Netflix Motivation Analysis

CS 287 Final Project

Drew Jepsen

Christian Rhodes

ABSTRACT

As of Q1 2023, there are 74.4 million paid Netflix subscribers in the U.S. [1]. Of the estimated 334 million people in the U.S. [2], 1 in 4.5 people subscribe to and watch Netflix content. This study is interested in analyzing popular content and identifying the motivations behind individuals' choices in the content they consume. We utilized a Kaggle dataset that housed Netflix movie and tv show metadata, IMDb data, and The Movie Database data to construct predictive linear models on The Movie Database popularity to gain insight into what programs are the most popular. And to assess which genres have the biggest impact on a show or movie's reception. The final model lacked significant accuracy, but important inferences were able to be pulled from them.

Keywords

Hedonic, Social Influence, Netflix, Linear Models

1. INTRODUCTION

With such a large user base, Netflix has large internal efforts to provide the very best content to their subscribers so as to maintain them for months and years to come. Although this sort of RnD is not publicly available, there is value in trying to understand some of Netflix's business practices via open data sets containing program information. As seen by the rise of video streaming in the last decade, it's clear that any media company would want to leverage this monumental global market. Thus, several companies have poured millions of dollars into creating these services, selecting and creating "popular" content, and supporting their services via personalized algorithms that suggest the very best content to each user.

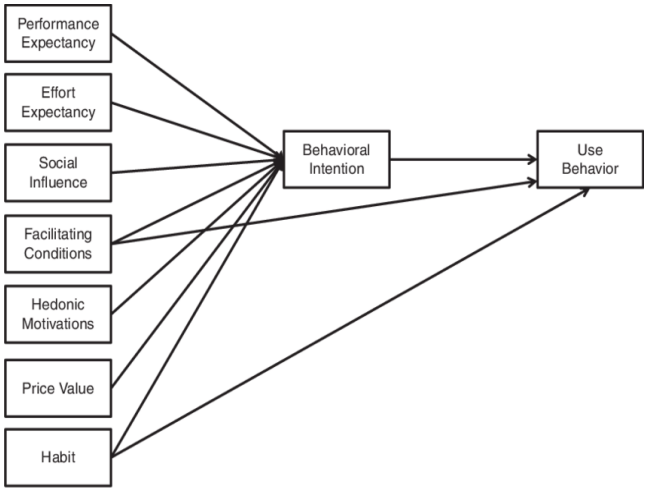
Our research takes a peek into the type of content on Netflix (US) that is most appealing to viewers by analyzing Netflix metadata (such as title, runtime, rating, genre, description, etc.) in conjunction with popularity and ratings from IMDB (The Internet Movie Database) and TMDB (The Movie Database). Using industry-standard data science practices, we attempt to predict the popularity and rating of a given movie or TV show. This kind of prediction is intended to replicate a slice of analysis and decision-making that Netflix performs internally for every program. From something as immature as a pitch using an incomplete script, it may be possible for Netflix (or anybody who uses some of our pipeline and prediction models) to predict how "popular" the finished, fully produced, final product will be.

Beyond the goal of predicting popularity, the results of this project can be used to make inferences about both the business side of Netflix and streaming and the consumer side. There are many applications of this, including identifying population consumption behaviors, predicting market trends, assuming the social influence or other facilitating conditions of a given movie or TV show, and others.

2. RELATED WORK

As discussed in the abstract and introduction, Netflix is an incredibly successful company that became synonymous with home entertainment. *The Netflix Effect*, by Kevin McDonald and Daniel Smith-Rowsy [3] reviews the success story of Netflix through the 21st century. Because of the large user base and diverse demographics that Netflix boasts, a lot can be learned about user expectations and motivation when using the service. This book identifies social, political, psychological, and economic aspects as mandating the relationship between content provision and consumption. This is a bidirectional relationship in which each influences the other. An example of this that was examined in the book is the popular Netflix show, Orange is the New Black. Studies referenced inside the book showed that the show improved attitudes towards homosexual and transgender people. These sorts of results are interesting, but not covered in this project.

