**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. They are 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. These big 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.

**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. We analyze the data and build a model to predict movie revenue.

**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.

**TMDB Box Office Prediction:**

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. The columns are as follows

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

**belongs\_to\_collection**: The name of the 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.

**4.1 Pre-Processing of Data:**

To apply models to the dataset, the data must be in a format that is acceptable by the model. We have data in the JSON format for some columns. Below are the data pre-processing steps are taken for each column.

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

**belongs\_to\_collection**: Data is present only for a few movies, movies that have a null value are considered as an individual movie and don’t belong to any collection. This column is converted to binary, 1-if the movie belongs to a collection, 0-if the movie does not belong to a collection

**budget**: No changes

**genres**: Data is given in JSON format as name of the genre and genre id, we calculate the number of genres for each movie and add a different column to the data and drop the original one

**homepage:** We don’t use this column as it has no use for regression.

**imdb\_id:** As the id is unique for every movie, we don’t use the column, hence we drop it.

**Original\_language:** We compare how each language is contributing towards the revenue and we observe that languages ‘en’ and ‘zh’ contribute a lot towards revenue and hence we take 1-en and 0-zh.

**Original\_title:** This is the name of the film as it is unique to every movie and hence, we drop it.

**overview:** We convert the variable to binary if the movie has an overview if label as 1 and if there is no overview we label as 0.

**popularity:** No changes

**Poster\_path:** The poster path is unique for every column and hence we drop it

**Production\_companies:** Data is in JSON format as id and name, we calculate the number of production companies for each movie and add to a new column and delete the original column

**production\_countries:** Data is in JSON format as id and name, we calculate the number of production companies for each movie and add to a new column and delete the original column

**Release\_date:** Data type converted to a date.

**Run\_time:**  No changes.

**Spoken\_languages:**  This column has the languages spoken in the film, which can be more than one.

**Release\_day:** Extracted day of the week from release\_date

**5. DEMONSTRATING PROCEDURES**

Decision tree learning is a predictive modeling approach used in biodata technologies and machine learning. The decision tree modeling can be explained in two entities, namely decision node and leaves node.

Leaves nodes are the decisions, or the outcomes and Decision nodes are where the data is stored as metadata and its application outcome to be split. There are two main types of

**Decision trees:**

**1. Classification Trees:** Classification is the process of predicting the class of given data points. Classes are sometimes called targets or categories.

**2. Regression Trees:** In regression trees, all regression techniques contain a single output variable and one or more input variables. The output is often numerical. The regression tree building methodology allows the input variable to be contiguous.

**Limitations of Multilinear regression:**

1. In multilinear regression, there is a linear relationship between both the 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.

4. It is too highly correlated.

5. Its analysis is difficult to fit in a single line through scatter plots.

6. To forecast the effects and impacts of changes it needs previous dependent variables.

**Limitations of Random Forest:**

1. Consists of many individual decision trees that operate as an ensemble.

2. Each entity in a random forest splits out as class predictions and the class with the most correlation becomes our model prediction.

3. There needs to be some actual signal in the RF feature so that the models built using those features do better than random guessing.

**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

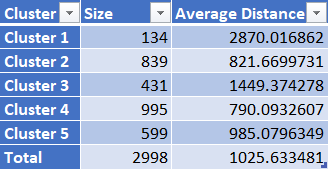
**Model 1: Clustering**

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.

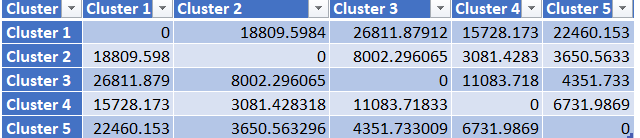
We take K ( no of clusters) as 5 and 30 iterations for this model, Value of K is generally selected using elbow plot, for this case study we use 5 as the K value and then group the movies based on the distances between the clusters.

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



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

****

From the above table, 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.

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.

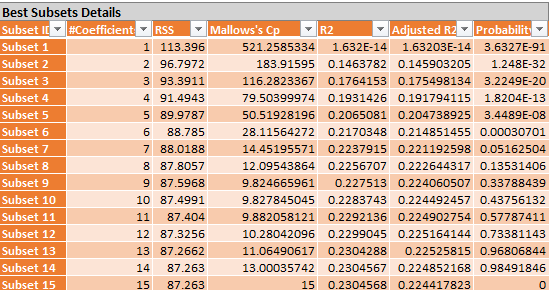
We use Multi Linear Regressor, Decision Tree Regressor and Ensemble-Random Forest Regressor to find out the factors that affect the revenue the most.

