Harvard X Capstone Project Movielens

Motolani Ojo-Bello

2022-10-29

1 INTRODUCTION

INTRODUCTION

The purpose of this project is to build a model that predicts the rating of a movie. The edx data set provided is used to train the model and the final evaluation of the model using the Root Mean Squared Error(RMSE) is carried out on the validation data set. Both data sets(edx and validation) have the same variables:

- rating: The rating of the movie by a particular user on a scale of 0.5 to 5 at intervals of 5
- user Id: The unique ID of the user who gave the rating
- movie Id: The unique ID of the movie that was rated
- title: The title of the movie accompanied by the year of its release
- time stamp: Unix time stamp of the time the movie was rated
- genre: A string of all the genres the movie belongs to separated by a '|'

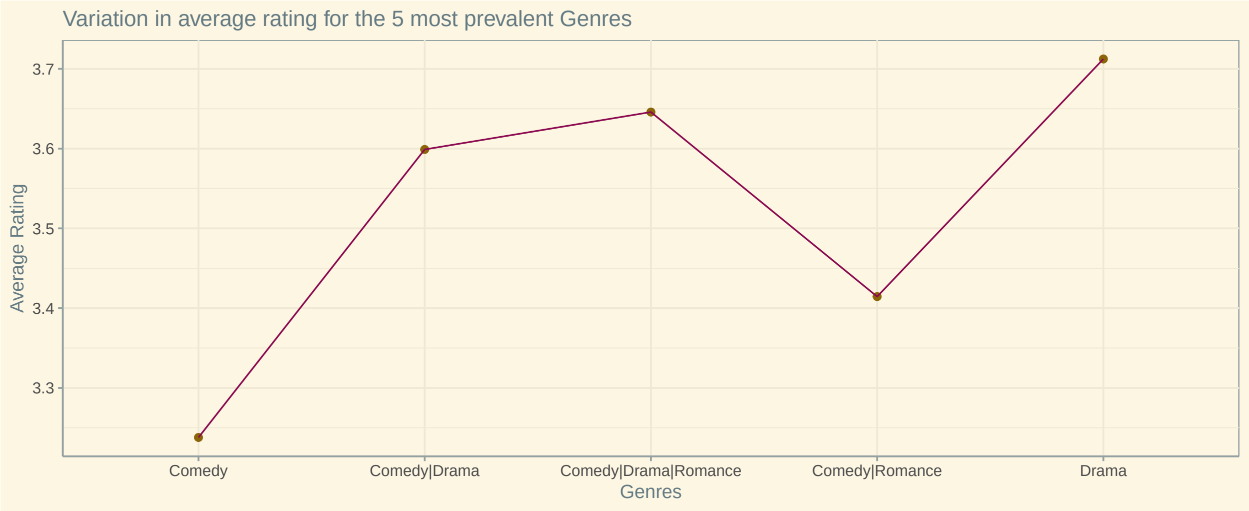
KEY STEPS

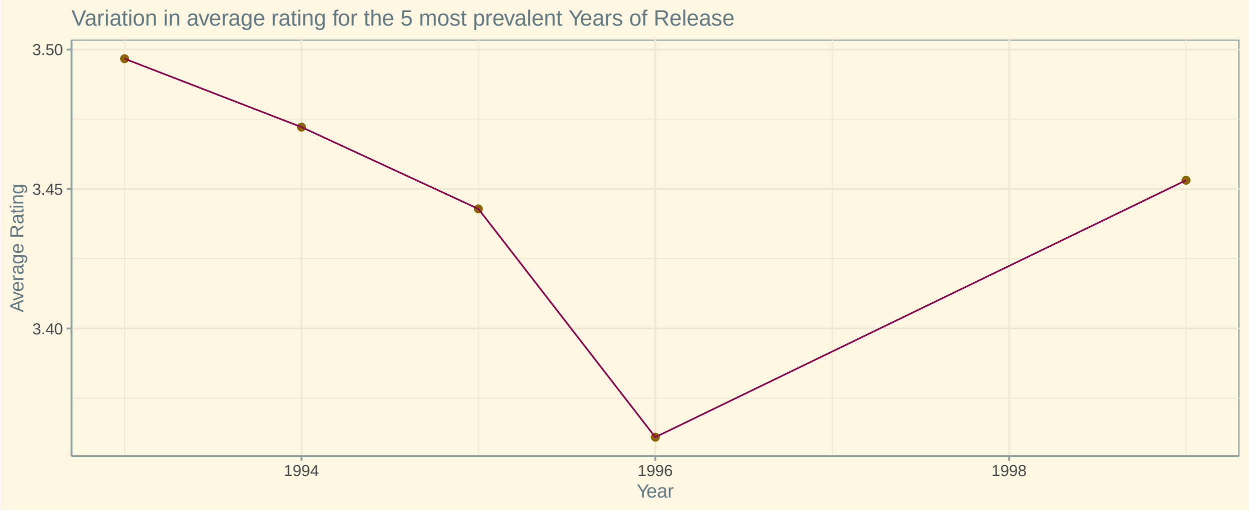
The key steps to completing this project included; separating the year of release from the title variable, extracting the year, month, day and nearest hour of rating from the time stamp variable and then separating the genres into variables for each unique genre possible. After this I carried out training and testing of linear regression and xgboost on the edx data set and evaluated their performance using RMSE. Finally, I used the xgboost method on the validation test and evaluated my final RMSE

