Predicting Individual Media Consumption with Passive Behavioral Data

Tomer Zur

The Harris Poll
tomer.zur@harrispoll.com

Introduction

- Predicting and recommending TV shows and movies for users to watch is integral to the success of any streaming service and a big driver of revenue and retention.
- In this project, we take the watch histories of Netflix subscribers and use them to:
  - recommend shows and movies for users to watch
  - predict which shows/movies each user will watch next.

Data

We use two data sources:

Samba TV data (user activity):

- Netflix users from before 9/26/2021
- User id, title, start time, end time, duration, season, episode, genre
- >38 million rows, almost 900,000 users

Netflix show data – from Kaggle:

- Information about over 8000 Netflix shows and movies
- Type (TV Show/Movie), title, country, date_added, release_year, rating (PG/R etc.), duration (# minutes for movies, # seasons for shows), listed_in (list of genres), description

Find this poster online!

Objective #1: Recommending Movies and TV Shows

Content Based Recommendation

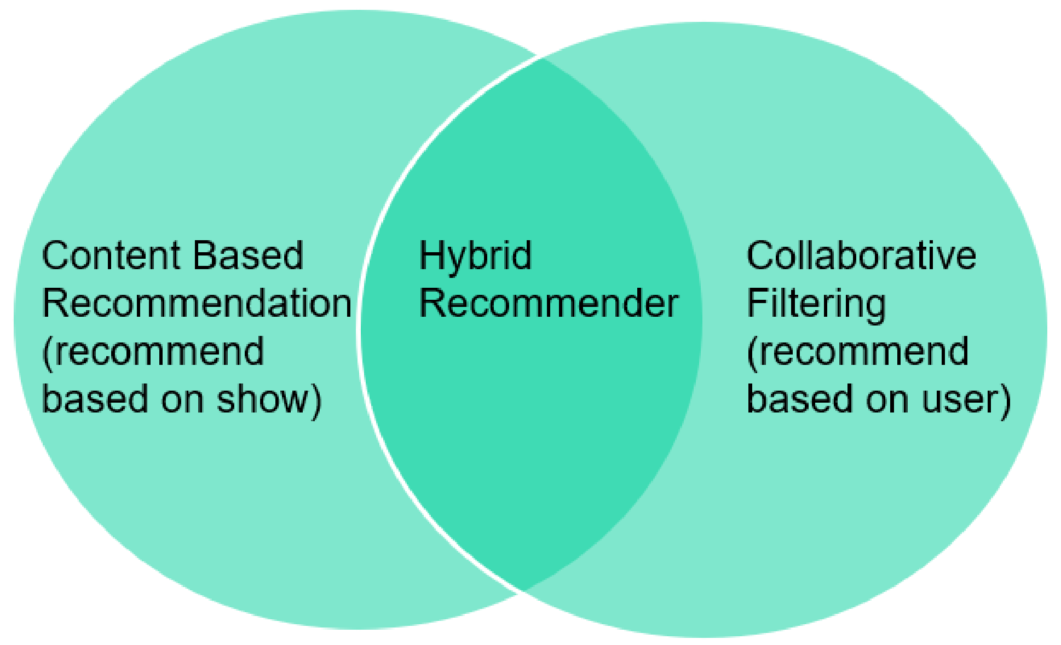
- Recommend shows that are most similar to a given show
- Use cosine similarity and NLP to choose these shows
- Ex. – shows most similar to Stranger Things are:
  - Chilling Adventures of Sabrina
  - Hemlockgrove
  - Beyond Stranger Things
  - Twin Peaks
  - Manifest

Collaborative Filtering

- Recommend shows that similar users to a given user watched/enjoyed
- Use SVD (singular value decomposition) and linear algebra to find similar users
- Ex. - many users watched both Stranger Things and Emily in Paris:
  - User x watches Stranger Things -> gets recommended Emily in Paris

Hybrid Recommendation

- Combination of Content Based Recommendation and Collaborative Filtering
- Predict shows that are recommended by both recommendation methods


Content Based Recommendation (recommend based on show)

Hybrid Recommender

Collaborative Filtering (recommend based on user)

Methods would need to be implemented in real-time to determine which recommendation method is best.