**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.

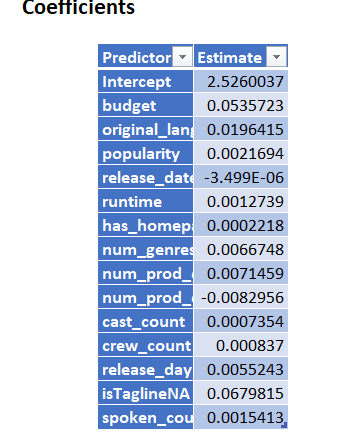
We apply the model on the data, which is portioned into Training and Validation, they contain 60% and 40% of the data respectively. Attribute Revenue is the target variable and remaining data is given as input except for the column Id.

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



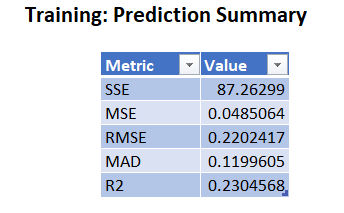
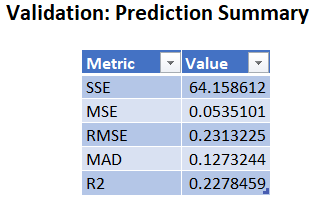
The best subset is chosen depending on the Cp and Adjusted R2 value, In this case, Subset 15 is the best.

Below are the coefficients are given by the model.



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

Below are the metrics for that subset.

 '

**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.

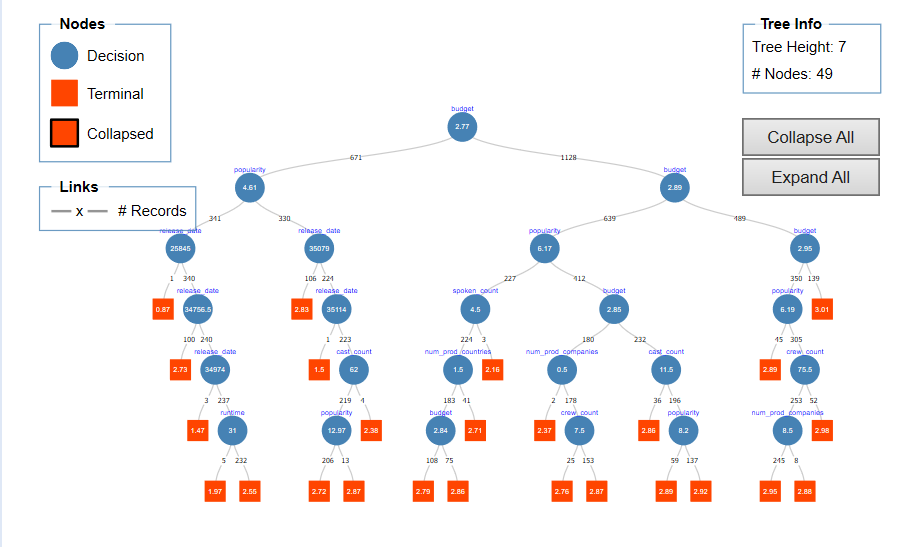
3. It is normally distributed in all nodes which makes regression complex in terms of relating to another independent variable.

**Model 2 – Decision Tree Regressor**

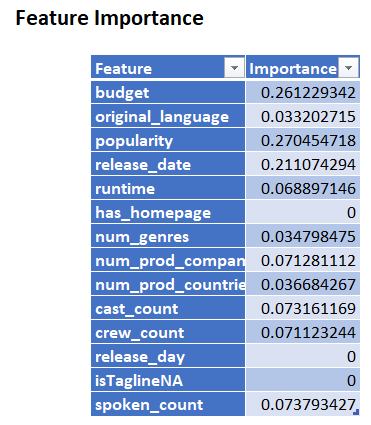
Decision Tree Regressor uses Standard Deviation as a metric to find the best attribute which is homogeneous and further follows the same process to divide all the attributes.

