IE7275 Fall 2021 Term Project Proposal (or Report)

Raghavi Rajurkar
Rohan Manish Gupta

1. Prediction of Box Office Revenue using scrapped TMDB Data

Name	Roles	Responsibilities
Raghavi Rajurkar, Rohan Gupta	Project Topic Discussion	Research project ideas and finalize one project topic
Raghavi Rajurkar, Rohan Gupta	Project Proposal	Discuss key information and draft the proposal as per the given instructions
Raghavi Rajurkar	Cleaning and Normalization of Data	Scrapped data contains bad data, like nested dicts etc., Prepare the data for Mining.
Raghavi Rajurkar	Dimensionality reduction	For different goals, only important features are needed. Exploratory analysis will be conducted.
Rohan Gupta	Data Mining Techniques	Attempts to predict the revenue, using a variety of Mining methods, will be made. Final accuracy must be high enough to be used for investment analysis.
Rohan Gupta	Results	Results will be a view of the accuracy, and real-world use of mined data for decision making for investment purposes.
Raghavi Rajurkar, Rohan Gupta	Visualization, Presentation	The results will be presented along with an interactive dashboard.

2. Introduction

The film industry earned $42.2 billion in box office revenue in 2019. For large studios like Disney and Sony, each film represents a short-term investment worth hundreds of millions of dollars with a large amount of risk. Thus, a predictive model that can assist in the selection of these investments with accuracy would be a tool with applications in higher level decision making. Previous models and research have shown that the box office revenue for a film can be predicted using factors such as cast, crew, plot keywords, budget, posters etc. However, films are also a creative medium, and thus their success and box office revenue also depend on subjective opinion, and thus prediction of this revenue presents a unique challenge. Approaches like SVM regression, and Linear Discriminant Analysis have been commonly used to make this prediction and will be considered for this project.

3. Problem Definition

Due to the nature of films as high risk, high reward investments, this project will focus on accuracy of prediction of box office revenue, which represents the return on the investment. To serve this goal, the importance of various features like crew, posters, budget, release dates etc., will be considered. These features all represent potential fields where studios can create direct impact with capital to hopefully generate a large box office return. Then predictive models will be built using this data to attempt to predict box office revenue with a high degree of accuracy.

4. Data Description

The dataset is about a film database to predict the box office collection. This dataset has been collected from TMDB. The film details, credits and keywords have been collected from the TMDB Open API. There are 22 variables and 7398 rows, and a variety of metadata obtained from the database. Films are labeled with an id. Data points include cast, crew, plot keywords, budget, posters, release dates, languages, production companies, and countries.

5. Proposed Methods

First, data will be cleaned and transformed into standard format for mining. Following this, summaries, visualizations and correlation analysis will be used to identify important features of a film with regards to its revenue. PCA analysis may be applied to the important features to utilize in predictive models. Mining techniques such as Linear Regression Analysis, SVM regression, LDA, along with classification techniques like Bayesian classifier and Linear Discriminant Analysis will be used to identify box office bracket. Performance of the models will be examined using various accuracy measures, lift charts, ROC curves and any other appropriate methods to ensure high accuracy.

6. Expecting Results (for proposal) or Experimental Results (for final report)

The model will produce a value of Box Office revenue or classification result to quickly compare films based on a variety of features. These results will be instrumental in higher level decision making to analyze films as high-risk high reward investments. Secondary results will attempt to create ideal values of features to maximize revenue. This will be done by using the final selected model to predict revenue and constantly iterating over values to produce ideal values.

7. Conclusion (for final report)

Summarize what you have in this project.

References

1. Sangjae Lee, Bikash KC, Joon Yeon Choeh, 2020, "Comparing performance of ensemble methods in predicting movie box office revenue", Helion Cell Press,
2. Rhee, Travis & Zulkernine, Farhana. (2016). Predicting Movie Box Office Profitability: A Neural Network Approach. 665-670.