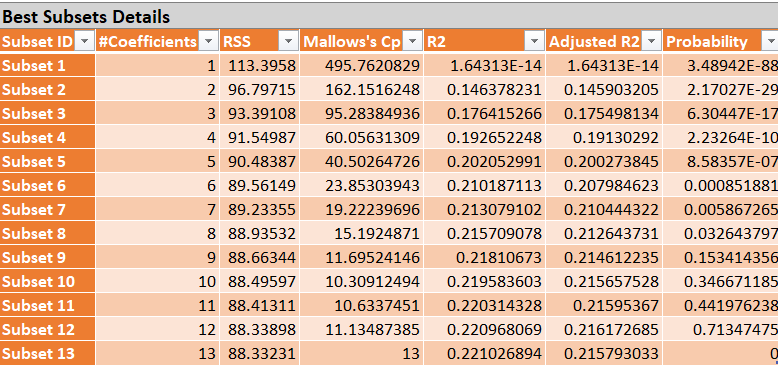


Figure 15. Subsets of feature selection

Subset 12 is chosen as the best predictor depending on the R2 and Cp value and this model gives the following output.

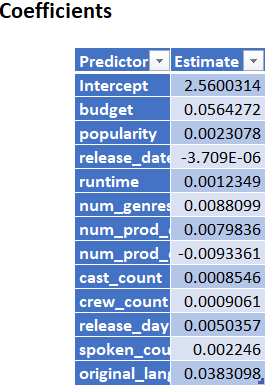


Figure 16.Coefficients of multilinear regressor

Error metrics for the data are as follows

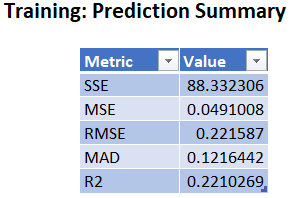
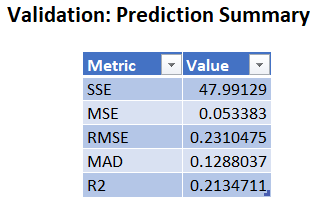
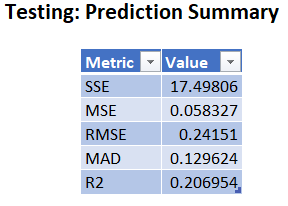
  

Figure 17. Prediction summary

**Model 3 -KNN**

K-Nearest Neighbor calculates the avg distance of N neighbors and then tries to measure the similarity between a data point and the target variable.

We apply the KNN model on the portioned data and give revenue as the target variable. For this study, we search the optimal K value between 1 and 15

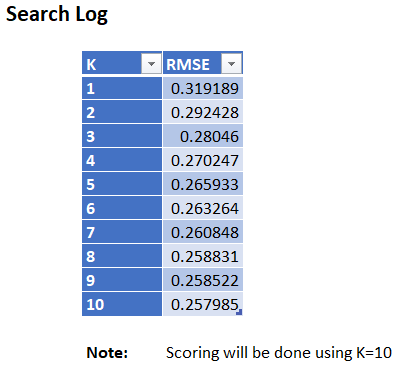
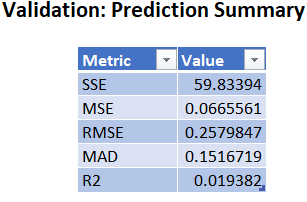
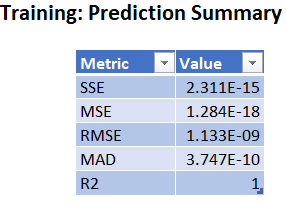
.

Figure 18. Optimal value of K

The model is built using 10 nearest neighbors.

Error metrics for the data are as follows



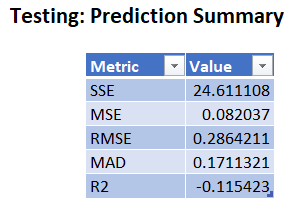


Figure 19. Prediction summary

From the above metric, we see that the KNN model overfits the training data as the error is very low for the training dataset. This model is not good for predictions.