The study *Netflix, Who Is Watching Now?*, by Cristóbal Fernández-Robin et al. [4], uses the UTAUT2 (Unified Theory of Acceptance and Use of Technology) model to model Netflix user behaviors. This study relates to advertising and creating profiles for users to provide them with content they are most likely to watch and enjoy. Hedonic Motivation and Social Influence are relevant features of this research. In a sense, our research project aims to simplify the results of this study by Cristóbal Fernández-Robin et al. A survey was performed with 415 Chileans and found that: "considering that Chile is a country with high-stress rates in the population, the need for distractive and hedonic elements is very high." (pg. 211). These results are similar to what we are trying to achieve with Netflix US. Where *Netflix, Who Is Watching Now?* used the UTAUT2 model and a survey that utilized a 7-point Likert scale to base its results upon, we are using program metadata and popularity + rating scores with various data science models to achieve a similar kind of result.



Source: Venkatesh *et al.* (2012)

Figure 1: UTAUT2 Model

3. METHODS

3.1 Data Exploration

The Netflix dataset used in this study was collected by Kaggle user, Victor Soeiro, and includes a total of 5850 entries spanning from the 1970s to 2022, each with 15 features. It should be noted that the dataset is heavily skewed toward modern movies and shows (2015-today). This bias could be attributed to Netflix's emphasis on "Netflix Originals" and the relative cost-effectiveness of acquiring the rights to modern productions compared to older classics. Additionally, the ease of accessing auxiliary information for modern productions may be a contributing factor.

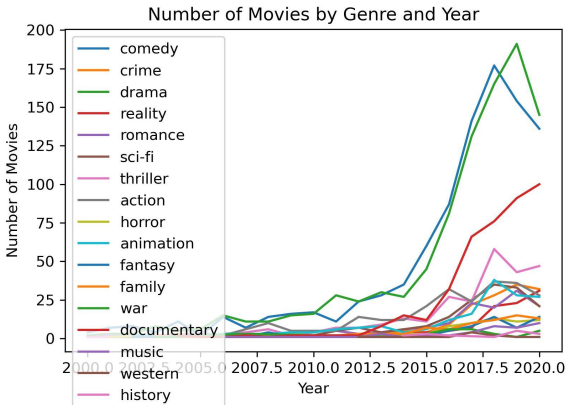


Figure 2: Graph showing the modern skew of movie and show releases.

Nine of the 15 features in the dataset were pulled directly from Netflix's metadata, including the id (a unique Netflix identifier), title, type (show or movie), description, release year, age certification, runtime, genres, production countries, and seasons. The remaining features were obtained from other sources. Three quantitative columns, namely imdb_id, imdb_score, and imdb_votes, were obtained from IMDb's user review scores for each movie. The scores and votes are of high interest for the research, as they should be directly correlated with viewer engagement. Two other features, tmdb_popularity, and tmdb_score, were obtained from The Movie Database. The tmdb_score is similar to the imdb_score but is on a different scale. To utilize both features, some form of regularization is necessary.

It should be noted that not all features in the dataset are useful for predicting popularity in future programs. Fields like title and description may be too unique to be helpful in this regard, and some features may be similar in concept, leading to multicollinearity. Therefore, it is crucial to inspect correlations among the features to identify which ones are most likely to contribute to predicting viewer engagement.

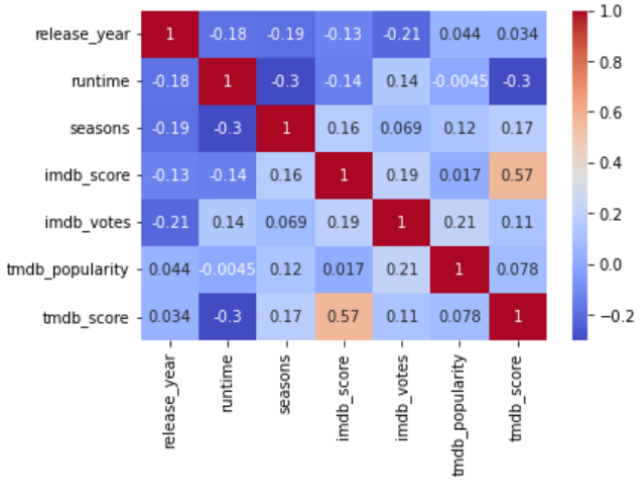


Figure 3: Correlation Matrix, comparing all the quantitative features of the data set. Only tmdb_score and imdb_score are shown to have any significant correlation. Note: these features have been normalized.

