Predicting IMDb Movie Ratings

By: Owen Doyle and Justin Self

1: Introduction

At their best, movies are an art form that make people feel some sort of emotion. Some movies entertain us, some make us sad, and others make us question the world around us. Some of the best movies just have something special about them that is difficult to describe, let alone quantify. Yet many people still attempt to describe what makes a good or bad movie, and there are a wide range of tastes and opinions about movies. These are often expressed in movie reviews. We can likely learn a lot about common characteristics of good movies by examining these reviews and data about the making of the movies these reviews are based on. Understanding this could help those who are involved in the movie-making process use their resources more effectively in order to create movies that people will like. Making movies people like also increases profits for these companies.

Previous projects have been done using the same IMDb database. One study used K-means clustering to predict the success of movies. They found that the nature of the data proved difficult to clean and find useful results from. They predicted that word of mouth might be a key component to the success of movies, which is difficult to quantify and is not well represented in the IMDb data (Meenakshi). Another study compared the reliability of ratings between the movie rating websites IMDb, Rotten Tomatoes, Metacritic, and Fandango. Among these sites, they concluded that Metacritic was the most reliable, but IMDb was close behind (Olteanu). Since our study will include the metascore from Metacritic, these claims mean that our data for the movie ratings should accurately represent the true quality of movies.

IMBd allows ordinary people to review movies they have watched. Aside from being a place to review the movies, IMDb contains vast information about the more than 6 million movies it hosts including plot descriptions, release dates, and where to watch the movie. Public websites like these are great resources for data analysis projects as they contain large amounts of data. One such use of this large dataset is to predict average IMDb rating using characteristics of the movie. The goal of our project is to create a model that can best predict the average ratings that viewers will give a movie, in the United States.

2: Data and Methods

The data set was published on Kaggle by Stefano Leone. It was created via web scraping, containing movies with more than 100 votes on the IMDb website. The repository contains four files, each describing various characteristics of movies, actors, and reviewers. For the purposes of this project, only the dataset consisting of variables that pertain to the movie was used.

After obtaining the data set from Kaggle, additional data cleaning was needed. All data cleaning was done in RStudio. To begin with, the data set was trimmed to only contain movies made in the USA, had English as its main language, and more than 500 votes on IMDb. Then the data were clipped to only contain observations, or movies, that did not have any null values. Finally, any movies produced before 1990 were clipped. Clipping the data, while it may introduce bias

into the model, was necessary due to the size of the data set. Aside from what we learned in class, that being that the more data the better, the sheer amount of data was too much to be able to process efficiently.

While the choices that were made in clipping the data may have introduced bias, it helped narrow our data into a useful data set. In general, the goal is to predict movies made in the USA so clipping the data to only include movies that were made in the U.S. in English made sense. The other clipping points, movies made after the 90s and had more than 500 votes, helped create a dataset that would contain the recent preferences of movie viewers. This will help in predicting future movies as the data comes from current movies.

Aside from the necessary clipping of the data, some data manipulation was needed. Specifically, a few of the variables, namely usa gross income, world wide gross income, budget, and year, were read as character variables rather than integers. Having string variables makes data analysis and modeling techniques impossible as we need numerical data to conduct research. To correct this, the variables were coerced into integer form, using rStudio. The simple coercing of the variables' types made all the following data analysis possible.

Another form of data manipulation that was performed was the creation of a new variable titled "votecat", for vote category. This variable creates labels for each movie, namely low, middle, and high, based on the IMDb average vote. The ten point scale was broken into thirds for this categorical variable. The sole purpose of this variable was to help break up the data for visualization purposes. Being able to distinguish between the categories, using color, in visual representations made a world of difference.

The final form of data manipulation that was performed on the data set was the creation of a new variable, director average. This variable sought to capture the quality of each director via the average score of all their movies. The idea was that directors will make a wide range of quality movies. By averaging all of their films, we would get a best guess at the quality of movies they produce. This can then be used as a predictor in predicting a movie's quality rating.

After all the data cleaning and manipulation, exploratory data analysis was performed. The goal of the exploratory data analysis was to determine which variables will be most important in determining the average score that reviewers give a movie. To figure out which variables were most important in predicting the average score, correlation coefficients were calculated and put into a heatmap. In addition, pairwise plots were created to give a good visualization of each variable's distribution. The results of this exploratory analysis will be discussed further in the results section.

In modeling the average vote for movies, regression techniques were required. Predicting any continuous variable can benefit from regression analysis as the output is a continuous quantitative variable. Thus, the methods used in modeling movie ratings were multivariate regression and a few extensions, namely lasso and bagging, an ensemble method of regression. This modeling technique would allow the input of both the continuous and

categorical variables deemed most important from the exploratory analysis as well as output of continuous quantitative variables, average vote.

In order to compare across the regression models that were built, a few metrics were used to assess the quality of the models. The two main metrics that were used in order to assess which models outperformed the rest were root mean squared error and r-squared values. Both of these metrics aim to quantify the amount of variation in the data that is accounted for by each model. With model creation and evaluation in place, the models (multivariate regression, lasso regression, and bagging) were tested.

