Movie Rating Prediction with Python

A Data Science Approach to Predicting Movie Ratings

Created by Ajay Khanna

Data Science Intern at Afame Technologies

TABLE OF CONTENTS

Project Overview

Project Description

The Movie Rating Prediction project involves analyzing a dataset containing information about Indian movies. The dataset includes details like movie name, year, duration, genre, rating, votes, director, and three main actors. The data will be used to build a predictive model for movie ratings and extract valuable insights from the movie industry.

Business Understanding

The film industry relies on understanding the factors that influence movie success. Accurately predicting movie ratings can aid in decision-making, such as choosing the right actors, directors, and genres, as well as determining marketing strategies

OBJECTIVES

- Develop a predictive model: Create a machine learning model to predict movie ratings based on the provided dataset. This is essentially a regression problem, where we aim to estimate the numerical movie ratings based on various features.

- Identify influential factors: Analyze the dataset to determine which factors (e.g., genre, director, actors) have the most significant impact on movie ratings.

- Provide actionable insights: Offer insights to the film industry stakeholders to make informed decisions about movie production, casting, and marketing.

Data Understanding

- Dataset Source:

Kaggle (IMDb India Movies)

- Key Features:

- Name, Year, Duration, Genre
- Rating (Target), Votes, Director
- Actor 1, Actor 2, Actor 3

- Target Variable: Rating

TOOLS AND TECHNOLOGIES USED

- **Programming Language**

- Python: The core language for the analysis and modeling.

- **Libraries for Data Analysis and Preprocessing**

- Pandas: For data manipulation and cleaning.

- NumPy: For numerical computations.

- Matplotlib/Seaborn: For data visualization.

- **Machine Learning and Model Building**

- Scikit-learn: For preprocessing, model training, and evaluation.

- XGBoost or LightGBM: For gradient boosting techniques.

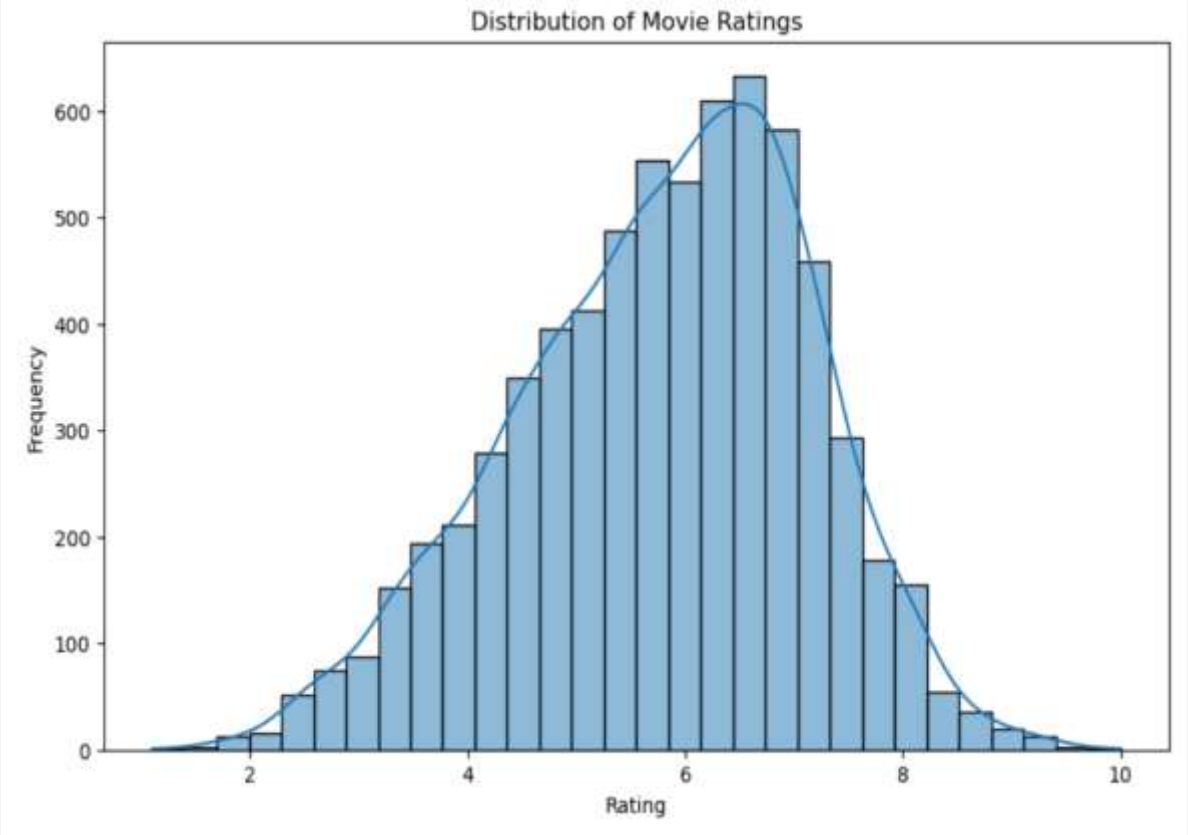
- **Others**

- Jupyter Notebook: Environment for writing code and documenting the project.

- SciPy/Statsmodels: For statistical analysis.