3.2 Data Cleaning

In our case, the Netflix dataset had only a few missing values (NaN values). We removed these values as we determined they would not significantly affect the overall data. After removing these rows, we were left with a clean dataset that could be used for further analysis with 5131 total entries and nine fields.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 5131 entries, 1 to 5849
Data columns (total 9 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   type             5131 non-null   object
 1   release_year     5131 non-null   int64
 2   runtime          5131 non-null   int64
 3   genres           5131 non-null   object
 4   seasons          5131 non-null   float64
 5   imdb_score       5131 non-null   float64
 6   imdb_votes       5131 non-null   float64
 7   tmdb_popularity  5131 non-null   float64
 8   tmdb_score       5131 non-null   float64
dtypes: float64(5), int64(2), object(2)
memory usage: 400.9+ KB
```

Figure 4: Table displaying the number of non-null values found in the finalized set of features. There are none as should be expected from the cleaning process.

However, we recognized that the dataset was complex, with 15 features, and wanted to simplify it to improve the accuracy of our models. Therefore, we decided to remove many of the features, including the title and description fields, which are unique to each entry and are unlikely to be helpful in predicting the popularity of future programs. We also removed some features that were conceptually similar, such as tmdb_score and imdb_score, to avoid multicollinearity issues in our models.

However, we retained all the quantitative features as we found little correlation among them during the exploration process. These features, including the imdb_score and imdb_votes, were of high interest for our research, as they should be directly correlated with viewer engagement.

Moreover, we decided to create 2 new features, the major one is called "genreMain," to simplify the dataset further. This feature holds the main genre for each entry, which is obtained from the first genre in the "genre" column. By focusing on the primary genre of each entry, we could reduce complexity and minimize the number of dummy variables required for one-hot encoding. The other engineered feature is tmdb_category which is just tmdb_popularity put into three bins named low, medium and high. This feature is used when creating the classification tree.

To ensure the quality and accuracy of the data, we performed data normalization and standardization to ensure that all the features were on the same scale. Additionally, we conducted a thorough analysis of the dataset to identify and address any inconsistencies, errors, or outliers.

In summary, the data cleaning process involved removing missing values, simplifying the dataset by removing unnecessary features, and constructing a new feature for easy analysis. We also normalized and standardized the data to ensure its quality and accuracy. The result was a high-quality dataset ready for further analysis and modeling.

4. RESULTS

4.1 Predictive Models

