Netflix Analysis

Group One

Andrew Chu, Ashley Krause, Geneway Tang, Jason Santucci

5 May 2022

DeSales University

Executive Summary

Our statistical analysis of Netflix was formed to assess how the media streaming and video rental company utilizes data to drive its customer engagement and business strategies. The purpose of this report is to present the methods which teams at Netflix use to improve their customer experience and which features shown within our dataset have the greatest effect on the trends at Netflix.

The sections of this report illustrate our process in data preparation, analysis, and our findings on how Netflix uses data for their predictive modeling in their business strategy to improve user engagements. We found that Netflix's platform provides users with personalized experiences and promotes intuitive interactions that create immersive experiences with exceptional design. Additionally, we found the long-term strategic objectives that Netflix have to improve its modeling. This, most importantly, led us to take a broad view of the methodological analyses at Netflix that drive them to become one of the largest and most successful media streaming companies on the market.

Project Motivation

There are various reasons why we have been motivated to analyze Netflix and the data that has been provided about Netflix. This media streaming and video rental company has always been a company that utilizes data to drive them to be better. They state on their website, “Our analytic work arms decision-makers around the company with useful metrics, insights, predictions, and analytic tools so that everyone can be stellar in their function. Partnering closely with business teams in product, content, studio, marketing, and business operations, we perform context-rich analysis to provide insight into every aspect of our business, our partners, and of course our members’ experience with Netflix” (Netflix Research). Their intense research driven by data has allowed us to become curious about the way Netflix can predict future trends based purely on analysis. This is truly what has motivated us.

After some brief research, we have noticed that Netflix uses large amounts of data to answer the questions to show how trends are moving along. Throughout this project, we have strived to find out just how Netflix is doing this. We think it is important to analyze this data because we can see how a successful company is utilizing its data analytic skills. The main objective of our project was to see how Netflix uses the data they are given from previous films and tv shows to predict the popularity of other films to add to their selection. We have been inspired to see if there is any method to their show selection within the company. This is on a basis of ratings, directors, genre, etc.

Another interesting thing that Netflix has done to predict was their recommendation section of the company. This has also made us motivated to look into it. Similarly, to how they select what films and television show to show, they also use data to recommend shows to users that they might like on a basis of their preferences. This is another factor that has prompted us to analyze the ways of their company analysis. The methods of recommendations based purely on customer preferences are fascinating and we want to make sense of their methodology.

Overall the major reason why it is important and we are motivated to start analyzing the Netflix data is to try and decipher how Netflix utilizes its data in a meaningful way to expand and promote its company in the best way it can. As well as to get to the bottom of their methodological analysis.

Data Description

Within this project, we have used secondary data found by Kaggle. The dataset is composed of fourteen columns and nine thousand eight hundred and twenty-five rows. Below you will find a list of all variables found within the dataset with a description to follow.

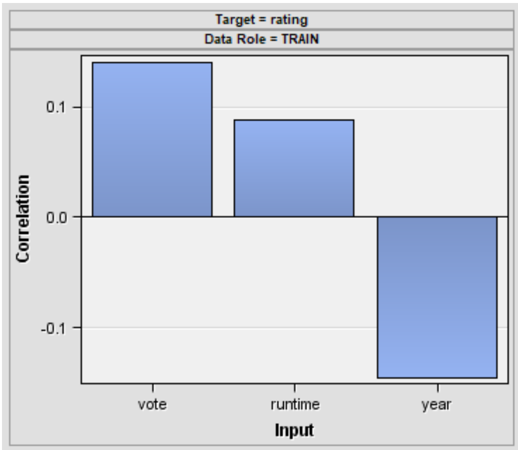
- Column 1
  - This Variable is a count of the data and allows for the dataset to be put back into order if mixed up.
- Title
  - The name of said television show or movie.
- Year
  - The year the television show had been released. This dataset has films released between 1905-2022.
- Kind
  - The description of what the television show or movie goes by. These selections include: episode, movie, tv mini series, tv movie, tv series, tv short, video game, video movie
- Genre
  - The said genre of the television shows or movies. This could be Action, Documentary, Family, Sci-fi, Comedy, Thriller, etc.
- Rating
  - The rating of the television show or movie was given by IMDB.
- Vote
  - The vote for the television show or movie was given by IMDB.
- Country
  - The country of which the television show or movie originated in.
- Language
  - The language of the television show or movie when it was released.
- Cast
  - The cast that had acted within the television show or movie.
- Director
  - The director of the television show or movie.
- Composer
  - The composer (writer or music) of the television show or movie.
- Writer
  - The writer of the script for the television show or movie
- Runtime
  - The amount of time it takes for the television show or movie from start to end.

