Predictive Analysis of Movie Dataset

Over the last ten years, the proliferation of cinema count has risen from 2,500 to 150,000, according to estimates. With the advancement of information technology, viewers' information and reviews may now be collected, stored and retrieved. It is now possible to draw conclusions from these datasets using today's computational capacity. This research focuses on a dataset which provides information about many such movies, views the dataset from an analytical standpoint, and forms conclusions over the dataset using machine learning techniques. In general movies are intrinsically entertaining and are cut into wide ranges based on their demographic scope. The dataset which has been used for the purpose of this report is a public dataset which has been retrieved from Kaggle.com. The dataset contains 19 columns (19 descriptive attributes about the movies) about 500 movies. The aim of this research is to answer different research questions regarding the dataset using predictive analysis through regression and classification algorithms.

Research Questions:

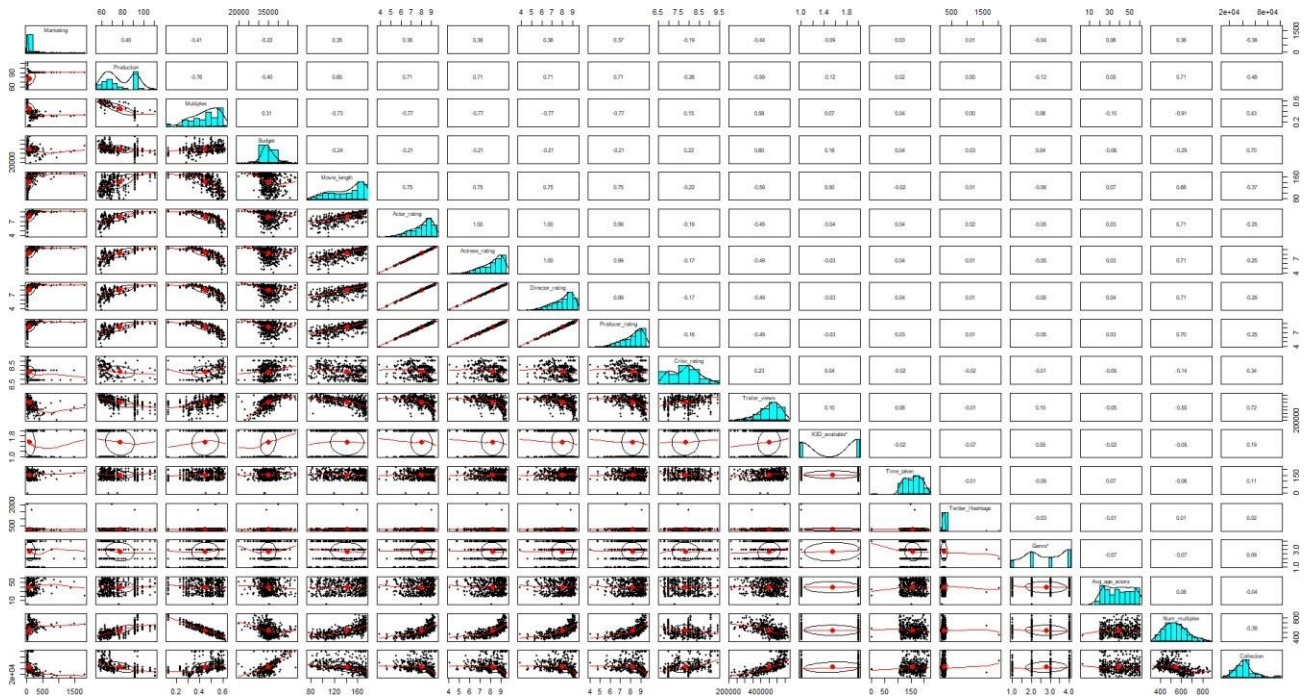
1. Is the success of a movie, which is measured by the Collections in this dataset, affected by other features of a movie, namely the budget of the film, ratings given by critics, number of views on the trailer, the expenses spent on Marketing and Production and so forth?
2. Can the expenses incurred on a movie be estimated from the collections made by the movie, and other such variables?
3. Can a new picture be classified into a particular Genre based on the movies already present, using the variables in the dataset as predictor variables?
4. Does Budget depend on Genre, does a particular Genre incur budget?