3: Results

Exploratory Data Analysis

The goal of the exploratory data analysis was to determine which variables will be most important in determining the average score that reviewers give a movie. Figure 1 shows the variables that are likely most highly correlated with the average vote. These variables include duration, budget, USA gross income, and metascore (a weighted average of reviews from reputable critics). Duration, USA gross income, and metascore all have moderate correlation with average vote, while the correlation between budget and average vote is surprisingly low.

Figure 1: Heatmap of Correlations

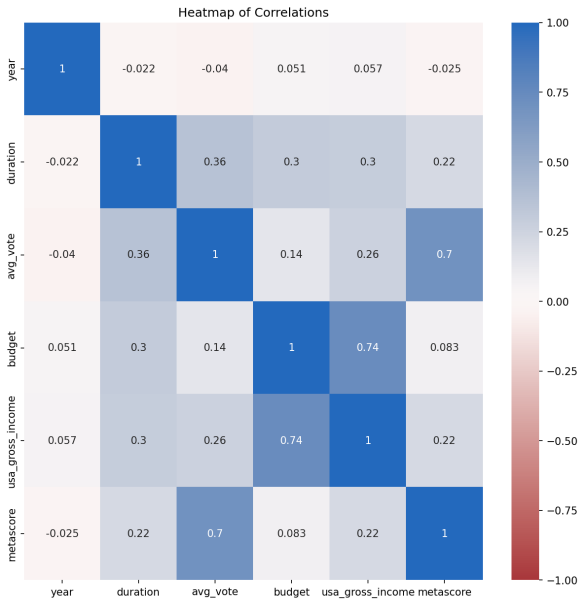
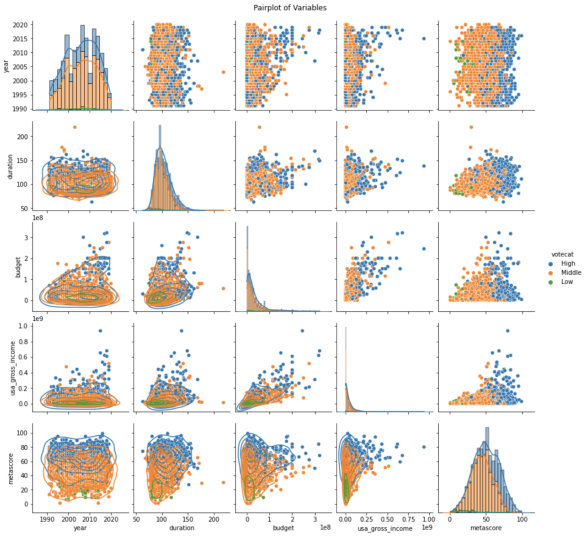


Figure 2 shows relationships between these variables of interest. The class used to separate the colors is based on a range of average votes. High votes are considered any average vote over 6.6. Medium votes are average votes between 3.3 and 6.6. Low votes are average votes below 3.3. The distribution shows that most movies receive average ratings, while fewer receive high ratings and even fewer receive low ratings. From the figure, it appears that the variable that provides the most distinction of ranges of scores is the metascore. Duration and USA gross income also provide some distinction. These three variables will likely be most important to a successful modeling effort.

Figure 2: Pairwise Plots



Modeling

	Coefficient
duration	7.185162e-03
budget	-3.546222e-09
usa_gross_income	1.339059e-09
metascore	2.192859e-02
director_avg	6.653231e-01

Figure 3

Our most successful model was a linear regression model using duration, budget, USA gross income, metascore, and director average as the predictor variables. The goal was to predict the average rating variable. Figure 3 shows the regression coefficients calculated for each of these variables. Metascore and director average were given the most weight, while the coefficients for budget and USA gross income were relatively small. The root mean squared error for this model was approximately 0.52, and the R squared value was approximately 0.71. We created other models that removed a variable, and they did not perform as well as the model with all the variables. A lasso regression model was also tested, and it produces a root mean squared error of approximately 0.64 and an R squared value of approximately 0.55. Finally, we used an ensemble modeling technique, creating a bagging regression model. The root mean squared error for this model was approximately 0.68, and the R squared value was approximately 0.51.

6: References

Meenakshi, K, et al. “A Data Mining Technique for Analyzing and Predicting the Success of Movie.” *Journal of Physics: Conference Series*, IOP Publishing, 1 Apr. 2018, https://iopscience.iop.org/article/10.1088/1742-6596/1000/1/012100.

Olteanu, Alex. “Whose Ratings Should You Trust? IMDB, Rotten Tomatoes, Metacritic, or Fandango?” *FreeCodeCamp.org*, FreeCodeCamp.org, 10 Apr. 2017, https://www.freecodecamp.org/news/whose-reviews-should-you-trust-imdb-rotten-tomatoes-metacritic-or-fandango-7d1010c6cf19/.