**Advantages of KNN:**

1. KNN algorithm is a Lazy learner, does not learn during training on the data instead learns when while making predictions
2. KNN is easy to implement.
3. New data can be added seamlessly.

**Dis-advantaged of KNN:**

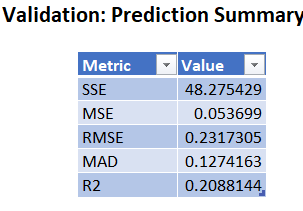
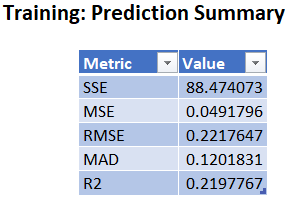
1. KNN does not work well with huge datasets.
2. If a data set a high number of dimensions, then KNN is not an optimal solution.
3. Sensitive to outliers and noise present in the data

**Model 4 -Ensemble Bagging**

Bagging is a process where we create the number of models for the same data by changing the input variables and rows for each model. It also uses weak learners to create a strong learner.

For this study, we use linear regression as the weak learner and we use 10 models to create the ensemble.

Below are the error metrics.



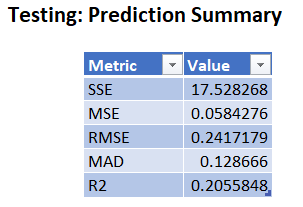


Figure 20. Prediction summary for Ensemble bagging

Ensemble Bagging does not overfit or underfit the data and also has a less RMSE score. This model is good for predicting movie revenue.

**Advantages of Bagging:**

1. As it uses multiple weak learners, it has less variance and avoids overfitting

**Dis-advantages of Bagging:**

1. The model is tough to interpret.

**Best-Model - Hypothesis 3:**

The prediction models used in this study are Decision tree regressor, Multi Linear Regressor, KNN Regressor, and Bagging Regressor. Ensemble Bagging can be chosen as the best model to predict, as it gives the lowest error rate on all the three partitions of the data. And also, being a powerful ensemble method using 10 linear regression models input reduces variance and avoids overfitting. Performance of the model is also evaluated using ROC curves, refer section 9.2

**7. STRATEGIC RECOMMENDATION**

Analysis from the model had revealed that the TMDB data set that is used in this study takes a count of genres present in the film, But the film's revenue might get effected depending on the production company’s fame. Taking the count of the data will give us only a part of the information and might not be suitable for every case.

Secondly, the analysis had shown that movies which in English and Chinese are doing good and collecting more revenue when they are released on the third day of the week.

As the budget of the movie increases the revenue generated from those movies is also increasing proportionally.so the production can use these recommendations which will have a significant impact on the movie revenue they generate.

**8. CONCLUSION**

Based on the Exploratory data analysis done on the data using python libraries and based on the machine learning analysis done on XL Miner we conclude the above hypothesis as below:

1. Based on the clustering analysis the movies can be categorized into 3 clusters, preferably movies that performed well, movies that did not perform well and movies that performed averagely in the box office
2. Budget, orginal\_language, popularity, and release\_date are the most important factors that influence the revenue of a film.
3. To predict the revenue of a film with the given data we use Bagging Regressor as it has shown to give the least error and also being an ensemble method, it is more bankable.

**9. APPENDIX**

**9.1 MetaData:**

**id**: Id of the movie in the data

**belongs\_to\_collection**: The name of the is represented collection that the movie belongs to.

**budget**: This column gives the amount spent on making the movie.

**genres**: Depicts which genre the movie belongs to. One movie can have multiple genres.

**homepage:** Weblink to the movie's home page.

**imdb\_id:** IMDBid of the film.

**Original\_language:** The original language in which the film was made.

**Original\_title:** This column contains the original name given to the movie.

**overview:** This column gives a brief overview of the film.

**popularity:** This column gives the popularity of the film.

**Poster\_path:** This column has a web link to the Movie Poster.

**Production\_companies:** This column has the id’s and names of the movies production company, there can be multiple production companies

**production\_countries:** This column has the id’s and names of the movies production countries, there can be multiple production countries.