Dataset:

The dataset which has been used for the purpose of this report is a public dataset which has been retrieved from Kaggle.com. The dataset contains 19 columns (19 descriptive attributes about the movies) about 500 movies. The attributes are as listed below:

Marketing Expense, Production Expense, Multiplex, Budget, Movie Length, Lead Actor Rating, Lead Actress Rating, Director Rating, Producer Rating, Critic Rating, Trailer Views, 3D availability, Time taken, Twitter Hashtags, Genre, Average age of actors, Number of Multiplexes, Collections.

Viewing the correlation that the variables in the dataset have with each other through the below plot. This helps in identifying which variables are statistically significant as compared to others in estimating and predicting a particular target variable.



CORRELATION PLOT BETWEEN COLLECTION AND MOVIE DATASET

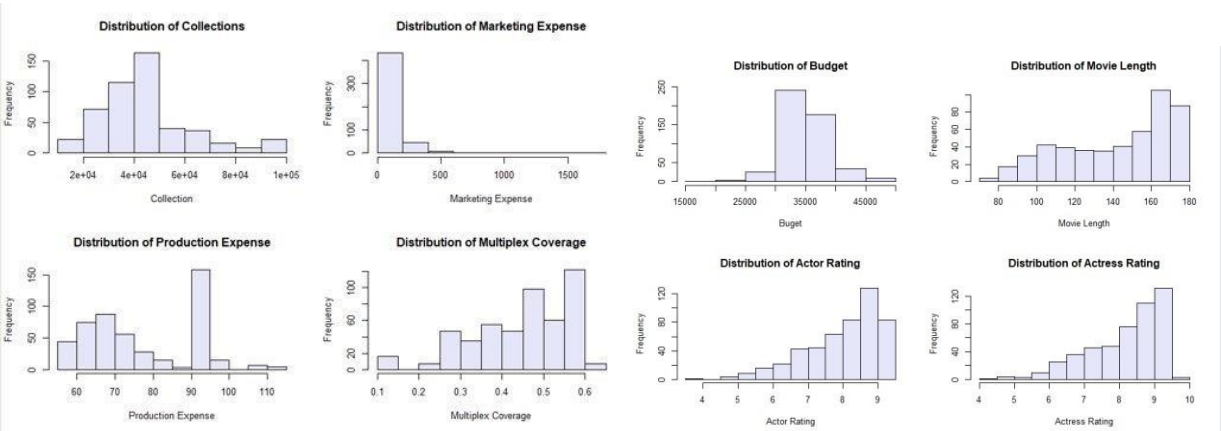
Proportional function is a utility function for calculating the proportions of a dataset's column's values. This helps in identifying which value in a column is more dominant as compared to others.

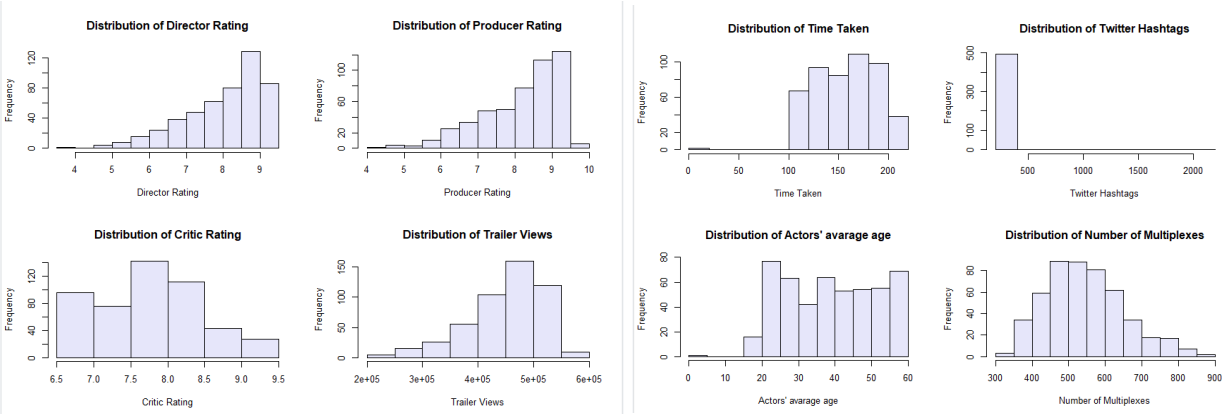
```
> # Proportions of categorical variables:
> prop.table(table(Movie_data$X3D_available))

       NO       YES
0.4493927 0.5506073
> prop.table(table(Movie_data$Genre))

   Action    Comedy     Drama  Thriller
0.1417004 0.3056680 0.1902834 0.3623482
```