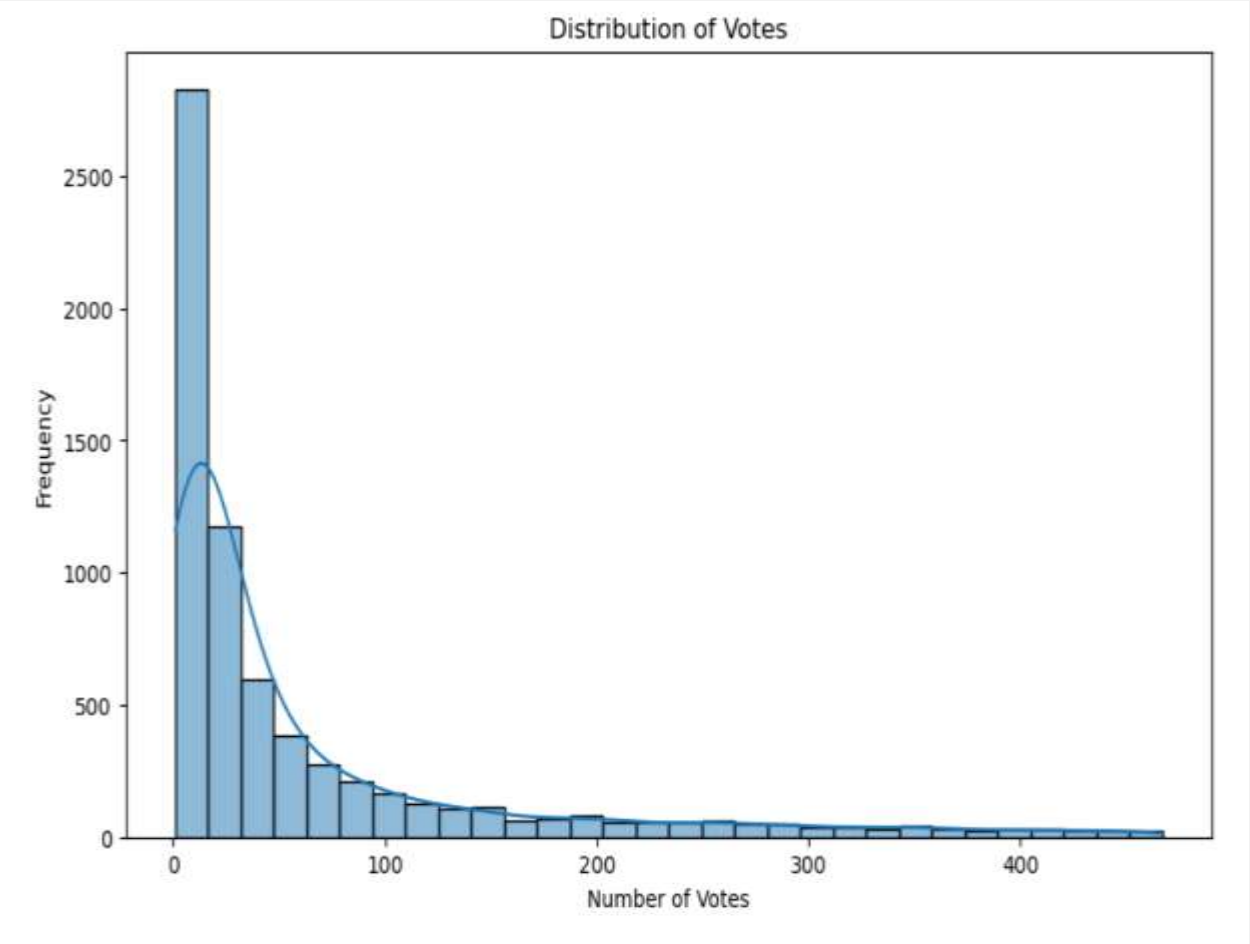
EXPLORATORY DATA ANALYSIS (EDA)

RATING ANALYSIS



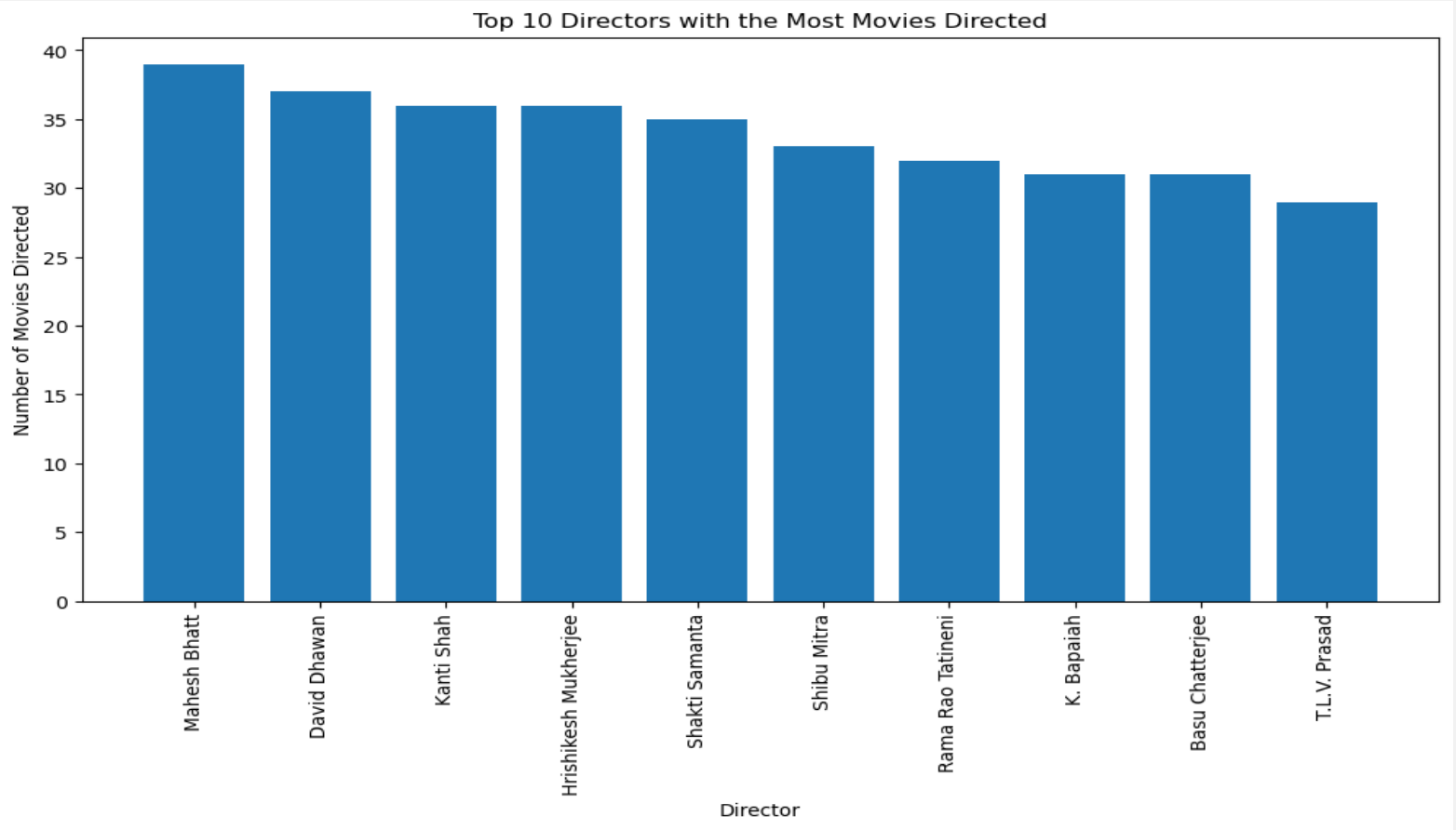
- Rating Distribution: The histogram shows a slightly normal distribution with a peak around a 7 rating value.
- Common Rating Range falls between 5 to 7.
- All ratings are positive
- There are few movies rated below 2 and above 8

