



# IMDB User Rating and Movie's Gross Income Prediction


Rachit Thirani (rthiran)

Arun Jaganathan (ajagana)

Vivek Varma Nadimpalli (vvnadimp)



# Contents

- Introduction
  - Data Set Description
  - Preprocessing Techniques
  - Methods
  - Results
  - Conclusion
- 



# Introduction

- IMDB – Internet Movie Database
- Movie ratings and gross income are based on various factors like
  - popularity of the actors and the directors,
  - the genre of the movie,
  - reviews by critics etc.
- Producers -> Can utilize it to predict how well the movie may perform at the box office and if it is a wise decision for him to invest in the movie.
- Directors -> Can utilize which combination of actors and genre is better for them.
- Public -> Can utilize it to decide what movies to watch.

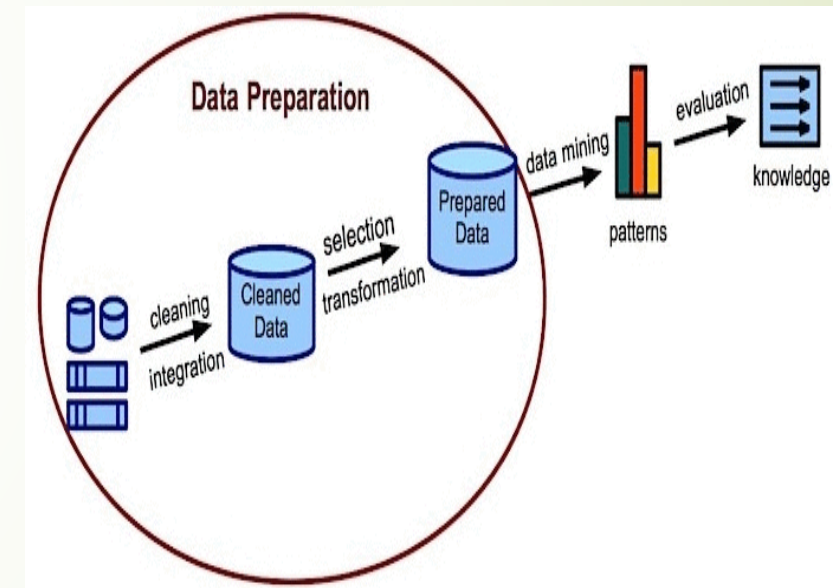
# Data Set Description

- IMBD 5000 Dataset
- 5043 rows, 28 attributes
- Categorical and Numerical attributes

```
'data.frame': 5043 obs. of 28 variables:
 $ color          : Factor w/ 3 levels "", "Black and White",...: 3 3 3 3 1 3 3 3 3 ...
 $ director_name  : Factor w/ 2399 levels "", "A. Raven Cruz",...: 924 796 2023 375 602 101 2026 1648 1220 551
 ...
 $ num_critic_for_reviews : int 723 302 602 813 NA 462 392 324 635 375 ...
 $ duration       : int 178 169 148 164 NA 132 156 100 141 153 ...
 $ director_facebook_likes : int 0 563 0 22000 131 475 0 15 0 282 ...
 $ actor_3_facebook_likes : int 855 1000 161 23000 NA 530 4000 284 19000 10000 ...
 $ actor_2_name     : Factor w/ 3033 levels "", "50 Cent", "A. Michael Baldwin",...: 1407 2218 2489 534 2433 2548
 1228 801 2440 653 ...
 $ actor_1_facebook_likes : int 1000 40000 11000 27000 131 640 24000 799 26000 25000 ...
 $ gross           : int 760505847 309404152 200074175 448130642 NA 73058679 336530303 200807262 458991599 30
 1956980 ...
 $ genres          : Factor w/ 914 levels "Action", "Action|Adventure",...: 107 101 128 288 754 126 120 308 126
 447 ...
 $ actor_1_name    : Factor w/ 2098 levels "", "50 Cent", "A.J. Buckley",...: 266 978 351 1965 524 439 784 218 33
 4 32 ...
 $ movie_title     : Factor w/ 4917 levels "#Horror\345\312",...: 397 2726 3275 3706 3328 1957 3285 3457 398 16
 27 ...
 $ num_voted_users : int 886204 471220 275868 1144337 8 212204 383056 294810 462669 321795 ...
 $ cast_total_facebook_likes: int 4834 48350 11700 106759 143 1873 46055 2036 92000 58753 ...
 $ actor_3_name    : Factor w/ 3522 levels "", "50 Cent", "A.J. Buckley",...: 3439 1392 3131 1765 1 2711 1967 216
 0 3015 2939 ...
 $ facenumber_in_poster : int 0 0 1 0 0 1 0 1 4 3 ...
 $ plot_keywords   : Factor w/ 4761 levels "", "10 year old|dog|florida|girl|supermarket",...: 1320 4283 2080 34
 84 1 652 4745 29 1134 2007 ...
 $ movie_imdb_link : Factor w/ 4919 levels "http://www.imdb.com/title/tt0006864/?ref=fn_tt_tt_1",...: 2965 272
 1 4533 3756 4918 2476 2526 2458 4546 2551 ...
 $ num_user_for_reviews : int 3054 1238 994 2701 NA 738 1902 387 1117 973 ...
 $ language       : Factor w/ 48 levels "", "Aboriginal",...: 13 13 13 13 13 13 13 13 13 13 ...
 $ country        : Factor w/ 66 levels "", "Afghanistan",...: 64 64 63 64 64 64 64 64 63 ...
 $ content_rating  : Factor w/ 19 levels "", "Approved",...: 9 9 9 9 1 9 9 8 9 8 ...
 $ budget         : num 2.37e+08 3.00e+08 2.45e+08 2.50e+08 NA ...
 $ title_year     : Factor w/ 91 levels "1916", "1920",...: 84 82 90 87 NA 87 82 85 90 84 ...
 $ actor_2_facebook_likes : int 936 5000 393 23000 12 632 11000 553 21000 11000 ...
 $ imdb_score     : num 7.9 7.1 6.8 8.5 7.1 6.6 6.2 7.8 7.5 7.5 ...
 $ aspect_ratio   : num 1.78 2.35 2.35 2.35 NA 2.35 2.35 1.85 2.35 2.35 ...
 $ movie_facebook_likes : int 33000 0 85000 164000 0 24000 0 29000 118000 10000 ...
```

# Data Preprocessing Techniques

- As part of pre-processing, we did the following:
  - 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 a 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.





# Methods



- For IMDB User Rating Predictions:
  - Multivariate Linear Regression
  - Artificial Neural Networks
  - Support Vector Regression
- For Movie's Gross Income Prediction:
  - Decision Tree
    - CART
    - Conditional Inference Tree
  - Clustering
    - K Means

