Netflix Data Analysis

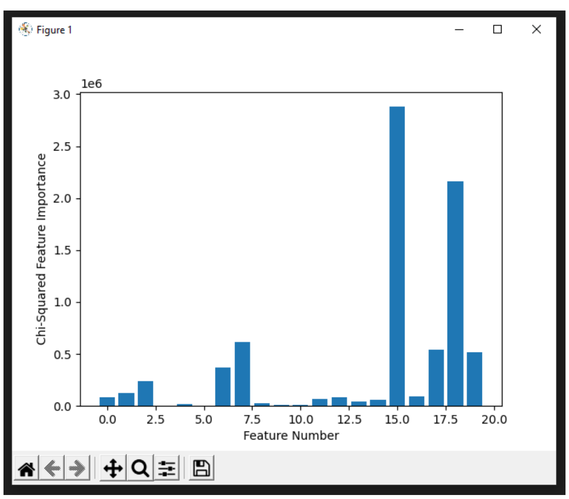
1. Introduction

The plethora of data found in the Netflix 2021 dataset can be used in many ways. My analysis on the Netflix dataset involves the hidden gem score of each movie or show, and figuring out which features are correlated with that score. I used multiple feature selection techniques to find the best features related to the hidden gem score, and then trained a machine learning algorithm on those features to learn which features best predict the score.

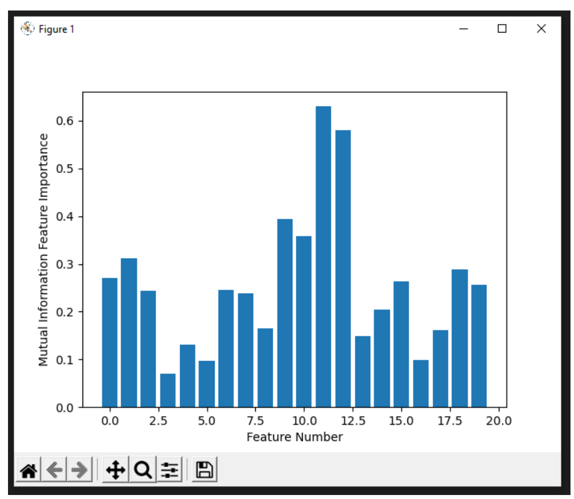
2. Pre-processing and Feature Selection

Initially, there were a couple features that weren't useful in finding the hidden gem score because the features were unique to each movie (title, links, poster, and trailers). There were also a lot of missing data within several features. To solve this, I deleted the useless features from the dataset, and put a value of 0 or -1 to cells that were missing a value. I also converted all categorical data into numerical data as this would be necessary to perform feature selection.

After completing all this, I performed three feature selection techniques on the dataset: Chi-Squared Feature Selection, Mutual Information Feature Selection, and the Recursive Feature Elimination. The results of all three are shown in the pictures below.



(a) Chi-Squared Feature Importance



(b) Mutual Info Feature Importance

```
Feature 0: 8 | Feature 1: 5 | Feature 2: 10 | Feature 3: 20 | Feature 4: 14 | Feature 5: 19 | Feature 6: 16 | Feature 7: 9 | Feature 8: 13 | Feature 9: 6
Feature 10: 4 | Feature 11: 2 | Feature 12: 3 | Feature 13: 18 | Feature 14: 12 | Feature 15: 15 | Feature 16: 11 | Feature 17: 7 | Feature 18: 17 | Feature 19: 1
```

(c) Recursive Feature Elimination Ranking

Feature 0: Genre	Feature 1: Tags	Feature 2: Languages	Feature 3: Series and Movies
Feature 4: Country Availability	Feature 5: Runtime	Feature 6: Director	Feature 7: Writer
Feature 8: Actors	Feature 9: View Ratings	Feature 10: IMDb Score	Feature 11: Rotten Tomatoes Score
Feature 12: Metacritic Score	Feature 13: Awards Received	Feature 14: Awards Nominated For	Feature 15: Box Office
Feature 16: Release Date	Feature 17: Netflix Release Date	Feature 18: Production House	Feature 19: IMDb Votes