VOTES ANALYSIS



- The distribution of votes is right-skewed, with the majority of movies receiving a relatively low number of votes. This suggests that many movies in the dataset may not be widely recognized or popular, as they have received fewer votes
- The long tail towards higher vote counts indicates that there are a smaller number of movies that have garnered a significant number of votes
- Popularity Range is between 0 and 50

TOP 10 DIRECTORS WITH THE MOST MOVIES DIRECTED



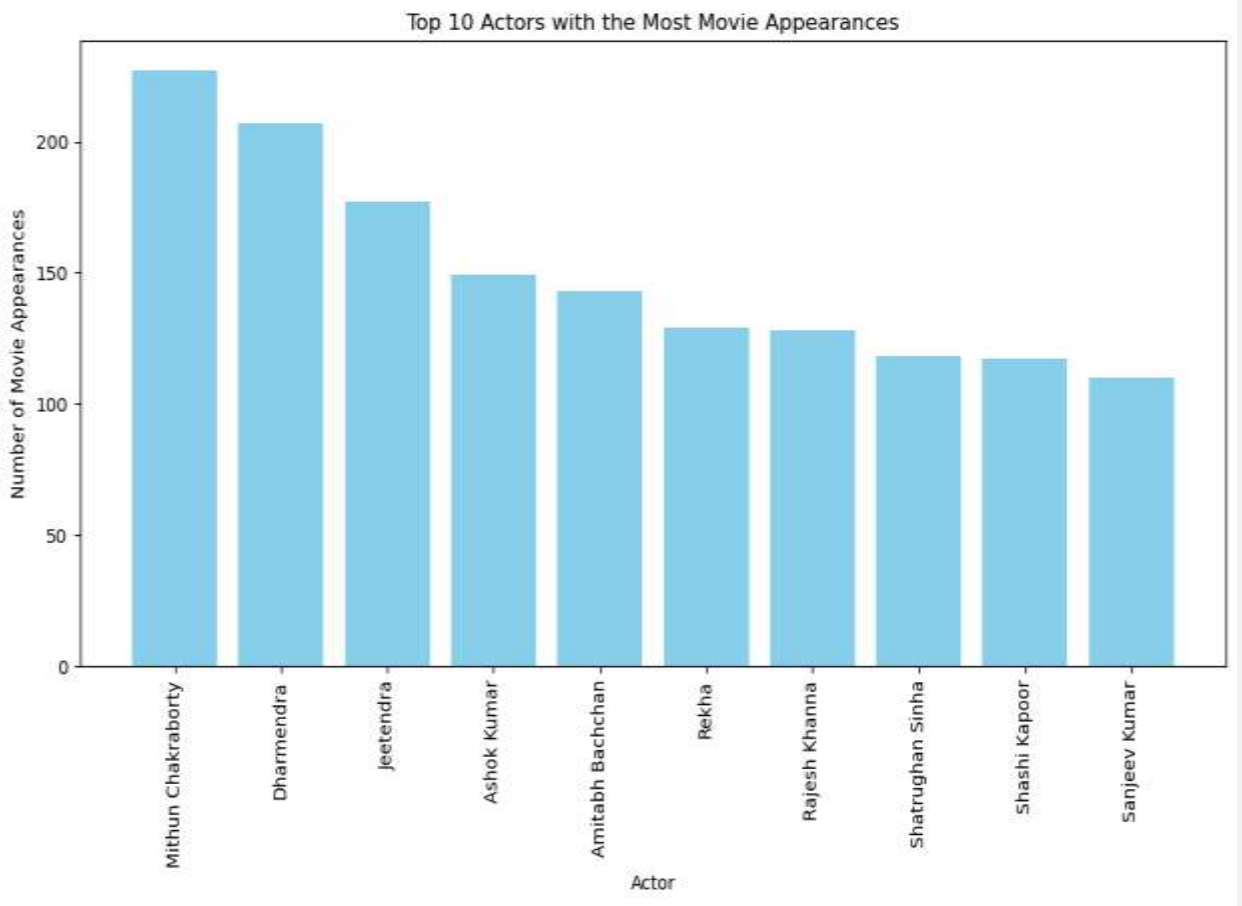
- Mahesh Bhatt directed the most movies

- The above directors are quite prolific in their careers.