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

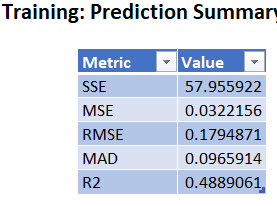
Below is the decision tree.

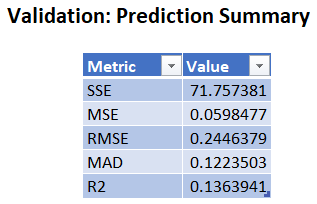


The most important variable as per the model is budget, followed orginal\_language, popularity and release\_date.



Below are the metrics for the decision tree



.

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.

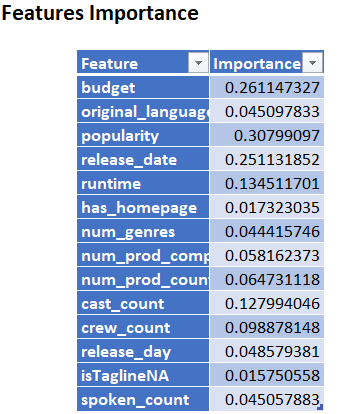
**Dis-Advantages of Decision Tree:**

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.

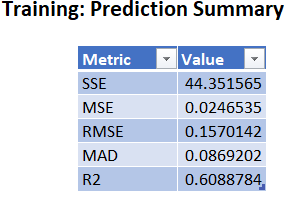
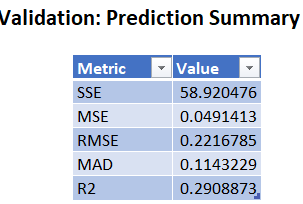
**Model 3 – Random Forest Regressor:**

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



Below are the metrics for the Random Forest Regressor.

**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.

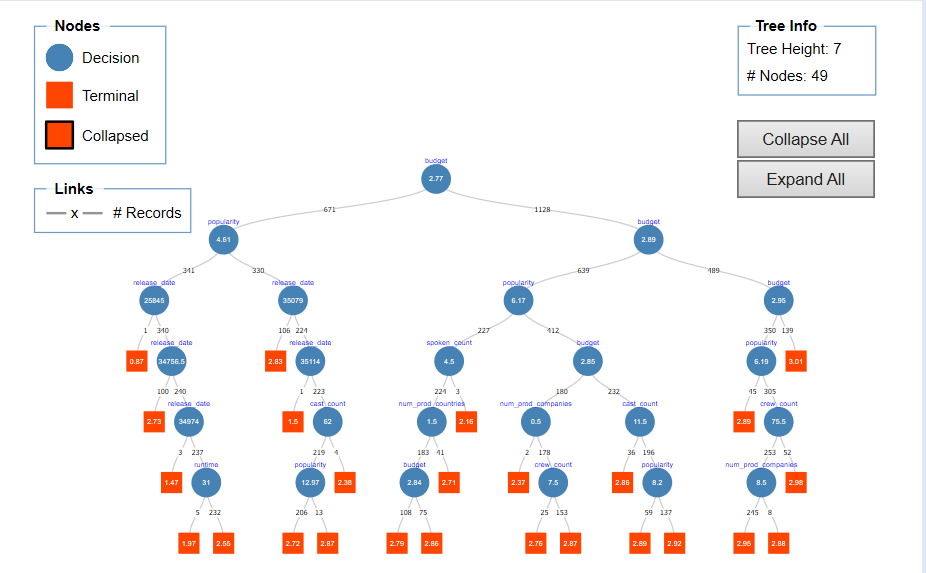
**Model 1 -** **Decision Regression Tree**

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.

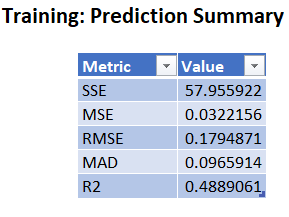
We divide the data into Training, Test and validation, they contain 60%,10% and 30% respectively. Attribute revenue is considered as the target variable and variables such as is\_tag\_line\_na and has\_homepage is neglected as could see that they do not provide any vital information to the model.