PROPORTIONAL FUNCTION

Histograms for numerical attributes:

The hist() function in the R programming language can be used to build a histogram. The frequency of values of a variable grouped into ranges is represented by a histogram. This function accepts a vector of values as input and plots the histogram. The histogram representations of the different attributes in the dataset are as seen below.

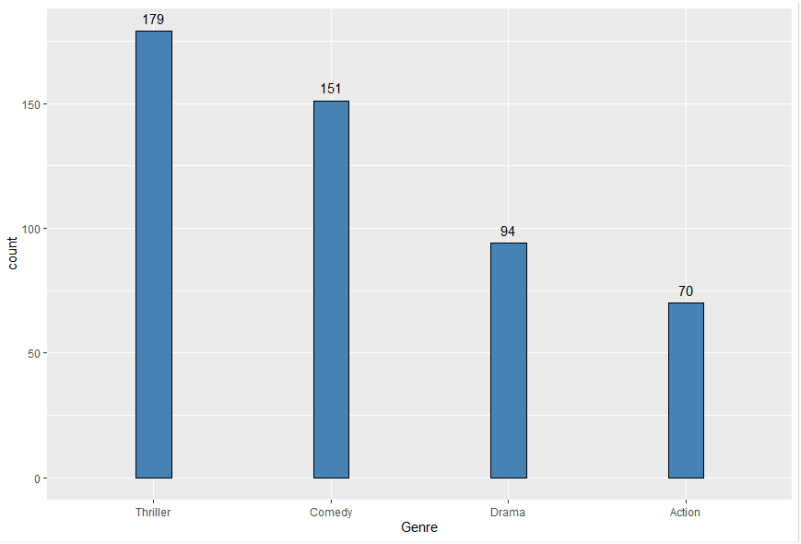




It can be observed that most of the variables have a left skewed histogram and the variables collection and number of multiplexes have a normal distribution. Some variables such as Actor's average age and Time taken have a uniform distribution.

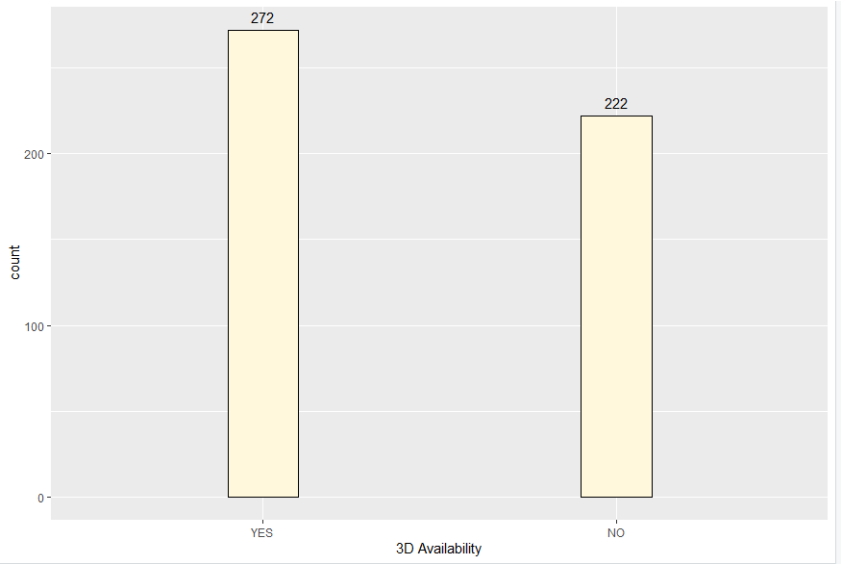
Bar Plots for Categorical Attributes:

A bar chart is a graph that uses rectangular bars with heights or lengths proportionate to the values they represent to convey categorical data. The bars can be plotted either horizontally or vertically. A column chart is another name for a vertical bar chart.



