MOVIE RATING PREDICTIONS FOR NETFLIX

Final Project - Introduction to Data Science
Instructor: Daniel Gutierrez
Author: Glory Kim

SPRING 2020

Table of Contents

1 - Introduction

- **Project Description:** Netflix is an American media-services provider and production company who initially built their business through DVD sales and rentals. Through a powerful on-demand streaming platform, Netflix has become an everyday part of life and evolved the way we watch the media, from movies to programs. Netflix created the opportunity to connect directly with the viewers and gather millions of data from viewing preferences, viewer information, and program information, spawning endless possibilities in data analytics
- **Data Set**: The dataset used for this project was acquired through Kaggle and comes from two sources:
  - https://www.kaggle.com/netflix-inc/netflix-prize-data: The first dataset comes from a Netflix competition for the best algorithm to predict user ratings for films. This set includes the movie title and the corresponding customer ratings.
  - https://www.kaggle.com/shivamb/netflix-shows: The second dataset comes from a third-party search engine called Flixable and includes details specific to the movie, i.e. Director, cast, category, etc.
- **Goal:** My goal is to provide a step-by-step analysis of identifying the algorithm that will predict Customer Ratings by expanding on the Netflix competition data to integrate feature variables or predictors from the Flixable data.
- **Hypothesis:** This project is to validate if Customer Rating can be predicted based on Genre, Movie Duration, and MPA Rating.

File Contents

FIle	Description
Netflix Prize Data: combined_data_1.txt, combined_data_2.txt, combined_data_3.txt, combined_data_4.txt	Over a million customer ratings, scaling from 1-5, of 17,770 movies from 480k randomly chosen, anonymous customers. Data was collected between October 1998 to December 2005.
Netflix Prize Data: movie_titles.csv	17,770 movies and their date of release
Flixable Data: netflix_titles.csv	6,234 tv shows and movies available on Netflix as of 2019 and their details such as movie description, cast, etc.

2 – Data

2.1 – Data Access

- Copied directly from the Kaggle data repository:
    - **Netflix Prize Data: “**combined_data_%.txt”
        - The first line of each file contains the movie id followed by a colon
            - MovieIDs range from 1 to 17770 sequentially.
        - Each subsequent line in the file corresponds to a rating from a customer and its date in the following format: CustomerID, Rating, Date
            - CustomerIDs range from 1 to 2649429, with gaps. There are 480189 users.
            - Ratings are on a five-star (integral) scale from 1 to 5.
            - Dates have the format YYYY-MM-DD.
    - **Netflix Prize Data: “**movie_titles.csv”
        - Format (Attributes): MovieID, YearOfRelease, Title
        - MovieID do not correspond to actual Netflix movie ids or IMDB movie ids.
        - YearOfRelease can range from 1890 to 2005 and may correspond to the release of corresponding DVD, not necessarily its theaterical release.
        - Title is the Netflix movie title and may not correspond to titles used on other sites. Titles are in English.
    - **Flixable Data:** “netflix_titles.csv”
        - Format (Attributes): show_id, type, title, director, cast, country, date_added, release_year, rating, duration, listed_in, description

2.2 – Data Transformation

Final results of the transformation have been provided below for dataframe ‘df’:

- The final dataframe has 13 columns and 9,076,397 rows.

```
> dim(df)
[1] 9076397       13
```