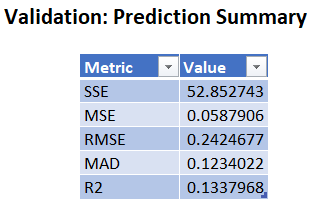
The model considers budget, popularity, release\_date, and run\_time as important features.

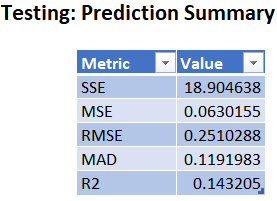
Below is the decision tree of the model built.



**Metrics for Regression Tree:**

****

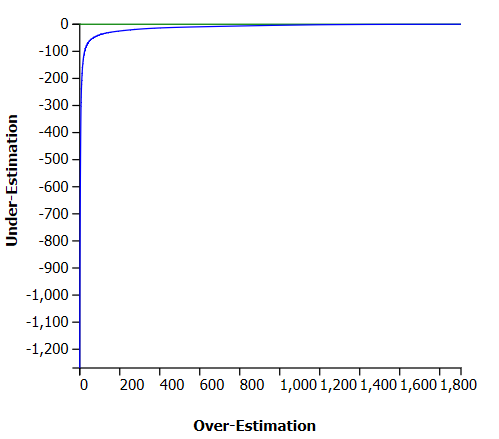
****



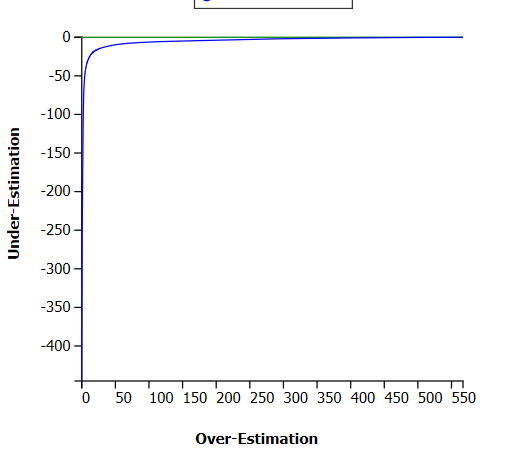
The model is performing well as it is giving a low error value and also the RMSE value for validation and test is close.

ROC curves are used to access the model performance

**Validation ROC Curve:**



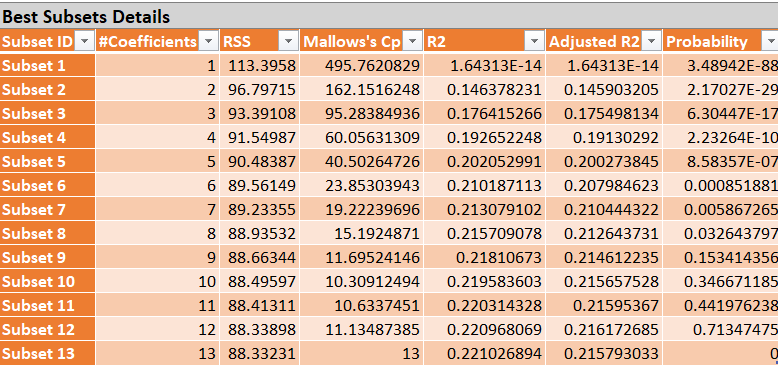
**Test ROC Curve:**



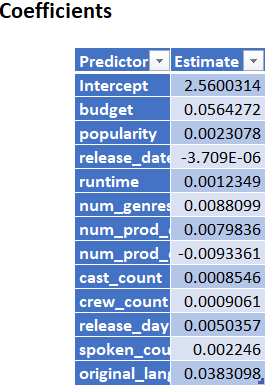
**Model 2 - Multiple Linear Regression:**

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

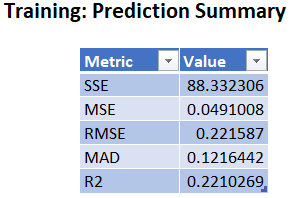
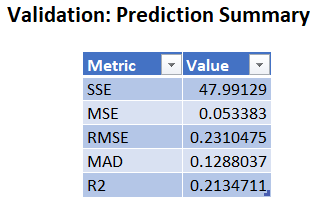
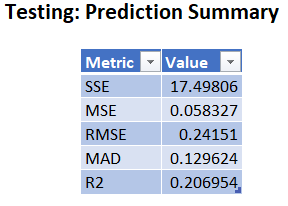
The model is applied to the portioned data and we use the best subset for feature selection.