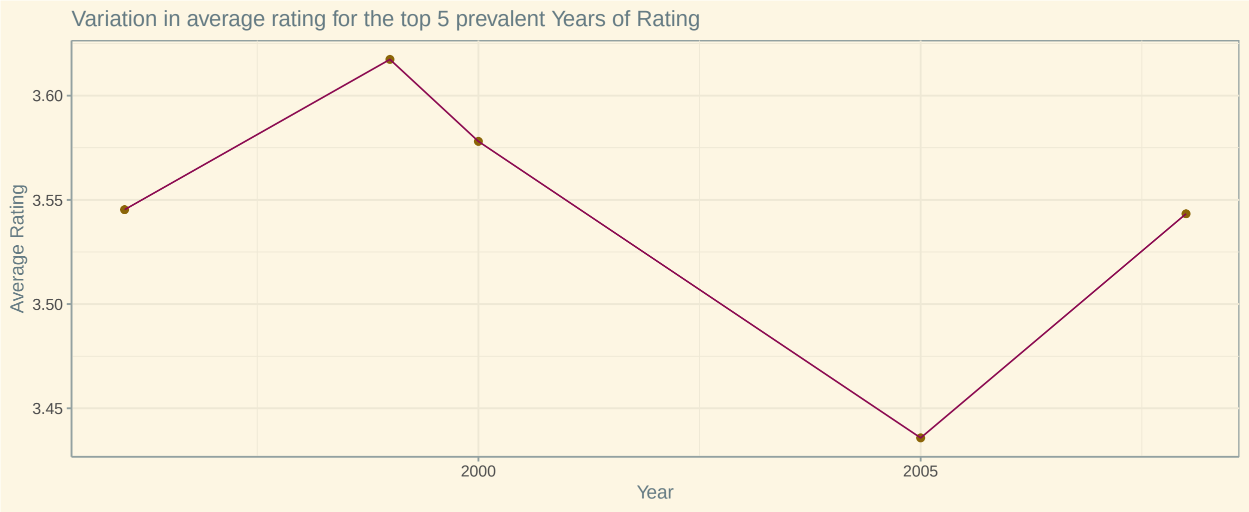
2 VISUALISATION AND METHODOLOGY

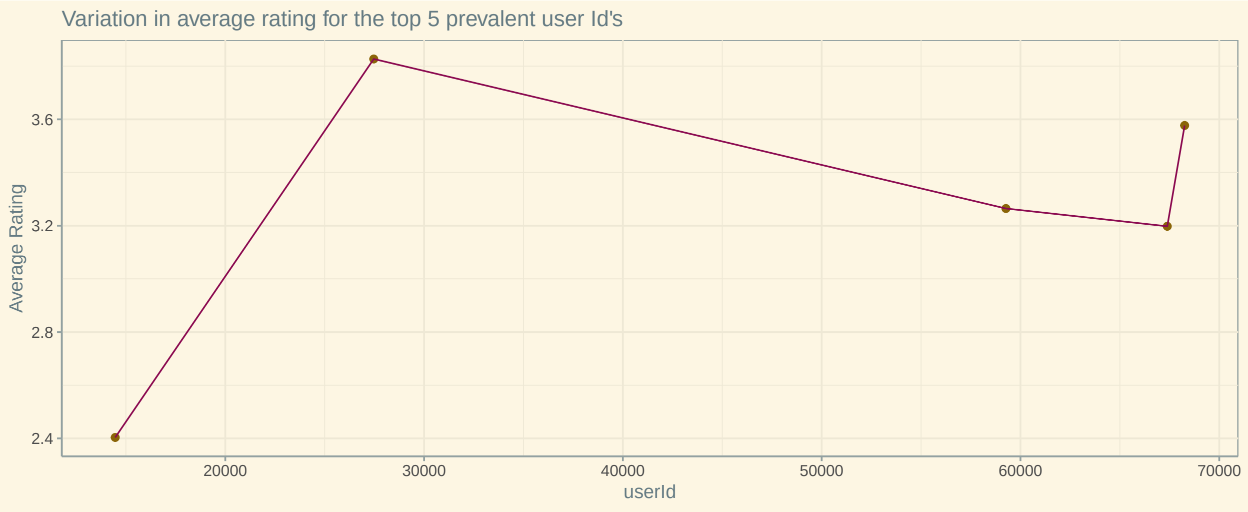
VISUALISATION

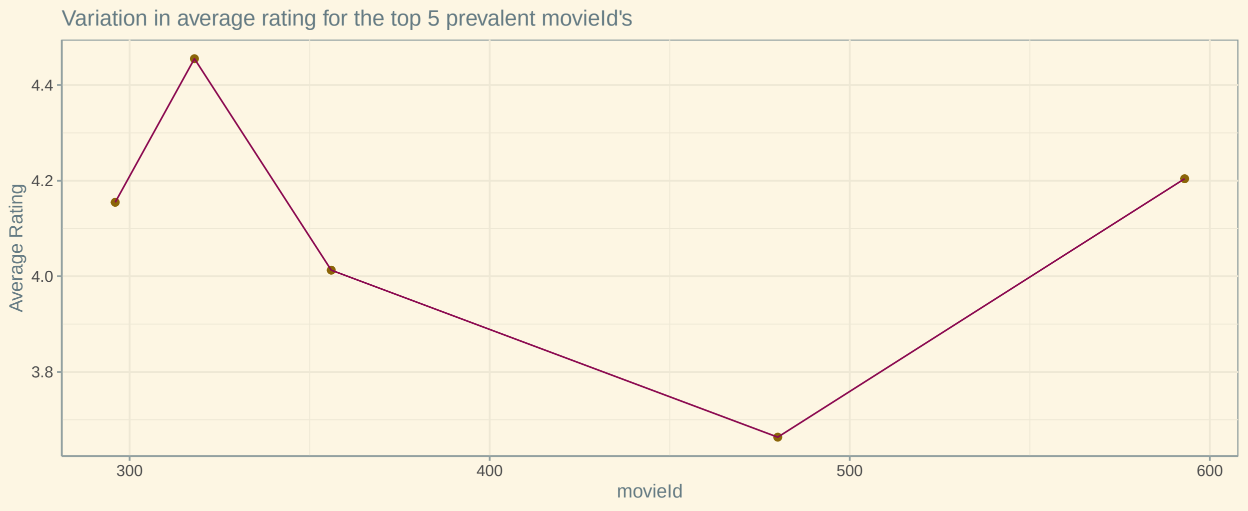
For my visualization, I looked at the variation of the average ratings with the predictors genre, year of release, year of rating, user ID and movie ID for the 5 most prevalent values of these predictors and observed general variability based on these predictors as seen in the plots below