- Summary provides basic characteristics of each column.

```
> summary(df)
MovieID       Movie_Title     Year_Of_Release Age_Rating Movie_Duration_min
Min.   :   30 Length:9076397    Min.   :1942 R      :4078662  Min.   : 18.0

1st Qu.: 5730 Class :character  1st Qu.:1994 PG-13 :3501070  1st Qu.:101.0

Median :10774 Mode  :character  Median :1999 PG    :1220285  Median :118.0

Mean   : 9928                   Mean   :1997 G     : 214597  Mean   :118.1

3rd Qu.:14454                   3rd Qu.:2003 TV-14 :  24136  3rd Qu.:131.0

Max.   :17697                   Max.   :2005 TV-PG :  22928  Max.   :224.0

                                             (Other):  14719

Genre                                                     Director
Action & Adventure                              : 841363   Length:9076397
Dramas                                          : 730274   Class :character
Action & Adventure, Comedies                    : 604952   Mode  :character
Dramas, Thrillers                               : 440659
Comedies, Romantic Movies                       : 428820
Action & Adventure, Comedies, Sci-Fi & Fantasy: 426773
(Other)                                         :5603556

Date_Added_Netflix   Listed_Country     CustomerID        Cust_Rating
Min.   :2011-09-27   Length:9076397     Length:9076397     Min.   :1.000
1st Qu.:2019-07-01   Class :character   Class :character   1st Qu.:3.000
Median :2019-10-19   Mode  :character   Mode  :character   Median :4.000
Mean   :2019-08-07                                         Mean   :3.653
3rd Qu.:2020-01-01                                         3rd Qu.:4.000
Max.   :2020-01-15                                         Max.   :5.000


Movie_Duration_Ranges_min Cust_RatingC
(90,120] :4602215         1: 349537
(120,150]:3118002         2: 849509
(60,90]  : 739024         3:2548891
(150,180]: 425399         4:3182842
(180,210]: 183683         5:2145618
(210,240]:   6085
(Other)  :   1989
```

- The title of each column can be seen below.

```
> names(df)
[1]  "MovieID"        "Movie_Title"            "Year_Of_Release"
[4]  "Age_Rating"     "Movie_Duration_min"     "Genre"
[7]  "Director"       "Date_Added_Netflix"     "Listed_Country"
[10] "CustomerID"     "Cust_Rating"            "Movie_Duration_Ranges_min"
[13] "Cust_RatingC"
```

- The first 6 rows can be seen below.

```
> head(df)
MovieID Movie_Title Year_Of_Release Age_Rating Movie_Duration_min       Genre
1       30 Something's Gotta Give   2003   PG-13   128 Comedies, Romantic Movies
2       30 Something's Gotta Give   2003   PG-13   128 Comedies, Romantic Movies
3       30 Something's Gotta Give   2003   PG-13   128 Comedies, Romantic Movies
4       30 Something's Gotta Give   2003   PG-13   128 Comedies, Romantic Movies
5       30 Something's Gotta Give   2003   PG-13   128 Comedies, Romantic Movies
6       30 Something's Gotta Give   2003   PG-13   128 Comedies, Romantic Movies

Director D ate_Added_Netflix  Listed_Country  CustomerID Cust_Rating   Movie_
Duration_Ranges_min
1 Nancy Meyers    2019-08-01  United States    1204833           4   120,150]
2 Nancy Meyers    2019-08-01  United States     124669           3  (120,150]
3 Nancy Meyers    2019-08-01  United States     694798           4  (120,150]
4 Nancy Meyers    2019-08-01  United States    2496005           4  (120,150]
5 Nancy Meyers    2019-08-01  United States     465897           4  (120,150]
6 Nancy Meyers    2019-08-01  United States     804510           4  (120,150]

Cust_RatingC
1   4
2   3
3   4
4   4
5   4
6   4
```

- Based on the observations above, I will explore the following variables, which includes my hypothesized variables, to predict Customer_Rating:

Feature Variables	**Description**	**Variation**
Age_Rating	MPA film rating to rate the movie's suitability for certain audiences based on content	10 (factor)
Movie_Duration_min	the length of the movie measured in minutes	8-224 min
Movie_Duration_Ranges_min	movie length bucketed into 30 min ranges	8 (factor)
Year_Of_Release	the year the movie was released to the public	42 discrete years (1942-2005)

The next step would be to perform some preliminary analysis to identify if there is a correlation between Customer_Rating and the feature variables listed.

3 – Exploratory Data Analysis

3.1 – Numeric Exploratory Data Analysis

In this section, I performed basic statistical techniques to analyze the whole dataset. The most insightful tools are displayed below:

1. The function prop.table(table(df$Age_Rating))*100 provides the distribution of MPA ratings. R-rated movies had the highest frequency at 44.9% followed by PG-13 movies at 38.5%.

