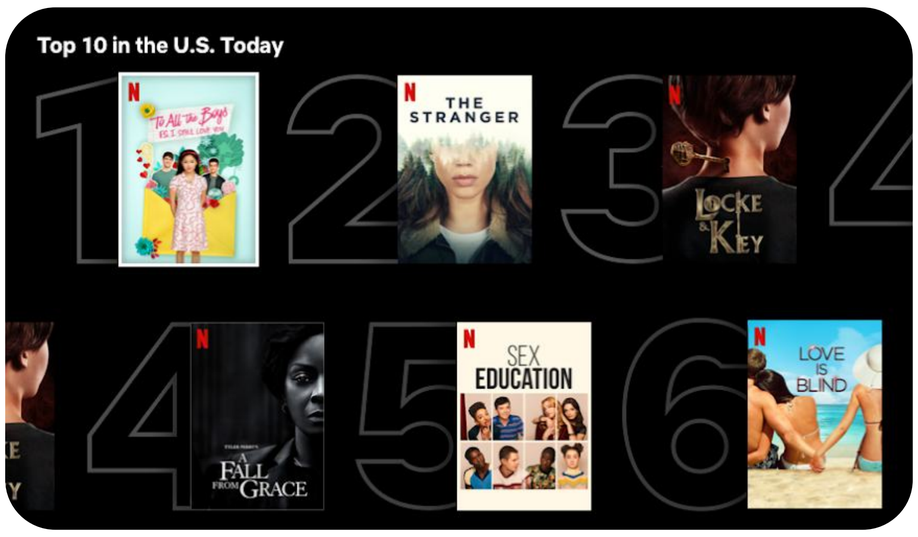
Examining the Netflix Top 10 Feature

Motivation/Introduction

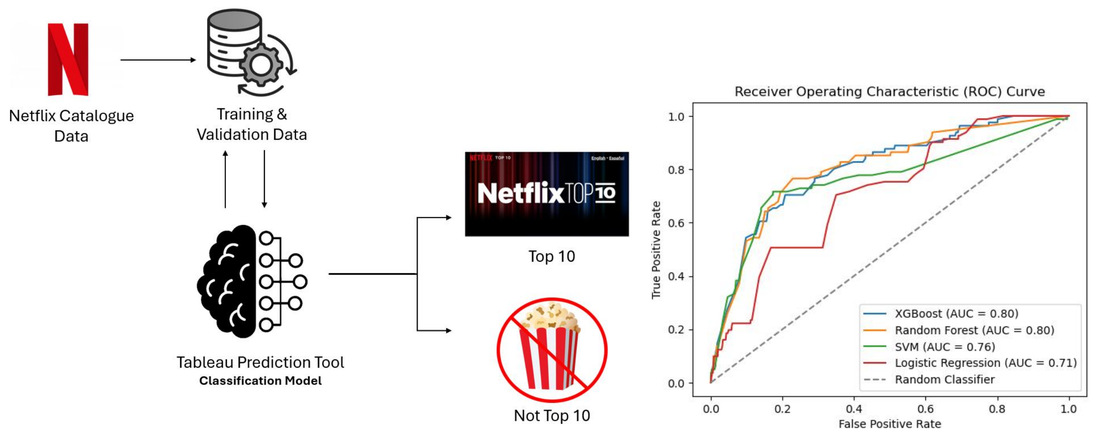
Welcome to our research project, aimed at dissecting the intricacies of Netflix's Top 10 rankings. While the streaming giant regularly updates its list, providing viewers with popular options, it lacks deeper insights into what truly drives a movie to the coveted Top 1 position. Recognizing the need for more informed consumption choices, we set out to delve into the key factors influencing these rankings.

Our project goes beyond surface-level popularity, focusing on deciphering the intricate workings of recommendation systems. By harnessing data analytics and visualization concepts, we've created a dynamic Tableau dashboard. This innovative tool empowers viewers to make informed decisions about their next watch, while also providing invaluable insights for business executives in the entertainment industry.