BAR PLOT ON GENRE

Based on the above bar graph of the variable 'Genre', it can be observed that Thriller genre has about 179 movies which is the highest count followed by comedy genre with 151. Drama and Action are the least occurring Genres according to the dataset.



BAR PLOT BASED ON AVAILABILITY

Based on the bar plot, observed that most of the movies have 3D availability with the count of about 272 from the dataset used. About 222 movies from the dataset do not have 3D availability.

Answers to the Research Questions:

1. **Is the success of a movie, which is measured by the Collections in this dataset, affected by other features of a movie, namely the budget of the film, ratings given by critics, number of views on the trailer, the expenses spent on Marketing and Production and so forth.**

Correlation between the target variable 'Collection' and the remaining predictor variables in the dataset:

```
     Marketing Production Multiplex    Budget Movie_length Actor_rating Actress_rating Director_rating
[1,] -0.3825955 -0.4833293 0.4296696 0.6982691   -0.3742048   -0.2496336     -0.2476027       -0.245265
     Producer_rating Critic_rating Trailer_views Time_taken Twitter_Hashtags Avg_age_actors Num_multiplex
[1,]      -0.2470019     0.3393278     0.7170357  0.1110919       0.02317686     -0.0449221    -0.3942704
```

CORRELATION VALUES BETWEEN COLLECTION AND MOVIE DATASET

It can be observed from the above counts that Collection has the highest positive correlation with Budget and the highest negative correlation with Production Expense.

Approach 1: Linear regression

A regression model that employs a straight line to explain the relationship between variables is known as linear regression. It searches for the value of the regression coefficient(s) that minimizes the total error of the model to find the line of best fit through your data. Initially a linear model was fit to the entire dataset to predict collections and the statistics of this model are as seen below:

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10460 on 478 degrees of freedom
Multiple R-squared:  0.687,     Adjusted R-squared:  0.6772
F-statistic: 69.95 on 15 and 478 DF,  p-value: < 2.2e-16
```

SUMMARY OF LINEAR MODEL

Based on the summary of the linear model, the model accounts for 68.7 % of the variation in

Collections variable.

In order to improve the performance of the linear model, only the statistically significant variables have been taken into consideration in the construction of the next model, which has brought up the Adjusted R-squared and F-statistic values of the model. This indicates a more reliable model than the first.

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10420 on 485 degrees of freedom
Multiple R-squared:  0.6849,    Adjusted R-squared:  0.6797
F-statistic: 131.8 on 8 and 485 DF,  p-value: < 2.2e-16
```

SUMMARY FOR LM(SUBSET)