Subset 13 is chosen as the best predictor and the model gives the following output.



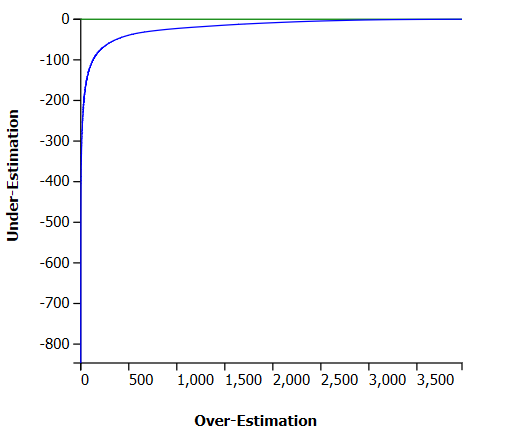
Error metrics for the data are as follows

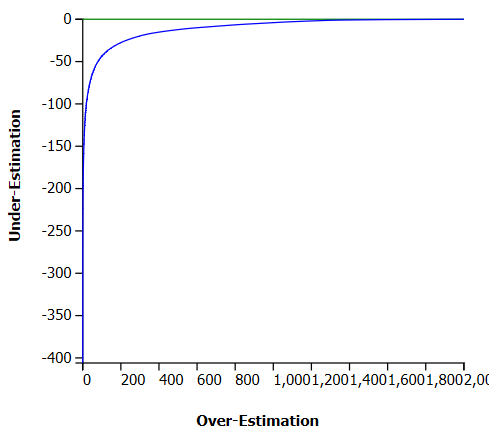
**ROC Curves:**

ROC curves are used to access the model performance

**ROC Curve - Testing**



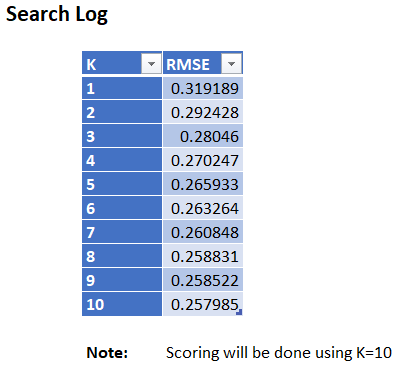
**ROC Curve - Validation:**



**Model 3 -KNN**

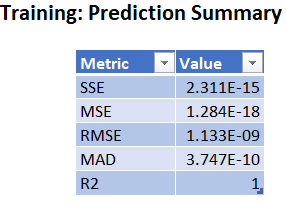
K-Nearest Neighbor calculates the avg distance of N neighbors and then tries to measure the similarity between a data point and the target variable.

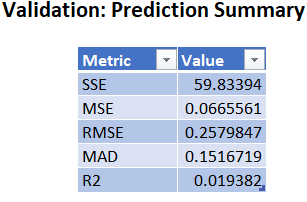
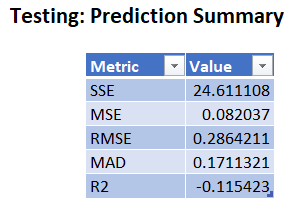
We apply the KNN model on the portioned data and give revenue as the target variable. For this study, we search the optimal K value between 1 and 15

.

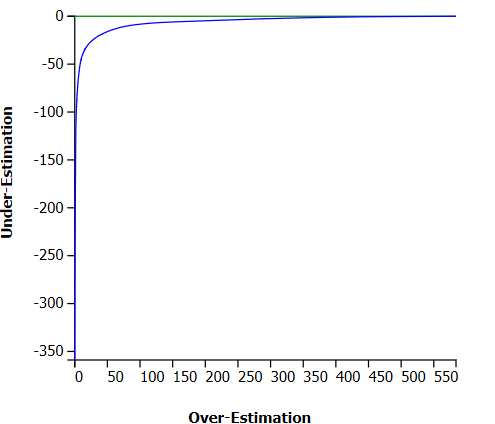
The model is built using 10 nearest neighbors.

