Predicting IMDB User Ratings and Movie Earnings

Arun Jaganathan (ajagana)
Rachit Thirani (rthiran)
Vivek Varma Nadimpalli (vvnadimp)
Group P-24

1 Introduction

Internet Movie Database or IMDb is a database of information related to films and T.V. shows. Users usually rate the movies or shows which they have watched. The collective ratings are stored in the database which can be used by other users to decide whether or not to watch the movie/TV show.

Movie ratings are based on various factors like popularity of the actors and the directors, the genre of the movie, review by critics etc. The data can also be used to predict how well the movie/TV show would be liked by the audience. Since ratings are continuous valued, we are inclined to use models such as linear regression, support vector regression and neural networks which would be giving a continuous output and thereby be used to predict the ratings.

2 Background

As part of the project, we are also predicting the Movies earnings. As it is difficult to predict the exact income a movie makes, we are dividing the movie's gross income into 5 categories and classify the earnings accordingly. Decision trees, Clustering Algorithms can be used to classify the movie's gross income into categories.

This analysis can be extended to help out the users to decide whether or not to watch a newly released movie. Also, Producers can use this analysis to predict how well the movie may perform at the box office and whether or not it is a wise decision for them to invest in the movie. In addition, even directors can use this to get knowledge of which combination of actors and genre suits them the best.

The reason that we chose to predict Gross Income and User Rating for a movie is because we were convinced that together, these parameters would give the best evaluation for any movies. Also we are using the prediction models in R programming language for this entire project.

The dataset is obtained from www.kaggle.com which is named as IMDb 5000. The unprocessed raw data, as the name indicates, has about 5043 rows of data with 28 different attributes both categorical and numerical. The features included are directors popularity, casts popularity, content rating, budget, language, country, number of users who voted, etc. For gross income, 3665 of the data points belong to the one category, i.e. 72% of data set belongs to one category.

3 Method

3.1 Preprocesing

As the first step towards prediction, the initial phase consisted of analyzing the data set and finding the minor flaws and how it could be overcome.
We did the following for data processing-
Normalizing the numerical data - To make all of the variables into proportion with one another.
Neglect the rows with missing values, those rows that does not affect output much - based on intuition.
Filling out the missing values with an average based on logical grouping.
In addition we also had to tweak the data to fit better for each models.
Correlation analysis, to find the most-influencing features, to make sure we did not neglect those missing values of the most-influencing features.
Also, certain categorical features, like color, country etc., are neglected as they had a very low correlation with the output.
The dataset was split randomly into 75% training set and 25% testing set.

3.2 IMDB Rating Prediction

Regression Algorithms are a form of supervised learning algorithms, that are used to model continuous valued output functions. A regression task begins with a data set in which the output values are known. A regression algorithm computes the value of the output as a function of the inputs for each data point in the dataset. The relationship between input features and output value is represented by a model, which can then be applied on a different data set in which the output values are not known.

3.2.1 Multivariate Linear Regression

Multivariate Linear regression is one of the most popular regression models. The output is continuous valued, which depends on one or many features. A linear function can be formed with all the features to predict the value of the output.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ + \beta_p x_p + \epsilon \tag{1}$$

where, x1, x2, ..., xn are the different attributes and 0, 1, ..., n are the constant.
In order to find the coefficients $\beta 0$, $\beta 1$, ..., βn, we must minimize the cost function

$$f(x, y, \beta, c) = \sum_{n=1}^{n} (y_i - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p))^2 \tag{2}$$

To do so we first partially differentiate the cost function with respect to β and equate it to zero. Once we do this for each β value and solve the resulting equations, we will get the corresponding values of the coefficients that minimize the cost function.

3.2.2 Artificial Neural Networks

Neural Networks (NN) is a data mining technique used for classification and clustering. It is modeled after the neural structure of the human brain. NN usually learns by examples. If NN is supplied with sufficient examples, it can perform classification and discover new patterns in data. Basic NN consists of three layers, input, output and hidden layer. Each layer can have multiple nodes. Nodes from input layer are connected to the nodes in the hidden layer. Nodes from hidden layer are connected to the nodes in the output layer. Those connections are assigned weights. One of the most popular NN algorithms is back propagation algorithm. After choosing the weights of the network randomly, the back propagation algorithm is used to compute the necessary corrections.
The algorithm can be decomposed in the following four steps:

i) Feed-forward computation

ii) Back propagation to the output layer
iii) Back propagation to the hidden layer
iv) Weight updates
The algorithm halts when the value of the error function has become negligible.

3.2.3 Support Vector Regression

Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm. The Support Vector Regression (SVR) uses the same principles as the SVM, with only a few differences. In the case of regression, a tolerance margin is set in approximation to the SVM. The main idea is to lower the error, creating the hyperplane which maximizes the margin.