Variation in average rating for the 5 most prevalent Genres
Average Rating
3.7
3.6
3.5
3.4
3.3
Comedy
Comedy|Drama
Comedy|Drama|Romance
Comedy|Romance
Drama
Genres

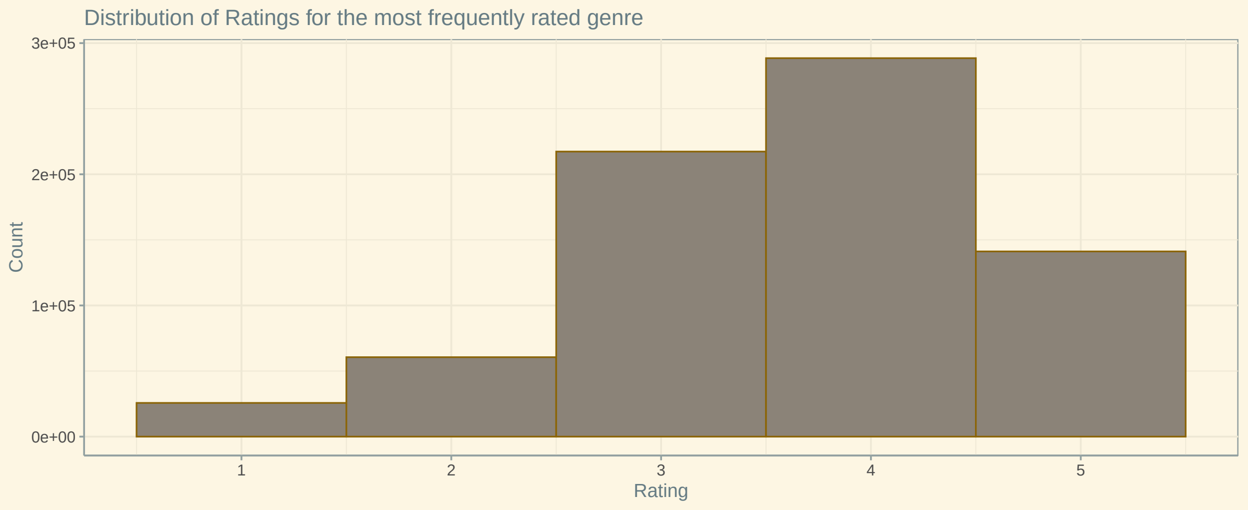

Variation in average rating for the 5 most prevalent Years of Release
Average Rating
3.50
3.45
3.40
1994
1996
1998
Year