Error metrics for the data are as follows

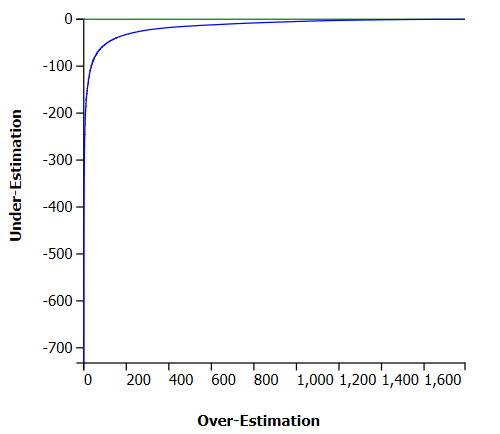


**ROC Curve - Testing**



**ROC Curve - Validation:**

****

**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:**

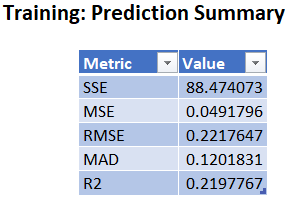
1. KNN does not work well with huge datasets.
2. If a data set a high number of dimensions, then KNN is not an optimal solution.
3. Sensitive to outliers and noise present in the data

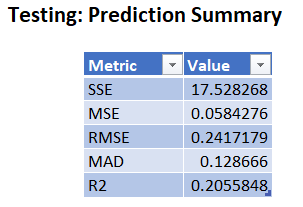
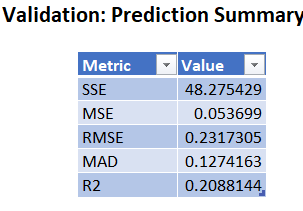
**Model 4 -Ensemble Bagging**

Bagging is a process where we create no 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.

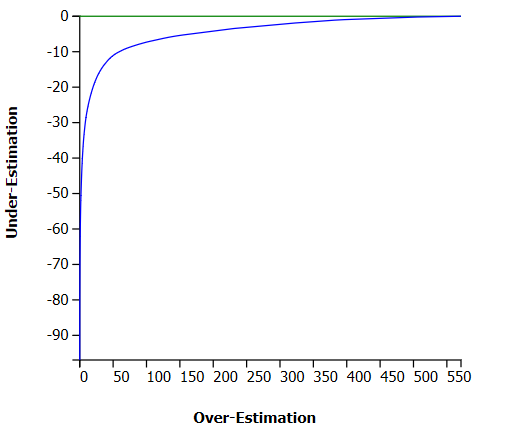




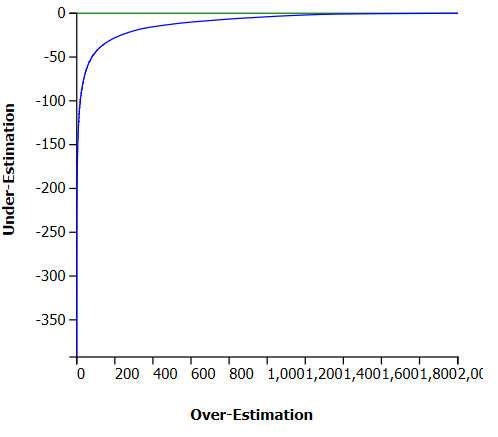
**ROC Curves:**

ROC curves are used to access the model performance

**ROC Curve - Testing**

****

**ROC Curve - Validation:**

****

**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.

**7. STRATEGIC RECOMMENDATION**

TMDB data set can be interpreted in different ways, in this study we opted to take count of genres present in the film, But the film's revenue might get effected depending on the production companies fame. Taking the count of the data will give us only a part of the information and might not be suitable for every case.

If the data set would have information such as the popularity of the cast or the popularity of the crew, it would be more useful as cast and crew are the main factors for any movie’s success. As that information is not provided taking the count of the cast and crew might be very useful,

Converting variables such as “belongs\_to\_collection”, “tag\_line” and “over\_view” into binary can be more useful as we need to perform NLP to understand variables tag\_line and over\_view.