Whether you're a viewer seeking the perfect movie or a creator/business looking to optimize production strategies, our project offers a comprehensive solution to navigate the complex world of streaming content.



Process Flow



Our Approach

Goal: Predict the likelihood of a show securing a position in the Netflix's Top 10.

Methodology: Utilized a **classification machine learning** model for predictive analysis.

Data Processing: Pre-processed data meticulously using **regex** for detailed extraction, **one-hot encoding** for categorical features, and **feature selection** for optimal model parameters.

Balancing Data: Addressed inherent imbalances within dataset by using the **SMOTE** technique for robust analysis.

Advanced Algorithms: Compared and employed **XG Boost, Random Forest, SVM, and logistic regression** for classification techniques.

Validation: Validated results by using **K-fold cross validation** and a series of metrics such as AUC and accuracy to assess model performance.

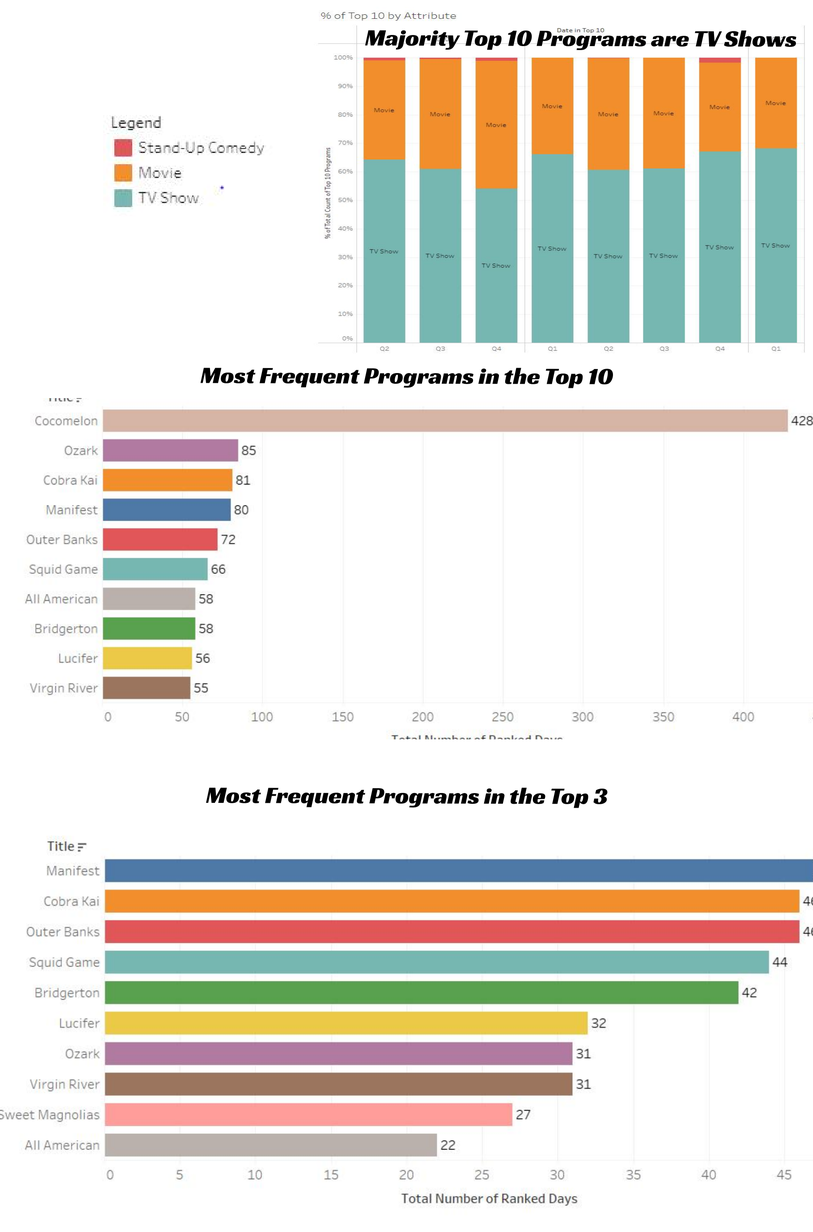
Data

The team acquired the data in the following ways:

