CinePredict: Leveraging Data Mining Techniques for Accurate IMDB Score Forecasting

Jaswanth Kranthi Boppana jboppana@iu.edu

Vineela Kunisetti vikuni@iu.edu

Vishnuvardhan Reddy Kollu viskollu@iu.edu

Abstract

The proposal aims to develop a predictive model for IMDB ratings and revenue generation for movies by incorporating novel factors such as actor influence, association rule mining and Implementing adaptive ML algorithms to get more accurate predictions. our approach integrates actor popularity metrics and story analysis, offering a more comprehensive and accurate prediction. This method can aid stakeholders in the film industry to make more informed investment and production decisions.

Keywords: Association Rules, Machine Learning, Data Pre-processing

Introduction

In recent years, the demand for accurate movie performance prediction has intensified, driven by the economic implications of box office success and audience approval. Traditional prediction models predominantly use straightforward parameters such as genre, and budget often falling short of the nuanced understanding necessary for accurate IMDB rating and revenue predictions. To address these limitations, we propose a data-driven approach that leverages novel parameters, such as actor popularity, storyline sentiment, and association rule mining, combined with Machine learning approach to provide more robust and accurate predictions.

Informed by recent studies, our project aims to improve prediction accuracy for IMDB ratings and revenue by incorporating actor influence metrics and association rule mining, offering a more nuanced understanding of movie success factors. Notably, **Eric (Kaggle)** explored data mining techniques to predict IMDB ratings but did not employ association rule mining, which could enhance feature interdependencies for improved prediction **[1]**. Building on this, we will employ association rules to capture hidden relationships between variables like genre, storyline elements, and cast pairings, inspired by insights from their work. Additionally, **Maitra et al. (IEEE)** developed a machine learning model for IMDB rating prediction, yet omitted actor-based metrics, an influential component our model aims to integrate **[2].**

In the realm of revenue prediction, **Balaji et al. (arXiv)** utilized machine learning models to predict movie revenue but focused on generic metrics, rather than personalized or dynamic features like actor popularity **[3].** We will extend this approach by incorporating metrics like social media following and actor awards, which can reflect public interest and engagement levels. Moreover, **Zhou and Ramesh (Springer)** demonstrated how storyline analysis impacts IMDB ratings, which we further build upon by using Natural Language Processing (NLP) to quantify storyline sentiment **[4].**

Our model combines these insights, using a robust set of features and association rules to predict IMDB ratings and revenue with greater precision than previous models. We are aiming to follow a proper data pipeline and test various decision centric models like Decision trees, K-NN, Random Forests and few more. We intend to also develop and ensemble method that which could be expected to turn out as a better approach. In the section below we propose our methodology that defines our technical workflow.

Methods

As we are looking to solve this problem with data driven approaches. It is important that we follow a conventional data pipeline to start of with and build up on it to get error free predictions. The primary stage starts with Data pre-processing of the collected data.

I. **Data Collection and Preprocessing:**

We are using the data collected from multiple sources, including IMDB for movie details, social media platforms for actor popularity metrics, and public databases for historical revenue information. The dataset contains over 5000 iterations with 20 various features. Data preprocessing steps will include handling missing values, encoding categorical data (e.g., genre, director), and normalizing numeric variables to ensure compatibility across machine learning algorithms.

II. **Feature Engineering:**

Our model will include:

1. **Traditional Features** such as budget, director, genre, runtime, and release date.
2. **Actor Influence Metrics:** Metrics like social media following, frequency of collaborations, and awards, which capture the actor's popularity and engagement.
3. **Association Rule Mining:** We will apply association rule mining algorithms to uncover relationships among multiple movie characteristics (e.g., genre and cast pairings) to identify high-impact feature combinations that correlate with higher ratings and revenue.
4. **Storyline Sentiment Analysis:** Using Natural Language Processing (NLP) techniques, we will analyse storyline descriptions to quantify sentiment, extracting features that represent tone, intensity, and thematic elements.

During the feature engineering we would also investigate various correlation and feature relation metrics to adjust the dataset. We would also intensely build multiple data visualization to understand the patterns and get clear understanding what the data is hiding. This helps to develop more robust models to enhance prediction accuracy.

5. **Machine Learning Algorithms:**

   To predict IMDB ratings and revenue, we will implement a combination of algorithms tailored to capture both linear and non-linear relationships:

   a. **Linear Regression and Decision Trees:** For initial modelling and establishing a baseline, these algorithms provide interpretability and efficient training on structured data. We intending to very deep into the data by creating classification tress and improving it with modulating its variations.
   b. **Random Forest and Gradient Boosting:** These ensemble methods, particularly Random Forest and Gradient Boosted Trees, are robust to overfitting and capture complex interactions between features. They will serve as our primary models for revenue prediction and ratings.
   c. **Association Rule Mining (Apriori Algorithm):** Its key to implement the Apriori algorithm to generate association rules, providing insights into significant relationships among variables like actors, genres, and storyline content, which are then fed into the model as high-impact predictors.
   d. **Support Vector Machines (SVM) and K-Nearest Neighbours (KNN):** These algorithms will be used in combination with feature selection techniques to refine prediction models based on storyline sentiment and actor metrics, complementing association rule insights with distance-based learning.
   e. **Neural Networks for Advanced Analysis:** Depending on dataset size, we may explore neural network architectures for deeper learning on textual features and actor influence, where complex patterns can be further refined through embeddings and feature interactions.

6. **Model Evaluation:**

   We will evaluate the model performance using metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for revenue prediction and accuracy and F1-score for IMDB rating prediction. Cross-validation will be employed to prevent overfitting and ensure model generalizability. We will explore various other evaluation metrics during the development.

Conclusion:

Till now, IMDB ratings and movie revenue predictions relied on basic parameters such as genre and budget, often lacking the complexity needed for accurate forecasting. We are aiming to address this gap by integrating novel metrics, including actor influence, storyline sentiment, and association rule mining, with adaptive machine learning models. This enhanced methodology improves prediction accuracy by capturing nuanced factors that contribute to a movie's success. Our model provides valuable insights for industry stakeholders, supporting data-driven decision-making in investments and production planning, and sets a new benchmark for predictive analytics in the film sector.

References:

1. https://www.kaggle.com/code/erikbruin/movie-recommendation-systems-for-tmdb/report

2. https://ieeexplore.ieee.org/document/8944604/references#references

3. https://arxiv.org/abs/2405.11651

4. Jose, A., Harikumar, S. (2022). Predicting IMDB Movie Ratings Using RoBERTa Embeddings and Neural Networks. In: Mathew, J., Santhosh Kumar, G., P., D., Jose, J.M. (eds) Responsible Data Science. Lecture Notes in Electrical Engineering, vol 940. Springer, Singapore. https://doi.org/10.1007/978-981-19-4453-6_13