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.384e+05  1.237e+04 -11.190  < 2e-16 ***
Marketing       -1.075e+01  3.150e+00  -3.414 0.000693 ***
Production      -1.423e+02  5.907e+01  -2.409 0.016372 *
Multiplex        2.331e+04  7.484e+03   3.115 0.001950 **
Budget           1.757e+00  1.547e-01  11.360  < 2e-16 ***
Producer_rating  4.760e+03  7.378e+02   6.452 2.67e-10 ***
Critic_rating    4.124e+03  7.538e+02   5.471 7.17e-08 ***
Trailer_views    1.044e-01  1.050e-02   9.941  < 2e-16 ***
Time_taken       3.512e+01  1.516e+01   2.317 0.020926 *
```

COEFF FOR SUMMARY(SUBSET)

To confirm if the model is apt for prediction or not, random rows were selected from the dataset and the Collection value for that row has been predicted using the dataset and then compared with the already existing value for collection. The values observed are as below:

Row number	Actual Value	Predicted Value	Predicted Lower Limit	Predicted Upper Limit
14	40800	38178.5	17572.27	58784.73
98	55000	53664.3	33048.96	74279.63
201	45200	45842.27	25312.29	66372.24

It can be observed that for the randomly selected rows, the predicted values of collection are approximately equal to the actual values and the actual value falls in the range predicted by the model for all three cases.

Approach 2: Ridge Regression

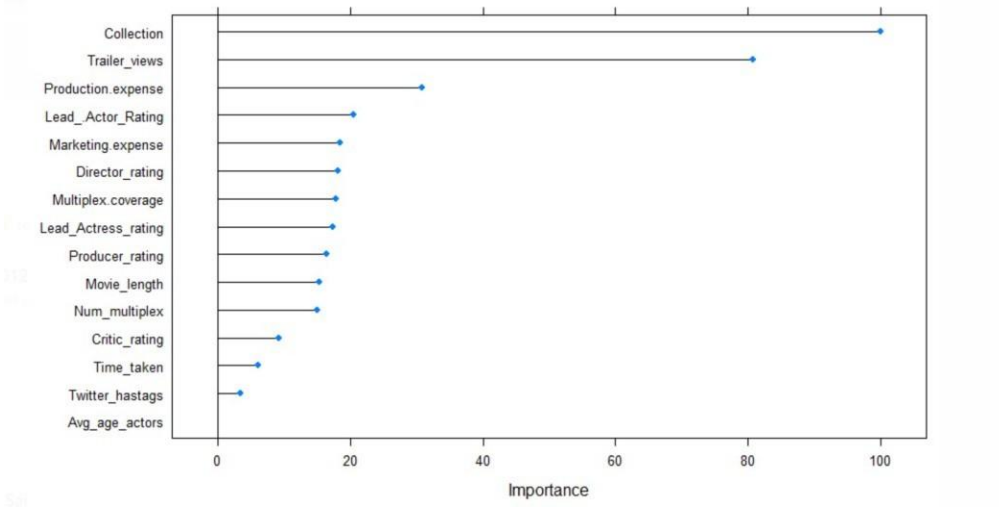
In order to have a different approach to predicting Collection in this dataset, Ridge Regression has been applied to the same for the prediction purpose.

Row number	Actual Value	Predicted Value
14	40800	38246.418
98	55000	53542.637
201	45200	45797.763

For the same data, ridge regression has gotten an accuracy of 72.3%, which is better than the

performance obtained by linear regression.

The below plot describes the degrees of importance of variables in the dataset in the ridge regression model:



VARIABLE IMPORTANCE PLOT

The accuracies obtained indicate that Collections for new movies can be reliably predicted using these models.

2. Can the expenses incurred on a movie be predicted from the collections made by the movie, and other such variables?

There are three types of expenses in this dataset: Budget, Production expense and Marketing expense.

Budget:

Approach 1: Linear regression

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2688 on 478 degrees of freedom
Multiple R-squared:  0.5436,    Adjusted R-squared:  0.5293
F-statistic: 37.96 on 15 and 478 DF,  p-value: < 2.2e-16
```

LINEAR MODEL FOR BUDGET

Using best subset method, a model with 7 significant variables has been chosen as the best model to estimate budget based on the summary statistics.

This model has given an accuracy of 54.3 % in estimating Budget.

To confirm if the model is apt for estimation or not, random rows were selected from the dataset and the Collection value for that row has been predicted using the dataset and then compared with the already existing value for collection. The values observed are as below:

Row number	Actual Value	Predicted Value	Predicted Lower Limit	Predicted Upper Limit
14	33046.69	34449.18	29159.12	39739.23

98	37368.49	36941.17	31639.33	42243
201	32724.51	33778.85	28464.82	39092.89

It can be observed that for the randomly selected rows, the predicted values of collection are approximately equal to the actual values and the actual value falls in the range predicted by the model for all three cases.

Approach 2: Ridge Regression

Values estimated by Ridge Regression model:

Row number	Actual Value	Predicted Value
14	33046.69	34684.47
98	37368.49	36644.58
201	32724.51	33699.56

This ridge regression gave an accuracy of 60.4%, which is better as compared to the linear regression model applied to the same data.