1. Netflix Top 10 Dataset: This dataset was publicly available (Kaggle).
2. Additional Netflix Dataset: The team used two additional Netflix datasets to enhance the original top 10 dataset. (Kaggle)

Data Characteristics

1. Size on Disk: The size of the first dataset is a csv file of approximately 500 KB, containing top ten data. The second is a csv file of 3.4 MB containing additional characteristics per program.
2. Number of Records: The exact number of records in the first dataset is 7,100 rows. The second is 8,808 rows.
3. Temporal: The dataset is time-series as they are related to the Netflix top 10 feature, which changes over time.
4. Data Quality Issues: The team encountered several data quality issues, such as:
   - The genre field had all genres listed in a single column, separated by commas.
   - There was a many-to-many relationship between titles and genres and countries (of production).
   - The duration field had different value types depending on whether the content was a TV show or a movie.
   - There were null values that required handling.
5. Data Enrichment: The team tried to use IMDB data to further enhance the Netflix top 10 dataset, but found it to be too "dirty" to incorporate at all.

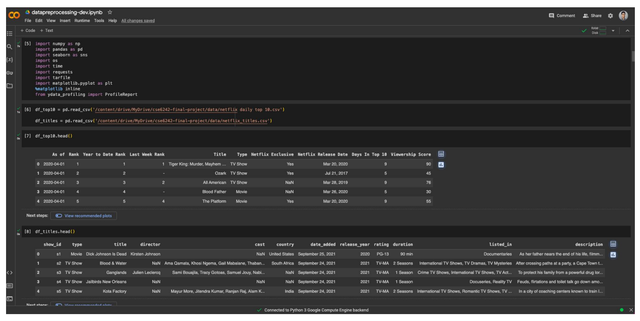


Data Sources

kaggle CSV — netflix-daily-top-10-in-us.csv

datacamp CSV — dataset-python-netflix-movie-data.csv

Data Pre-processing



Pre-processing Methods

- Data Joining
- Data munging
- One-Hot Encoding Countries and Genres, respectively

Experiments and Results

Findings:

The following attributes have the greatest correlation with success of landing in the Netflix Top 10:

a. **TV Show is more likely than a Movie**
   i. If the program is a TV Show, then a Miniseries or a TV Show with a smaller number of seasons is more likely to reach Top 10.
   ii. If the program is a Movie, then having a longer runtime correlates with success

b. **Genres with higher likelihoods are Crime, Reality, Drama, Action, Mystery, Sci-Fi, or Fantasy**

Manual cross-validation was conducted on the 10 programs on the visualization on the left.

- Our tool predicts high chances of all the programs on the left ending up in the Top 10 with the exceptions of Cocomelon (3%) and Squid Game (19%).
  - Cocomelon and Squid Game are simply unicorns within their genres.
  - Our tool's output should not be seen as a guarantee of success or failure, but rather an educational guide.
  - Cocomelon does not appear in the most frequent programs in the Top 3 (bottom left visualization)

Drawbacks/Future Development:

- Our tool allows for unrealistic scenarios such as:
  - Having no genre at all can often lead to 50% chance of success.
  - Having a unrealistically large number of genres will sometimes lead to 90+% chance of success.
  - Creating a thousand-minute movie will lead to 100% chance of success.

Conclusion:

- For realistic scenarios, our tool does an excellent job of predicting a program's chance of entering the Netflix Top 10.
- Our tool should be used in parallel with the normal creative process for a new program and should not be used as a process replacement.