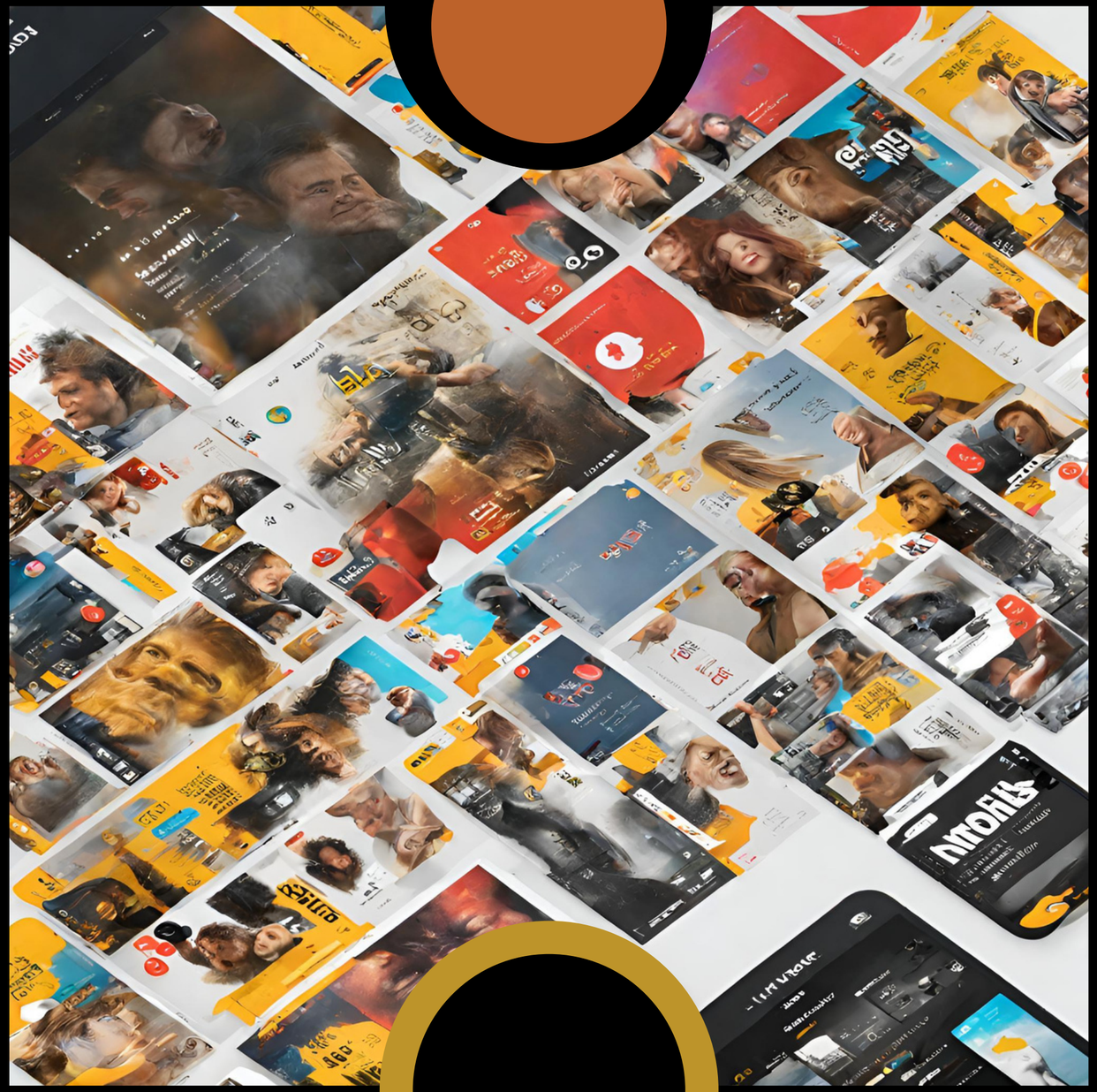

PREDICTING IMDB SCORES

By using Applied Data science

Overview

Introduction

This paper presents a comprehensive overview of predicting IMDb scores using data science techniques. IMDb, as an influential platform for movie ratings, provides valuable insights into the audience's perception of a movie's quality and popularity. By accurately predicting IMDb scores, filmmakers, producers, and movie enthusiasts can make informed decisions regarding marketing strategies and content improvements. This paper delves into the problem definition, design thinking, data processing, data sources, model selection, training sets, and provides real-time examples.