After careful consideration of what data we had access to, we chose to predict tmdb_popularity as it is the closest to our focus on what types of shows and movies are popular and why. The fields we kept to train this model were 'release_year', 'runtime', 'seasons', 'imdb_score', 'imdb_votes', 'tmdb_score', and our engineered feature, 'genreMain.' GenreMain was transformed into 13 one-hot encoded fields for each of the genres: 'documentary', 'drama', 'fantasy', 'war', 'comedy', 'thriller', 'crime', 'romance', 'action', 'western', 'history', 'music', 'horror', 'sci-fi', 'animation', 'family', 'reality', 'unknown', and 'sport'.

Firstly, we created a linear model passing in all of these fields. We made this first model due to its simplicity and worry that a non-linear regression model would be over complicated and too high-dimensional. After creating the model, the RMSE was 137.59 with an R-squared of 0.27 when scoring based on the testing data.

We next created a Lasso regression model due to our fear of model overfitting that came up when inspecting the scatter plot and coefficient importance. This model's accuracy was similarly bad, but the coefficient importance gave us great inference into what aspects contribute to popularity in shows and movies. It produced an RMSE of 18929.44 and an R-Squared of 0.27.

Next, we wanted to see if some other model would improve the accuracy of our models, so we created a Random Forest Regressor, with a max depth of 3 to maintain readability. The RMSE came out to be 80.23 with an R-squared of 0.17 based on the testing data.[1]

The final attempt at a model that we wanted to create utilized another engineered feature, which we called tmdb_category, which subsist of tmdb_popularity separated into 3 bins named "low"(0 to 3), "medium"(4-7) and "high"(8-10). With this feature, we created a classification tree with a max depth of 3. This model was the best by far with an accuracy of 61%.[2]

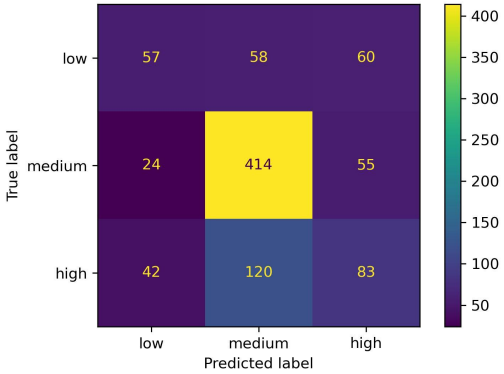


Figure 5: Confusion Matrix of the Classification Tree predicting tmdb_category

4.2 Feature Importance

In addition to using linear models to predict Netflix program popularity, we analyzed feature importance by examining the coefficients of the models. While our linear models may have limitations in providing accurate predictions, analyzing feature importance allows us to gain valuable insights into the factors that influence program popularity.

Our findings indicated that the genre of a show or movie has the most significant impact on its popularity, surpassing other metadata such as runtime or release year. Specifically, we observed that the action and sci-fi genres had the most significant positive impact on a program's reception, while the reality and sports genres had the most significant negative impact.

Furthermore, we visualized the coefficient importance graph for the lasso regression model, which helped us identify which genres had the most significant impact on a program's popularity. The graph revealed that the sci-fi genre had the most positive impact, while the reality genre had the most negative impact.

Our analysis of feature importance highlighted the crucial role of genre in determining program popularity on Netflix. Content creators and producers can leverage this insight to tailor their content to the preferences of Netflix subscribers, increasing their chances of success on the platform.

