Linear and Logistic Regression for Star Wars Movie Review Predictions

Riley D. Fiske

Concordia College

DATA-318: Data Mining

Dr. Gregory Tanner

February 24, 2022

Linear and Logistic Regression for Star Wars Movie Review Predictions

The ranking of any movie is of course subjective, but there is something special, or sometimes even controversial, regarding how an individual chooses to rank the Star Wars movies. This project sought to create models using linear and logistic regression to predict how an individual would rank the original six Star Wars films using their gender, age, household income, education, and location as predictors. These variables all have categories that every person could sort themselves into, which makes this model's results that much more fascinating.

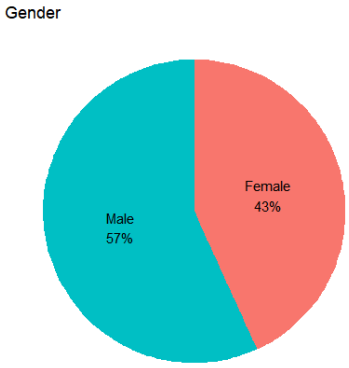
Finding the Data

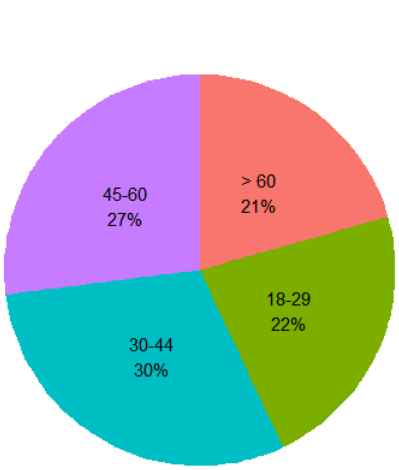
The star-wars-survey dataset from ABC News' Fivethirtyeight.com, collected in 2014, consists of originally 1186 respondents on how they would answer a variety of questions relating to the Star Wars films and some personal information. According to Fivethirtyeight.com's article titled *America's favorite 'star wars' movies (and least favorite characters),* the survey was conducted via SurveyMonkey from June 3rd to June 6th 2014. The manner in which this data was collected should add a shroud of skepticism to our analysis as it's a survey taken at the leisure of the respondent, as well as an internet survey that is prone to vulnerabilities.

Before being able to glean anything from the data sample, there are some empty responses that had to be pruned through. Samples were removed that said they had not seen Star Wars or were not fans of the movies since those had many blank responses, and rows that had blanks in them for specific movie rankings or for any of the personal information categories were also removed to include only full rows. This could potentially skew the data but was better than leaving in empty responses for categorical data. Certain categories of data were also merged to

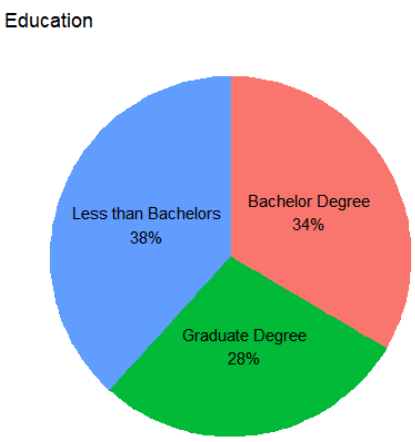
remove certain categories that had few responses. After cutting it all down, 433 observations remained to draw conclusions from.