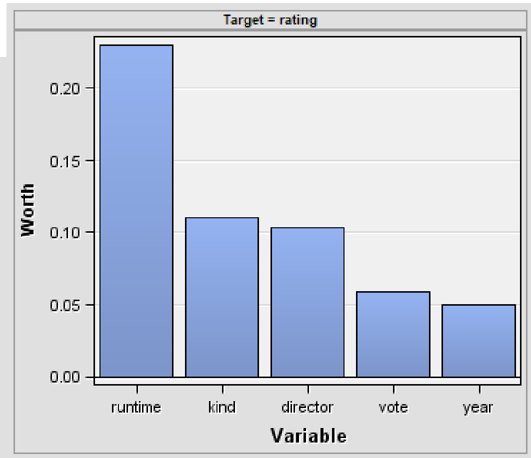
Data Prep Activities

Initially, we retrieved the dataset from Kaggle. Our data was already fairly clean and well organized. However, we needed to remove entries that had missing values. After that, we began working with the data.

Models and Diagrams Used

- Stat Explorer





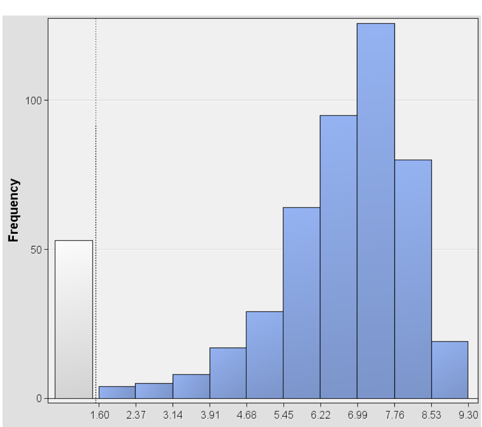

```
Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN
```

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	director	INPUT	513	138		20.26	['Akira Kurosawa']	0.59
TRAIN	kind	INPUT	8	0	movie	56.80	video movie	14.54

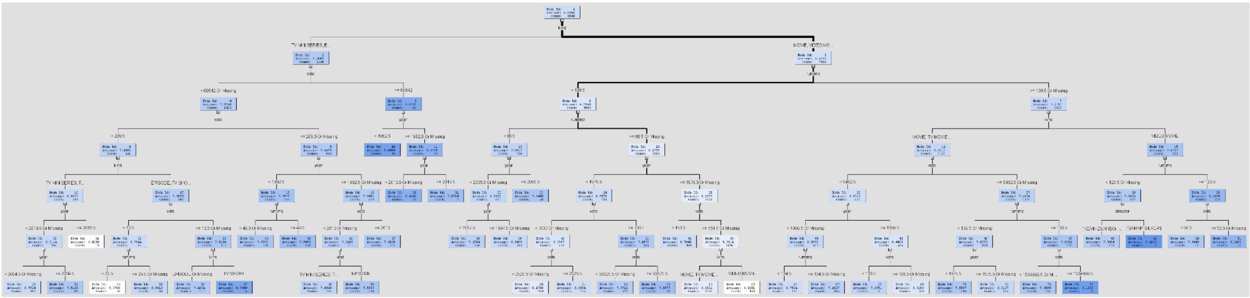
```
Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN
```

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
rating	INPUT	6.680635	1.285113	8949	807	1	6.9	9.6	-0.81107	0.757476
runtime	INPUT	98.11594	63.68129	8763	993	1	94	1620	8.410935	141.9943
vote	INPUT	21218.21	98048.73	8949	807	5	1535	2462087	12.20194	201.5009
year	INPUT	1994.74	16.24509	9756	0	1905	1999	2023	-1.46315	3.023835

