MIS545 Data Mining Project

To Predict a Movie's IMDb Rating

1. Problem Statement

Since the onset of Hollywood in the 19 centuries, movie industry has held a significant market share of modern entertainment industry. However, with fierce competitions from gaming and online streaming services, movie industry needs to be ever more sensitive about market demand. Stakeholders including investors, directors and actors need to better understand their customers and make investment decisions with caution.

One figure that indicates whether or not a movie is favored by audience is the IMDb rating. According to the internet site, it takes all the individual ratings cast by IMDb registered users and use them to calculate a single rating with the algorithm of weighted average (IMDb, 2018). Many factors affect one's opinion towards a movie and it is valuable to understand what factors may lead to a high or low rating.

This project is going to evaluate the relationship between movie ratings and multiple movie attribute. The goal is to identify patterns in movie industry and predict movie trend. The analysis is aimed to help stakeholders to explore market opportunities and make business decisions. Three specific business problems will be addressed with this analysis:

No.1: What is the trend of movie rating? Do movie ratings have high volatility through years?

No.2: What are some of the most significant attributes that contribute to a moving rating?

No.3: Predict the rating of a future movie with certain attributes, for example genre, number of voting and movie runtime.

2. Dataset description

The data being used is from IMDb datasets, which is available through IMDb website (IMDb, 2018). The entire IMDb datasets are consisted of six relational databases. Each subset is contained in a tab-separated-values (tsv) formatted file in the UTF-8 character set. A '\N' is used to denote that a particular field is missing or null for that title/name. Two subsets are selected for this study: title.basics and title.ratings. Table 1. lists the dataset composition (IMDb, 2018). Main fields of interest include averageRating, genres, and startYear.

IMDb Subset	Field Name	Data Type	Data Format	Description
title.basics.tsv.gz	tconst	Independent	string	alphanumeric unique identifier of the title
	titleType	Independent	string	the type/format of the title (e.g. movie, short, tvseries, tvepisode, video, etc)
	primaryTitle	Independent	string	the more popular title / the title used by the filmmakers on promotional materials at the point of release

	originalTitle	Independent	string	original title, in the original language
	isAdult	Independent	boolean	0: non-adult title; 1: adult title.
	startYear	Independent	YYYY	represents the release year of a title. In the case of TV Series, it is the series start year.
	endYear	Independent	YYYY	TV Series end year. '\N' for all other title types
	runtimeMinutes	Independent	eMinutes	primary runtime of the title, in minutes
	genres	Independent	string array	includes up to three genres associated with the title
title.ratings.tsv.gz	tconst	Independent	string	-
	averageRating	Dependent	numeric	weighted average of all the individual user ratings
	numVotes	Independent	numeric	number of votes the title has received

Table 1. IMDb Datasets Composition

3. Data preprocessing

Data preprocessing include the following steps:

Data consolidation

Step 1: Merge the two databases by using "tconst" as the primary key.

Data selection

Step 1: The original datasets have included categories other than movie. As I am only interested in movie, a subset of the dataset is created by only selecting "movie" under the "titleType" field. Once the subset of movie is created, the field "titleType" is no longer needed, thus excluded.

Step 2: Based on data screening, the following fields are excluded due to irrelevance to analysis: "tconst", "endYear", "primaryTitle", and "originalTitle".

Data Transformation

Step 1: "\N" is transformed to NA, and any record that contains null value is eliminated.

Step 2: The data type of "startYear" is changed to factor.

Step 3: Categorization:

3.1 "genres" is in the type of string array, with up to three genres separated with comma. The assumption is that genres are placed in the order of

importance, with the first one being most important. Therefore, the first value is kept.

3.2 A new field “numGenre” is created to capture the factor that indicates the variety of the movie. “1” means only one type of genre is listed in the original “genres” field, “2” indicates two, and “3” represents three.

3.3 A new field “ratingRank” is created to categorize a rating. A rating value lower than 5 is Low; greater than or equal to 5 is considered as High, with 10 being the highest rating score.