TOP 10 ACTORS WITH THE MOST MOVIE APPEARANCES

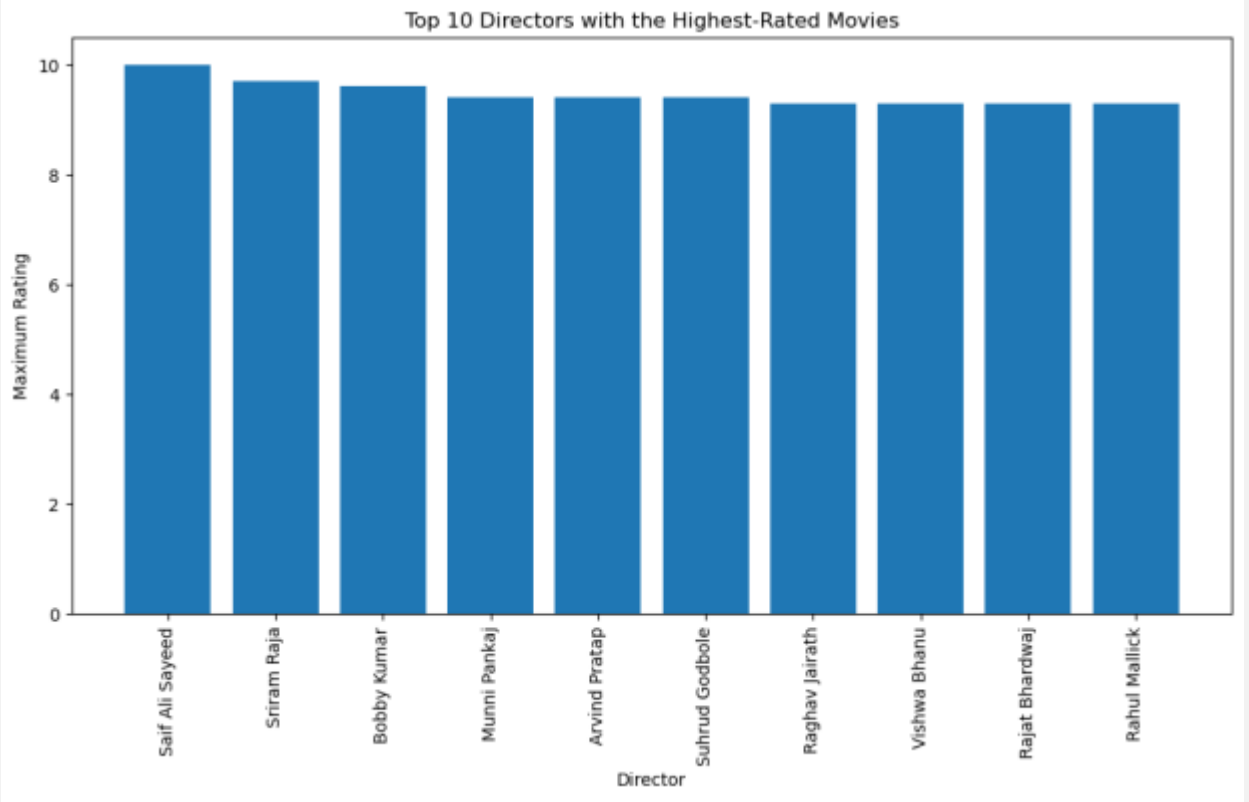
- Prolific Actors: The top 10 actors have made a substantial number of movie appearances, indicating their prolific careers in the film industry

- The actor who appeared the most in these movies is Mithun Chakraborty.