Approach 3: Lasso Regression:

Lasso regression can be used when there is multicollinearity in the data, as there is in this particular dataset.

Lasso regression to estimate Budget in this dataset gave an accuracy of 60.4 %, which is on par with the accuracy obtained by Ridge Regression.

Production Expense:

Approach 1: Linear Regression:

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.994 on 478 degrees of freedom
Multiple R-squared:  0.674,     Adjusted R-squared:  0.6638
F-statistic:  65.9 on 15 and 478 DF,  p-value: < 2.2e-16
```

LM FOR PRODUCTION EXPANSES

Using best subset method, a model with 7 significant variables has been chosen as the best model to estimate Production Expense based on the summary statistics.

This model has given an accuracy of 67.4 % in estimating Production Expense.

To confirm if the model is apt for estimation or not, random rows were selected from the dataset and the Collection value for that row has been predicted using the dataset and then compared with the already existing value for collection. The values observed are as below:

Row number	Actual Value	Predicted Value	Predicted Lower Limit	Predicted Upper Limit
14	71.28	75.96	60.24	91.69
98	72.12	78.27	62.54	93.99

201	76.18	73.72	58.05	89.39

It can be observed that for the randomly selected rows, the predicted values of collection are approximately equal to the actual values and the actual value falls in the range predicted by the model for all three cases.

Approach 2: Ridge Regression

Values estimated by Ridge Regression model:

Row number	Actual Value	Predicted Value
14	71.28	75.95
98	72.12	75.21
201	76.18	77.58

This ridge regression gave an accuracy of 67.6%, which is slightly better as compared to the linear regression model applied to the same data.

Approach 3: Lasso Regression

Lasso regression to estimate Production expense in this dataset gave an accuracy of 68.7 %, which is slightly better than the accuracy obtained for ridge regression and Linear Regression.

Marketing Expense:

Approach 1: Linear Regression:

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 148.9 on 478 degrees of freedom
Multiple R-squared:  0.2724,    Adjusted R-squared:  0.2496
F-statistic: 11.93 on 15 and 478 DF,  p-value: < 2.2e-16
```

LM FOR MARKETING EXPENSES

Using best subset method, a model with 7 significant variables has been chosen as the best model to estimate Marketing Expense based on the summary statistics. This model has given an accuracy of 27.2 % in estimating Marketing Expense.

Row number	Actual Value	Predicted Value	Predicted Lower Limit	Predicted Upper Limit
14	32.59	49.23	-243.12	341.59
98	22.97	95.22	-197.2	387.65
201	22.72	61.77	-230.22	353.77

It can be seen that the accuracy is very low for this model in estimating Marketing Expense and the Predicted values are far off from the Actual values. This confirms that for this dataset, Marketing Expenses cannot be accurately predicted from the predictor variables.

Approach 2: Ridge Regression

Values estimated by Ridge Regression model:

Row number	Actual Value	Predicted Value
14	32.59	52.74
98	22.97	89.79
201	22.72	75.26

This ridge regression gave an accuracy of 34.9%, which is slightly better as compared to the linear regression model applied to the same data, but still not a reliable value to estimate Marketing Expense values.

Approach 3: Lasso Regression

Lasso regression to estimate Marketing expense in this dataset gave an accuracy of 27.4 %, which is slightly better than the accuracy obtained Linear Regression but less than that of Ridge Regression.

3. Can a new picture be classified into a particular Genre based on the movies already present, using the variables in the dataset as predictor variables?

Naive Bayes is a classification method based on Bayes' Theorem and the assumption of predictor independence. Naive Bayes is a Supervised Non-linear classification algorithm.