**8. CONCLUSION**

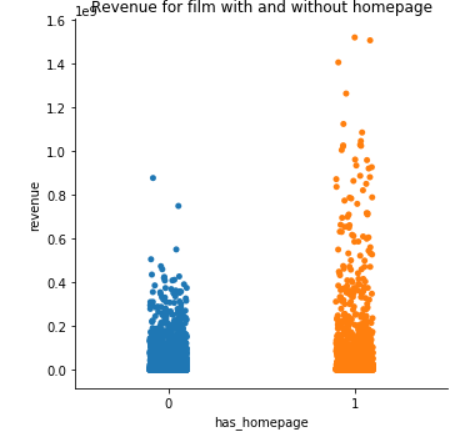
Based on the Exploratory data analysis done on the data using python libraries and based on the machine learning analysis done on XL Miner we conclude the above hypothesis as below:

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 Visualizations:**

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

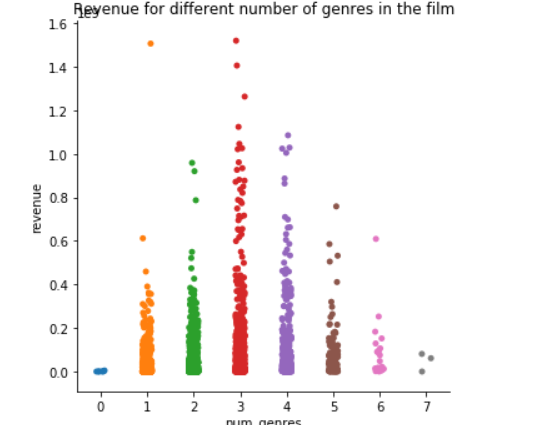
****

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

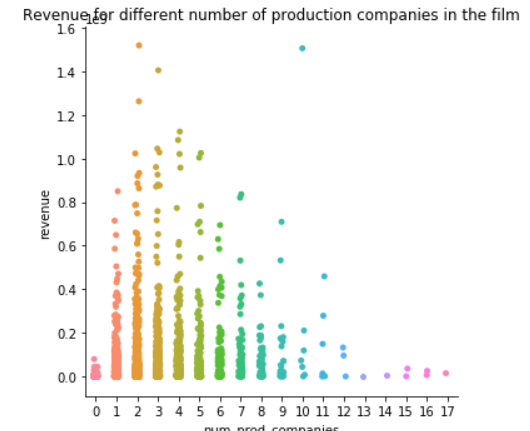
**ii. Revenue vs no\_of\_genres**

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

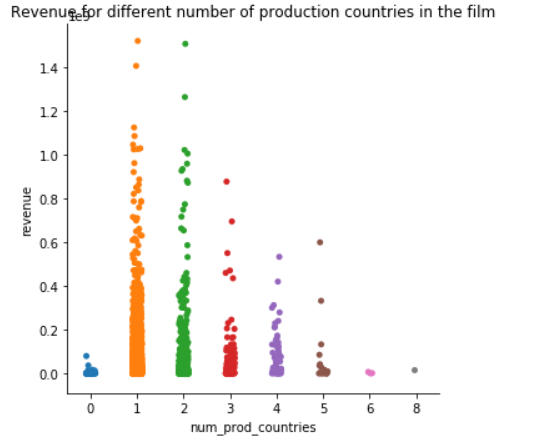
****

**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:**

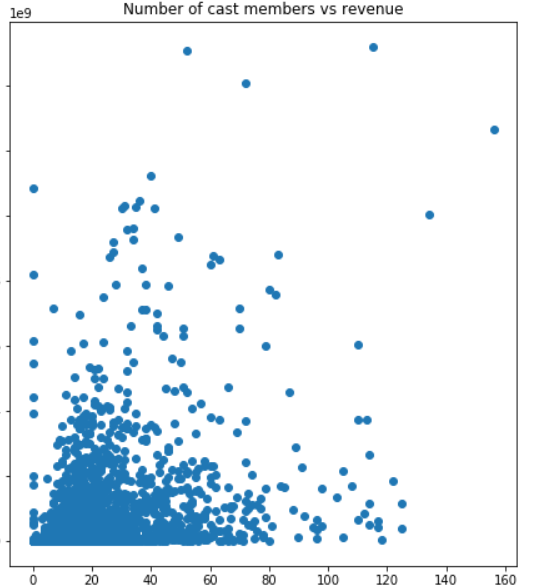
****

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

**v. Revenue Vs Ca**

****

**vi. Revenue Vs Cars**

**:**