



Movie Rating Prediction



Introduction

Introduction

In this project, we aim to develop a machine learning model to predict movie ratings based on various characteristics of the movie. Accurately predicting movie ratings can have a variety of applications, such as:

- Recommending movies to users

- Helping movie studios and distributors decide which movies to invest in and promote

- Providing insights into what factors influence a movie's rating

To achieve this goal, we will use the IMDB 5000 Movie Dataset, which contains information on over 5000 movies. The model we develop will be trained on this data and will be evaluated based on its performance on a test set.

In the following slides, we will describe the data set we are using, the methodology we followed to develop the model, the results we obtained, and our conclusion."




Data



Data Set Overview

IMDB 5000 Movie Dataset


The IMDB 5000 Movie Dataset is a collection of movie data compiled by Carol Zhang. It includes information on over 5000 movies, including the movie's title, cast, crew, budget, and box office earnings, as well as various metadata such as the movie's genre, language, and release date. The dataset is available on Kaggle.





Data Set Statistics


The IMDB 5000 Movie Dataset contains information on 5,624 movies. The dataset includes the following features:

- Movie title
 - Cast
 - IMDB score
 - Budget
 - Box office earnings
 - Genre
 - Language
 - Release date
 - Various metadata (e.g., Crew, number of votes, etc.)
 - Movie poster
- 



Data Preprocessing

Before using the IMDB 5000 Movie Dataset to train a machine learning model, we need to clean and preprocess the data. This includes tasks such as:

- Removing any missing or corrupted data
 - Handling any inconsistencies or errors in the data
 - Normalizing or scaling the data
 - Encoding categorical variables
 - Splitting the data into training and test sets"
- 



Methodology

Feature Selection

Before training our machine learning model, we need to select the features that will be used to predict IMDB score. We selected the following features based on their potential relevance to movie ratings:

- Genre
- Cast
- Crew
- Budget
- Box office earnings
- Language
- Release date
- Various metadata (e.g., number of votes, etc.)

We encoded categorical variables using one-hot encoding and normalized the data using min-max scaling."



Model Training

We trained three different machine learning models to predict movie ratings:

- Decision Tree Regressor
- Random Forest Regressor
- Gradient Boosting Regressor

We also used Lasso regularization to improve the generalization performance of the models.

We used a 5-fold cross-validation approach to tune the hyperparameters of each model, and selected the model with the best performance on the validation set.

Finally, we evaluated the selected model on a test set to measure its performance.



Evaluation Metrics

To evaluate the performance of our model, we used the following evaluation metrics:

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)

These metrics allow us to quantify the difference between the predicted movie ratings and the true ratings, and give us an idea of how accurate our model is.



Results



Model Performance

The table below shows the performance of our model on the test set, using the mean squared error (MSE), and root mean squared error (RMSE) as evaluation metrics:

Model	MSE	RMSE
Decision Tree	0.74	1.002
Random Forest	0.53	0.76
Gradient Boost	0.52	0.77

The Gradient Boosting model had the lowest MSE, and RMSE, indicating that it performed the best among the three models we trained.



Model Limitations

While our model was able to achieve good performance on the test set, there are a few limitations to consider:

- The model is only as good as the data it was trained on. If the IMDB 5000 Movie Dataset is not representative of the overall movie market, the model's predictions may not generalize well.
-
- The model only uses a limited set of features to make predictions. There may be other factors that influence movie ratings that are not captured in the data set.
-
- The model is not able to predict ratings for movies that have not yet been released, as it only uses data on movies that have already been released.

Despite these limitations, our model provides a good starting point for predicting movie ratings and could be useful in a variety of applications.



Conclusion

Conclusion

Key Findings:

In this project, we developed a machine learning model to predict movie ratings based on various characteristics of the movie. Our key findings are as follows:

The Gradient Boosting model performed the best among the three models we trained, with the lowest mean absolute error, mean squared error, and root mean squared error on the test set.

The model's performance was influenced by the choice of features. Genre, cast, crew, budget, and box office earnings were among the most important features in predicting movie ratings.

The model's predictions were generally accurate, but there were some cases where the model's errors were larger. This could be due to a variety of factors, such as the limited set of features used by the model or the limitations of the data set



Future Work:

There are a few directions we could take to improve the performance of our model in the future:

- Use a larger and more diverse data set to train the model. This could help the model generalize better to the overall movie market.

- Explore other machine learning models and feature selection methods to find a more accurate and robust solution.

- Incorporate additional features that may be relevant to movie ratings, such as the movie's plot summary or reviews from critics and users.

- Develop a real-time prediction system that can predict ratings for movies that have not yet been released, using data such as trailers and marketing materials.

We believe that these steps could help to further improve the performance of our movie rating prediction model.