Tik Tok video popularity prediction

Nikolina Trivunić

1. Motivation

The rapid growth of TikTok as a social media platform has made predicting video popularity crucial for content creators and marketers. This project aims to develop a predictive model for estimating the number of views (`playCount`) based on various video and author-related features.

2. Research questions

The primary objective is to address the following questions:

- Can machine learning techniques accurately predict the popularity of TikTok videos?
- Which features, such as likes, shares, comments, and author metadata, significantly influence video popularity?
- How effective are regression models in capturing the variability in `playCount`?

Dataset Description: The dataset consists of 1000 rows and 44 columns, including numerical metrics (`diggCount`, `shareCount`, `commentCount`), textual data, and metadata about video authors (`authorMeta.fans`, `authorMeta.heart`).

3. Related work

This project was undertaken without direct reference to existing academic literature on predicting TikTok video popularity. As such, the methodology and approach were primarily guided by practical considerations and experimentation with available data. While the specific techniques and models used in this project were not drawn from prior studies, the general concept of leveraging machine learning for predicting social media engagement served as a guiding principle.

4. Methodology

Data Preprocessing:

- Cleaned the dataset by removing irrelevant columns (`videoUrlNoWaterMark`, `videoApiUrlNoWaterMark`, `musicMeta.musicAlbum`).
- Imputed missing values using K-Nearest Neighbors imputation and standardized numerical features (`diggCount`, `shareCount`, `commentCount`) using StandardScaler.
- Created additional features (`likes_per_view`, `comments_per_view`, `shares_per_view`) to capture relative engagement metrics.

Feature Selection:

- Identified key features (`authorMeta.fans`, `authorMeta.heart`, `diggCount`,

shareCount, commentCount) based on correlation analysis and domain knowledge.

Model Development:

- Utilized a RandomForestRegressor for its ability to handle non-linear relationships and feature importance analysis.
- Conducted hyperparameter tuning using GridSearchCV to optimize model performance (n_estimators, max_depth, min_samples_split, min_samples_leaf).

5. Discussion

In developing a predictive model for TikTok video popularity, several key methodologies and decisions were explored and evaluated. The approach involved data preprocessing, feature engineering, model selection, and performance evaluation.

- **Data Preprocessing**: The dataset underwent comprehensive preprocessing steps to ensure data quality and model readiness. This included:
    - Removing irrelevant columns (videoUrlNoWaterMark, videoApiUrlNoWaterMark, musicMeta.musicAlbum) to focus on relevant predictors.
    - Imputing missing values using K-Nearest Neighbors imputation to preserve data integrity.
    - Scaling numerical features to a standardized distribution using StandardScaler(). Initially, alternative scaling methods such as MinMaxScaler() were considered but StandardScaler() was chosen for its better performance in conjunction with the regression models.
- **Feature Engineering**: Additional features were engineered to capture relative engagement metrics:
    - likes_per_view, comments_per_view, and shares_per_view were calculated to provide insights into viewer interaction normalized by playCount.
- **Model Development**: The choice of model was critical in achieving accurate predictions of playCount. Initially, a range of models including Linear Regression and RandomForestRegressor were tested. GridSearchCV was preferred over RandomizedSearchCV due to its ability to exhaustively search hyperparameter space and consistently yield superior results in terms of performance metrics such as MSE, RMSE, MAE, and R2 score.
- **Feature Selection**: Two approaches to feature selection were compared:
    - Initially, a broader set of features (authorMeta.fans, authorMeta.heart, diggCount, shareCount, commentCount) was considered.
    - However, through iterative testing and analysis, it was found that focusing on a subset of features (diggCount, shareCount, commentCount, likes_per_view, comments_per_view, shares_per_view) yielded significantly improved model performance. This subset demonstrated higher predictive accuracy with an RMSE of 0.188 and an R2 score of 0.950, compared to RMSE of 0.322 and R2 score of 0.854 when using a broader feature set.
- **Implications and Insights**: The results highlight the importance of selecting relevant features

and appropriate preprocessing techniques in improving model accuracy for predicting TikTok video popularity. Features related to direct viewer interaction (`diggCount`, `shareCount`, `commentCount`) proved to be crucial predictors, reflecting their strong correlation with `playCount`.

In conclusion, by leveraging advanced preprocessing, thoughtful feature selection, and rigorous model optimization techniques, this study demonstrates a robust framework for predicting video engagement on TikTok. These findings provide valuable insights for content creators and marketers seeking to optimize their strategies on social media platforms.

6. References

The dataset used in this project was sourced from Kaggle:

- TikTok Trending Videos Dataset. Available at: https://www.kaggle.com/datasets/marqueurs404/tiktok-trending-videos