IMDb

PROJECT PROPOSAL

PREDICTING IMDB RATINGS
USING LINEAR REGRESSION

PROPOSED BY:
SULTANAH ALDOSSARI
RAWAN ALHARBI

COMPANY BACKGROUND

IMDB is an online database of films and television programs. IMDb stands for Internet Movie Database. It is an extremely detailed and rich source of film data that features top movies, news, free movies, reviews, movie trailers, showtimes, DVD movie reviews, celebrity profiles, and more. It is the most authoritative source of entertainment information, with features designed to help fans explore the world of movies and shows and decide what to watch.

PROBLEM STATEMENT

We will predict IMDb ratings for movies based on features available in the dataset. The model uses linear regression and features gathered by scraping IMDb movie information with BeautifulSoup.

THE AIM

Our goal is to create a linear regression model with machine learning (ML), to predict movie ratings using various features (e.g., run time, Meta score, year, etc.).

QUESTIONS/NEEDS

- What are the features that have the strongest correlation with our target, number of voters?
- Are movie ratings predictable from our features?

DESCRIPTION OF DATASET

IMDb has an API available to download bulk data, but a primary requirement for this project was to obtain data through web scraping. Our data consist of 16 features and 1000 observations sourced from IMDb website. The features are (PosterLink, Series Title, Released Year, Certificate, Runtime, Genre, IMDB Rating, Overview, Meta score, Director, Star1, Star2, Star3, Star4, No of Votes, Gross). Our target variable is the number of votes and the features that might help us in predicting our target variable are (Released Year, Runtime, IMDB Rating, Meta Score, Number of Votes).

TOOLS

- Python
- Jupyter Notebook
- SQLite3

LIBRARIES

- Matplotlib
- Pandas
- Numpy
- Seaborn
- SciKit Learn
- StatsModels