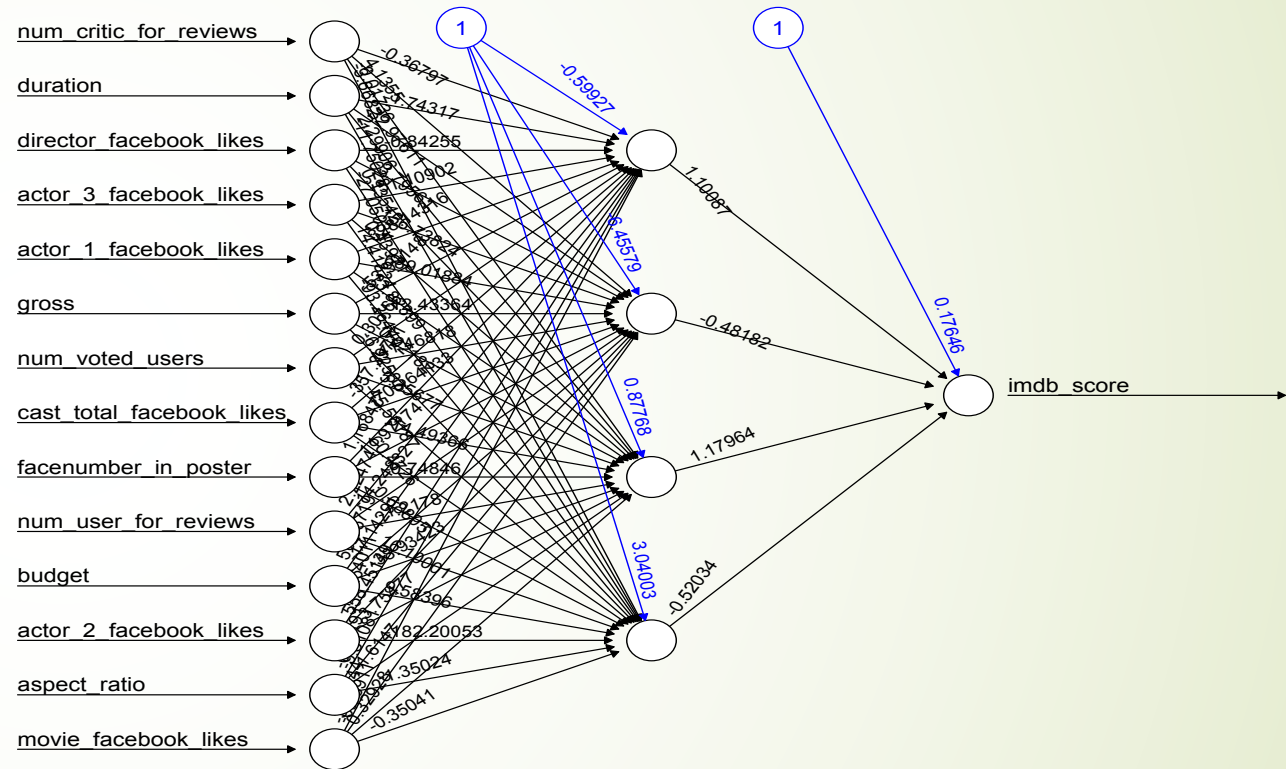


# Multivariate Linear Regression

- Supervised Learning Algorithm
- Popular Regression Model
- Continuous output which is dependent on 1 or more inputs.
- $Y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_px_p + c$ 
  - where  $x_1, x_2, \dots, x_n$  are the different attributes and  $\beta_0, \beta_1, \dots, \beta_n$  are the constants.
- The numerical attributes were **taken** and used to create a model to predict the user-rating for a movie.
- The obtained MSE value for this model was 0.008015355539.

# Artificial Neural Networks

- Neural Networks are trained using back propogation.
- Can approximate almost any function.
- All numeric attributes were taken as inputs and 4 hidden layers were chosen as it reduced the error.
- The MSE value obtained was 0.005837021496.