[1] Image of the Random Forest Regressor, Section 10.1

[2] Image of the Classification Tree, Section 10.2

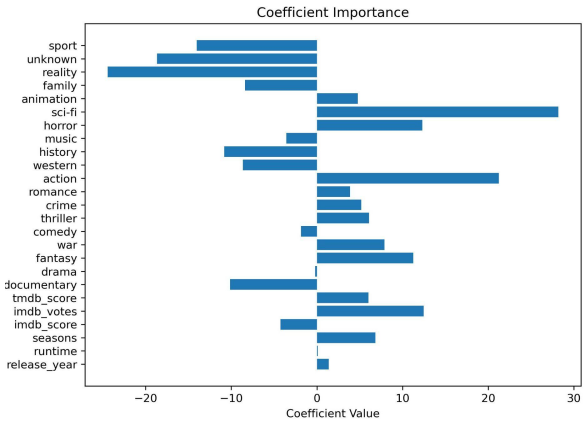


Figure 6: Graph of Coefficient Importance based on the constructed Linear Model.

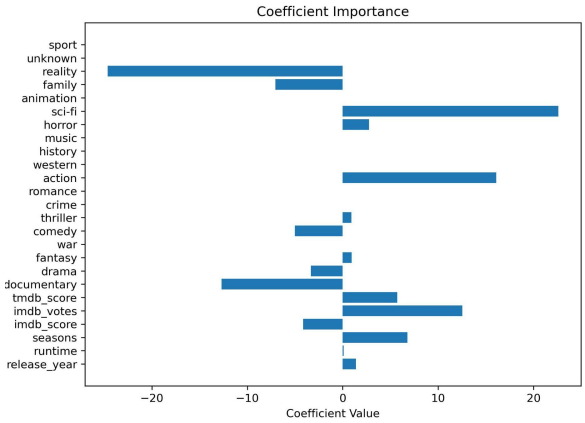


Figure 7: Graph of Coefficient Importance based on the constructed Lasso Model.

5. DISCUSSION

Our results show that genre is the most significant determinant of popularity, demonstrating that runtime, season count, or age rating have a low correlation to popularity. Assessing popularity of media is complex and accounts for all sorts of factors that we did not have access to in our research. For example, a global or national event may direct millions of people to a particular show or movie. When the 2020 Covid-19 virus ravaged the United States and the world, suddenly millions of people were stuck at home and free to consume more media than ever before. The Netflix original series Tiger King released at the onset of the pandemic on March 20th, it exploded in popularity and had an estimated 19 million viewers in the first 10 days of its release [6]. The reason for Tiger King's incredible popularity could not be explained in our data, since the Covid-19 pandemic is not reflected therein.

A limitation of our research was connecting our dataset to our goal. Either more data collection would have been necessary, or much more robust feature engineering with validation would have been required. Because of this, we should have limited our scope and focused strictly on popular content from the start. Something else we tried but did not have much success with was extracting features from the 'description' feature. There did not seem to be any terms or phrases with a high correlation to popularity (IMDB or TMDB).

Our implementation of the data science lifecycle into our project needs improvement. Feature engineering is a crucial stage of the data science pipeline that would benefit our project. After our data preparation was complete and aligned with our expectation of a great dataset, we did not work backward after we began modeling. Although we engineered some features via one-hot encoding and scaling, we did not experiment in depth with imputing null or suspicious values or try to compute features relating to "motivations", such as 'escapism', or 'education'. After iterating through the data science lifecycle once, we thought it was best to make connections and conclusions to these motivations via our model's results. However, in retrospect, it may have been beneficial to spend more time engineering features like these.

6. FUTURE WORK

Due to the time restraints, our best model's predictive power was not very strong. Beyond doing more EDA, possibly aggregating more data, and improving the models, future work involves collecting data from human research. Survey data could directly collect user motivations, which would be incredibly valuable for our research. Simply collecting their motivations with what they watched during a period of time would allow us to combine that data with our dataset of program info and begin feature engineering for other titles not mentioned by the survey. Even with a sample of 100 Netflix users, we are confident that we could use multiple imputation and regression imputation to populate existing titles with motivations.