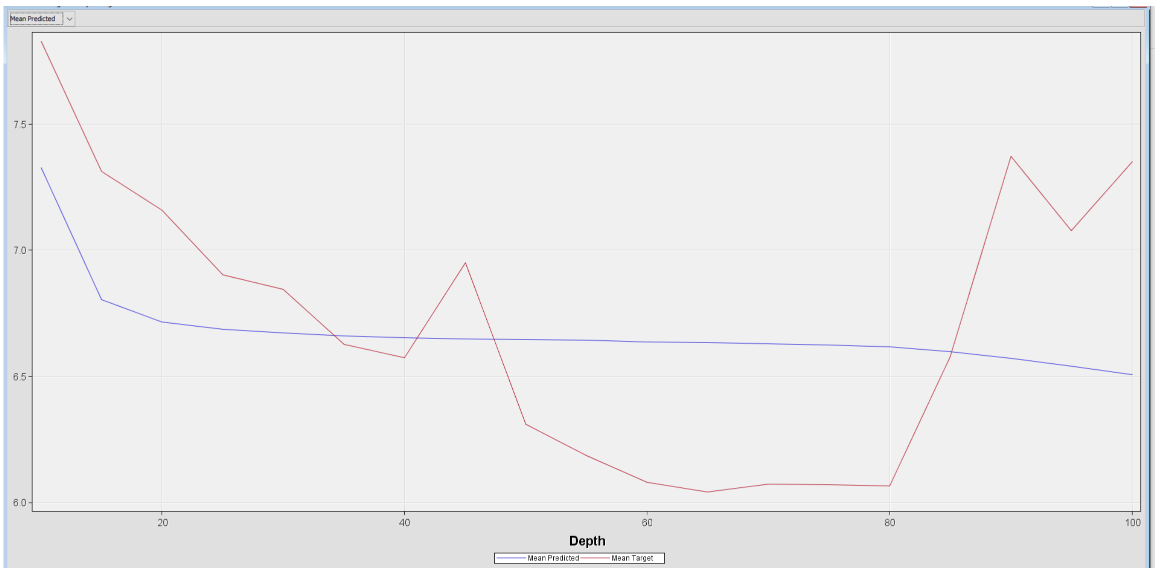
- Graph Explorer

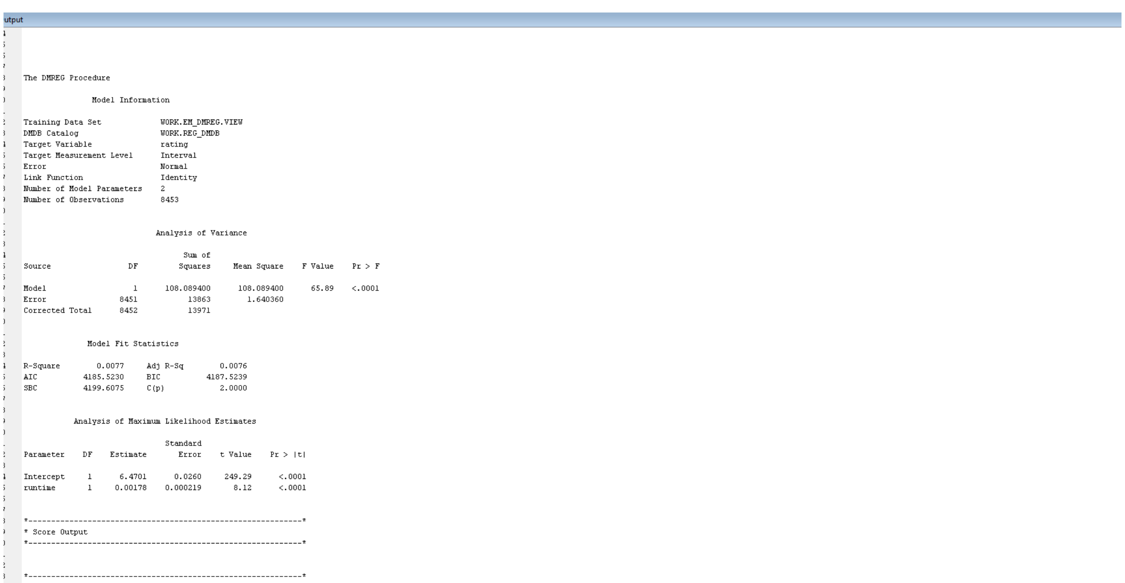


- Decision Tree



- Regression Analysis





```
The DMREG Procedure

             Model Information

Training Data Set            WORK.EM_DMREG.VIEW
DMDB Catalog                 WORK.REG_DMDB
Target Variable              rating
Target Measurement Level     Interval
Error                        Normal
Link Function                Identity
Number of Model Parameters   2
Number of Observations       8453


                        Analysis of Variance

                              Sum of
Source              DF       Squares    Mean Square   F Value   Pr > F

Model                1    108.089400     108.089400     65.89   <.0001
Error             8451         13863       1.640360
Corrected Total   8452         13971


             Model Fit Statistics

R-Square        0.0077    Adj R-Sq       0.0076
AIC          4185.5230    BIC         4187.5239
SBC          4199.6075    C(p)           2.0000


           Analysis of Maximum Likelihood Estimates

                                Standard
Parameter   DF   Estimate          Error   t Value   Pr > |t|

Intercept    1     6.4701         0.0260    249.29     <.0001
runtime      1    0.00178       0.000219      8.12     <.0001


*------------------------------------------------------------*
* Score Output
*------------------------------------------------------------*


*------------------------------------------------------------*
* Report Output
```