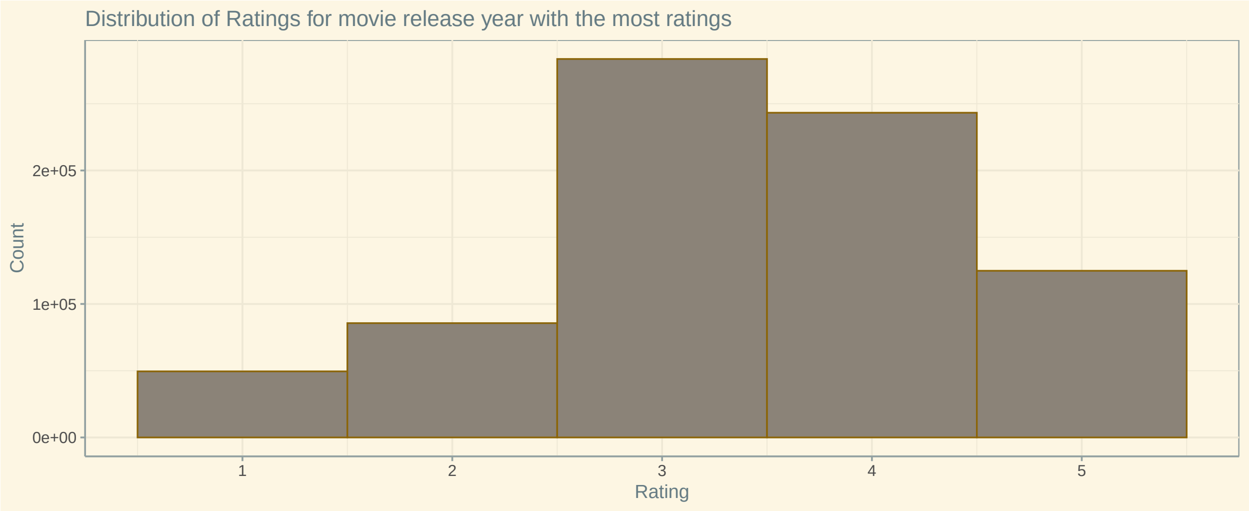

Variation in average rating for the top 5 prevalent Years of Rating
Average Rating
3.60
3.55
3.50
3.45
2000
2005
Year



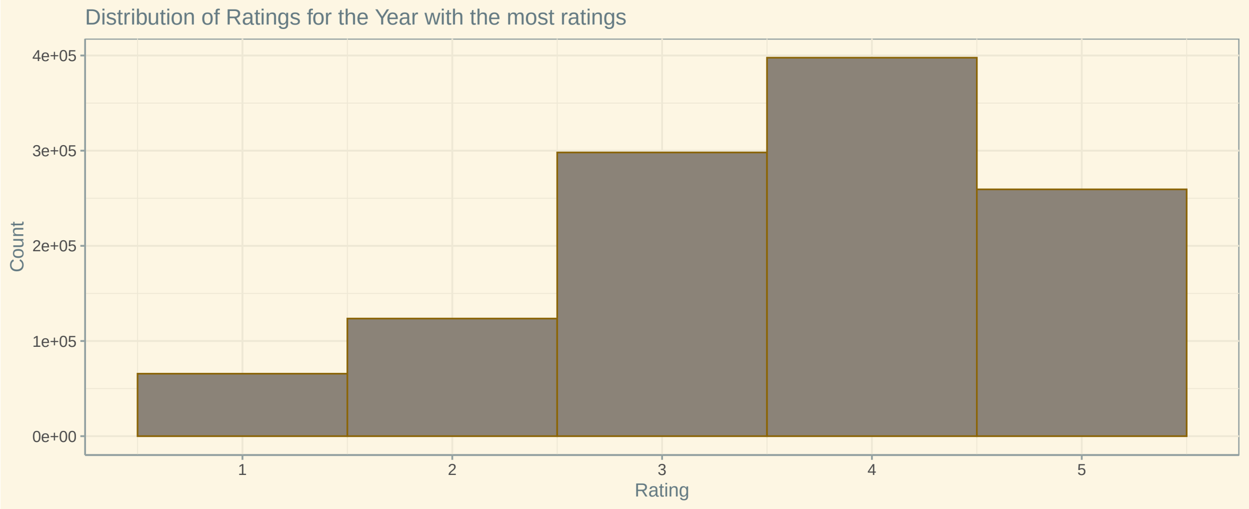


I also plotted the distribution of ratings for the most prevalent of each of these predictors and from the plots below, none of the distributions appear normal. This suggests that a linear regression model might not provide the best estimates