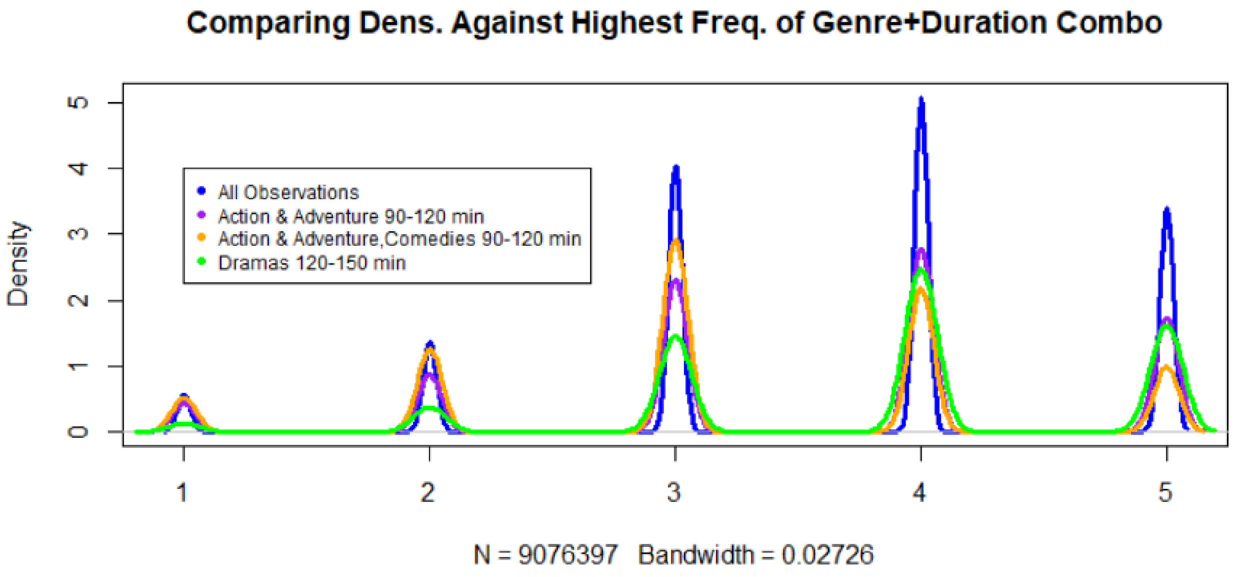
```
> prop.table(table(df$Age_Rating))*100
```

Freq. % by Age_Ratng	G	NR	PG	PG-13	R	TV-14	TV-G	TV-MA	TV-PG	TV-Y7
	2.36%	0.01%	13.44%	38.57%	44.93%	0.26%	0.05%	0.08%	0.25%	0.01%

2. Per the prop.table function, about 17% of the data is related to Action movies that range between 90-150 min. The 6 combinations make 25.6% of the total distribution. The density of the 3 highest results were then overlaid against the total Customer_Rating density on the graph below. In all categories, most of the Customer Ratings resided between 3 and 4.

```
> t5 <- as.data.frame(prop.table(table(df$Genre,df$Movie_Duration_Ranges_m
in))*100)
> t6 <- subset(t5, prop.table(table(df$Genre,df$Movie_Duration_Ranges_min)
)*100 > 3)
```

```
                                          Var1      Var2     Freq
277                         Action & Adventure  (90,120] 5.395930
286               Action & Adventure, Comedies  (90,120] 4.532140
291 Action & Adventure, Comedies, Sci-Fi & Fantasy  (90,120] 3.194252
369                         Action & Adventure (120,150] 3.873861
443                                     Dramas (120,150] 4.358778
451                         Dramas, Thrillers (120,150] 4.276796
```



3. In order to evaluate whether the predictors were dependent on the response variable, Customer Ratings, the Chi-Squared statistical test was performed below. The p-values scored below 0.05, indicating a strong dependency.

```
> X1 <- chisq.test(df$Age_Rating,df$Cust_Rating)
> X2 <- chisq.test(df$Movie_Duration_Ranges_min,df$Cust_Rating)
> X3 <- chisq.test(df$Genre,df$Cust_Rating)
> X4 <- chisq.test(df$Year_Of_Release,df$Cust_Rating)
> X5 <- c(X1$p.value,X2$p.value,X3$p.value,X4$p.value)
> X5 < .05
[1] TRUE TRUE TRUE TRUE
```