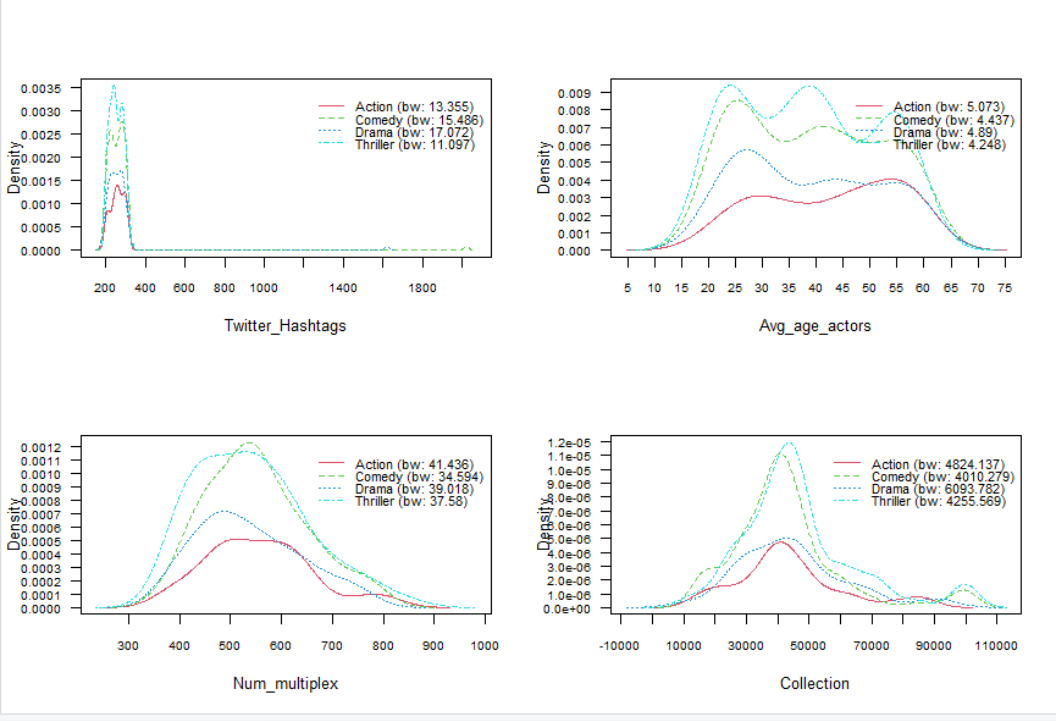
The Naive Bayes model is simple to construct and is especially good for huge data sets. It is simple to use and forecast the test data set class quickly. In comparison to numerical input variables, it performs well with categorical input variables. For better results, it requires independent predictor factors.

The Naive Bayes algorithm is called "Naive" because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. So, Naive Bayes is widely used in Sentiment analysis, document categorization, Email spam filtering etc.

In R, a confusion matrix is a table that categorizes predictions in relation to actual values. It has two dimensions, one of which will represent anticipated values and the other will represent actual values. The 2x2 matrix in R was visible in the majority of the resources.

For this dataset and this research question, applying Naïve Bayes algorithm to classify 'Genre' in order to know if a new picture be classified into a particular Genre based on the movies already present, using the variables in the dataset as predictor variables.

For this model, the misclassification rate of Genre in the training data is 54.4%, which equals an accuracy of 45.6%, and the misclassification rate of Genre in the testing data is 61.2%, which equals an accuracy of 38.8%.

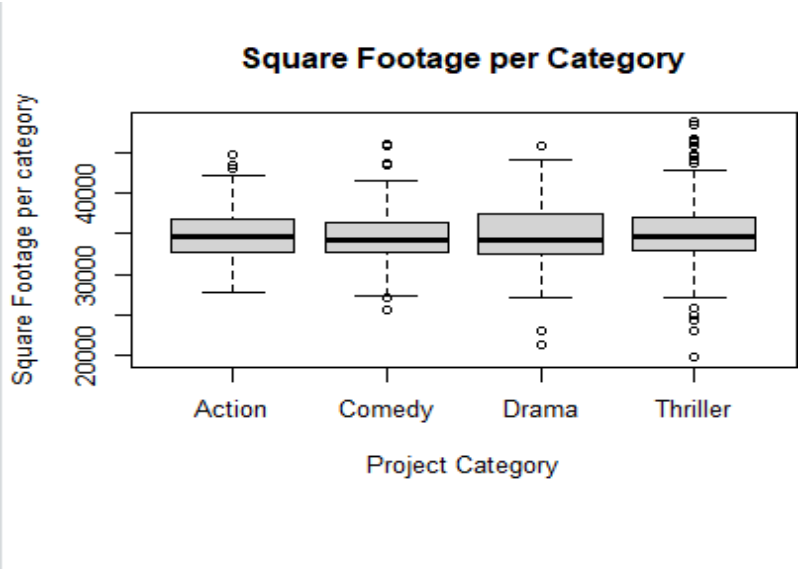


NAIVE BAYES MODEL PLOT

This indicates that the classification of Genre of new movies based on the details of the existing movies is not error free, but it can be reliable to some extent.