Findings

Throughout the analysis process, very interesting things have been discovered. To start we looked into the stat explorer to get a basic understanding of the data we are dealing with. It is interesting to see that vote and runtime both have a positive correlation in comparison to the rating. This means that as the vote or runtime increases, the rating of the film increases as well. We also noticed that year actually has a negative correlation when compared to the rating. This means that as the year increases, the rating of the film decreases. Along with the graphics we could also look at the statistics for the variables of the data that involve numbers. This allowed us to truly understand the data. The statistics allowed us to look into the means of certain variables, showing a mean rating of 6.6. This means the average rating for films on Netflix is 6.6 ratings. We also see a mean of about 98 minutes for a film found on Netflix. An average vote score of 21218 and an average year of 1994 were found for the films found on Netflix.

The next step we took was to look into the graph explorer. Here it gave us a graphic that took a look into the frequency of the ratings occurring. Here we noticed that most of the films were rated between 6 and 7.5. It is interesting to find out that many of the films that are put on Netflix aren't all highly rated. The graphic actually shows a slight left skew. This shows that a good amount of the data is more highly rated, however, there are a decent amount of films that are lower-rated. This is very interesting due to the fact that Netflix doesn't always have highly rated films in their selection.

Next, we took a look at the decision tree node to try and understand how the ratings are related to the other data in the dataset. We notice that the first split is on a basis of kind, splitting it by tv-series and movies. This is interesting to know that the first split of the data with targeting the rating is the kind of film we are looking at. The second split related to the runtime and the vote of the said film. Throughout the entire tree, we can see that the majority of the tree is described by vote, runtime, kind, and year. This is interesting to note because out of the variables given the ones that are describing the ratings the most are vote, runtime, kind, and year. I would understand if the vote and rating were similar in comparison. However, runtime and year are interesting discoveries in describing the ratings.

The next step was looking into a regression analysis between the runtime and the rating variable. The focus point of this analysis was the R-Squared value. The R-Square value was 0.077. This was an unexpected value for this variable due to the higher correlation and worth compared to the other variables. The second part of this analysis was the predicted mean chart given. We noticed the difference in the target mean when the depth reached 45 and at 60. Taking a look at this model, we took a step back and evaluated the significance of the runtime variable.

Managerial Implications

After the data was analyzed, different other factors in the operations must be considered while looking at Netflix. One of these factors is the price that Netflix charges for its streaming services. We felt the need to look at this because the ratings of their movies will affect the price people are willing to pay. If the average movie has a lower rating people will not be willing to pay more for the service. This data was important to improve ratings on upcoming movies to not only keep subscribers engaged but to attempt to persuade new consumers to purchase Netflix. The different variables that were analyzed were the year, runtime, and vote. The rating is used as the target variable.

The first variable that was looked at was the runtime of the film being observed. The runtime was the first variable that was considered due to the worth value given. Compared to the other variables, runtime had the largest value of worth giving more significance to this variable. The runtime of the film has a slight positive correlation with the rating. With a mean of 98.12 minutes for runtime, Netflix might adopt a policy to purchase or produce movies that meet the requirements for runtime. An example of this would be to focus on films that are under 161.80 minutes. This would mean that Netflix's focus on films is less than one standard deviation above the mean. This would imply that consumers enjoy shorter films, using this information Netflix could adopt policies to strengthen their ratings and keep Netflix at a competitive price point.

The second variable that was considered was the year that the film was released. Looking at the Stat Explore portion of the analysis, the year had a slight negative correlation and a lower worth value. As a manager for Netflix, there should be a focus on creating more movies rather than purchasing the rights to older movies. This strategy would help improve ratings.

The data should aid Netflix in creating a long-term strategic plan for the company. With the number of subscribers each month decreasing, Netflix needs to take action to correct this decline. This data will improve the average ratings Netflix offers in its library. This plan should include an action such as adding movies with short runtimes. As well as adding new movies and creating new content to add to the library. This data is not meant to serve as a quick solution but rather as a long-term guide to generating more revenue.

Works Cited

Guna, Akash. “Netflix Prize Shows Information (9000 Shows).” Kaggle, 24 Oct. 2021, https://www.kaggle.com/akashguna/netflix-prize-shows-information.

“Netflix Research.” Netflix Research, https://research.netflix.com/research-area/analytics.