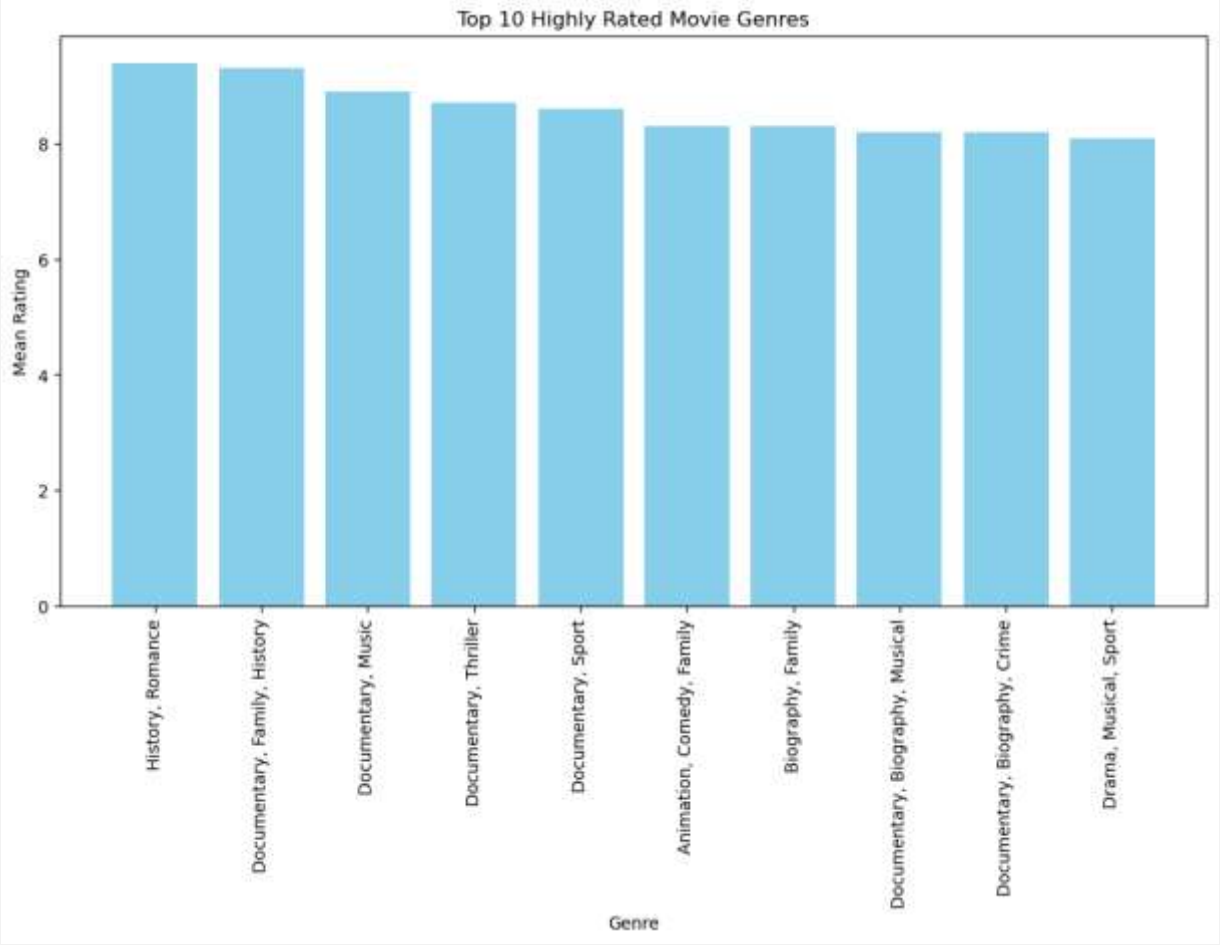


TOP 10 DIRECTORS WITH THE HIGHEST-RATED MOVIES

• Saif Ali Sayeed has directed the most succcessful Movies among other directors. This shows that the likelyhood of a movie to be rated high if directed by him is high.

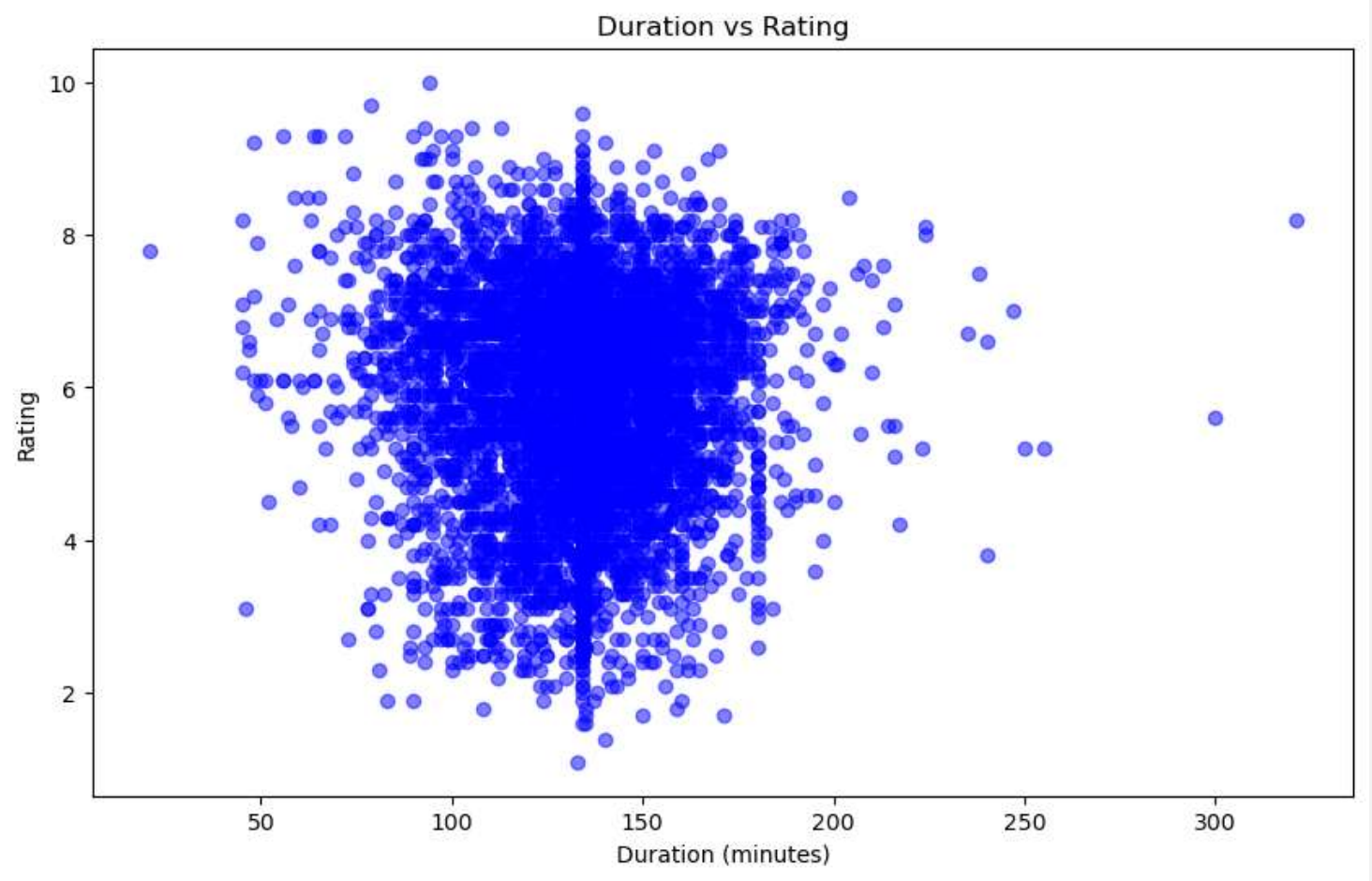


TOP 10 HIGHLY RATED MOVIE GENRES



- Genre Popularity: The top 10 highly rated genres are likely to be popular among audiences, as reflected in their mean ratings
- History, Romance genre has the highes rating hence wise to investing in the genre
- Critical Acclaim: High mean ratings often indicate that these genres receive positive reviews and critical acclaim from both audiences and critics.
- Filmmakers, actors, and production companies may consider collaborating within these genres to create well-received movies.