**Release\_date:** This column gives the movie’s release date.

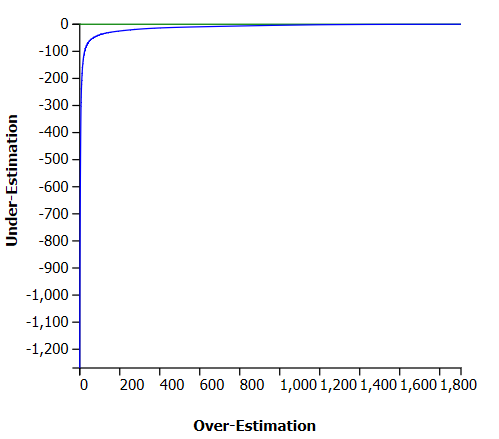
**Run\_time:**  This column gives the movie run time in minutes.

**Spoken\_languages:**  These columns have the languages spoken in the film, which can be more than one.

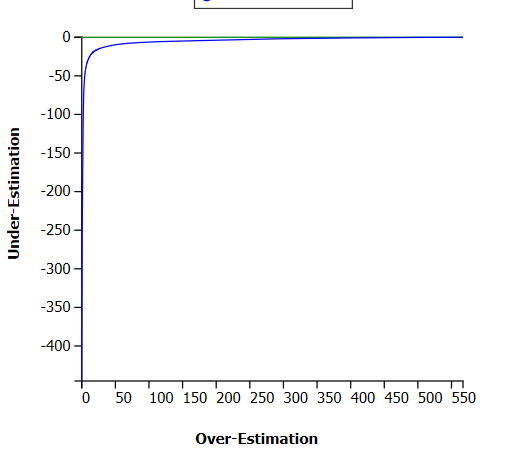
* 1. **ROC curves for prediction models:**