From the summary of preprocessed data, Table 2, we can see that there are 177,100 records in total. It has a wide range in terms of year, run time and genres. The dataset has six independent variables, with “averageRating” and “ratingRank” being two dependent variables. “averageRating” will be used for statistical and descriptive analysis, while “ratingRank” will be used for predictive analysis.

```
> summary(dataFull)
    isAdult           startYear    runtimeMinutes        numVotes
 Min.   :0.00000   Min.   :1894   Min.   :    1.00   Min.   :      5
 1st Qu.:0.00000   1st Qu.:1973   1st Qu.:   82.00   1st Qu.:     17
 Median :0.00000   Median :1999   Median :   91.00   Median :     64
 Mean   :0.01673   Mean   :1990   Mean   :   93.59   Mean   :   3774
 3rd Qu.:0.00000   3rd Qu.:2011   3rd Qu.:  102.00   3rd Qu.:    347
 Max.   :1.00000   Max.   :2018   Max.   :14400.00   Max.   :1912210

         genres         numGenre     averageRating    ratingRank
 Drama       :50597   Min.   :1.000   Min.   : 1.000   High:150141
 Comedy      :41912   1st Qu.:1.000   1st Qu.: 5.500   Low : 26959
 Documentary:20767   Median :2.000   Median : 6.400
 Action      :18327   Mean   :1.773   Mean   : 6.269
 Crime       : 9234   3rd Qu.:2.000   3rd Qu.: 7.200
 Adventure   : 7405   Max.   :3.000   Max.   :10.000
 (Other)     :28858
> str(dataFull)
'data.frame':    177100 obs. of  8 variables:
 $ isAdult       : int  0 0 1 0 0 0 0 0 0 0 ...
 $ startYear     : int  2015 2016 1972 2017 2011 2015 2016 2015 2010 2015 ...
 $ runtimeMinutes: int  98 66 39 82 143 76 136 96 81 70 ...
 $ numVotes      : int  278 11 6 82 257 11 10 176 337 292 ...
 $ genres        : Factor w/ 27 levels "Action","Adult",..: 1 1 2 3 3 3 3 5 6 6 ...
 $ numGenre      : int  3 3 2 1 3 2 3 3 1 1 ...
 $ averageRating : num  1 1 1 1 1 1 1 1 1 1 ...
 $ ratingRank    : Factor w/ 2 levels "High","Low": 2 2 2 2 2 2 2 2 2 2 ...
```

Table 2. Summary of Working Dataset

4. Statistical summary & descriptive analysis

With the processed data, statistical and descriptive analyses are conducted. The scatter plot between rating and year shows that movie ratings over the year

tend to have tiers. While the middle group is consistently ranging from 5 to 8, which is consistent with the statistical analysis, more and more recent movies are being rated low. The box plot between rating and genre, Table 5, does not show a clear association between the two. However, by comparing the different genres, documentary and sports have the highest average rating, while horror is rated the lowest on average. The plot in Table 6 shows a positive relationship between rating and movie run time, with more voters leading to higher ratings.

Rating Statistical Sumamry	
Mean	6.27
Median	6.40
Standard Deviation	1.31
Skewness	-0.55
Range	9.00
Min	1.00
Max	10.00
Count	177,094

Table 3. Statistical summary of “averageRating”

