Predicting IMDb Ratings from Movie and TV Metadata Using Machine Learning
Project Description Report

Kaan Baykal 221101028
Beril Aydın 231101002
Anıl Özişler 211101081

1. Project Description

This project aims to build a machine learning model that predicts the IMDb rating of movies and TV shows based on their metadata features such as genre, runtime, certificate, description, cast and directors. By analyzing these diverse attributes, the model seeks to uncover the underlying relationships between content characteristics and audience evaluations. The ultimate goal is to provide an analytical framework that can estimate the expected IMDb score of newly released content and quantify how certain features such as the combination of genres, the reputation of the director, or the complexity of the description contribute to the overall reception of a production.

2. Short Literature Review

1. **R. K. Bansal, S. Kumar, and M. Kumar, "Movie Rating Prediction Using Machine Learning Techniques," IEEE International Conference on Computational Intelligence and Data Science (ICCIDS), 2019.**
   Abstract: This study presents a comparative analysis of various regression and classification algorithms for predicting IMDb movie ratings. The authors evaluate algorithms such as Linear Regression, Decision Tree, Random Forest, and Support Vector Machines, concluding that ensemble methods yield superior accuracy. The paper emphasizes the importance of preprocessing and feature selection when handling movie metadata, which directly aligns with the objectives of this project.

   URL: https://ieeexplore.ieee.org/abstract/document/8944604

2. **F. Geng, C. Gao, and Y. Zhu, "A Data Mining Approach to the Factors Affecting Movie Box Office and Popularity," in *Data Science and Knowledge Engineering for Sensing Decision Support*, Springer, 2014.**
   Abstract: This research explores how movie-related variables such as genre, cast, and director influence popularity and performance metrics including box office and user ratings. Using data mining and regression analysis, the authors demonstrate that structured metadata can successfully explain audience preferences. Their findings reinforce the potential of predictive modeling techniques in understanding the correlation between film characteristics and audience evaluation.

   URL: https://link.springer.com/chapter/10.1007/978-3-319-11740-9_41

3. **H. S. Patel and P. J. Patel, "Movie Rating Prediction Using Machine Learning," *International Journal of Innovative Science and Contemporary Technologies (IJISCT)*, vol. 2, no. 3, pp. 9–14, 2016.**
   Abstract: This paper applies supervised learning techniques to predict IMDb ratings based on movie metadata. The authors focus on regression models and highlight the significance of preprocessing steps such as text vectorization and normalization. Their

results indicate that machine learning models can effectively capture non-linear relationships between movie attributes and audience scores, supporting the feasibility of metadata-based rating prediction systems.

URL: https://journals.cfrit.com/index.php/ijisct/article/view/62/43

3. Dataset Source

- **Source:** Kaggle – TV and Movie Metadata with Genres and Ratings (IMDB)

- **Collection Method:** The dataset was compiled from IMDb's public database and verified using multiple entertainment data sources. It is freely available as a structured CSV file and does not require scraping.

- **Possible Backup Source:** IMDB Movies Dataset

4. Dataset Description

- **Domain:** Entertainment Analytics / Media Studies
- **File Type:** CSV (~40MB)
- **Number of Records:** 100,000+ entries
- **Columns:**
  - `movie`: Movie name
  - `genre`: Main and secondary genres
  - `runtime`: Duration in minutes
  - `certificate`: Age rating (e.g., PG, R, TV-MA)
  - **Target Variable:** `rating` (numerical value between 0–10)
  - `star`: Actors of the movie
  - `description`: Short brief about the movie
  - `votes`: Number of IMDb user votes
  - `director`: Director of the movie

5. Tools, Frameworks, and Technologies

- **Programming Language:** Python
- **Main Libraries:**
  - *Data Analysis:* pandas, numpy

    - *Visualization:* seaborn, matplotlib
    - *Preprocessing & Encoding:* scikit-learn, category_encoders
    - *Modeling:* scikit-learn (Random Forest, XGBoost, Ridge Regression)
    - *Evaluation Metrics:* RMSE, MAE, R^2

- **Platform:** Visual Studio Code, Google Colab
- **Version Control:** GitHub repository
- **Hardware/Cloud:** Google Colab GPU runtime for faster training and hyperparameter tuning

6. Planned Workflow and Task Division

Team Member	Tasks
Beril Aydın Kaan Baykal Anıl Özişler	Data cleaning, encoding, and preparing the dataset for modeling.
Kaan Baykal	Random Forest, Ridge Regression training
Beril Aydın	XGBoost, Random Forest training
Anıl Özişler	XGBoost, Ridge Regression training
Beril Aydın Kaan Baykal Anıl Özişler	Model validation and hyperparameter evaluation, visualization, and reporting. Evaluating results using previously mentioned evaluation metrics and integrating results into the final report.