1. **Linear Regression ROC curve**

**Validation ROC Curve:**

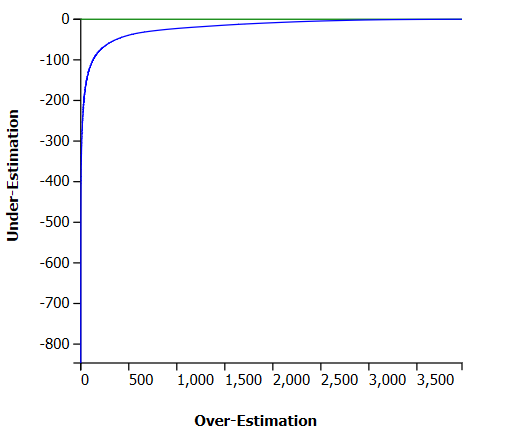


**Test ROC Curve:**

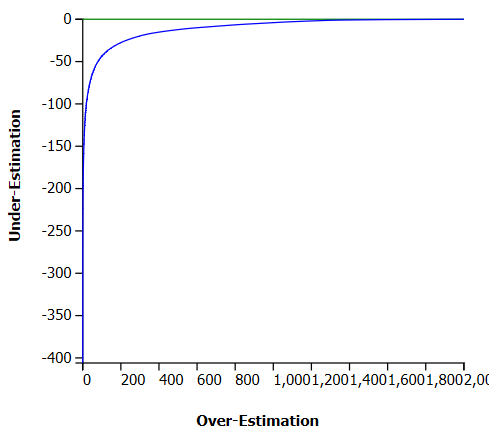


1. **Decision Tree ROC curve**

**ROC Curve - Testing**

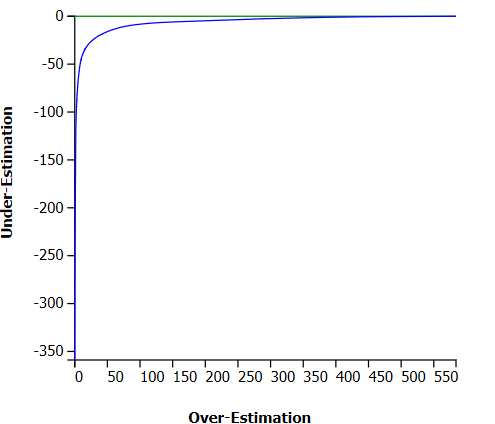


**ROC Curve - Validation:**

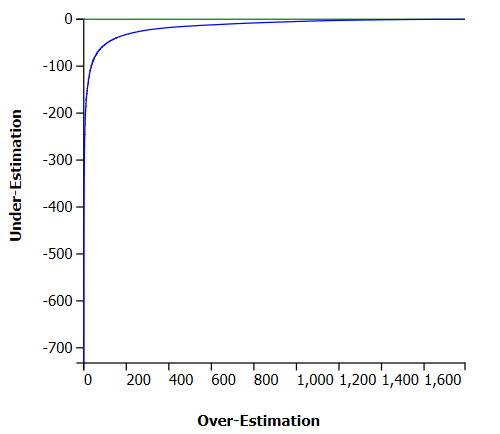


1. **KNN ROC Curve**

**ROC Curve - Testing**

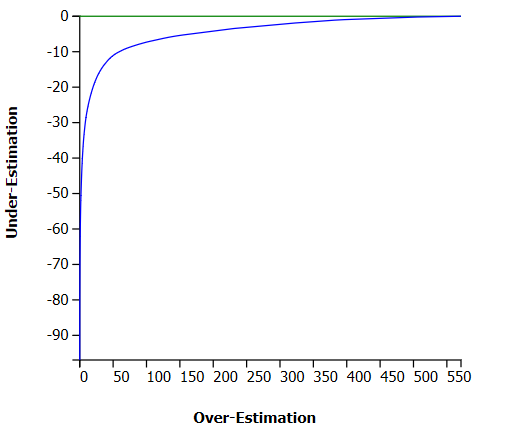


**ROC Curve - Validation:**

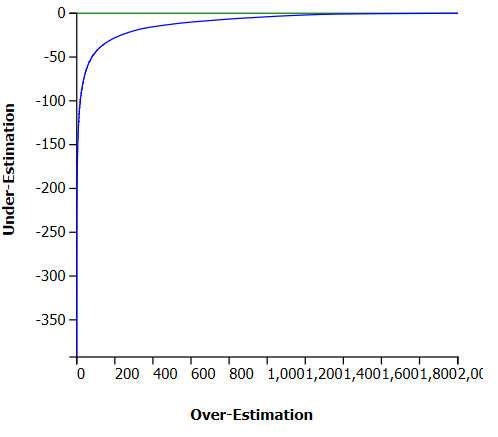


1. **Ensemble Bagging Regressor ROC curve**

**ROC Curve - Testing**



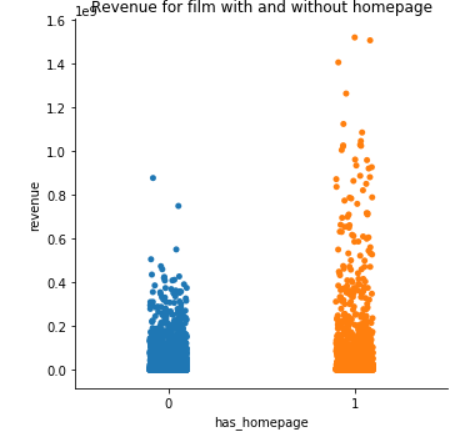
**ROC Curve - Validation:**



**9.3 References:**

**9.4 Visualizations:**

**i. Revenue Vs has\_homepage:**

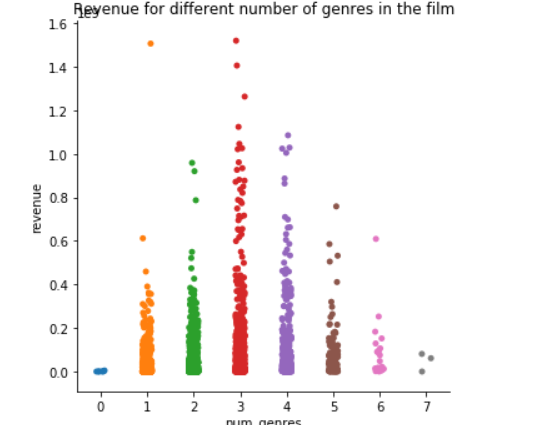
****

From the above visualization, we can conclude films having an home\_page tend to get higher revenue.