4. Does Budget depend on Genre, does a particular Genre incur budget?

Based of the analysis with respect to the Budget variable in the dataset, obtained the result as even though there are many Genres the budget is around same amount for all the Genres.



BOX PLOT OF BUDGET BASED ON GENRE

Findings:

Based on the above research questions and all the prediction methods it is possible to say that the dataset, which has been chosen provides ample scope for various predictive methods. These also indicates that this dataset can be used for further research processes in the future experiments too.

Since the analysis is complete, the observations made are, Thrillers are mostly released movies. Collections for the comedy movies are more compared to the other genre movies. There are more movies which are available in 3D compared than normal 2D movies. The action genre movie actors age is less compared to the comedy movie genre actors.

The success of the movie, as measured by Collection in this dataset, could be estimated with an accuracy of 68.7% using all the predictor variables in the dataset, and ridge regression obtained a slightly better score than that which is 71.2%. After conducting best subset method and considering only the statistically significant variables for the analysis, the overall performance of the linear model improved, as in better F-statistic and adjusted R^2 values were obtained.

The expenses incurred on a movie, as provided by Budget, Production and marketing expenses in the dataset, can be estimated with the below provided accuracies:

		Accuracy		
		Budget	**Production Expense**	**Marketing Expense**
Model	**Linear Regression**	54.3%	67.4%	27.2%
	Ridge Regression	60.4%	67.6%	34.9%
	Lasso Regression	60.4%	68.7%	27.4%

It can be observed from these counts that for this dataset, Marketing Expenses cannot be accurately predicted from the predictor variables present.

When the classification algorithm Naïve Bayes was applied to classify the Genre for unseen movies, an accuracy of 45.6% was observed for training data and 38.8% for testing data.

Based on the box plots for Budget with respect to Genre, it can be observed that the Budget of a movie does not depend on the Genre, and all the movies have similar budgets irrespective of the Genre.

Future Research:

For future research the dataset can be further utilized and further regression methods with multiple views of data can be applied to the dataset. The dataset itself can be updated with more information since many movies in various Genres are being released globally. There by, multiple subsets can be created from the dataset and analyzed using various algorithms that were not applied in this research.

Conclusion:

Based on the research done on the above dataset, it can be said that analysis determines several trends and patterns of data in the dataset, which plays a major role for future predictions, so the later users of the data can follow them by eliminating the risks which were already faced or observed in this dataset and step up with more accuracy in success than failure.

Appendix:

APPENDIX A - ATTRIBUTES DESCRIPTION

Below are the initial set of attributes and the description from the data source.

Marketing Expense: Describes about the expenses for marketing of each movie.

Production expense: Describes about the expenses for production of each movie

Multiplex coverage: Describes about the coverage that is done through multiplex

Budget: Describes about the Budget of the movie.

Movie_length: Describes about the length of movie.

Lead_ Actor_Rating: Rating of the Lead Actor.

Lead_Actress_rating: Rating of the Lead Actress.
Director_rating: Rating of the Director.

Producer_rating: Rating of the Producer.

Critic_rating: Rating of the Critic.

Trailer_views: Number of views for the trailer.

3D_available: Determines whether the movie has the 3D availability or not.

Time_taken: Describes about the time taken

Twitter_hastags: Describes about how many times that movie appeared in hastags.

Genre: Describes about which genre the movie belongs.

Avg_age_actors: Average of the age of the actors worked in the film.

Num_multiplex: Count of the multiplexes the movie released.

Collection: Describes about the collections the movie got.

Appendix B: Box plot for each attribute