DATA 5610 - Project #0

Predicting YouTube Video Popularity Using Machine Learning by Dallin Moore

The Problem

YouTube is interested in getting even more views on its platform so we are analyzing a dataset of trending videos to see what affects trending videos getting the most views. In this study, we utilize publicly available YouTube trending video data to predict the number of views a video will receive. Our goal is to construct machine learning models and analyze the importance of different features to inform future recommendation system improvements.

The Dataset

We used the publicly available US YouTube trending dataset here, which includes information on video attributes such as title, description, category, tags, likes, and trending dates.

Preprocessing Steps:

- **Text Processing**: Titles, descriptions, and tags were cleaned and vectorized using TF-IDF.
- **Feature Engineering**: Created a new feature, ‘elapsed_time_trending’, representing the time difference between video publishing and trending.
- **Handling Missing Values**: NaN values in numeric columns were replaced with zeros, and text data was filled with empty strings.
- **Encoding Categories**: The categorical feature ‘categoryId’ was mapped to its respective category name for better interpretability.

Modelling

We trained and evaluated four regression models to predict video view counts The following were chosen based on their ability to run within the constraints of the dataset and in a timely manner:

- AdaBoost Regressor
- Random Forest Regressor
- Decision Tree Regressor
- K-Nearest Neighbors Regressor

Model Evaluation

Each model was trained using an 85-15 train-test split with a further validation split for hyperparameter tuning. Performance was measured using Root Mean Squared Error (RMSE) and R-squared (R^2) scores. As seen in the following table, RandomForest Regressor was the best in performing terms of both RMSE and R2 score.

Results Summary

Model	**RMSE**	**R^2 Score**
AdaBoost Regressor	14,952,744	-5.1231
RandomForest Regressor	2,683,623	0.8028
DecisionTree Regressor	3,879,223	0.5879
K-Nearest Neighbors Regressor	3,910,120	0.5813

Feature Importance Analysis

To understand the influence of different variables on video views, we used feature importance scores from the Random Forest model. The most influential features were:

- **likes:** 0.7517
- **elapsed_time_trending:** 0.0462
- **times_trending:** 0.0107
- **description_bts:** 0.0105
- **tags_kpop:** 0.0040
- **description_films:** 0.0039
- **title_mv:** 0.0037
- **description_https:** 0.0036
- **description_ad:** 0.0035
- **description_facebook:** 0.0032

Conclusion

This study demonstrates the feasibility of using machine learning to predict which trending videos on YouTube will get the most views, and in turn generate more revenue. Our findings suggest that recommendation algorithms should prioritize likes, videos that are trending for a long duration of time, videos with ‘bts’ in the description, videos with ‘kpop’ in the tags, and videos with ‘films’ in the description. By focusing on these key features, YouTube can improve its recommendation systems, ensuring that users are presented with content that is more likely to increase platform engagement.