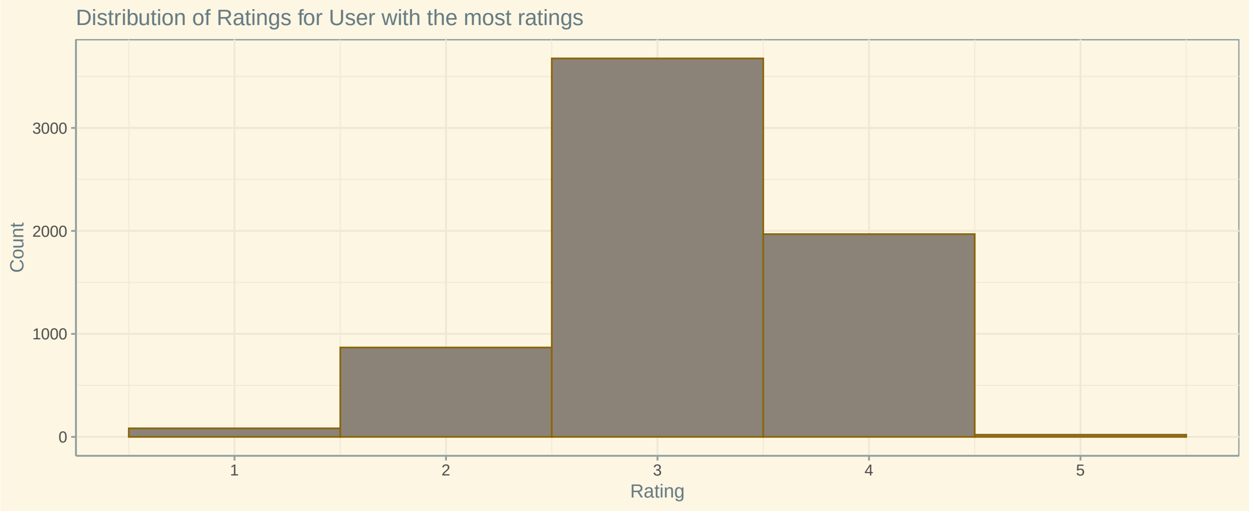




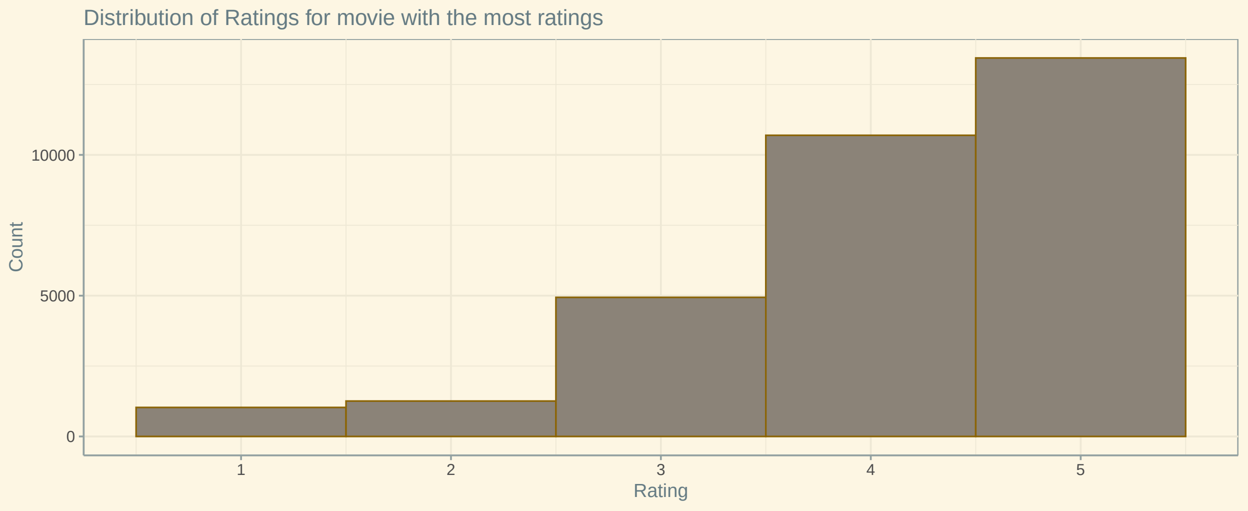
Distribution of Ratings for movie release year with the most ratings