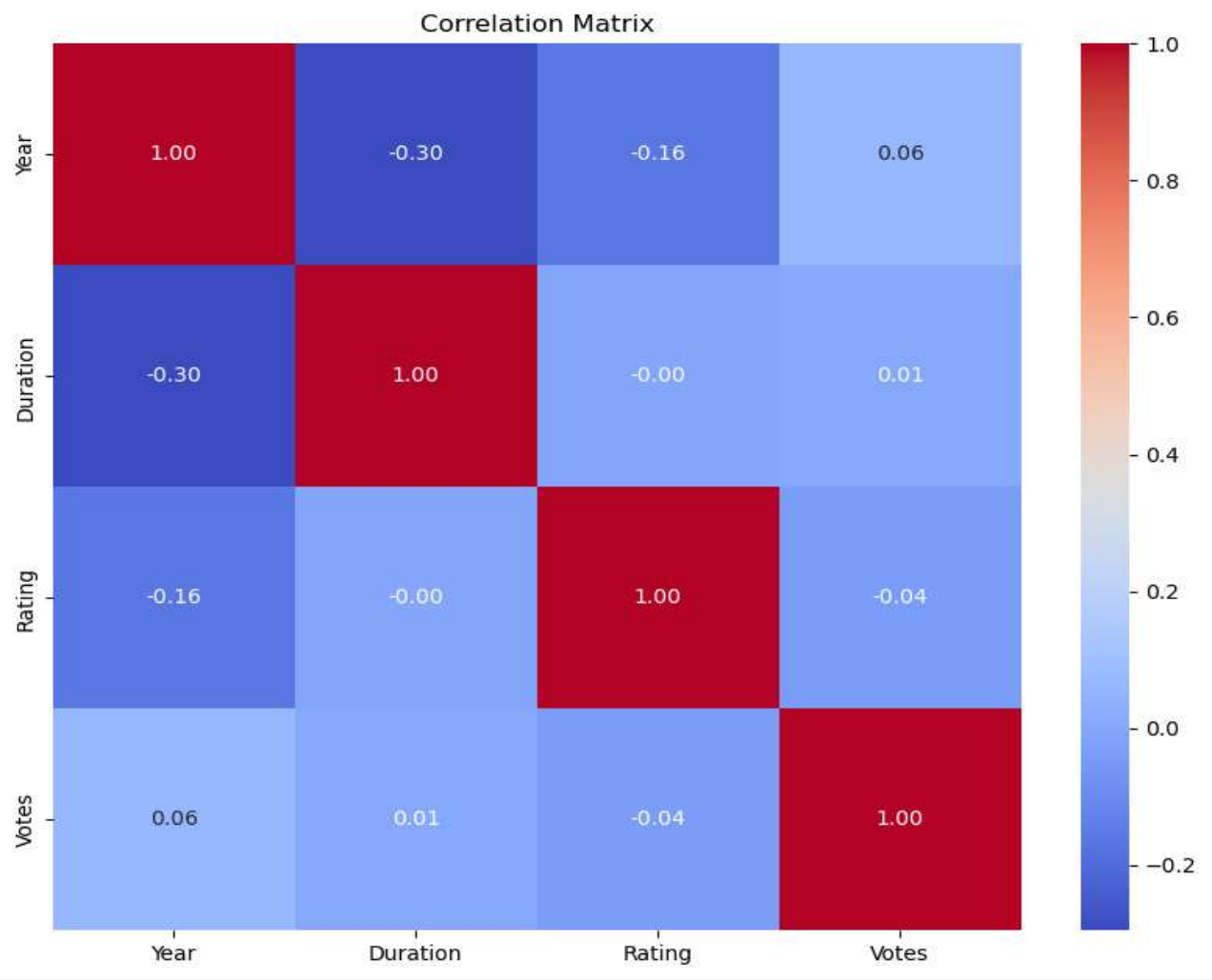
DURATION VS RATING



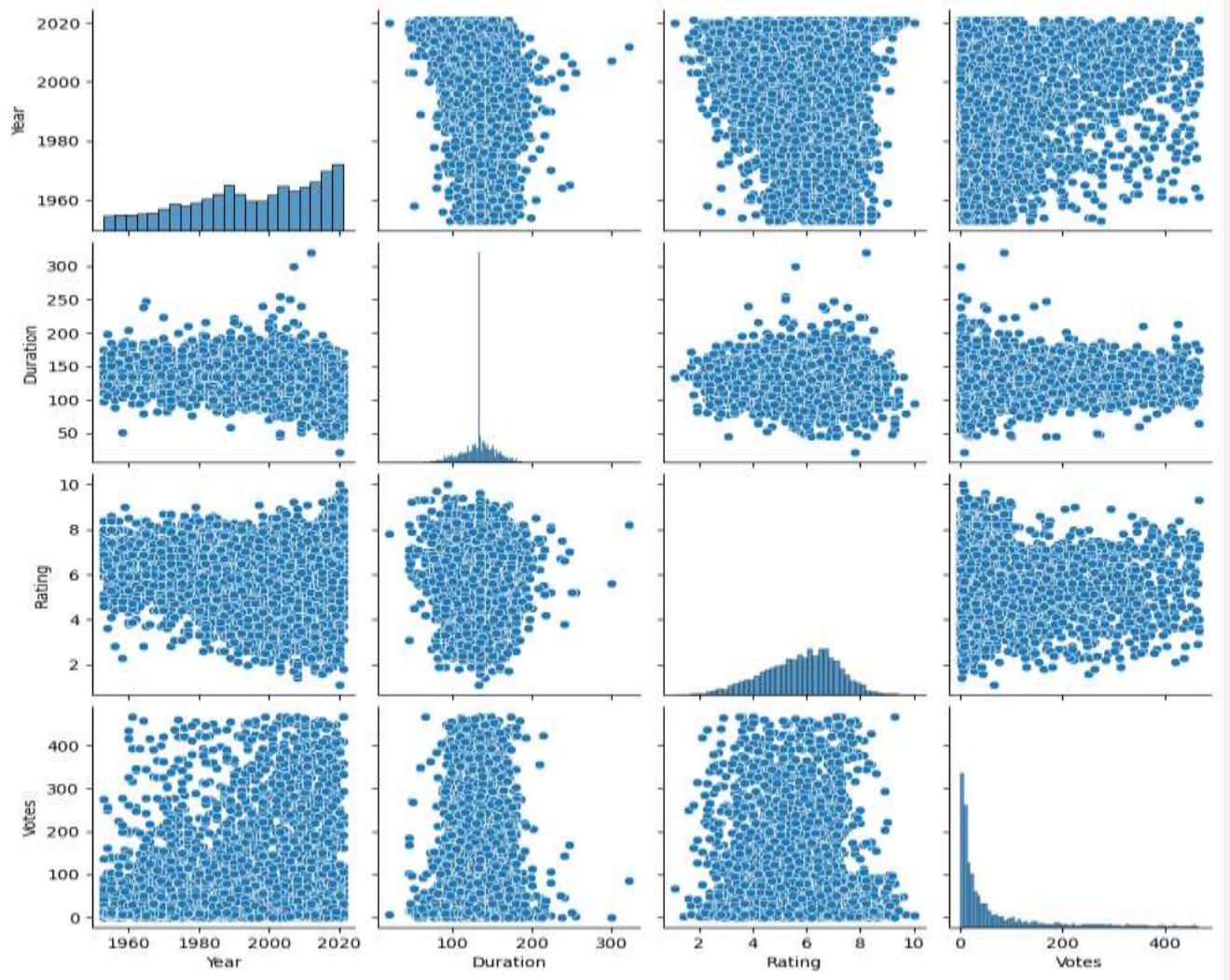
- There doesn't appear to be a strong linear relationship between movie duration and ratings. you can't easily predict a movie's rating based solely on its duration.

