Nikita Belii

Final Project

CAP4773

04/29/2024

Name: Netflix Movies and TV Shows.

Source: https://www.kaggle.com/datasets/shivamb/netflix-shows

Dataset Description:

- The "Netflix Shows" dataset, available on Kaggle, contains detailed information about movies and TV shows available on Netflix as of 2019. This dataset includes over 6,000 records representing a unique show or movie. It features various attributes such as the show's title, director, cast, country of production, date added to Netflix, release year, rating, duration, and genre. The dataset is useful for exploring content trends on Netflix, performing content-based analyses like genre popularity, and predicting attributes such as the maturity rating of a show based on its metadata.
- Number of Instances: 8807
- Number of Attributes: 12
- Attribute Information: show_id object type object title object director object cast object country object date_added object release_year int64 rating object duration object listed_in object description object

Problem Being Addressed:

The primary objective of this analysis was to predict the content rating (Y) of Netflix shows based on various features (X) such as the director, genre, duration, and release year of the shows. Content rating

prediction would help in classifying shows into appropriate age and maturity groups, ensuring viewers are presented with content that is suitable for their age group.

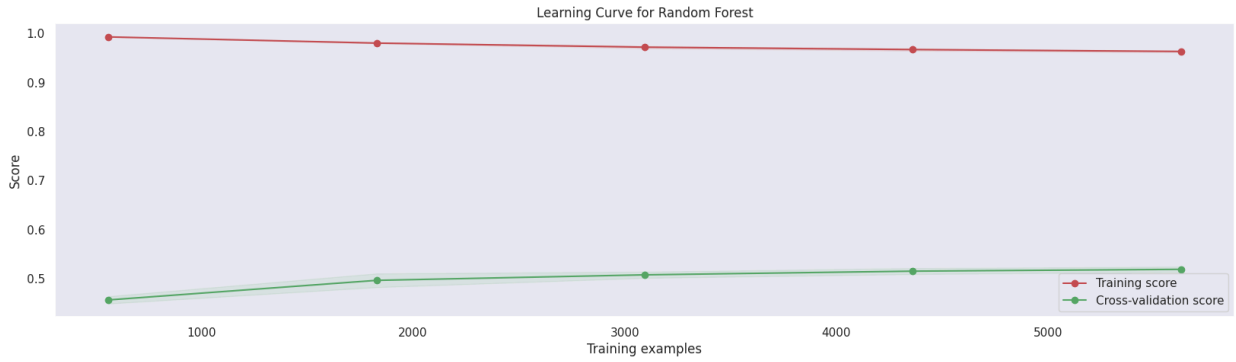
Approach Taken to Address the Problem:

To work on this prediction task, I employed two machine learning models: Logistic Regression and Random Forest. These models were chosen due to their distinct characteristics; Logistic Regression for its simplicity and interpretability, and Random Forest for its ability to handle nonlinear relationships and its robustness against overfitting with complex datasets.