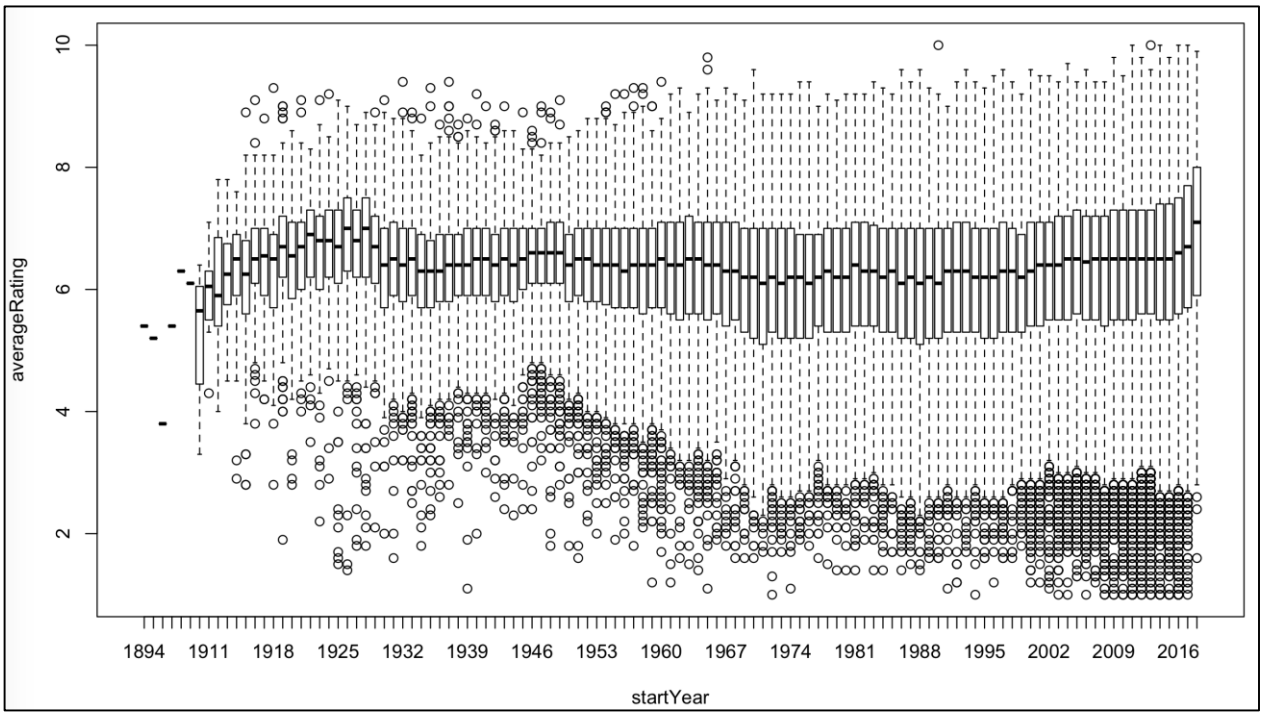

Table 4. Scatter plot: averageRating vs Year

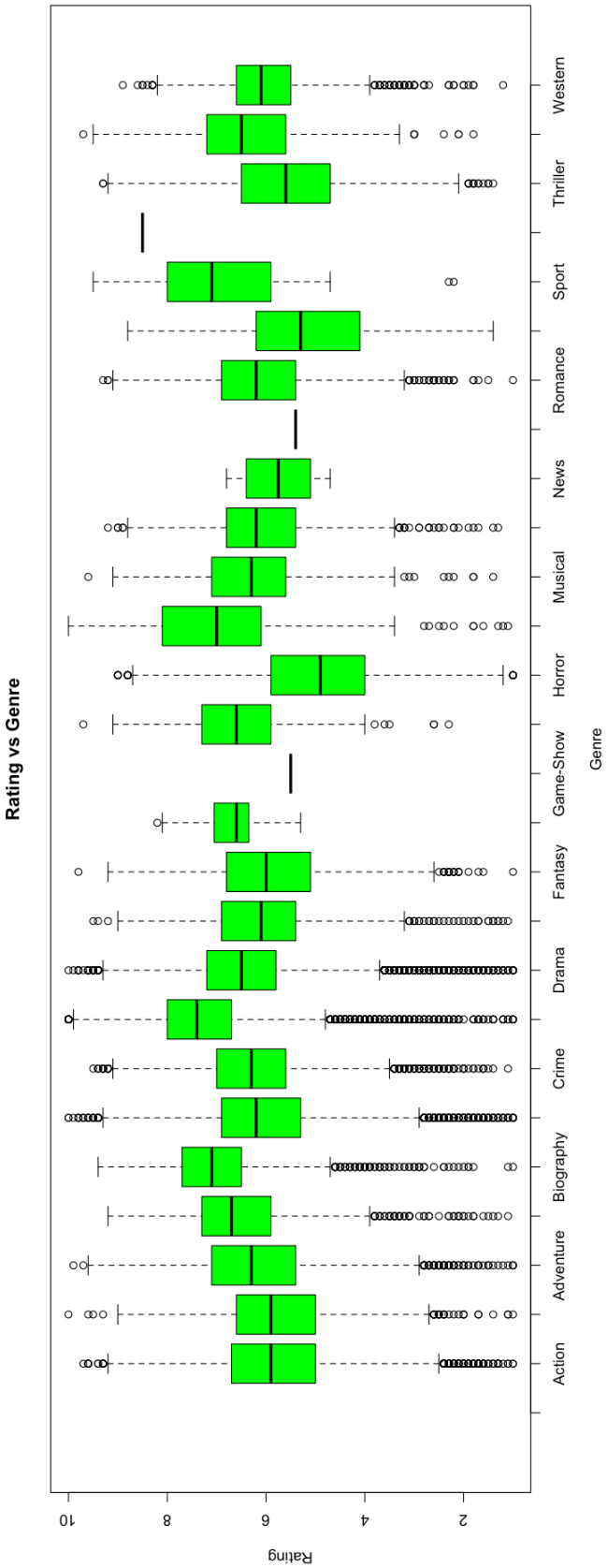

Table 5. Boxplot Rating: averageRating vs Genre

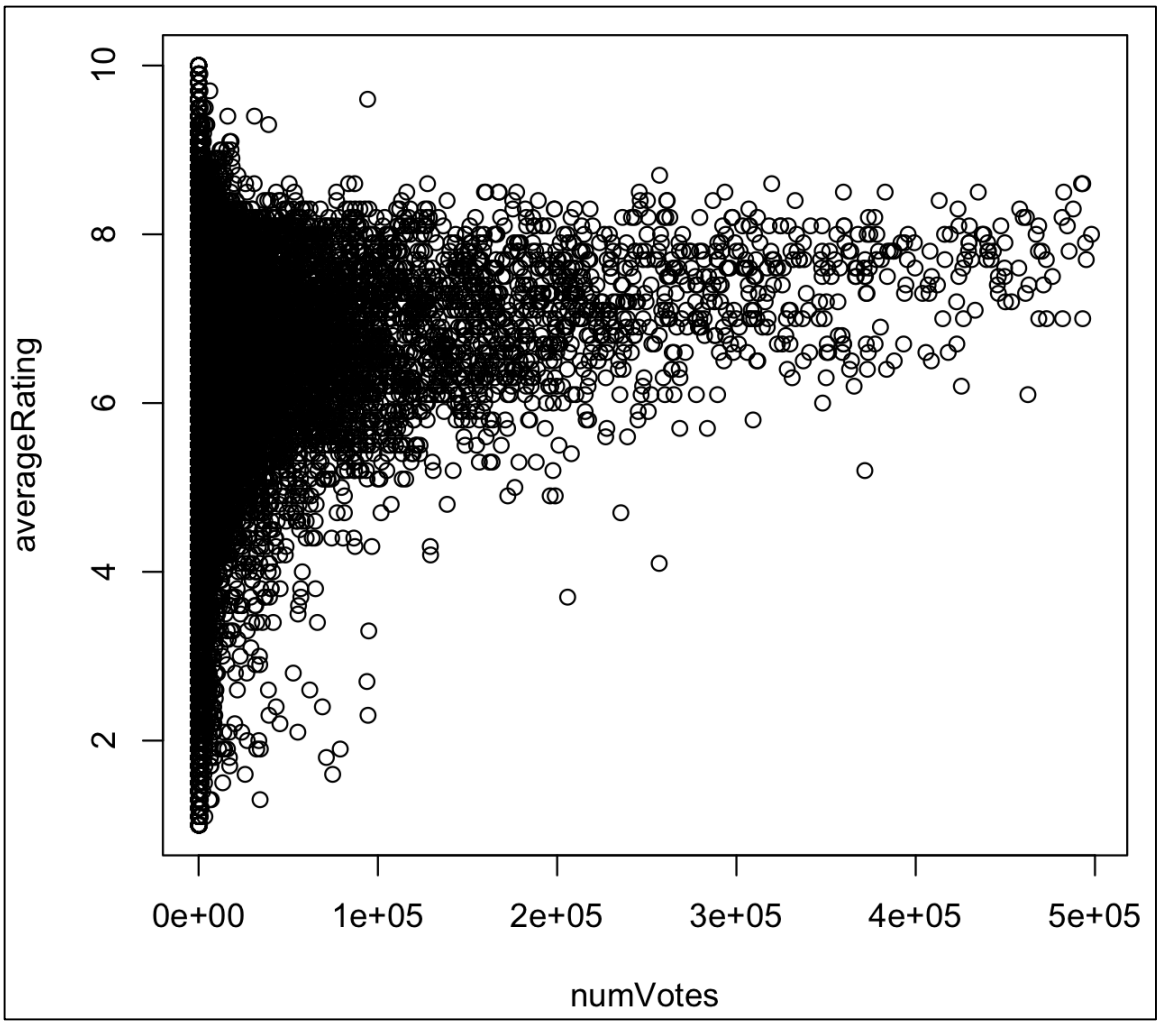

Table 6. Plot Rating: averageRating vs numVotes

5. **Algorithm Slection**

Decision tree has been selected to predict whether a movie will be rated high or low. The dependent variable is ratingRank, not the actual rating value, as a categorical result is more valuable than continuous variable in terms of identifying patterns and relation. Based on previous statistical and descriptive analysis, the following independent variables will be used for predictive analysis:

- isAdult: whether it is adult title or non-adult title
- Genres: the type that the movie is classified
- numGenre: how many different genres the movie is classified
- numVotes: the number of online voting the movie received
- runtimeMinutes: the length of time a movie runs

Naïve Bayes is also employed to predict and test movie rating using the same set of variables.