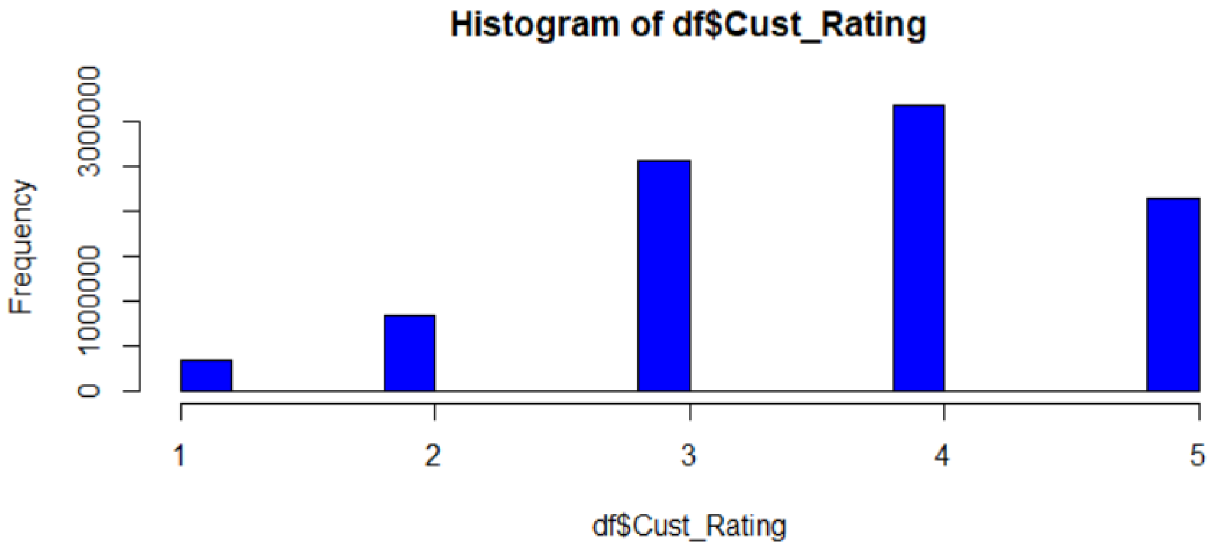
4. The final numerical evaluation was using Pearson correlation between the continuous variables. A perfect correlation would be defined as 1. Based on the results, “Year_Of_Release” and “Movie_Duration_min” had a low correlation to Customer Ratings. In particular, the “Year_Of_Release” predictor had conflicting findings against the Chi-Squared test, which proves to show high dependency.

```
> cor(df[,c(3,5,11)])
                    Year_Of_Release Movie_Duration_min Cust_Rating
Year_Of_Release          1.00000000         0.03773258  -0.1158073
Movie_Duration_min       0.03773258         1.00000000   0.1277536
Cust_Rating             -0.11580729         0.12775362   1.0000000
```

