# IMDB-analysis

## Dataset

This analysis was performed on the IMDB datasets found here: dataset

In this dataset we can find the following tables:

- title.akas.tsv.gz: This table gives us a list of regions that a movie was featured in. The worldwide appeal of a movie can indicate how popular it is and can also correlate with its quality.
- title.basics.tsv.gz: Gives basic information like runtime, number of votes, etc. This table will form the basis of our dataset.
- title.crew.tsv.gz: Gives the directors and writers that have worked on a title.
- title.episode.tsv.gz: this table is irrelevant for us since we are only considering movies
- title.principals.tsv.gz: This gives us an ordered list of actors, directors, and other crew members who have worked on a film
- title.ratings.tsv.gz: This table maps a title to its IMDB rating.
- name.basics.tsv.gz: Gives the names of actors and some titles that they are known for. This data will be used later for imputation.

## Notebooks

This repository contains the following notebooks

- Data Sampling: This notebook features a first look at the data and subsamples the data for the movie and TVMovie categories
- Data Imputation and Feature Engineering: This notebook attempts to impute any missing features and created a new set of features that can be used in the dataset
- EDA: This performs exploratory analysis on our created dataset. Here we explore the relationship of our features with our target variable
- Modeling: This notebook features data preprocessing, model building, and model tuning.

## Summary and Key Insights

### Sampling

- Out of the 1,219,870 unique titles, ~ 328,000 were either movie or TV movie titles

### Data Imputation and Feature Engineering

- Imputed values for genres using the distribution of genres that the crew has worked on in the past
- Imputed runtime and release yer using the TMDB API
- Created features for professional quality pertaining to both the cast and the crew
- Created features for the number of regions that a movie was featured in
- One-hot encoded the genre values

### EDA

- Found relationships between each of our features and our target (IMDB ratings)
- There is a strong positive relationship between the professional quality features and the IMDB rating
- There is a weak positive relationship between the number of votes and the number of votes with the IMDB rating
- Runtime has no effect on the IMDB rating
- We see certain genres have ratings distributions deviating from the average

  Better than average:

  - War
  - Sport
  - Music and Musicals
  - Documentaries

  Worse than average:

  - Sci-fi
  - Horror
  - Adult
- There is a strong correlation between `Adult` and `isAdult` (as expected) so isAdult was dropped while modeling

### Modeling

- Tried four different regression models
  - Linear Regression (Baseline)
  - Random Forest (Overfits)
  - XGBoost (Best Performer)
  - Neural Net
- The XGBoost model performed the best with an R2 Score of ~0.83 on test data.
- Looking at the coefficients of the linear model we can see that the `crew_mean` and `cast_mean` features have a very positive relationship with the rating. Certain genres like `film-noir`, `animation`, `short` and `western` also have a strong positive relationship with the IMDB rating. Other genres like `sci-fi`

and `adventure` show a negative relationship

- Looking at the feature importance of the XGBoost model, we can see that the professional quality features play a very important role in the model decision. This indicates that the strongest predictor for the IMDb rating of the movie is the quality of the crew. Another interesting insight is that the CREW quality matter more than the CAST quality when it comes to IMDB rating. Most production houses will weigh the cast more when it comes to "bankability" but the IMDb rating depends more on the directors, writers, etc.
- Performed hyperparameter tuning using hyperopt which uses bayesian search to find an optimal set of parameters.
- The largest mispredictions occur when the cast of the movie is decent but the movie overall is bad.

# Future Scope

Given more time, we can improve and augment this dataset to produce better predictions. These are the few ideas that can be implemented in the future.

## Dataset

- Augment the dataset with external sources like the TMDB API. I did use the TMDB API for data imputation, but due to its rate limiter for the free tier, I wasn't able to use it to enrich the entire dataset within the given timeframe.
- We can see that the model's predictions heavily depend on the performance of the cast and crew. Finding a way to better define the performance index of the cast and the crew may impact the model's predictive power.
- Adding temporal features for the cast and the crew may aid in the model's performance. How well a movie doesn't necessarily depend only on the cast's and crew's past performances but also on their performance trajectory. Having a feature that encodes this temporal aspect may help in increasing predictive capability.
- Film budgets are also a strong indicator of quality. Finding a way to add in the budget could help improve the model's predictive performance. This information was available on the TMDB API, but I couldn't incorporate it due to the rate-limiting constraints
- We could better impute the genres if we also had reviews for each of the movies. Performing topic modeling or text classification can help us map certain keywords to certain genres.
- We can add external ratings like Metacritic score which correlate with the IMDB ratings

## Models

- More hyper-parameter tuning to gain better model performance.
- More model analysis to understand what distribution of features have the lowest predictive power

## Inference

- Build an API service that can do real-time inference for new movies.

# Takeaways

- Augment the dataset with external sources and APIs (TMDB)
- Since the model relies on cast and crew metrics, build more features related to professional quality.
- Crew metrics matter the most relative to other features
- Certain genres of film have a higher average rating than others