Introduction

As industries transform and evolve from sitting in the stands viewing information (looking at information or knowing that it exists) to applying relevant and reliable knowledge on the stage of innovation and change, data science and machine learning techniques play an integral role to facilitating this transition.

In recent years, consumer preference data has been highly influential on society as it has shown to impact investment behaviors, which has reflected in the success and failure of many organization, entrepreneurs, and social figures. A clear example of this situation would be the affect that film review websites (e.g., IMBD, Rotten Tomatoes, Fandango, and Google Reviews) have impacted the film industry’s box office and associated revenues. These rating data blended with consumer reviews have birthed algorithmic recommendation systems, which are applied across various streaming media platforms (e.g., Netflix, Hulu, Disney+, YouTube, Amazon Prime) and thus serve as an avenue for organizations, entrepreneurs, and social figures to understand consumer preferences and strategize to gain competitive advantage.

The following data study was conducted to provide a surface-level understanding the various factors associated with film rating and provide a baseline recommendation for projecting future ratings.

Setting and Data Preparation

The study utilized a public dataset of film rating, titled MovieLens. The big data file contained over 10 million film rating for more than 9,000 film across the 1930s into the 2000s. The data for the study included the factors of film id number, film title, film release year, film genre, and film rating, which used a five-point scale (1 being the lowest or poor quality rating and 5 being the highest or top quality rating).

Following data modeling protocol, the dataset should be partitioned into two datasets. One dataset, which is a smaller subset of the whole dataset is used to build and train the data model. Once the data model is trained, it is tested against the second partitioned dataset for validation. For the purposes of this study, the MovieLens dataset was parsed into a 10/90 structure with 10% of the data being used for training and 90% of the data being used for model validation testing.

```
## -- Attaching packages ---------------------------------------------------------------------
---------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
```

```
## -- Conflicts -------------------------------------------------------------------------------
---- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
## Warning: package 'DescTools' was built under R version 3.6.3
## Warning: package 'caret' was built under R version 3.6.3
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following objects are masked from 'package:DescTools':
##
##     MAE, RMSE
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
## Warning: package 'data.table' was built under R version 3.6.3
```

```
##
## Attaching package: 'data.table'
```

```
## The following object is masked from 'package:DescTools':
##
##     %like%
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
## Warning in set.seed(1, sample.kind = "Rounding"): non-uniform 'Rounding' sampler
## used
```

```
## Joining, by = c("userId", "movieId", "rating", "timestamp", "title", "genres")
```