CORRELATION MATRIX

- The variables are not Highly correlated. The absence of strong correlations between numerical variables is a positive sign, as it reduces the risk of multicollinearity in regression analysis



PAIRPLOT



No strong correlation between variables

MODELING

- Baseline Model - Linear Regression
- Second Model - Random Forest Model
- Third Model - Gradient Boosting Regressor
- Hyperparameter Tuning of Random Forest Model
- The tuned Random Forest performed better than the tuned model in terms of model generalization and avoiding overfitting, due to its more balanced performance between training and test data. This will be the final model used for prediction.

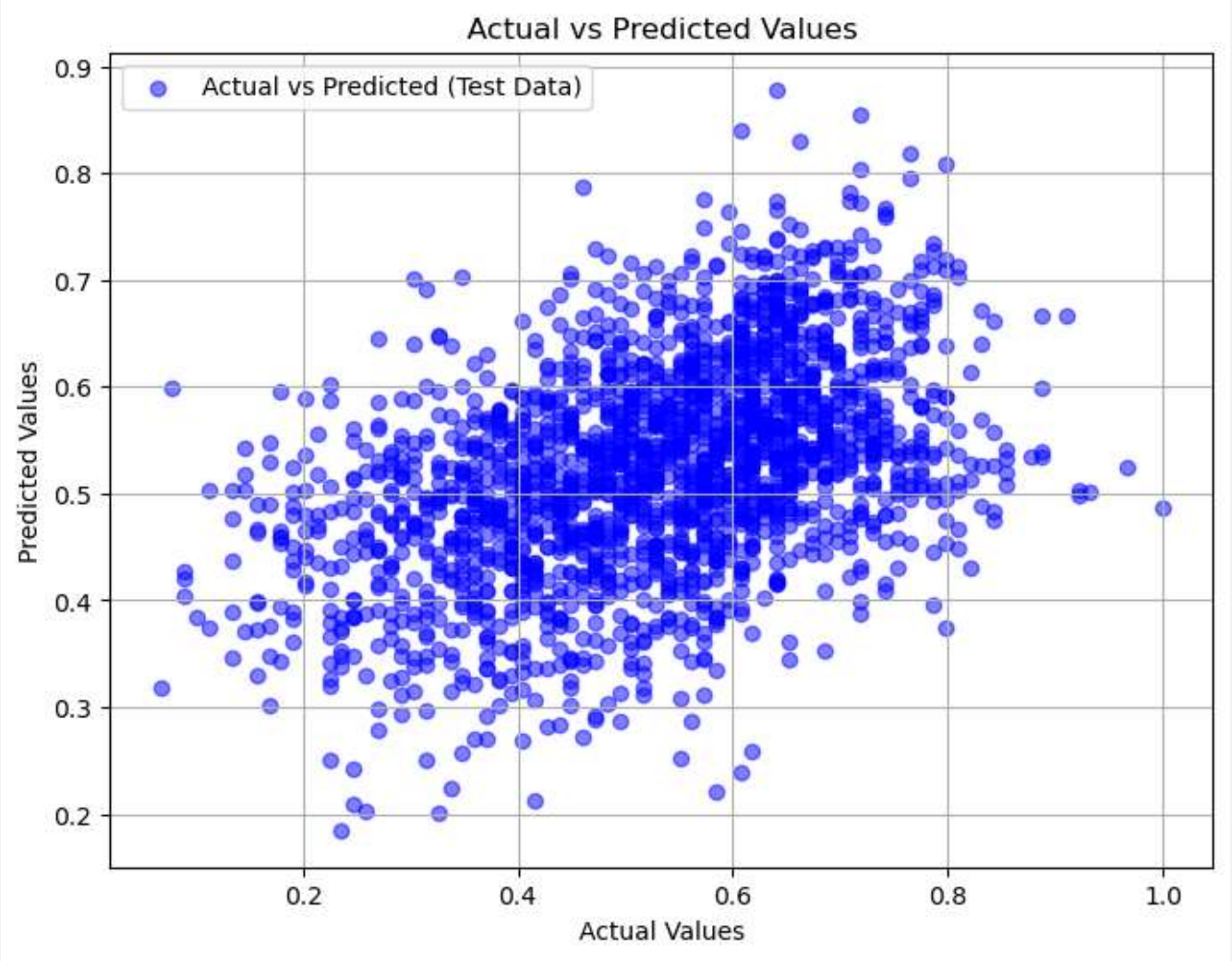
MODEL EVALUATION

- Metrics:

- Mean Squared Error (MSE)

- R-squared (R^2)

- Visualization of predicted vs actual ratings.