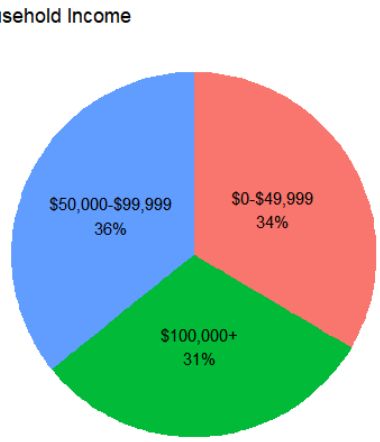
Exploratory Analysis

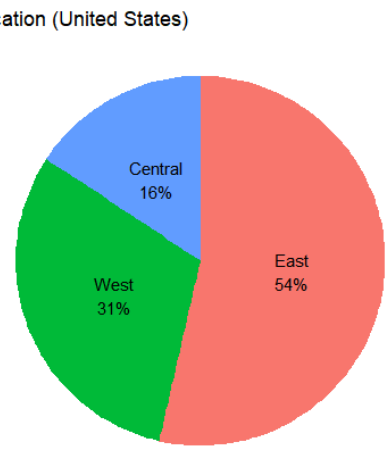




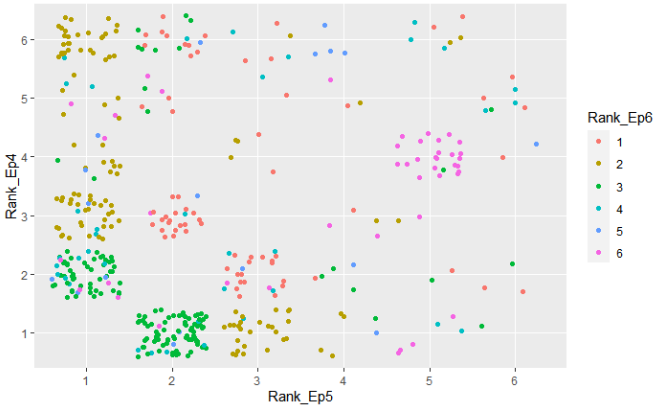
To begin, fans of Sci-Fi are typically thought to be male, and this sample does reflect a male majority, however there is a decent number of female samples. When looking at the percentages of each age range participating in the survey, it can be seen that the largest range is from 30-44, second being 45-60, and the oldest and youngest being fairly close (by 8 respondents). Star Wars first came out in 1977 and this survey came out in 2014, which would put the kids that saw Star Wars as it was coming out in the larger two categories. The three ranges for household income are pretty equivalent, as are the three categories of education, meaning individuals at different economic levels can enjoy Star Wars and it is not catered to one or the other. The location of sampled individuals does seem to be skewed more to the east coast with few samples in central US.







When creating representations of how different groups would rank different Star Wars films, patterns of voting emerged quite clearly. If someone were to rank one of the prequel films (Episodes 1-3) low, they are more likely to also rank the other prequel films low. Conversely, if someone were to rank one of the original films (Episodes 4-6) high, they are more likely to also rank the other original films high. This holds true for the majority of the predictors.



Data Models

Linear Regression – Ranking 1-6

After the data was cut down and the preliminary graphics were made, data models for predicting how an individual would rank any Star Wars film using linear regression were created. The larger dataset was cut down into a training set of 80% of the data and a test set of 20% to have data for the model to be tested against. The predict function was then cast on the test set using the linear regression model created for each of the six films, and then a confusion

```
[1] "The Phantom Menace Confusion Matrix"
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3  4  5  6
         1  0  0  0  0  0  0
         2  0  0  0  1  0  0
         3  4  1  3  5  2  0
         4  3  2  2 14  3  4
         5  2  2  4  9  7 15
         6  0  0  0  1  1  4

Overall Statistics

               Accuracy : 0.3146
                 95% CI : (0.2203, 0.4217)
    No Information Rate : 0.3371
    P-Value [Acc > NIR] : 0.7094

                  Kappa : 0.1377
```

```
[1] "Attack of the Clones Confusion Matrix"
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3  4  5  6
         1  0  0  0  0  0  0
         2  0  0  0  0  0  0
         3  0  0  0  0  0  0
         4  1  4  3  6 14  2
         5  3  4  6 10 24 12
         6  0  0  0  0  0  0

Overall Statistics

               Accuracy : 0.3371
                 95% CI : (0.2403, 0.4451)
    No Information Rate : 0.427
    P-Value [Acc > NIR] : 0.967

                  Kappa : -0.01
```

```
[1] "Revenge of the Sith Confusion Matrix"
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3  4  5  6
         1  0  0  0  0  0  0
         2  0  0  0  0  0  0
         3  0  0  0  0  0  0
         4  4  2  4 10  6 12
         5  1  2 10 11 13 13
         6  0  0  0  0  0  0

Overall Statistics

               Accuracy : 0.2614
                 95% CI : (0.1734, 0.3659)
    No Information Rate : 0.2841
    P-Value [Acc > NIR] : 0.7188

                  Kappa : 0.046
```

matrix was used to test the accuracy of the predictions. If one were to blindly guess at the predictions for each movie, it would be a roughly 1/6 chance of guessing correctly, so success is measured by being greater than 16.67% in accuracy. The confusion matrices for Episodes 1-3 were all around 30% accurate, which is a little less than double the chance of guessing at random. It should be noted that in 2014, the public opinion on the Star Wars prequel trilogy was quite low, so it makes sense that the model was predicting lower ranks for all three of these films and getting the ranking correct.