Additional data collection involving demographic data could be useful for linking motivations with types of content. In the *Netflix, Who Is Watching Now?* study, a survey of Chilean Netflix users found that escapism is a motivation due to the stressful conditions of the country [4]. This demonstrates that more factors than demographics are in play, such as lifestyle and geopolitical conditions.

In conclusion, future work includes capturing more data (direct motivations via survey, demographic data, lifestyle data, etc.), aggregating it with our existing program metadata data, or using it in concordance with it to better formulate and present results, and analyze + visualize those results.

7. CONCLUSION

After investigating popular content on Netflix, we have determined that popularity is difficult to predict. That being said, with our dataset of over 5500 Netflix titles, some of the most popular genres are reality, sci-fi, action, comedy, and documentary. This is reflected in section 4.2 and figures 4 + 5 when predicting tmdb_popularity. Since these genres are generally popular over time, little information was gained regarding user motivations.

8. ACKNOWLEDGMENTS

Our thanks to the class, CS 287 Data Science 1, especially Nick Cheney for their feedback. And our great thanks to Victor Soeiro for the construction of the database.

9. REFERENCES

[1] Stoll, J. (2023, April 20). U.S.: Quarterly Netflix paid subscribers count 2023. Statista. https://www.statista.com/statistics/250937/quarterly-number-of-netflix-streaming-subscribers-in-the-us/

[2] Bureau, U. C. (2022, December 29). New Year's Day 2023: January 1, 2023. Census.gov. https://www.census.gov/newsroom/stories/new-years-day.html

[3] McDonald, K., & Smith-Rowsey, D. (2016). The netflix effect: Technology and entertainment in the 21st Century. Bloomsbury Academic, an imprint of Bloomsbury Publishing Inc.

[4] Fernández-Robin, C., McCoy, S., Yáñez, D., Hernández-Sarpi, R. (2019). Netflix, Who Is Watching Now?. In: Meiselwitz, G. (eds) Social Computing and Social Media. Design, Human Behavior and Analytics. HCII 2019. Lecture Notes in Computer Science(), vol 11578. Springer, Cham. https://doi.org/10.1007/978-3-030-21902-4_15

[5] Soeiro, V. (2021). Netflix TV Shows and Movies. Kaggle. https://www.kaggle.com/victorsoeiro/netflix-tv-shows-and-movies

[6] Maglio, T. (2021, November 23). *"tiger king 2" debuts at no. 2 on Netflix's top 10 list with 30 million hours watched*. TheWrap. https://www.thewrap.com/tiger-king-2-ratings-netflix-joe-exotic/

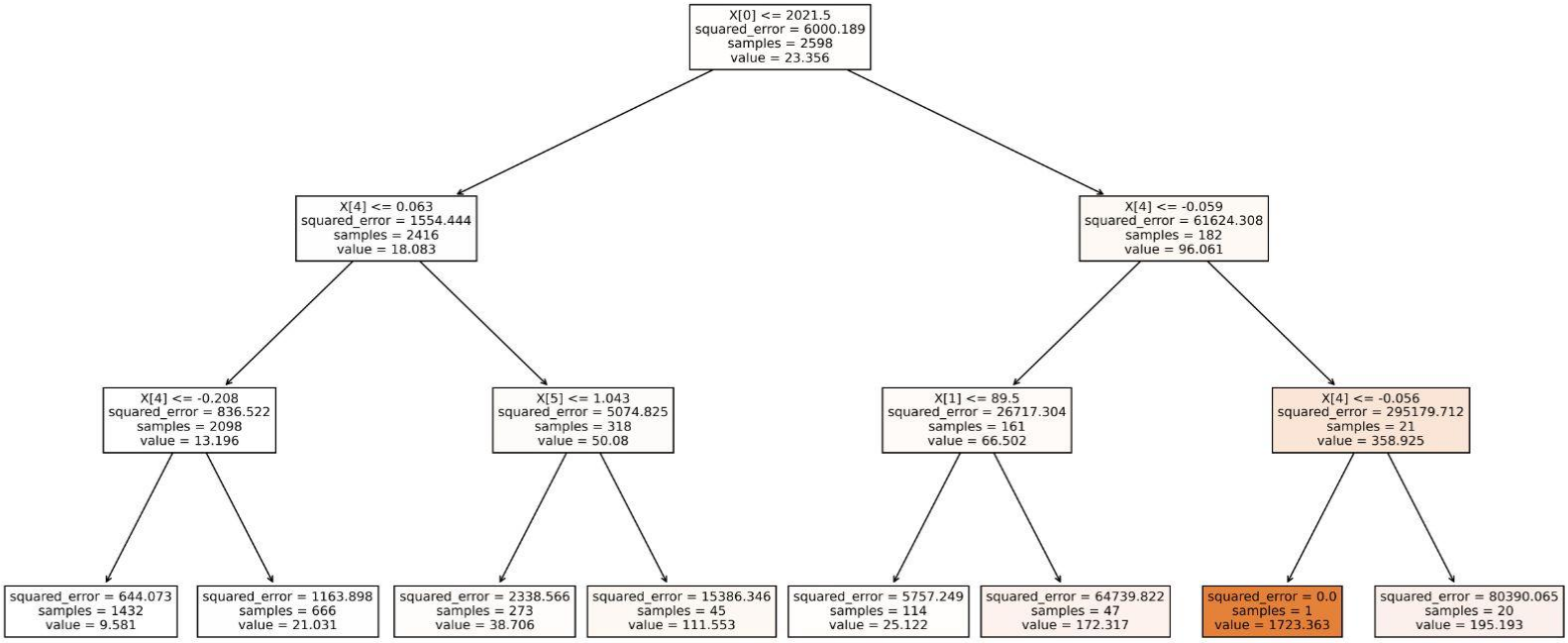
About the authors:

Drew Jepsen is a junior computer science and statistics student at the University of Vermont. He is interested in model design and backend server development.

Christian Rhodes is a senior computer science student at the University of Vermont. His focus is on web development, where he designs and develops full-stack web applications. Christian currently does contract web development work for various clients, ranging from small businesses to web-3 startup companies.

10. APPENDIX

10.1 Random Forest Regressor Tree Predicting tmdb_popularity



10.2 Classification Tree Prediction tmdb_category