Descriptive Data

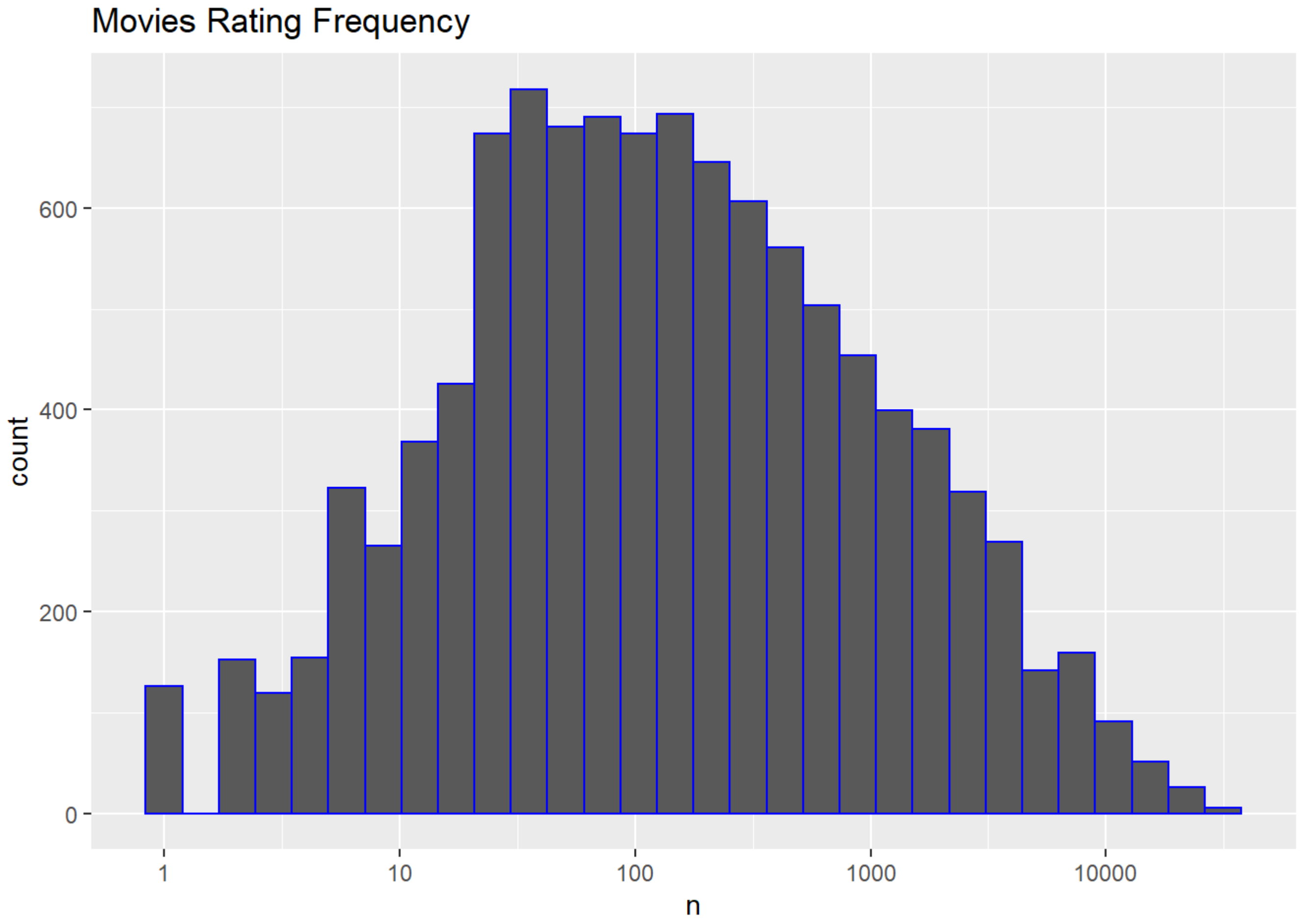
Prior to model development, descriptive statistics were conducted to better understand the data and distribution of the variables. The first assessment of the data was conducted by calculating the overall summary statistics of the training dataset.

```
##      userId         movieId          rating        timestamp
##  Min.   :    1   Min.   :    1   Min.   :0.500   Min.   :7.897e+08
##  1st Qu.:18124   1st Qu.:  648   1st Qu.:3.000   1st Qu.:9.468e+08
##  Median :35738   Median : 1834   Median :4.000   Median :1.035e+09
##  Mean   :35870   Mean   : 4122   Mean   :3.512   Mean   :1.033e+09
##  3rd Qu.:53607   3rd Qu.: 3626   3rd Qu.:4.000   3rd Qu.:1.127e+09
##  Max.   :71567   Max.   :65133   Max.   :5.000   Max.   :1.231e+09
##     title              genres
##  Length:9000055     Length:9000055
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

In addtion, the mean and standard deviation was calulated for the film rating data.