6. Execution Result & Evaluation

Decision Tree

```
Call:
C5.0.default(x = train[, var_names], y = train$ratingRank)


C5.0 [Release 2.07 GPL Edition]   Sun Mar  4 17:31:10 2018
-------------------------------

Class specified by attribute `outcome'

Read 35420 cases (6 attributes) from undefined.data

Decision tree:

genres in {Action,Adult,Adventure,Animation,Biography,Comedy,Crime,Documentary,
:          Drama,Family,Fantasy,Film-Noir,Game-Show,History,Music,Musical,
:          Mystery,News,Reality-TV,Romance,Sport,Talk-Show,Thriller,War,
:          Western}: High (33826/4725)
genres in {Horror,Sci-Fi}:
:...numVotes > 3364: High (149/28)
    numVotes <= 3364:
    :...numVotes <= 82: High (606/237)
        numVotes > 82: Low (839/288)


Evaluation on training data (35420 cases):

        Decision Tree
      ----------------
      Size      Errors

         4 5278(14.9%)   <<


       (a)   (b)    <-classified as
      ----  ----
     29591   288    (a): class High
      4990   551    (b): class Low


    Attribute usage:

    100.00% genres
      4.50% numVotes


Time: 0.1 secs
```

Naïve Bayes

```
> M.model

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
     High       Low
0.8480185 0.1519815

Conditional probabilities:
      isAdult
Y            [,1]      [,2]
  High 0.01488661 0.1210997
  Low  0.02574173 0.1583680

      runtimeMinutes
Y          [,1]     [,2]
  High 94.09703 59.40307
  Low  91.10700 20.40489

      numVotes
Y           [,1]      [,2]
  High 4269.0048 33340.837
  Low   828.7934  5594.882

      numGenre
Y          [,1]      [,2]
  High 1.773885 0.8161199
  Low  1.765989 0.8055423

      genres
Y          [,1]     [,2]
  High 7.646012 4.517432
  Low  8.086938 5.960119

> M.predict <- predict (M.model, test, type = 'class')
> results <- data.frame (actual = test [ , 'averageRating'] , predicted = M.predict)
> table(results)
      predicted
actual  High   Low
  High 10880 34133
  Low    909  7209
```

Accuracy = (10880+7209)/(10880+34133+909+7209)=18089/53131=34%
Precision = 10880/(10880+909)=92%
Recall = 10880/(10880+34133)= 24%
F-score = 2*0.92*0.24/(0.92+0.24)=0.4416/1.16=0.38

7. Implications & Conclusions

No.1: What is the trend of movie rating? Do movie ratings have high volatility through years?

Based on the study result, movie rating over the years are consistently volatile. A trend in the recent decade is that more movies are being highly rated, while at the same time more movies are getting low rating. Although the average rating for recent movies are higher compared to history.

No.2: What are some of the most significant attributes that contribute to a moving rating?

Descriptive analysis shows each genre has certain rating range, with horror movie rending to be the lowest rated and sports being the highest rated on average. It also shows that there is a positive relationship between amount of voting and rating. It makes sense in terms that a popular movie will attract more people to vote thus it may not be an independent variable.

Based on the result from decision tree, **genre** is a strong influencer on rating. While most genres do not significantly show high or low rating, Horror and Sci-Fi are differentiated from the group. It is an indicator that Horror and Sci-Fi are more critically viewed by audience, thus potentially have a higher risk in succeeding.

An implication from this study is that other factors should be included to predict a more accurate movie rating. For example, investing companies can heavily influence marketing, thus is likely to be a significant influencer on a movie's success. Other potential variances include gender ratio of the casting team, and whether released during holiday.

No.3: Predict the rating of a future movie with certain attributes, for example genre, number of voting and movie runtime.

With the IMDb dataset, the predictive model did not show a high accuracy rate, although the precision rate is 92%. Therefore, the existing dataset does not serve as a reliable base to predict whether a movie will receive a high or low rating based on genres, number of voting and runtime. Given the high stake of movie business, it is worthwhile to explore other factors in the effort of more accurately predicting movie ratings.

Reference

IMDb. (2018, 03 02). *IMDb Datasets*. Retrieved from https://www.imdb.com/interfaces/

IMDb. (2018, 03 02). *IMDb Help Center-FAQ for IMDb Ratings*. Retrieved from IMDb.com: https://help.imdb.com/article/imdb/track-movies-tv/faq-for-imdb-ratings/G67Y87TFYYP6TWAV?ref_=helpsect_pro_2_4#