Table 1: Feature Number to Corresponding Name

The Chi-Squared Test rated features 15, 18, 7, 17, and 19 or the box office, production house, writer, Netflix release date, and IMDb votes as the top five most important features related to the hidden gem score. Mutual Information deemed features 11, 12, 9, 10, and 1 as the top five, while Recursive Feature Elimination found features 19, 11, 12, 10, and 0 as the top five features. Mutual Information and Recursive Feature Elimination shared three features, which were Rotten Tomatoes Score, Metacritic Score, and IMDb Score. The Chi-Squared Test, on the other hand, only shared one feature with Recursive Feature Elimination, IMDb Votes, and shared no features with Mutual Information.

3. Machine Learning Algorithm

After learning which five features were preferred by each feature selection method, I wanted to learn which five features were the best at predicting the hidden gem score determined by FlixGem. I trained a Random Forest Classifier, a supervised machine learning algorithm, on each of the five features, and the data is shown below in the next three pictures.

Chi-squared Test

	precision	recall	f1-score	support
0	1.00	0.50	0.67	4
1	0.08	0.04	0.05	83
2	0.26	0.26	0.26	243
3	0.36	0.45	0.40	387
4	0.31	0.19	0.24	246
5	0.18	0.16	0.17	57
6	0.24	0.17	0.20	112
7	0.29	0.31	0.30	222
8	0.54	0.66	0.59	487
9	0.17	0.05	0.07	44
accuracy			0.38	1885
macro avg	0.34	0.28	0.29	1885
weighted avg	0.35	0.38	0.36	1885

(a) Random Forest with Chi-Squared 5 Features

Mutual Information

	precision	recall	f1-score	support
0	1.00	0.50	0.67	4
1	0.89	0.94	0.91	83
2	0.88	0.91	0.89	243
3	0.74	0.78	0.75	387
4	0.71	0.77	0.74	246
5	0.07	0.04	0.05	57
6	0.20	0.10	0.13	112
7	0.26	0.27	0.27	222
8	0.64	0.68	0.66	487
9	0.53	0.45	0.49	44
accuracy			0.64	1885
macro avg	0.59	0.54	0.56	1885
weighted avg	0.62	0.64	0.63	1885

(b) Random Forest with Mutual Info 5 Features

Recursive Feature Elimination

	precision	recall	f1-score	support
0	0.67	0.50	0.57	4
1	0.89	0.95	0.92	83
2	0.90	0.90	0.90	243
3	0.73	0.77	0.75	387
4	0.68	0.71	0.69	246
5	0.30	0.25	0.27	57
6	0.47	0.28	0.35	112
7	0.50	0.50	0.50	222
8	0.69	0.75	0.72	487
9	0.54	0.32	0.40	44
accuracy			0.69	1885
macro avg	0.64	0.59	0.61	1885
weighted avg	0.68	0.69	0.68	1885

(c) Random Forest with RFE 5 Features

The statistics in the three pictures reveal that the features preferred by the Recursive Feature Elimination scored better on average with respect to the f1-score ($0.61 > 0.56 > 0.29$) and the accuracy ($0.69 > 0.64 > 0.38$). Thus, IMDb votes, Rotten Tomatoes scores, Metacritic scores, IMDb scores, and the genre of a movie or show are correlated and the best predictors for the hidden gem score.

4. Implications and Future Works

Thinking about it logically, it makes sense that the hidden gem score correlates more with movie or show scores than with other features. After all, if the movie or show scores are all high across reviews, naturally, the hidden gem score will be high as well. In the future, it would be better to test features other than the scores and see how well those correlate with the hidden gem score. It would also be interesting to see what features correlate with box office numbers. Knowing what features correlate with higher box office numbers may actually help in the real world because movie producers can focus more on these features to produce higher box office numbers. These two new ideas are something I will likely test in the near future.