Brandon Parmanand

STA 5104 – Term Project

Predicting Movie Audience Score using various predictor variables

# Introduction

## Data Description

The dataset contains 651 instances of movies released in theaters between the period of 1970 – 2014. The variables of the data set include IMBD ratings, rotten tomatoes critics and audience ratings and scores, actors and directors, movie runtime, genre, releases dates, award nominations and wins, and more. Some variables will be deleted as it has no value to my intended statistical analysis such as the list of actors. After doing some initial data exploration, I removed rows with empty data and selected to keep 17 variables.

## Target Problem

Is there a correlation between the critic score and audience score? Or genre and the audience score? How about MPAA rating and audience score? My target problem is to find what variables affect the audience score and whether we can predict the audience score using the variables such as genre, critic rating, IMBD rating, release date, and runtime.

## Statistical Problem

My statistical problem will using the given variables such as genre, MPAA rating, release dates, IMBD rating, and critic scores to predict the audience score.

## Proposed Statistical Analysis

My proposed statistical analysis will be using a linear regression model to predict the audience score of the movies. Some initial research shows that there can be minimal correlation between certain variables to the audience score such as movie runtime.

# Methods

For data exploration methods, I plotted various histograms to see if there is any useful information such as if movies are released more during a specific time during the year. I created two new variables to show whether the movie was released during summer season or not and the same for the award season(see Results). I plotted box plots to visualize correlations between predictor variables and audience score. I also conducted a scatter plot matrix to see correlation between ratings and scores.

My next method was conducting a neural network prediction. I spent time converting all the variables with yes, no answers such as whether the movie was nominated for a best picture award, into binary variables with 1= yes and 0= no. Categorical variables with more than two answers such as genre were converted into dummy variables. This resulted in having a total of 36 predictor variables in my data set. After creating my dummy variables, I used the preprocess function to normalize my data by scaling everything between the range of [0,1]. I then split my 6o percent of my data into the training dataset and the remaining in the validation dataset. Using all of my predictor variables in my neural network, I set the number of nodes to 5. I graphed both the results from my training Neural Network and validation Neural Network into a scatter plot to the acutals.

My final method was using a linear regression model to predict the audience scores. I also had the data separated into a training dataset and validation dataset. However, I did not have the dummy variables or binary variables created as the linear model would take that into account. The data was also not normalized before running the regression. In search of building the best linear regression model for my predictor variables, I used three search algorithms: the forward selection, backward elimination, and the stepwise regression.

# Results and Discussion

Doing research, I know that movies come out during two seasons for the year. There is the summer season, usually in June and July, where producers are known for releasing blockbuster movies where many people are going to the theaters. Movies released during this season are known for bringing the most revenue to the box office. There is also the award season in November through December where producers are releasing movies in anticipation of the award season coming up.

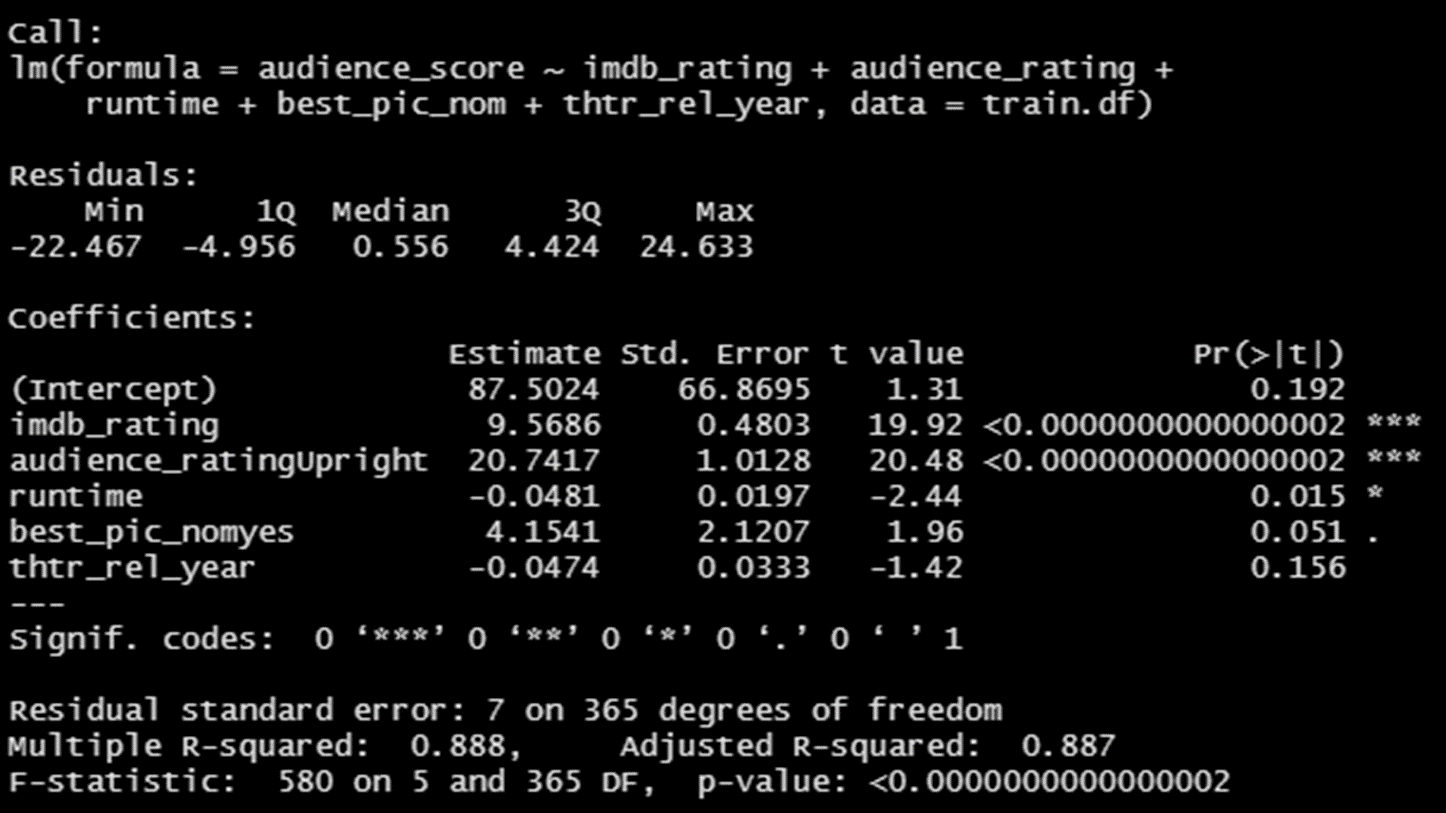
Plotting the results from running the neural network showed a linear relationship between the predicted audience score and the actual audience score on both the training dataset and the validation dataset. I was expecting to have to drop certain predictor variables, but the model did a good job in developing the results.

Chart, scatter chart

Description automatically generated Diagram

Description automatically generated

Using the linear regression model algorithms to determine the best subset of variables helped me narrow the variables down. However, using the forward selection, backward elimination, and stepwise regression all resulted in using the same variables: imdb\_rating, audience\_rating, runtime, best\_pic\_nom, thtr\_rel\_year.

 Timeline

Description automatically generated

Looking into the predicted values for a subset of the validation dataset, we see the residuals only span between about -10 and 10 which is small comparing the highest potential score of 100 and lowest of 0.

# Conclusion

I conclude that we can predict the audience score using the other predictor variables. A Neural Network using all variables on average concluded a predicted value that had little error to the actual value. I also conclude using a linear regression model with fewer variables according to running the stepwise algorithms proved to be a better model to predict audience scores.