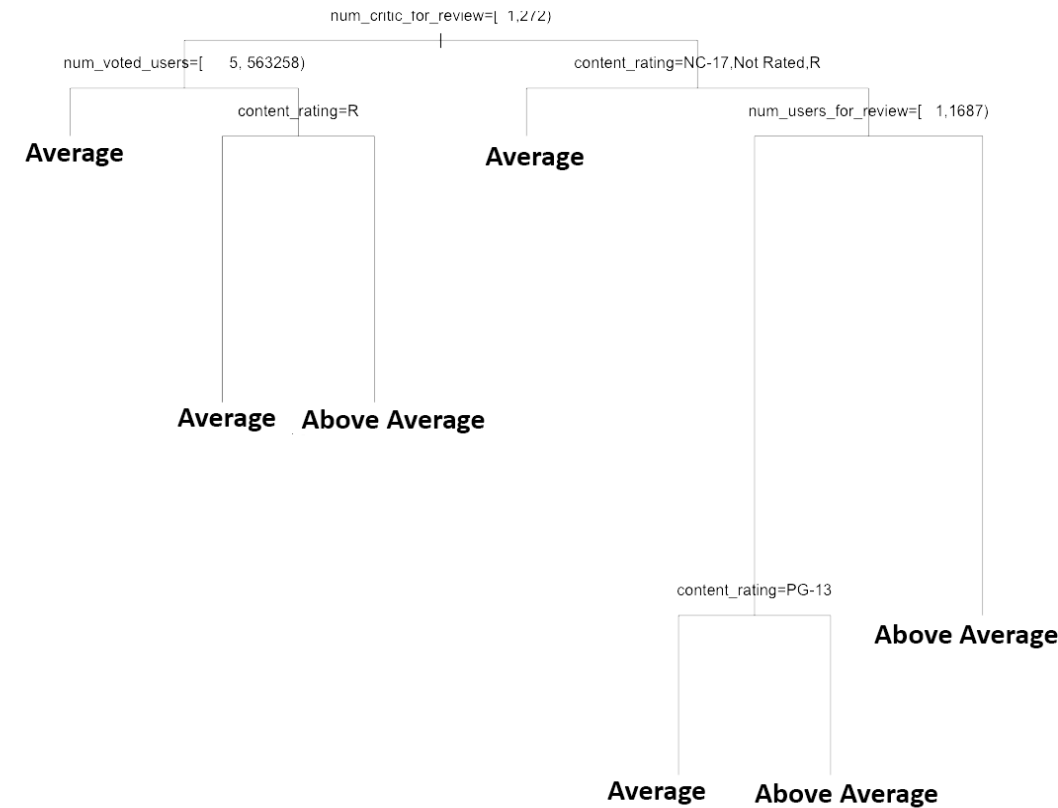


# Support Vector Regression

- SVR uses similar principles as Support Vector Machines.
- SVR maximizes margins, so its slightly more robust.
- It supports kernels so, even non-linear functions can be modeled.
- The MSE value obtained on the IMDB dataset was 0.006228478239.

# CART Modeling

- Used for categorizing gross income
- Binary Tree
- Obtained by recursively partitioning the data space and fitting a simple prediction model within each partition
- Graphical
- Simple and easy to implement