Objective #2: Predicting Next Show Watched

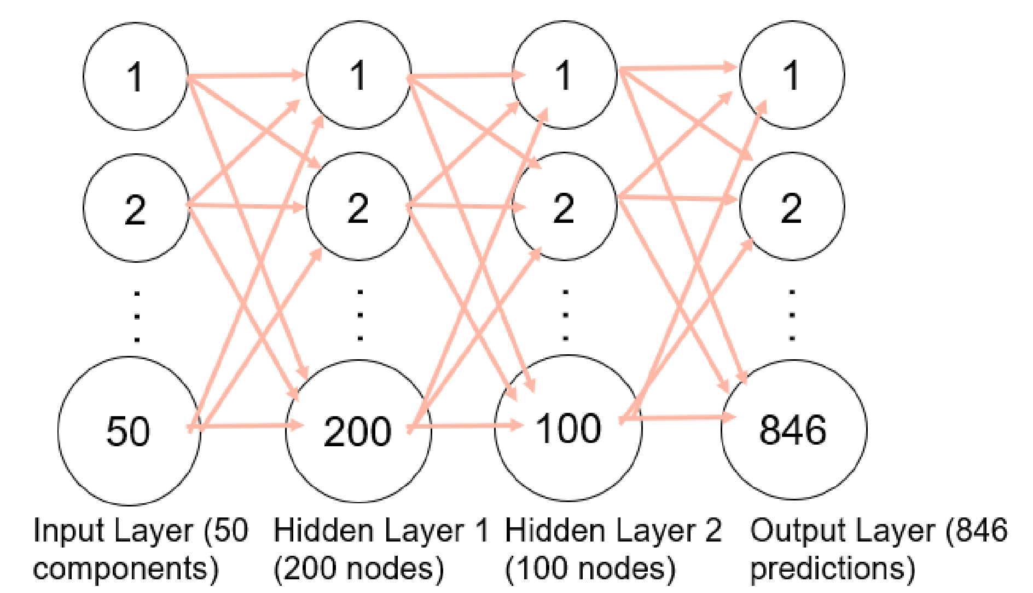
Attributes used to make predictions:

- watch time per show (1 hot encoded)
- # shows watched by genre, country, rating, movie/tv show

Model Used:

Neural Network (type of neural network)

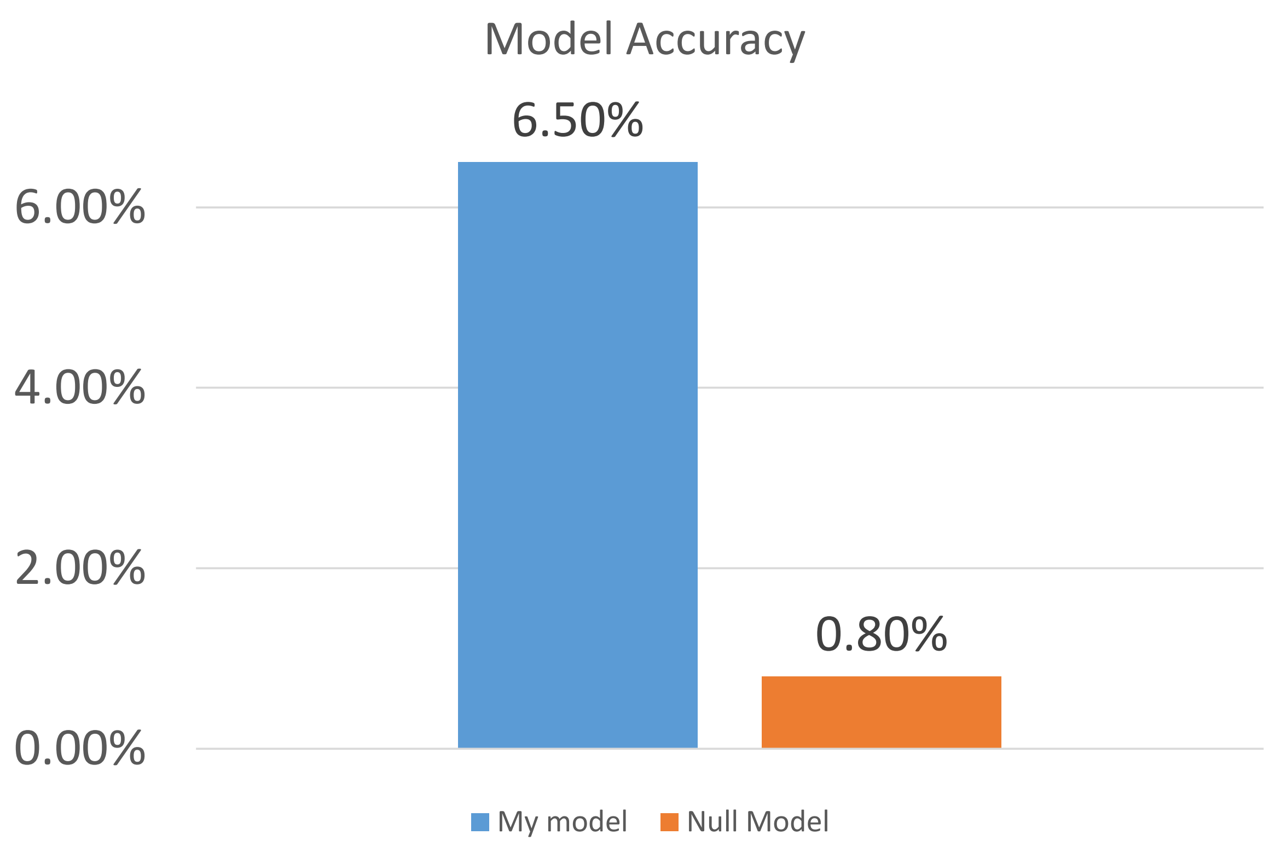
- We are predicting the next show a user will watch
- We have 1 model that predicts the probability of each show being the next show


