# Movie Ratings Prediction Model

## Project Overview

The Movie Ratings Prediction Project is all about predicting movie ratings based on a variety of characteristics like genre, budget, popularity, and release year. By exploring and processing movie data, we aim to understand the factors that influence a movie’s rating and ultimately create a model that can predict how well a movie might be received by audiences.

## Dataset Details

### Problem Statement

The goal of this project is to develop a model that can estimate movie ratings accurately using various movie attributes. We hope the insights from this project can be valuable to movie producers, distributors, and marketers to understand the elements that contribute to a movie’s success.

### Dataset Source

The data for this project comes from Kaggle, specifically from The Movie Database (TMDb), which has extensive information on over 1,000,000 (1M) movies. This data covers attributes ranging from genre and language to budget and popularity.

#### Kaggle Dataset Name:

Full TMDB Movies Dataset 2024 (1M Movies)

#### Kaggle Dataset Link:

(https://www.kaggle.com/datasets/asaniczka/tmdb-movies-dataset-2023-930k-movies)

### Quick Overview of the Dataset

This dataset is great for identifying trends and patterns in the movie industry. It has many details about each movie, so it’s packed with information we can use to find out what factors might be key to predicting ratings.

### Attributes Breakdown

Here’s a quick look at what’s in the dataset and why it matters:

|  |  |
| --- | --- |
| Attribute | Description |
| id | Unique identifier for each movie. |
| title | Movie title, the easiest identifier for us! |
| vote average | Average viewer rating, our target for prediction. |
| vote count | Number of ratings, adding weight to the average. |
| status | Release status (e.g., Released, Rumored). |
| release date | Movie’s release date, important for time-based patterns. |
| revenue | Total revenue, showing financial success. |
| runtime | Movie length, a factor in viewer preferences. |
| adult | Marks if the movie is for adult audiences only. |
| genres | List of genres, key for predicting appeal. |
| popularity | Popularity scores often reflecting interest levels. |
| production companies | Companies involved may indicate budget and style. |
| spoken languages | Languages spoken in the movie, linked to the target audience. |
| original language | Original production language, useful in cultural context. |

## Project Workflow

This project has a few stages, each building toward an effective prediction model:

### 1. Data Cleaning

* Handled any missing data and inconsistencies.
* Standardized dates, formatted text, and organized all data for easier analysis.

### 2. Exploratory Data Analysis (EDA)

* Created visualizations to look at genre distributions, language diversity, and popularity.
* Analyzed revenue over time and identified how various factors influence ratings.

### 3. Preprocessing for Modeling

* **Scaling:** Standardized numerical data to keep all attributes on a comparable scale.
* **Encoding:** Converted categorical data (like genres and languages) into formats our model can understand.

## Project Status and What’s Next

Right now, we’re at the preprocessing stage. Here’s what’s coming up:

* **Feature Engineering:** We’ll craft new features to better capture the nuances of movie data.
* **Model Selection:** We’ll test different models (like regression, decision trees, and ensemble methods) to see what fits best.
* **Model Evaluation and Optimization:** Fine-tuning and optimizing our model to boost accuracy.

**> Note:** This project is a work in progress, and new features will be added and refining of the model will be done as we go along.