```
## [1] 3.512465
```

```
## [1] 1.060331
```



Model Framework

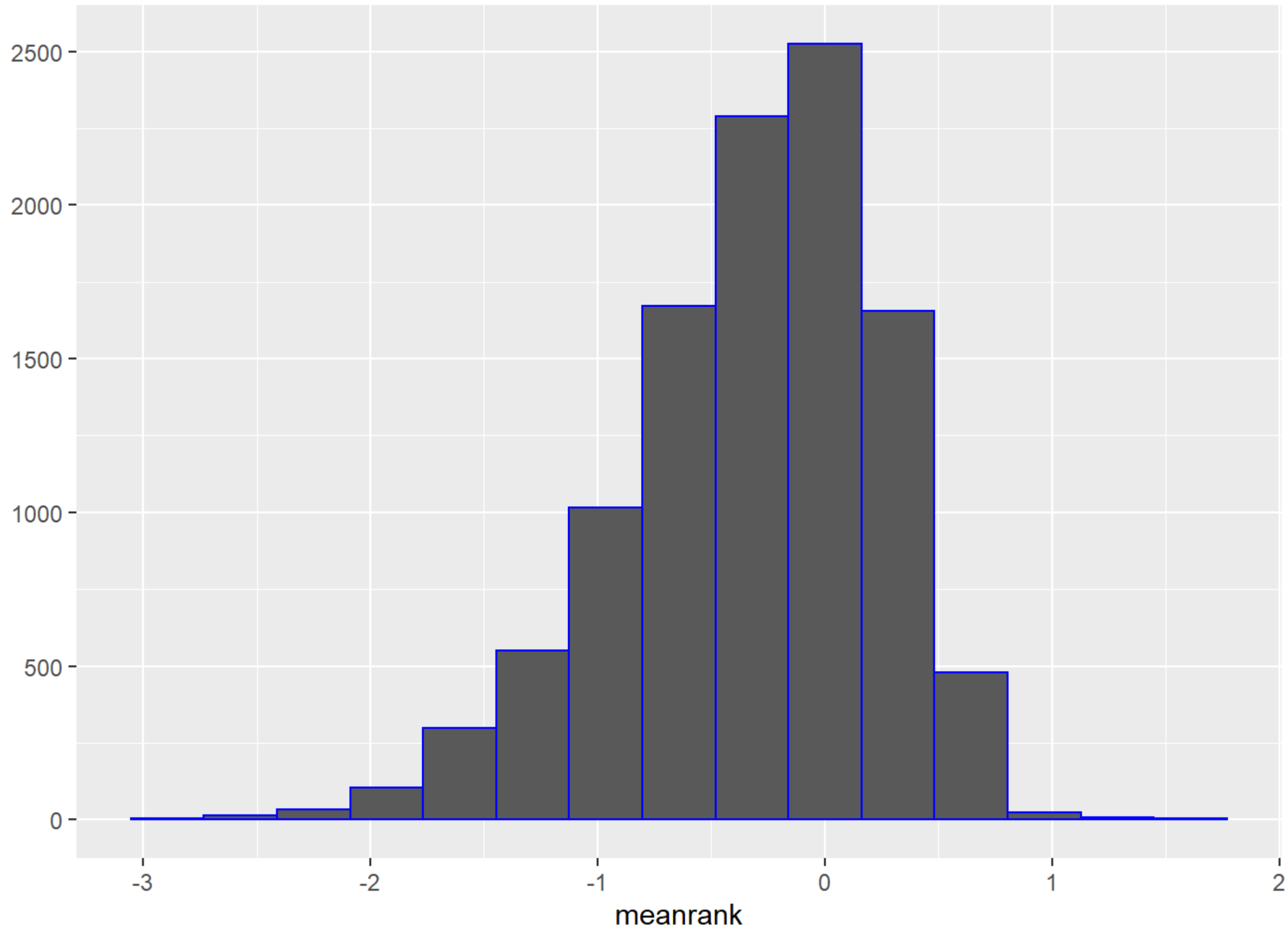
The first iteration of teh model used the ovreall average to estimate the unknown ratings. For this appraoch the mean mean of the training dataset was used to calculate Root Mean Square Error.

```
## [1] 1.061202
```

An average rating was developed and used to predict the unknown ratings. The MeanRate model was tested for validity to measure the effect.