1
2
50
1
2
200
1
2
100
1
2
846
Input Layer (50 components)
Hidden Layer 1 (200 nodes)
Hidden Layer 2 (100 nodes)
Output Layer (846 predictions)

Results:

Best accuracy I got was: 6.5%

- 8x better than null model



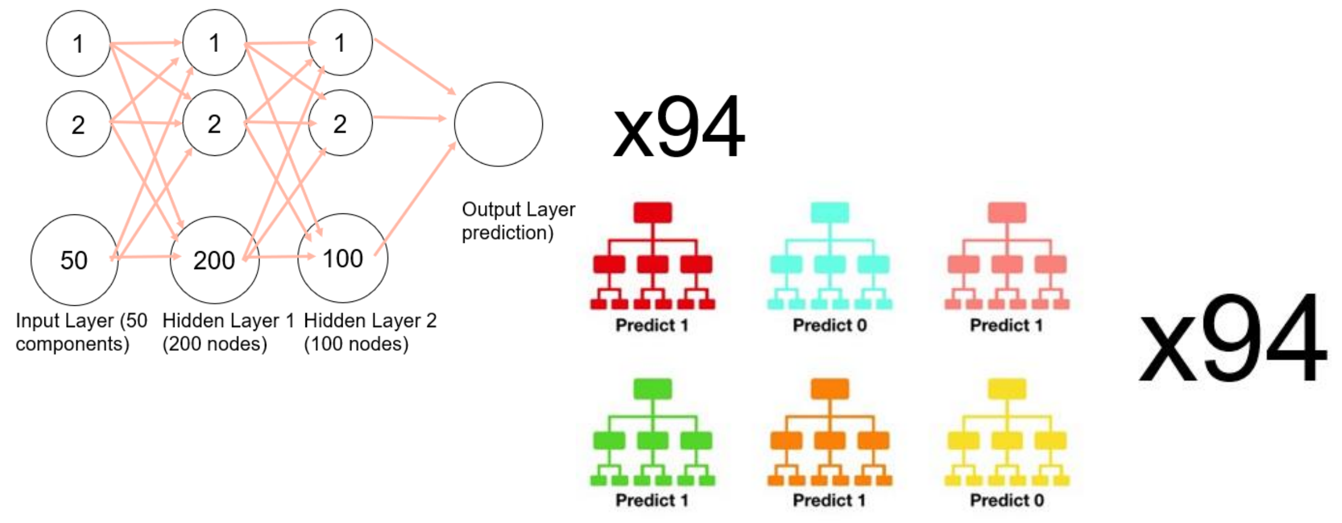
To increase the usability of the model, we looked at all shows watched in the next 60 days rather than the next show.

Objective #3: Predicting Shows Watched in Next 60 Days

Models Used (continued):

Neural Net vs Random Forest

- Predicting shows a user will watch in the next 60 days
- We have 1 model per show



Results:

Neural Model #2:

- Predicted whether each user will watch a show
- Made these predictions for all shows with enough viewer data
- Confusion matrix across all shows:

	Pred. Negative	Pred. Positive
Actual Negative	39755	9358
Actual Positive	1489	458

- Total accuracy: 78.76%, F1 score: 7.7%
- Accuracy (shows watched in next 60 days): 19.04%
- We can advertise to 9816 users (19.2% of users)
- For these users, our predictions have 4.7% accuracy

Random Forest:

- Same data used as Neural Model #2

	Pred. Negative	Pred. Positive
Actual Negative	47484	1629
Actual Positive	1839	108

- Total accuracy: 97.92%, F1 score: 1%
- Accuracy (shows watched in next 60 days): <1%
- (Overfitted more than the neural network)
- We can advertise to 1737 users (3.4% of users)
- For these users, our predictions have 6.2% accuracy

Conclusions

- Implemented multiple recommendation models
- Prediction model for next show was 8x more accurate than null model, but nominally still low (6.5%)
- When predicting over the next 60 days: Neural model reached a far larger audience than the random forest, but had less accuracy

Next Steps

- Include user ratings
- Incorporate more features into dataset
  - Ex. – show descriptions/tags
- Try predicting something less specific than watching a specific show
  - Ex. – watching a specific genre
- Change prediction timeframe
  - Instead of predicting shows watched over next 30 or 60 days, try next 75 or 90 days