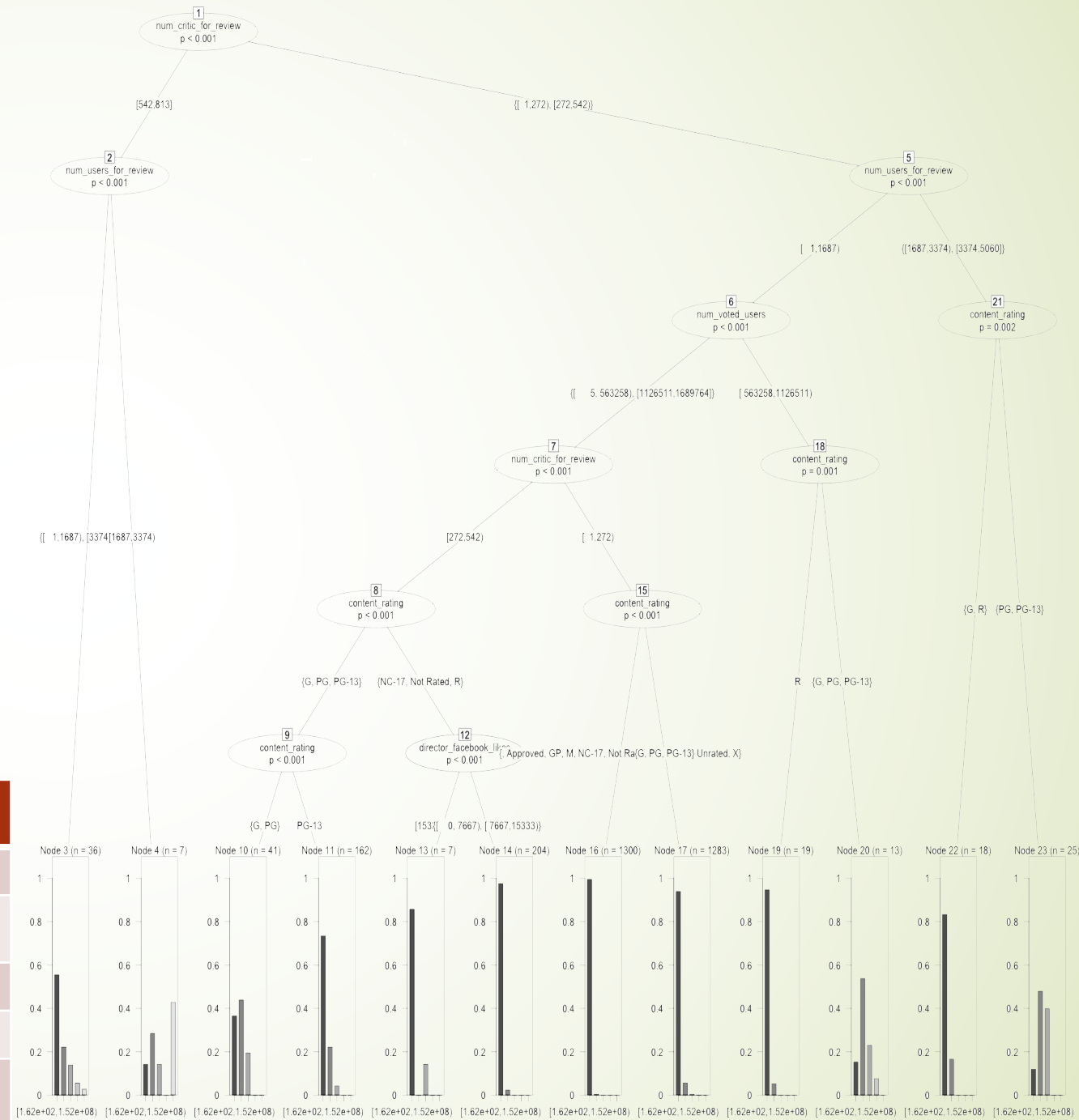


Predicted-> Actual v	Average	Above Average	High	Very High	Extremly High
Average	959	47	4	0	0
Above Average	7	12	7	1	1
High	0	0	0	0	0
Very High	0	0	0	0	0
Extremly High	0	0	0	0	0

# Conditional Inference Tree

- uses a significance test procedure in order to select variables (permutation tests)
- conditional inference procedures
- the algorithm tests if any independent variables are associated with the