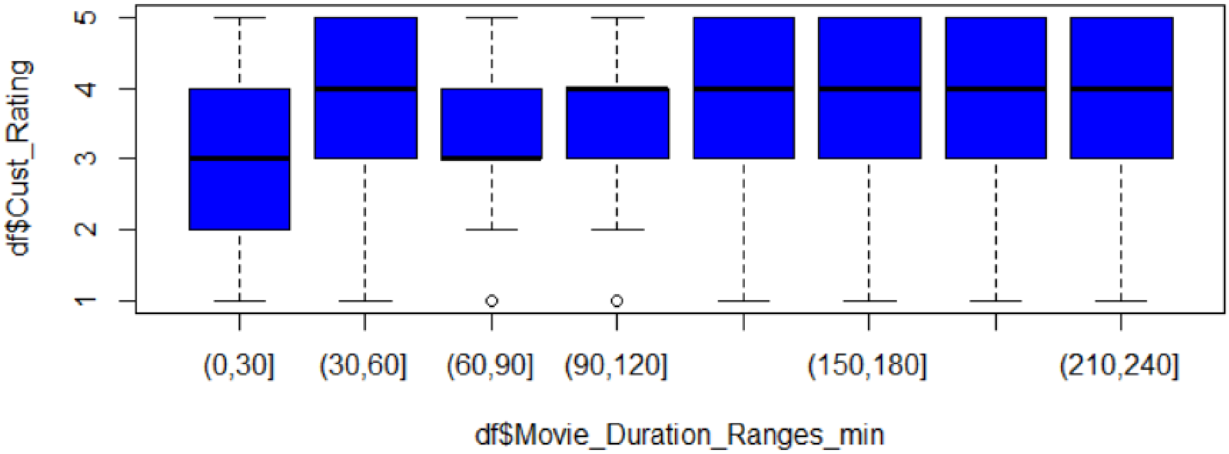
The proceeding sections takes exploratory analysis one step deeper through visualizations.

3.2 – Exploratory Visualization

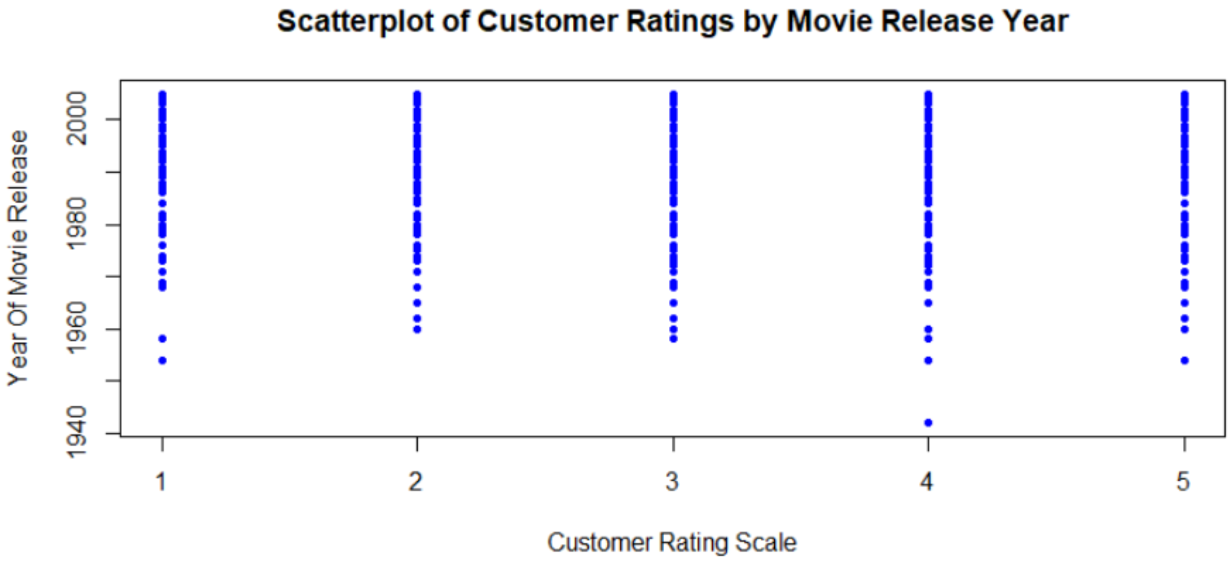
- The histogram below shows the number of observations for each Customer Rating (on a scale of 1-5). At about 3 million observations out of 9 million, customers gave the most ratings of 4.



- The boxplot below shows the distribution of Customer Ratings by Movie Duration by 30 min ranges. The following interpretations can be made:
  - Most ratings fell at 3 and above, and more consistently if the movie was over 90 min.
  - 75% of movies between 60-120 min had ratings 4 and below.
  - Overall, the longer the movie, the higher the rating.



- The scatterplot below evaluates the distribution of Customer Ratings based on Year of Movie Release. I used a sample set of 10,000 records for better visualization. Majority of the distribution sat above the year 2000 and in the higher ratings.