```
## [1] 0.9439087
```

A visual representation of the of the MeanRate model distribution was produced.

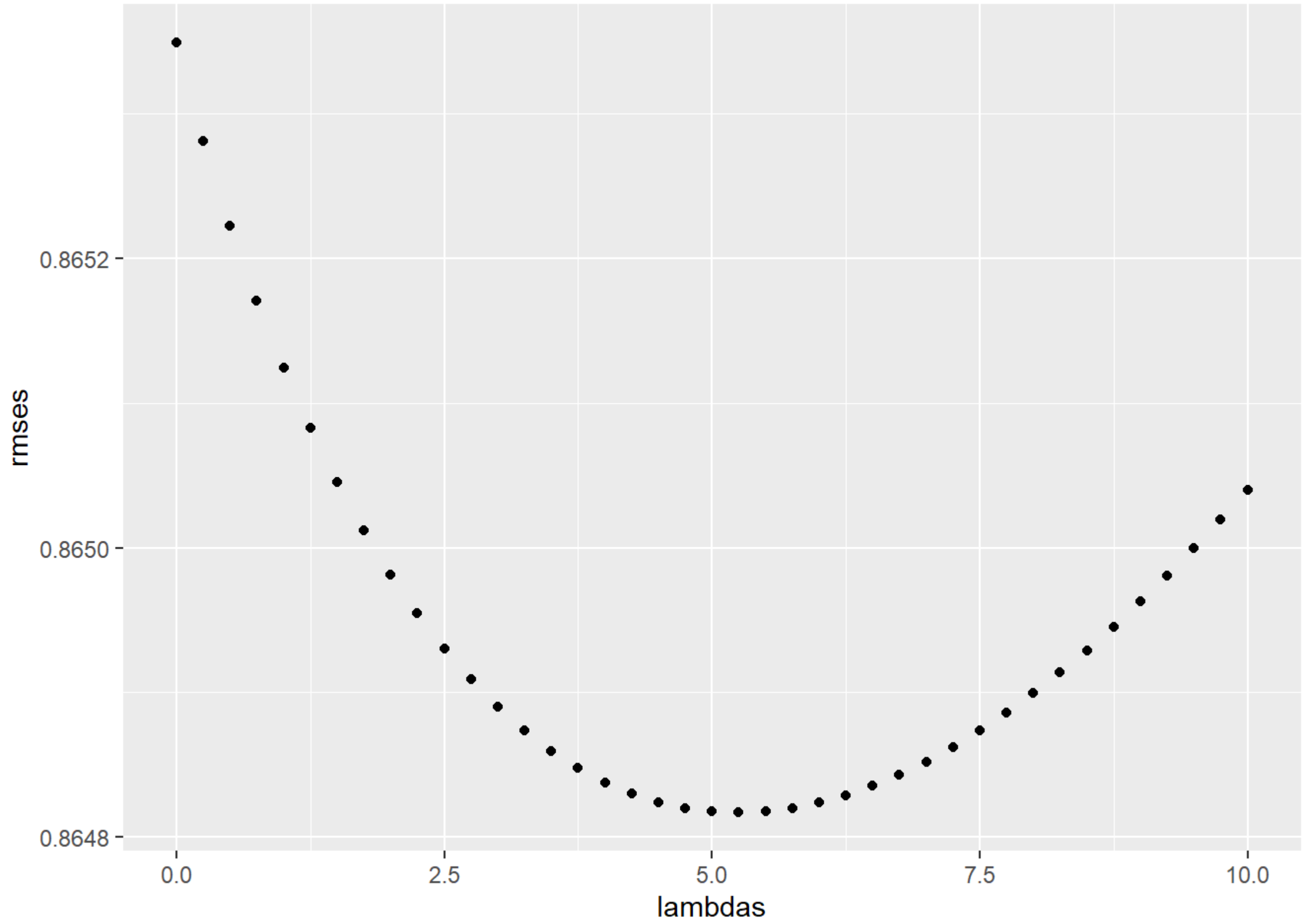


To strengthen the validity of the model, user bias was introduced as an additional factor to influence the model output.

The MeanRate and UserBias model was tested for validity to measure the effect.

```
## [1] 0.8653488
```

To enhance the model and reduce the effect of the errors, regularization was used to reduce overfitting related to outliers and other data anomalies. A visual representation using a quick plot of the Root Mean Square Error and lambdas.



Finalized Model

Once regularization was accounted for, the final data model was develoepd and tested. The results were a decrease in the Root Mean Square Error which yielded a more accruate prediction of future film ratings.

```
## [1] 0.864817
```

Conclusion

The study found that using data science techniques made significant improvements on the data model effectiveness. In summary, the subset of the MovieLens dataset was effectively modeled and trained to produce a baseline for accurate predictions for future film ratings. The final model shows that there were incremental enhancements to the data model as new concepts were introduced, which reflected a decrease in the Root Mean Square Error.

Model 1. Mean (mu) had RMSE of 1.061202 Model 2. MeanRate had RMSE of 0.9439087 Model 3. MeanRate an UserBias has an RMSE of 0.8653488 Model 4. MeanRate an UserBias has an RMSE of 0.864817

Therefore, the data model continued to improve as the variety of factors were being accounted for and included. Future analysis should be conducted with a recommendation to parse data by film year and genre type to best determine how genre or timeframe influences film ratings.