As for the accuracy of the models for the original trilogy, the accuracy tends to be slightly lower, and that is most likely due to predicting the exact ranking for each of these three beloved films. Surprisingly, the film that is popularly believed to be the best film was not the easiest for the model to predict, at only 22% accuracy. Return of the Jedi was probably the easiest of these three for the model to predict accurately as it is generally conceived as the weakest of the original films, ranking at number 3 on many peoples' charts, which is where the model placed it as well.

```
[1] "A New Hope Confusion Matrix"
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3  4  5  6
         1  0  0  0  0  0  0
         2 10 11  5  5  1  6
         3 12  8  7  4  1  7
         4  4  1  1  1  4  1
         5  0  0  0  0  0  0
         6  0  0  0  0  0  0

Overall Statistics

               Accuracy : 0.2135
                 95% CI : (0.1337, 0.3131)
    No Information Rate : 0.2921
    P-Value [Acc > NIR] : 0.9632

                  Kappa : 0.0465
```

```
[1] "The Empire Strikes Back Confusion Matrix"
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3  4  5  6
         1  1  1  1  0  0  0
         2 25 18 11  5  5  3
         3  6  8  1  0  4  0
         4  0  0  0  0  0  0
         5  0  0  0  0  0  0
         6  0  0  0  0  0  0

Overall Statistics

               Accuracy : 0.2247
                 95% CI : (0.143, 0.3255)
    No Information Rate : 0.3596
    P-Value [Acc > NIR] : 0.9979

                  Kappa : -0.0645
```

```
[1] "Return of the Jedi Confusion Matrix"
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3  4  5  6
         1  0  0  0  0  0  0
         2  4  4 10  2  1  2
         3 10 19 19  6  3  7
         4  0  1  1  0  0  0
         5  0  0  0  0  0  0
         6  0  0  0  0  0  0

Overall Statistics

               Accuracy : 0.2584
                 95% CI : (0.1714, 0.3621)
    No Information Rate : 0.3371
    P-Value [Acc > NIR] : 0.9563

                  Kappa : -0.0812
```

Logistic Regression – Top 3 or Bottom 3

```
[1] "The Phantom Menace Confusion Matrix"
Confusion Matrix and Statistics

          Reference
Prediction Top 3 Bottom 3
  Top 3        6        4
  Bottom 3    16       61

               Accuracy : 0.7701
                 95% CI : (0.6675, 0.8536)
    No Information Rate : 0.7471
    P-Value [Acc > NIR] : 0.36247

                  Kappa : 0.2577

 Mcnemar's Test P-Value : 0.01391

            Sensitivity : 0.27273
            Specificity : 0.93846
```

```
[1] "Attack of the Clones Confusion Matrix"
Confusion Matrix and Statistics

          Reference
Prediction Top 3 Bottom 3
  Top 3        1        0
  Bottom 3    19       67

               Accuracy : 0.7816
                 95% CI : (0.6802, 0.8631)
    No Information Rate : 0.7701
    P-Value [Acc > NIR] : 0.4584

                  Kappa : 0.075

 Mcnemar's Test P-Value : 3.636e-05

            Sensitivity : 0.05000
            Specificity : 1.00000
```

```
[1] "Revenge of the Sith Confusion Matrix"
Confusion Matrix and Statistics

          Reference
Prediction Top 3 Bottom 3
  Top 3        0        0
  Bottom 3    23       65

               Accuracy : 0.7386
                 95% CI : (0.6341, 0.8266)
    No Information Rate : 0.7386
    P-Value [Acc > NIR] : 0.5558

                  Kappa : 0

 Mcnemar's Test P-Value : 4.49e-06

            Sensitivity : 0.0000
            Specificity : 1.0000
```