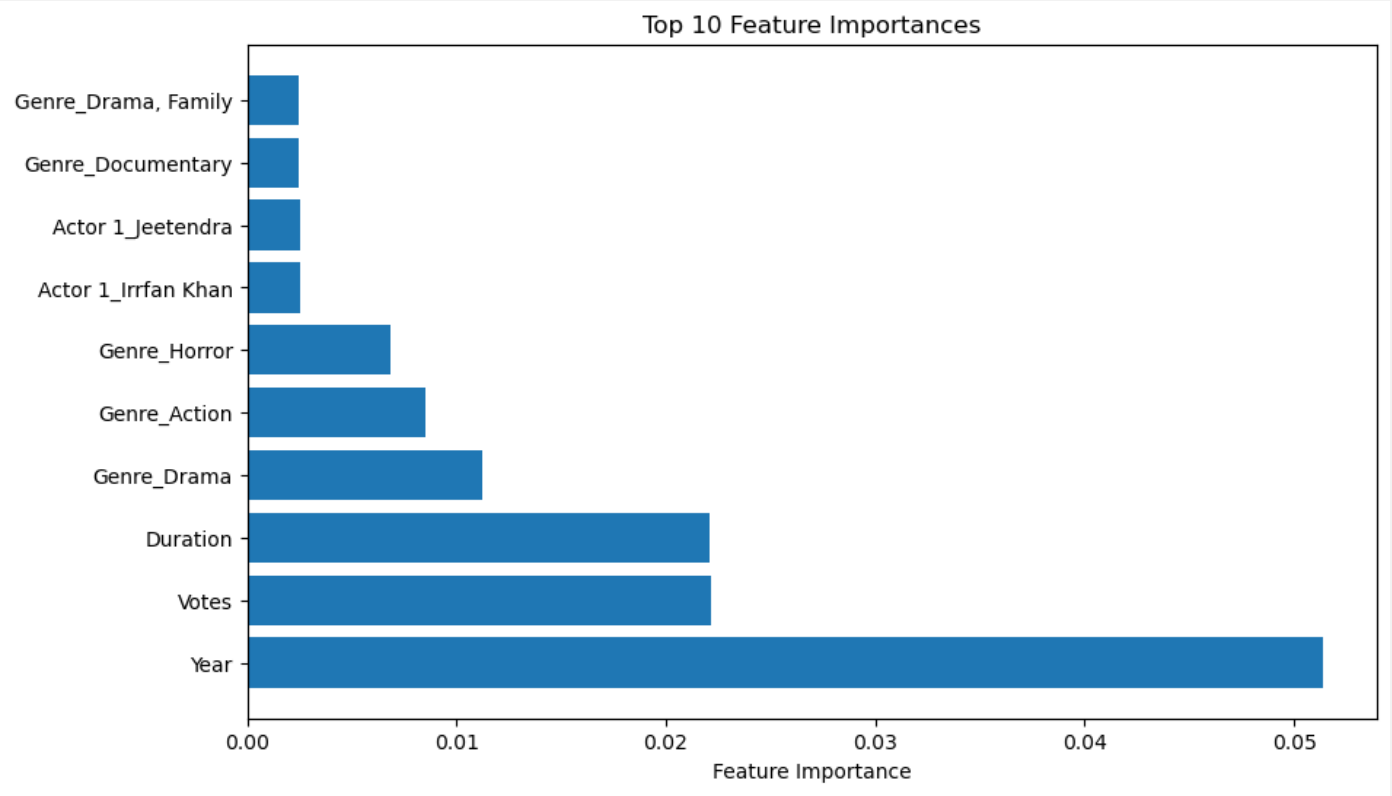


The scatter plot likely shows a scattered pattern of points away from the diagonal line, which aligns with the negative Test R^2 and non-ideal Test MSE. These results collectively suggest that the Linear Regression model is not a good fit for predicting movie ratings and may require improvement, such as exploring more complex models or refining feature selection and engineering.

MOST IMPORTANT FEATURES

Features that had most significant impact on the target variable in the above model:

- Year
- Votes
- Duration



RECOMMENDATIONS

- Year: Stay current with trends for successful movie releases.
- Votes: Encourage audience reviews for more reliable ratings.
- Duration: Align with audience preferences for movie length.
- Collaboration: Partner with established industry figures.
- Genres: Invest in highly-rated genres like History and Romance.
- Duration vs. Ratings: No strong duration-rating correlation.
- Critical Acclaim: Aim for positive reviews and acclaim

CONCLUSION

In conclusion, this project provides valuable insights and a predictive model for movie rating prediction. The film industry can benefit from these findings to make data-driven decisions regarding movie production, casting, and marketing. The most influential factors identified are the year of release, the number of votes, and movie duration.

NEXT STEPS

- Model Refinement: Improve the rating prediction model with advanced machine learning methods.
- Feature Engineering: Enhance the model's performance by experimenting with new features.
- User Reviews Analysis: Analyze user reviews and sentiments for deeper audience insights.
- Real-time Data: Implement real-time data updates to keep the model current with the latest trends

APPENDIX

For notebooks, datasets and scripts, follow this GitHub Repository link:

https://github.com/Ajay15Khanna/Afame-Technologies-Data-Science-Intern