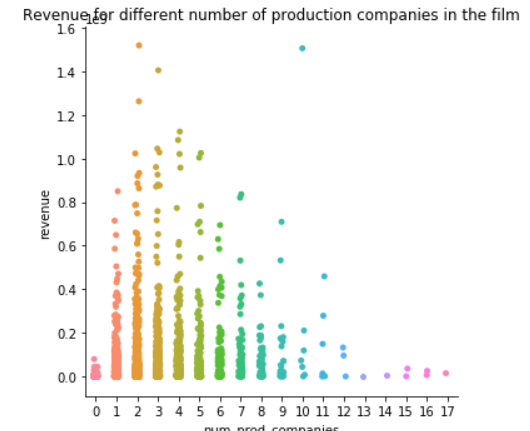
**ii. Revenue vs no\_of\_genres**

The below visualization shows how a movie with multiple genres performs at the box office.

Films having # or 4 genres tend to perform better than the rest

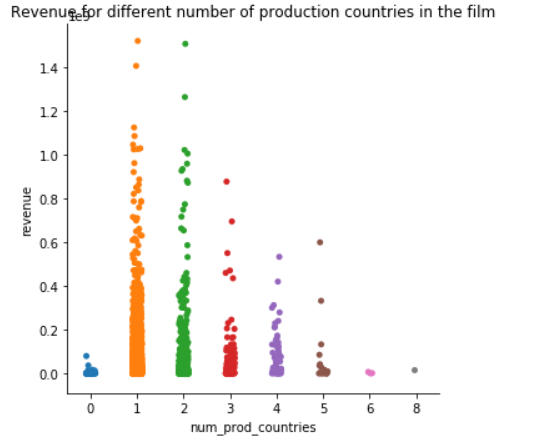
****

**iii. Revenue vs no\_prod\_companies**

****

We can see that movies having 2 or 3 production companies perform better than the other, as the number of companies increases the revenue goes down

**iv. Revenue vs no\_prod\_countries:**

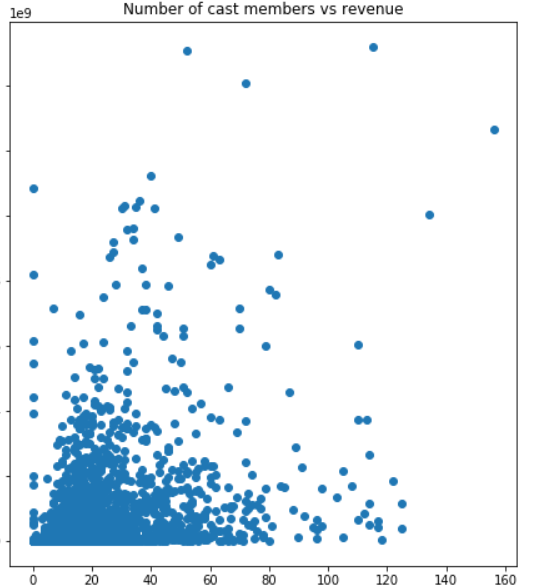
****

AS the number of production countries increases we observe that the revenue decreases.

**v. Revenue Vs Crew**

****

**vi. Revenue Vs Cast**

**:**

**REFERENCES**

**[1]** <https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1>

**[2]** [https://towardsdatascience.com/simple-and multiple-linear-regression-in-python-c928425168f9](https://towardsdatascience.com/simple-and%20multiple-linear-regression-in-python-c928425168f9)

**[3]** Patel, Neha and Singh, Divakar “An Algorithm to Construct Decision Tree for Machine Learning based on Similarity Factor,” International Journal of Computer Applications,2015.

**[4]** M. T. Lash and K. Zhao, “Early Predictions of Movie Success: The Who, What, and When of Profitability,” Journal of Management Information Systems, vol. 33, no. 3, pp. 874–903, Feb. 2016.

**[5]** <https://medium.com/@Mandysidana/machine-learning-types-of-classification-9497bd4f2e14>

**[6]** KNN Model-Based Approach in Classification[Gongde Guo1 , Hui Wang 1 , David Bell 2 , Yaxin Bi 2 , and Kieran Greer 1]