Distribution of Ratings for the Year with the most ratings

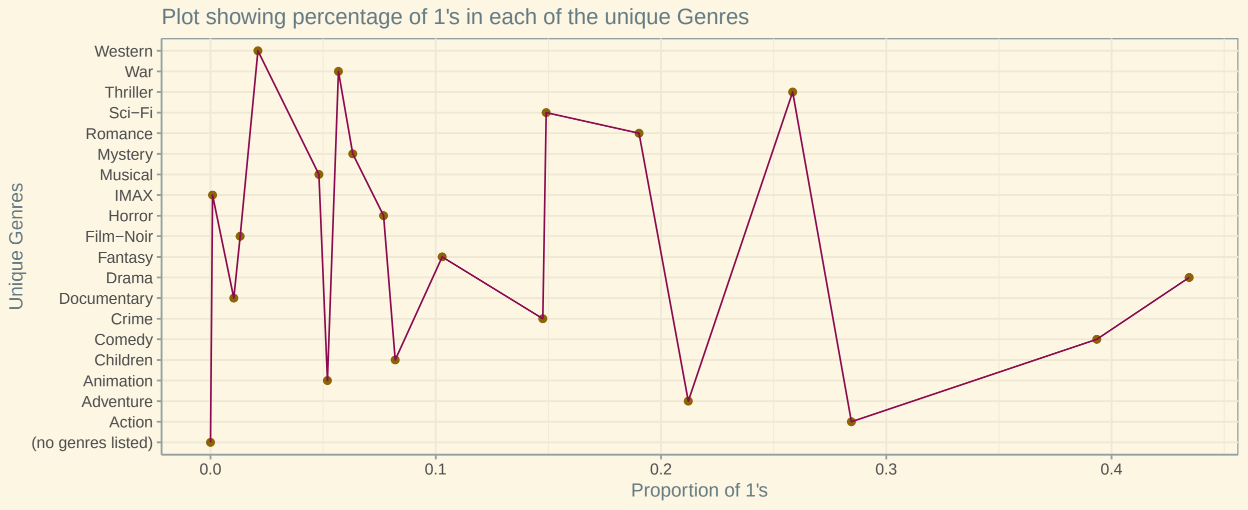


Distribution of Ratings for User with the most ratings



METHODOLOGY

To start with, using str_extract and str_replace, I removed the year of release in the movie titles and made it a separate predictor year_of_release, after this I removed the title predictor since intuitively it correlates to the movie ID. Then using the mutate and as_datetime functions, I converted the rating time stamps into readable date/time. Following from that, I used the lubridate package to separate the rating date/times into year_of_rating, month_of_rating, day_of_rating and nearest_hour_of_rating. Lastly, looking at the genres predictor it had 797 leves, because of this wide variability for a factor, I used str_extract and a nested for loop and created new predictors for all the 20 unique possible genres Comedy, Romance, Action, Crime, Thriller, Drama, Sci-Fi, Adventure, Children, Fantasy, War, Animation, Musical, Western, Mystery, Film-Noir, Horror, Documentary, IMAX, (no genres listed) and gave them values 1 or 0, 1 if the movie belonged to that genre and 0 otherwise. However, on observing the distribution of these genres in our data set, more than half of the genres are only represented in under 10 percent of our total data as seen in the plot below.This will have an adverse effect on the ability of our model to make accurate predictions



3 RESULTS

Despite the distributions of the predictors suggesting that a linear regression model might not provide the best predictions I still used that as my baseline model and got an RMSE of 1.0329523 then to improve this I deployed xgboost which uses an ensemble of decision trees to make predictions and got a slightly improved RMSE of 0.9531976. Finally, applying the xgboost model that was trained with the edx data on the validation set, I obtained a similar RMSE of 0.9537109

4 CONCLUSION

Due to the size of our data(9000055) and the number of predictors(27) it was difficult to train a more robust model like random forest on our data set with the limitations of a regular computer. Also, a data set with better distribution of genres would likely have yielded a better model. Despite all this, I was still able to estimate ratings on the validation set within an RMSE of 0.9537109 which is on average less than one rating point away from the actual rating.