Abstract

Film Industry is not only a industry or a centre of entertainment, rather it is now a centre of global business. All over the world is now excited about a movie's box office success, popularity etc. A huge data is available online about these movies success or popularity. We have used hollywood movie list from Wikipedia and their rating from IMDb movie rating website to create our data set. Then machine learning classification algorithms are applied of the data set. Lastly an efficient model is developed to predict a movie's IMDb rating. The model gives good classification measures with the data set.

Problem Definition

The objective of this project is to develop a predictive model that accurately estimates IMDb scores for movies. IMDb scores range from 1 to 10 and serve as a measure of a movie's overall quality and audience reception. By accurately predicting these scores, stakeholders can gain valuable insights into the potential success of a movie and make data-driven decisions to enhance its appeal.

Design thinking

To address this problem, a design thinking approach will be employed. This involves empathizing with the stakeholders, identifying their needs, ideating various solutions, prototyping models, and iterating based on feedback. By understanding the end-users' requirements, the model can be tailored to meet their expectations effectively.

Methodology

The working method for this work involves few steps. The methodology is shown in figure 1. The steps are described below.

- Data Extraction
- Data Processing
- Appling Machine Learning techniques

Data Extraction

Data is extracted from Wikipedia and IMDb movie rating website. We have merged data from two platforms for our data set. Note that, data about only Hollywood movies released on the year of 2018 is extracted from Wikipedia. About 250 Hollywood movies are released on the year of 20184 . The extracted data from wikipedia contains title of the movie, studio, cast and crew, genre, country, month and date of release, year.

Data Processing

Data preprocessing means to prepare the data for classification. Data is processed according to the requirements of classification. Here, for preprocessing the data, instances with missing attributes are removed. Finally we got data set of 242 movies. The features of the processed data set are Title, Studio, Director, Screenplay, Actor, Actress, Genre, Country, Year, Rating. Here, Rating is the class attribute. The details are described in table 1. The data set now produced is an imbalanced data set, as there are only 3 movies of flop class and 138 average movies.

Machine Learning Technique

We have used Weka 3.8.3 tool [16] and applied five machine learning algorithms [shown in table 3] to build the model. Among the classifiers Bagging and Random forest is ensemble method. Random forest starts classifying with multi decision tree. J48 also classifies using decision tree, it is often referred as statistical classifier [17]. IBK is K-Nearest Neighbour algorithm and it is a non-parametric method. Lastly a probabilistic classifier Naive Bayes, which is based on Bayes' Theorem.

Model Selection

Various machine learning algorithms, including linear regression, random forest, and gradient boosting, will be explored to predict IMDb scores. These algorithms can capture the relationships between movie features and IMDb scores, enabling accurate predictions. The performance of different models will be evaluated using appropriate metrics such as mean absolute error (MAE) or root mean squared error (RMSE). The best-performing model will be selected for the prediction task.

Data Source

The primary data source for this project is the IMDb dataset, which provides a comprehensive collection of movies along with their associated ratings. Other publicly available datasets or APIs that offer related information, such as movie reviews, social media sentiment, and box office performance, will also be explored. These additional data sources can provide valuable insights to augment the predictive power of the model.

Training Sets

To train the model, the dataset will be split into training and testing sets. The training set will be utilized to fit the model, while the testing set will evaluate its performance. Techniques such as cross-validation will be employed to prevent overfitting and ensure the model generalizes well to unseen data. Refinement of the model will occur iteratively, based on its performance on both training and testing sets, to optimize its predictive capabilities.

Conclusion

Predicting IMDb scores using data science techniques offers valuable insights into movie quality and popularity. By leveraging available data, employing appropriate model selection techniques, and refining models through training and testing, accurate predictions of IMDb scores can be achieved. This paper has provided a brief overview of the problem definition, design thinking process, data processing, data sources, model selection, training sets, and real-time examples. The potential of data science in predicting IMDb scores has been showcased, offering stakeholders a valuable tool for decision-making in the film industry.