In SVR the input is mapped onto an n dimensional feature space using a non-linear mapping following which the construction of a linear model is performed. The linear model is given by

$$f(x, w) = \sum_{i=1}^{n} w_i g_i(x) + b$$

where $g(x)$ is a set of non-linear transformations and b is the bias term.

3.3 Gross Income Prediction

As gross income is a continuous value and no one can predict accurately the gross income, we decided to divide the gross income into different categories - Average, Above average, High, Very High, Extremely High. We used different approaches to solve this classification which are:

i)Classification and regression tree
ii)Conditional inference tree
iii)Kmeans clustering

3.3.1 Classification and Regression Tree

Classification and regression tree is machine-learning method for constructing prediction models from data. The models are obtained by recursively partitioning the data space and fitting a simple prediction model within each partition. As a result, the partitioning can be represented graphically as a decision tree.

3.3.2 Conditional Inference Tree

The conditional inference tree works similar to classification and regression tree method. In addition, it uses a significance test procedure in order to select variables for branching. The algorithm tests if any independent variables are associated with the given response and chooses the variable that has the strongest association with the response.

3.3.3 K-Means Clustering

K-means Clustering is a Clustering algorithm which follows a simple way to classify a given data set by dividing it into k different clusters by defining k different centroids, which act as the cluster representatives and change with every iteration when a new cluster is formed, until no new cluster is formed.

4 Plan and Experiment

The objective of this project is to determine which features influence the IMDB User Rating and Movie's Gross Income. we are also interested in finding out which Regression and Classification Models perform better for this dataset.

Multivariate linear regression was used to predict the rating IMDB users would possibly give to a movie because the IMDB rating is a continuous valued variable. Since regression works only when the features are numerical, the categorical features were first converted into numerical values and then model was created. This did not yield better results for when only numerical attributes were taken. Therefore, only the numerical attributes were considered and normalized and the multivariate linear regression model was applied on 75% of the dataset which was selected as random to form the training set. The model obtained was the applied to the remaining 25% of the dataset which is the test set.

Neural networks were chosen for this dataset as neural networks can approximate almost any function regardless of its linearity. For this IMDB dataset we took the numerical attributes which were normalized as inputs to the neural network and created a model that predicts the IMDB user ratings as output.
In order to determine the number of hidden nodes required in the hidden layer we ran the model with by changing the number of hidden nodes and checked the corresponding error it produced. It was found that the error decreased as we increased the number of hidden nodes until 4 and then started to increase after 4. Hence the number of hidden nodes for which the error was minimum was determined to be 4.

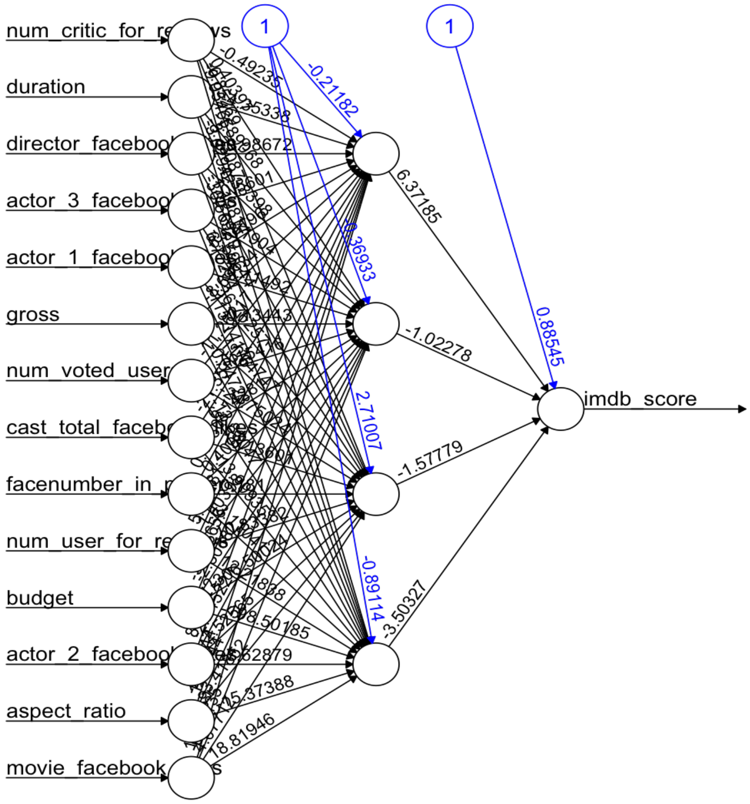

Figure 1: Artificial Neural Network

SVR was chosen for this dataset as it is advantageous in high dimensional space because SVR optimization is not dependent on the input space dimensionality. The numerical dataset created from the main dataset was used as input to the SVR to create the model. Then the created model was applied on the 25% of test data to determine the performance of this model.