given response variable, and chooses the variable that has the strongest association with the response.

Predicted-> Actual v	Average	Above Average	High	Very High	Extremely 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
Extremely High	0	2	2	0	0

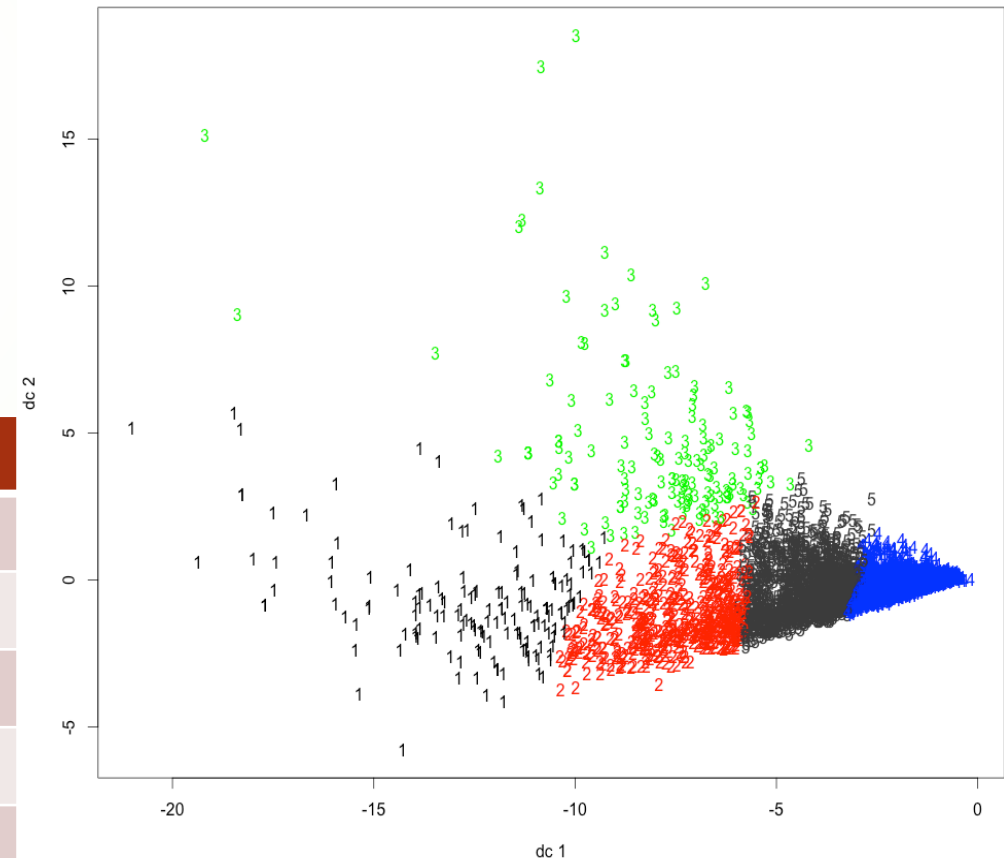


# K Means Clustering

- Unsupervised learning algorithm.
- Simple and easy way to cluster.
- Aims to partition  $n$  observations into  $k$  clusters.