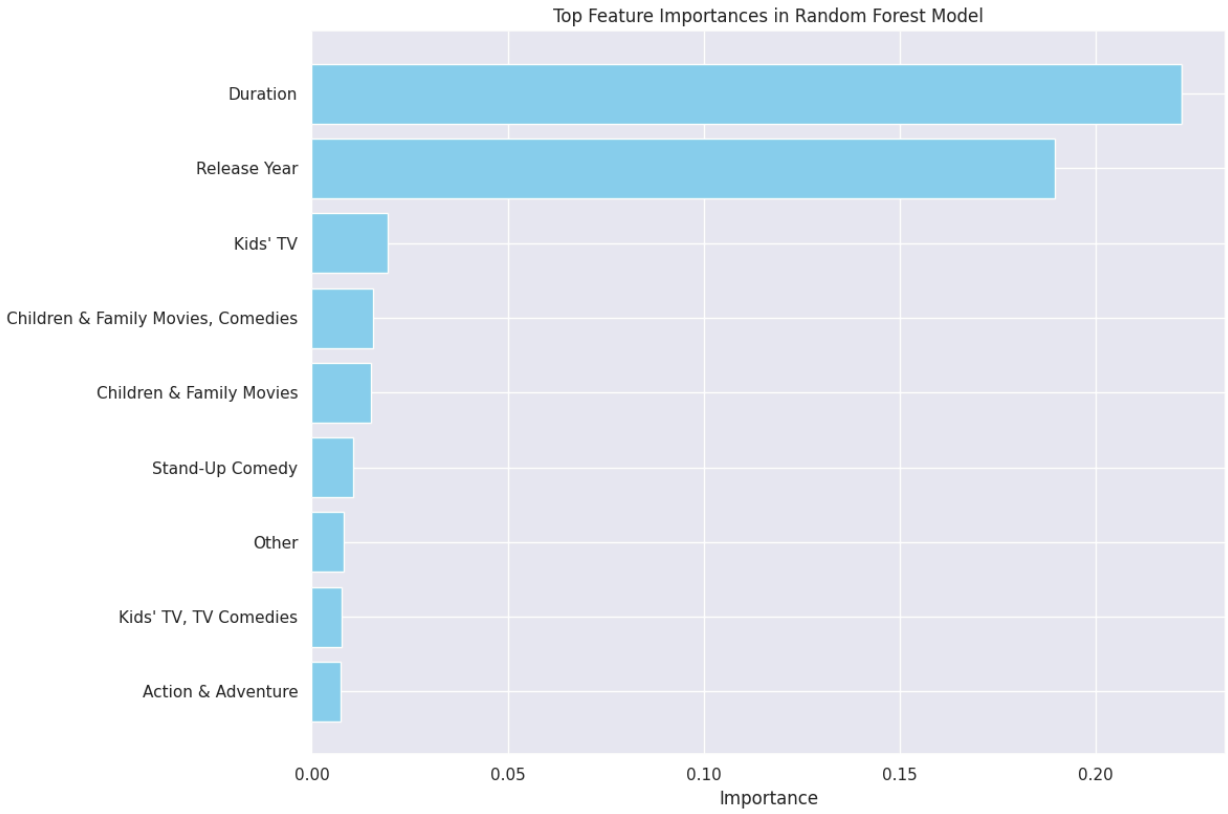
Findings of All Analyses:

The analysis began with data preprocessing, where categorical variables were one-hot encoded, and numerical features such as duration and release year were standardized. The Logistic Regression model achieved an average cross-validation accuracy of approximately 53%, suggesting moderate predictive power. The Random Forest model, on the other hand, showed similar accuracy but displayed signs of overfitting, as indicated by the significant gap between training and test accuracies in the learning curves.

Figure that Supports Investigations



A learning curve graph for the Random Forest model clearly illustrates the overfitting issue, where the training accuracy is almost perfect, but the validation accuracy is significantly lower. This gap suggests that while the model performs excellently on the training data, it fails to generalize effectively on unseen data.



The feature importance graph clearly shows that 'Duration' and 'Release Year' are key factors in predicting content ratings. This graph supports the findings about how time-related factors heavily influence show ratings. Specifically, the length of the shows and their release years stand out as critical elements.

Interpretations of Each Finding:

From the learning curve analysis, I inferred that the Random Forest model, despite its robustness in handling complex datasets, tends to learn excessively from the noise in the training data, leading to poor

generalization. On the other hand, the relatively consistent accuracy across cross-validation folds for Logistic Regression indicates a stable but limited ability to capture the complexities of the dataset.

From analyzing important features, I inferred that the duration and release year of shows significantly impact ratings. This could suggest that newer and perhaps longer shows tend to adhere to evolving standards and audience preferences.

Strengths and Limitations of the Approach:

The strengths of the Logistic Regression model lie in its interpretability and ease of implementation. However, its limitations are evident in its inability to model more complex relationships as effectively as tree-based methods. The Random Forest model, while powerful in capturing complex data relationships, showed a significant drawback due to overfitting, which might be mitigated by tuning hyperparameters or employing techniques like pruning.

The feature importance analysis revealed that 'Duration' and 'Release Year' were highly predictive, suggesting that the length of the show and more recent release years strongly influence the content rating. This might reflect changes in content regulation over time or differences in content length affecting viewership and rating.

Conclusion

The analysis demonstrated that while both models have their merits, they also possess significant limitations that could be addressed with further model tuning and exploring alternative machine-learning approaches. Improving the model's ability to adopt new data without losing its

interpretability will be critical for deploying this model in a real-world production environment, where it can be used to classify new Netflix shows automatically into appropriate content rating categories.