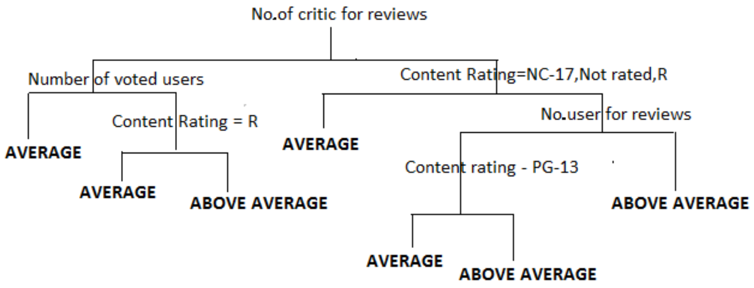

Figure 2: Classification and Regression Tree Plot.

Classification and regression was used as the base model for the category prediction for gross income. We began our search for an accurate model prediction by selecting different attributes. Discretization was required for continuous valued attributes for converting them into categories. Initially 3 categories were chosen, but as we proceeded further with model prediction, we realised that the data is highly concentrated and belongs to one category in particular. It was then decided to increase the categories. We tried different value of categories. As we increased the number of categories, the model started to overfit and the complexity of the model increased tremendously. we chose 5 categories as for the category prediction of gross income.
Different combination of discretised numerical data with categorical data led to different accuracy. The best accuracy which was achieved was .9354.
We realised the output belonged majorly form tow categories. Even though the accuracy was good, but the model was not favoured as the output when applied to real world data would not have given an accurate prediction.

Predicted–	Average	Above Average	High	Very High	Extremly High
Average	959	47	4	0	0
Above Average	7	12	7	1	1
High	0	0	0	0	0
Very High	0	0	0	0	0
Extremly High	0	0	0	0	0

Our quest led us in exploring more decision tree algorithms. We came up with conditional inference tree method

We used the same attributes which we used for the classification and regression tree. The accuracy which we got was slightly less that the previous method, but the output belonged to every category, i.e., we could get predictions from all the different categories of gross income. Hence this method was favoured as when this method would have been applied to the real world dataset, we would have got a better prediction. The data which was used in test and training was more concentrated into one category.

Predicted–	Average	Above Average	High	Very High	Extremly High
Average	959	47	4	0	0
Above Average	9	10	5	1	1
High	0	0	0	0	0
Very High	0	0	0	0	0
Extremly High	0	2	2	0	0

In search for some other method of prediction except trees, K-Means was implemented.

We tried clustering the data into clusters. Initially we clustered the data in to three clusters, but all the clusters pointed towards one category. Then the number of clusters