When using logistic regression, the predictions had to be scaled down to a top-3 ranking or a bottom-3 ranking. The accuracy is of course going to be higher as a result, but a linear model was also created using a top-3 bottom-3 scale and the accuracy ratings were within 1-2% of the logistic model. When analyzing these confusion matrices, it is important to note the sensitivity and specificity. The model was not very generous in handing out top-3 rankings to the prequel movies, specifically when looking at the sensitivity (0.00) and specificity (1.00) of Revenge of the Sith. This means the model did not report any false positives but predicted many false negatives; it reported 0% of the actual positives correctly and 100% of the negatives correctly. The high accuracy given off by the model is worth noting and this pattern continues when looking at the original trilogy matrices.

```
[1] "A New Hope Confusion Matrix"
Confusion Matrix and Statistics

          Reference
Prediction Top 3 Bottom 3
  Top 3       54       22
  Bottom 3     5        7

               Accuracy : 0.6932
                 95% CI : (0.5858, 0.7871
    No Information Rate : 0.6705
    P-Value [Acc > NIR] : 0.371315

                  Kappa : 0.1841

 Mcnemar's Test P-Value : 0.002076

            Sensitivity : 0.9153
            Specificity : 0.2414
```

```
[1] "The Empire Strikes Back Confusion Matrix"
Confusion Matrix and Statistics

          Reference
Prediction Top 3 Bottom 3
  Top 3       72       16
  Bottom 3     0        0

               Accuracy : 0.8182
                 95% CI : (0.7216, 0.8924)
    No Information Rate : 0.8182
    P-Value [Acc > NIR] : 0.5662957

                  Kappa : 0

 Mcnemar's Test P-Value : 0.0001768

            Sensitivity : 1.0000
            Specificity : 0.0000
```

```
[1] "Return of the Jedi Confusion Matrix"
Confusion Matrix and Statistics

          Reference
Prediction Top 3 Bottom 3
  Top 3       65       20
  Bottom 3     2        0

               Accuracy : 0.7471
                 95% CI : (0.6425, 0.8342)
    No Information Rate : 0.7701
    P-Value [Acc > NIR] : 0.7422920

                  Kappa : -0.0436

 Mcnemar's Test P-Value : 0.0002896

            Sensitivity : 0.9701
            Specificity : 0.0000
```

The Empire Strikes Back has the highest accuracy, but this is most likely because, as was stated earlier, it is popularly recognized as the best of the six films. It makes sense that the model would predict all positive results for it, the accuracy being high since that is more than likely true for the population, and the sensitivity being 1.00 and specificity being 0.00 as a result. The matrix for A New Hope is interesting to analyze as it has the lowest accuracy, and a fairly diverse spread of predictions.

Interpretation of Results

It is fairly impressive that linear regression can predict with two times the accuracy of random chance how the test set would rank each of the original six Star Wars films, and that logistic regression was very successful in guessing if one were to rank each one in their top three or bottom three. It may be less impressive, however, if you know information about how people rank them in-general, independent from the dataset. As a quite serious member of the Star Wars community, I would predict that most people would give the prequel movies ranking 4-6 and the original movies ranking 1-3, especially in 2014. The prequels were not as well received when they came out and the years afterwards, so it makes sense that the model created from the dataset also reflects that in both the linear and logistic regression models.

I attempted to optimize these models by removing the variables that were not statistically significant and by using preprocessing by combining logically compatible dummy variables for the categorical predictors, however the results stayed about the same before and after all the attempted optimizations.

If this survey were to be conducted again today, I believe the results would be quite different as the kids who grew up with the prequels as their Star Wars are now in a valid age

range now, and there are 3 new Skywalker Saga films to be added to the ranking list. The opinions of the prequel movies have also lightened after the new films' release so I believe the results would be more diverse and make the models more unique and interesting.

The cultural phenomenon that is Star Wars has a wide variety of opinions on each of its films by its passionate fans, and to create a model to predict this is quite difficult. However, the results of this study show that it is definitely possible to try and get a general feel for how a given demographic feels about a movie.

References

Fiske, R. (2022) Project1RMD [Source Code] https://github.com/rdfiske17/DATA-318/tree/main/Project1RMD

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2021). *Dplyr: A grammar of data manipulation.* R package version 1.0.7. https://CRAN.R-project.org/package=dplyr

Hickey, W. (2014, July 22). *America's favorite 'star wars' movies (and least favorite characters).* Fivethirtyeight.com, https://fivethirtyeight.com/features/americas-favorite-star-wars-movies-and-least-favorite-characters/

Flowers, A (2014) Star Wars Survey [Source Code] https://github.com/fivethirtyeight/data/tree/master/star-wars-survey

R Core Team (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Wickham et al., (2019). *Welcome to the tidyverse.* Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686