Predictive Cinematics

Harnessing Data to Predict Cinematic Success

Project Overview

Problem Statement

Streaming services and movie producers face the challenge of making data-driven decisions about content acquisition and production.

Proposed Solution

Develop a machine learning model to predict movie success, using movie metadata and user reviews.

Dataset Overview

Combined 3 separate datasets with extensive cleaning to form a comprehensive dataset of text reviews and movie metadata

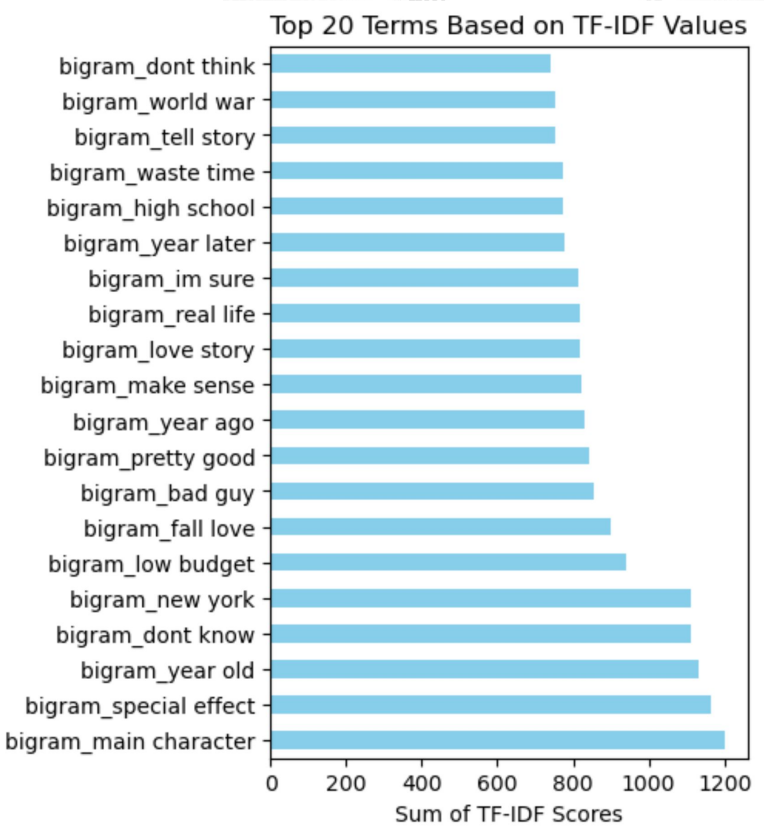
Data Pre-Processing

- ‘Entity score' to encode actors, directors, and production companies
- One-hot encoded movie genres
- Cleaned and tokenized review text transformed with TF-IDF

Exploratory Data Analysis & Modeling

Key Insights

- Lack of statistically significant frequent collaborators across movies
- Evenly distributed target variable
- Challenges encoding high cardinality features
- Two rounds of feature engineering
- “Low budget” text feature ranked highly in feature importance



Model Interpretation

- **Ensemble Model Performance**
  Top accuracy of 0.66
- **Text Features Dominance**
  Outperformed metadata features significantly
- **Strengths & Weaknesses**
  Excels at 'Poor' & 'Excellent', struggles with 'Average' & 'Good' classifications
- **Next Steps**
  Further feature engineering; exploration of neural networks