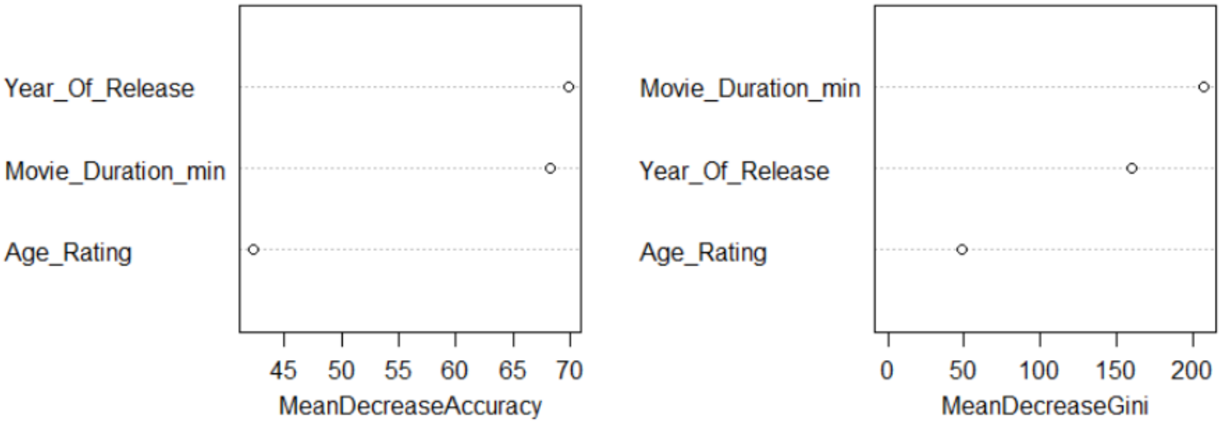
4 – Machine Learning

In this section, I used the randomForest algorithm to predict Customer Ratings and verify my hypothesis. I chose this algorithm because I wanted my predictions to yield Customer Ratings categorically. Because of the sheer volume of data, I continued to use my sample set of 10,000 observations to break out my training set (60%) and testing set (40%).

- Once I trained the algorithm using the randomForest function, I ran the test data to make predictions. Then I compared the predicted Customer Ratings against the actual test Customer Ratings, shown below in a confusion matrix. The highest accuracy rate was for Customer Rating = 4 at 53.3% (762/1,428 records). The calculation for the total misclassification error rate was 62%. Both represent a poor prediction rate, discrediting my original hypothesis.

Prediction	**1**	**2**	**3**	**4**	**5**
1	3	3	2	10	4
2	3	2	2	5	2
3	63	166	439	381	150
4	61	163	548	762	455
5	18	41	130	270	317
Total:	148	375	1121	1428	928

- Based on further investigation using the importance function and plot below, the feature variable “Age_Rating” was the cause of the high error rate (higher numbers indicate a more important predictor).



5 – Conclusion

I began with the exploratory statistical analysis to get a good understanding of the dataset. My first analysis was to evaluate the distribution of the different variables and their relationships to one another. I found that most of the observations were concentrated around Customer Ratings of 3 and 4 and more current movies. There was also indication for a positive relationship between movie length and Customer Ratings.

In the next step , I used some basic statistical concepts to evaluate the predictors against the response variable, which showed conflicting results for the "Year_Of_Release" predictor. The Chi-Squared test showed high dependency to the response variable whereas the Pearson function showed a low correlation. This leads me to believe that either (a) one of the tests are not conformed to the nature of the data, i.e. factors, or (2) dependency exists but not the type of correlation that Pearson expects.

The final step in the analytical process was train the randomForest algorithm and evaluate the accuracy of the predictions. Unfortunately, my hypothesis was disproved with a misclassification error rate was 62%, mostly due to the Age_Rating (MPA Rating) variable. My next steps to improve predictability would be to (a) eliminate this variable and re-run the algorithm or (b) use other predictors, i.e. Genre, to determine Customer Ratings.

Please refer to R Script for detailed analysis.