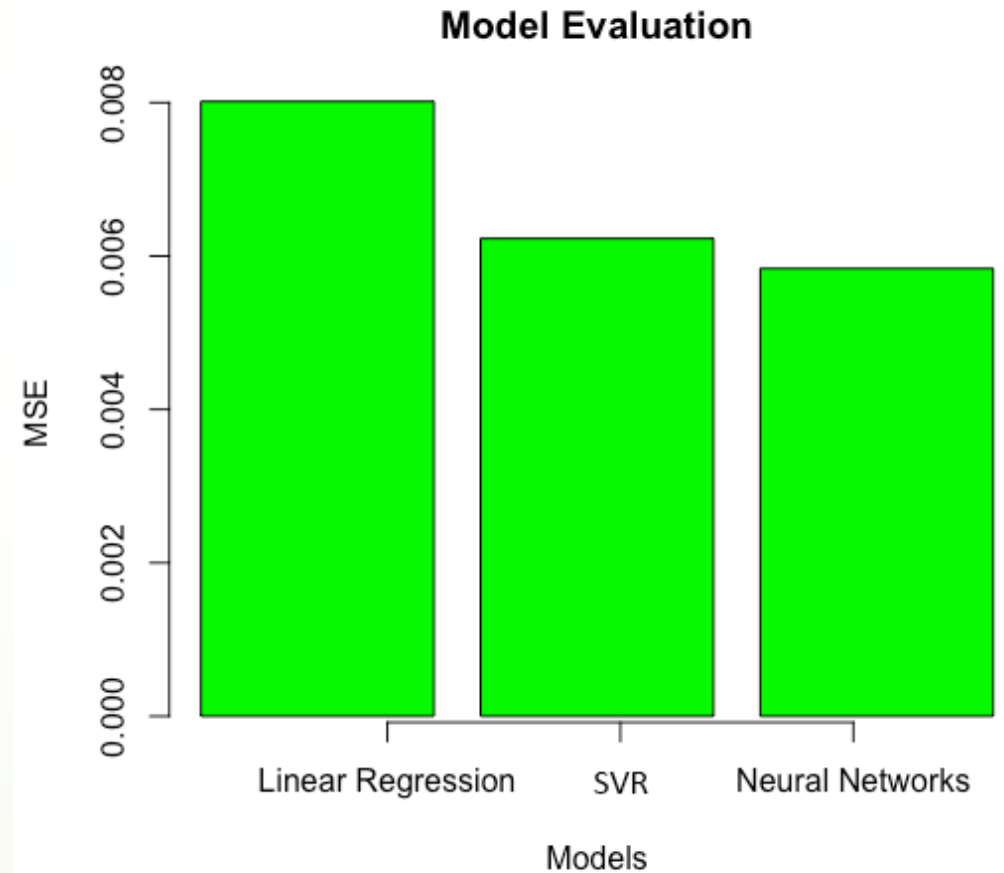
Cluster Number	Average	Above Average	High	Very High	Extremely High
1	96	43	21	1	4
2	517	53	5	0	0
3	81	45	21	3	1
4	1841	16	0	0	0
5	1130	72	4	0	0

K-means clustering with 5 clusters



# Results for IMDB User Rating Prediction

- Mean Squared Error is used to compare the different models created.
- From the plot, Neural Networks seem to perform better in predicting the user rating.
- This could be because it is more adept at modeling complex functions when compared to linear regression and SVR.







# Results for Gross Income Prediction

- ▶ CART Model gives a accuracy of 0.9354
- ▶ Conditional Inference tree gives an accuracy of 0.93352
  - ▶ Most of the data points belong to a single class (“Average” gross income).
- ▶ For K means clustering:
  - ▶ In every cluster, the majority of the data point belong to one class, therefore, we are predicting the same class for every data point.



# Conclusion



- From the result of our analysis, it was found that movie ratings were highly influenced by the numerical attributes and not the categorical attributes.
- Features like number of critics reviewed and number of Facebook likes for the cast and director played a major role in deciding the IMDB rating a user might give whereas attributes like country, language, etc did not play any role.
- From the gross income prediction, it was found that most of the movies lie in one category of gross income.
- Features such as Content rating, number of users who voted for the movie, number of reviews played a major role in predicting the gross income.