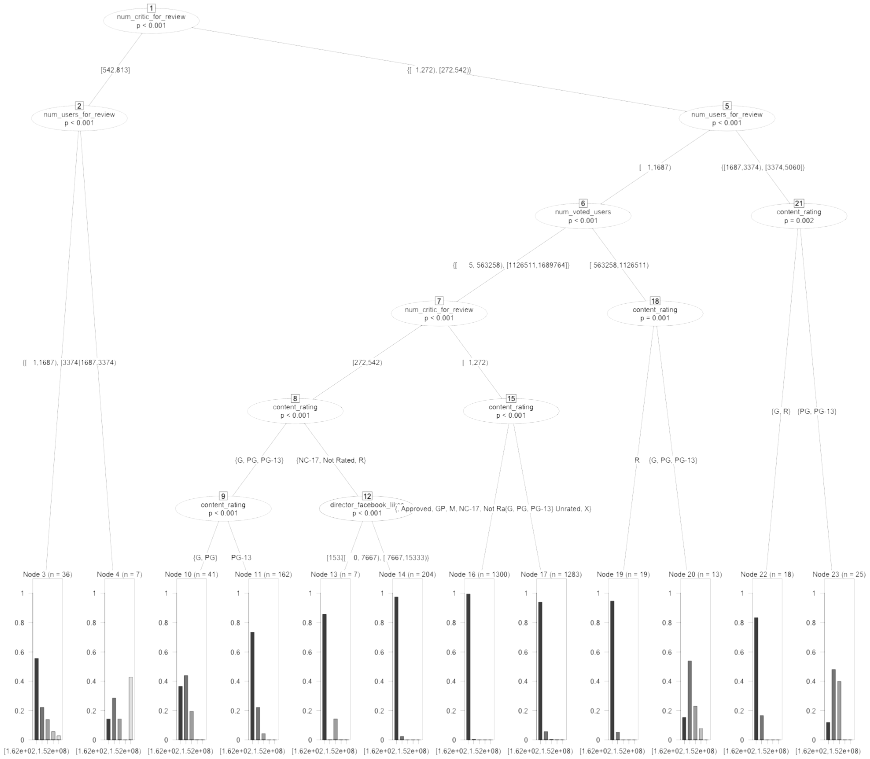

Figure 3: Conditional Inference Tree Plot

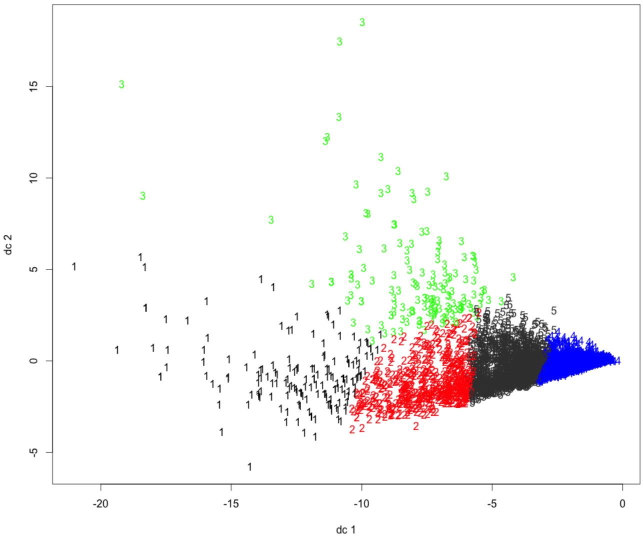

Figure 4: K-Means Clustering with 5 clusters

were increased, but again all the clusters pointed to one category of gross income. Hence this method was not suitable for categorizing gross income values.

In order to access the predicted ratings generated by the models, we determine the mean squared error(MSE). These errors are calculated for each model created using different preprocessed datasets. Following which, a comparison of the error values among all the models is done. The model with least MSE is determine to be the better model.

Mean Squared Error (MSE) is by far the most common measure of numerical model performance. It is simply the average of the squares of the differences between the predicted and actual values. It is a reasonably good measure of performance, although it does overemphasizes the importance of larger errors.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 \tag{3}$$

Accuracy is used to evaluate the decision tree model.

$$Accuracy = TruePositive + TrueNegative/(TruePositive + FalsePositive + TrueNegative + FalseNegative) \tag{4}$$

5 Results

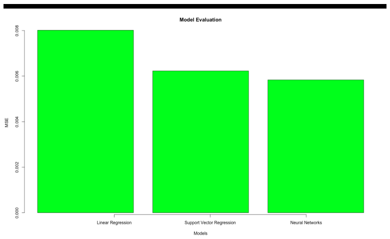

Figure 5: MSE results for different models

Figure 5 shows the MSE for the different models used to predict the IMDb rating.

Using correlation analysis we found out that numberical attributes From the above information, we can infer that the MSE values for neural networks, are the lowest out of the three models used for predicting the IMDb user ratings.

	Mean	Standard Deviation
Dataset	0.6466	0.1057
Linear Regression	0.6451	0.0622
SVR	0.6531	0.06505
ANN	0.6430	0.0788

CART Model gives an accuracy of 0.9354. But it fails to predict data for all the categories of gross income.

Conditional regression tree gives an accuracy of 0.933, which is a bit less than the accuracy of CART Model. But it classifies data into all the categories of gross income. This is due to the fact that most of the data point which were in the dataset belonged to average class in majority. hence this method id favored over the CART Model.

K-Means Clustering predicted all the data points to from one class(Average). this is due to the fact that all the data points belonged to one class majorly. Hence this method was discarded.

6 Conclusion

To predict the rating, we found out that the numerical values were more correlated to the rating and hence were crucial in predicting the rating. We were able to deduce that the most influential features were the number of users who voted, etc.
Combining certain categorical attributes did not help in improving the accuracy of the model.

Out of all the regression models, based on the mean squared error values it is clear that the neural network model was the most accurate in predicting the IMDB user ratings.

In predicting the categories of gross income, Conditional Inference Tree Model was favored, even though it has a little less accuracy than the CART Model. It is due to the fact that it would have a greater accuracy in case of data distributed into all the categories equally.

7 References

[1] M. Joshi, D. Das, K. Gimpel, and N. A. Smith. (2010) Movie reviews and revenues: An experiment in text regression, *NAACL-HLT*

[2] Mladen Marovic , Marko Mihokovic , Mladen Miksa, Sinisa Pribil, & Alan Tus. (2011) Automatic movie ratings prediction using machine learning , *MIPRO*

[3] IMDB 5000 Movie Data Set from Kaggle.com
`http://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset`

[4] Github Repository
`https://github.ncsu.edu/ajagana/ALDA-Project`