Project Presentation

REZA MARZBAN

IMDb movie Dataset

- Chose 3 tables out of 7 available tables, and joined them on Movie ID (inner-join).
- Dropped unuseful columns.
- Remaining raw features:

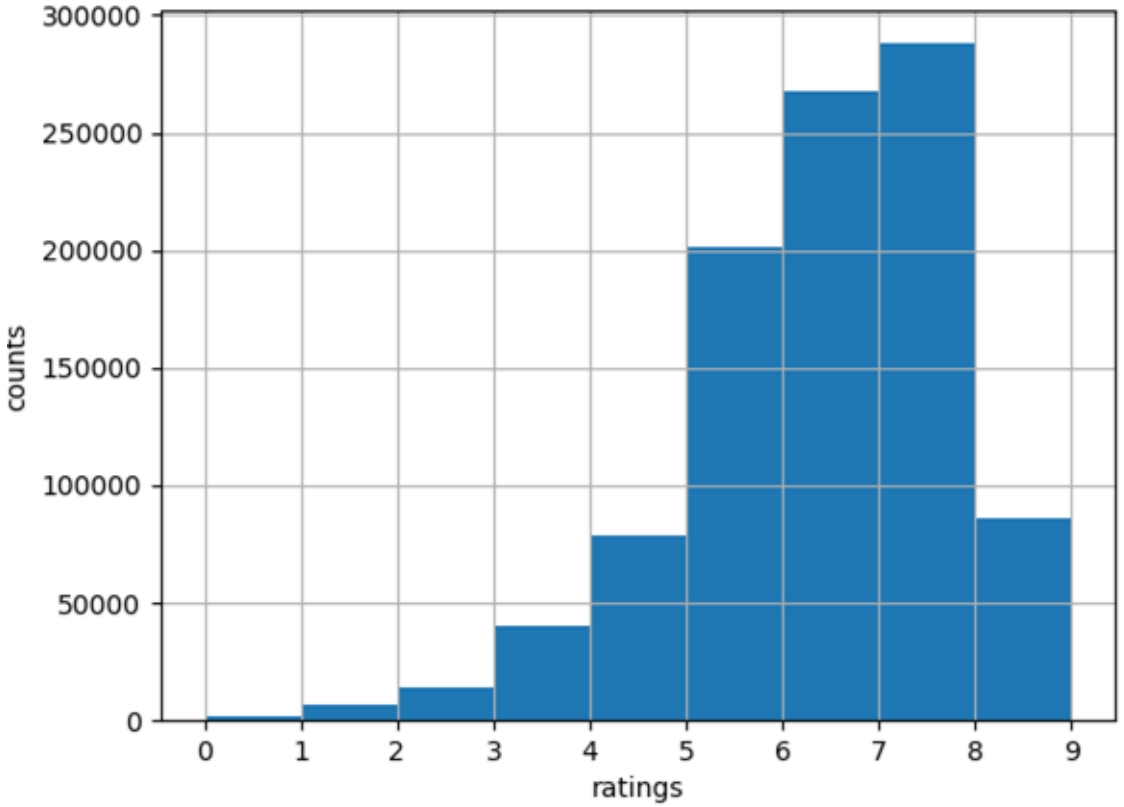
  Year, Genres, Title-Type, runtime, directors, average-rating

- **Goal:** Try to come up with either a regression or classification problem to predict rating.

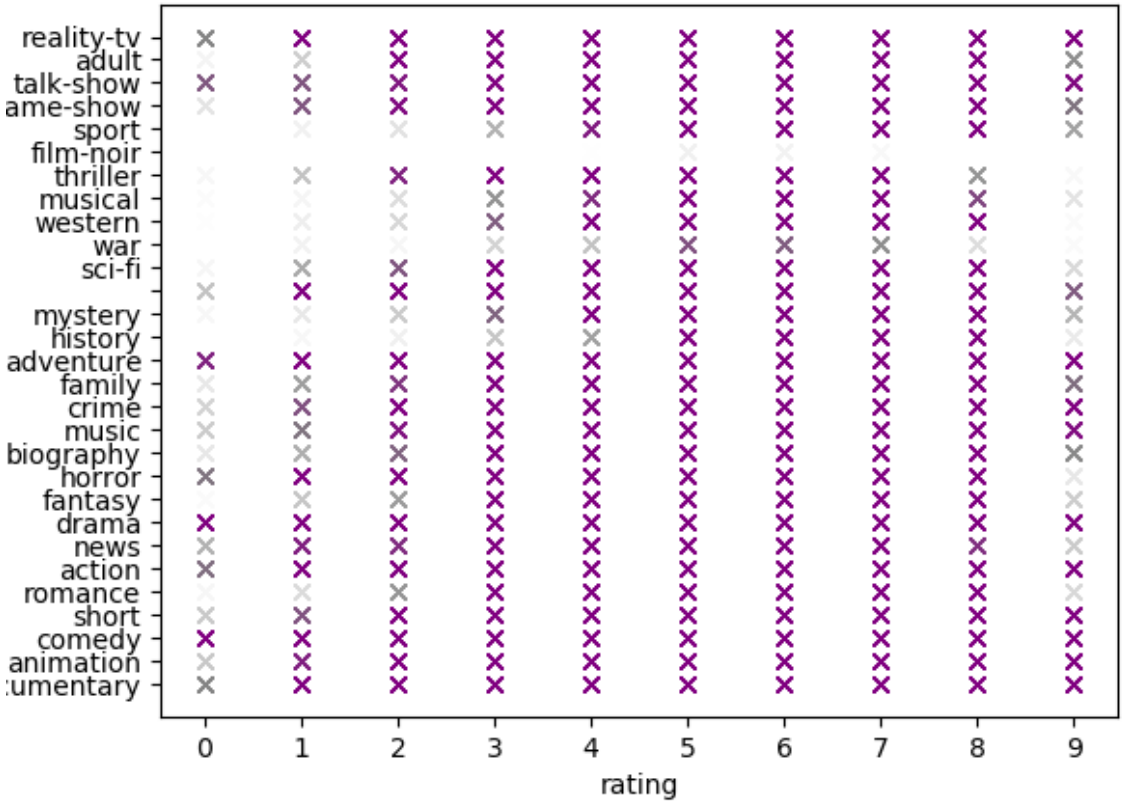
https://www.imdb.com/interfaces/

Initial Results

RATING DISTRIBUTION

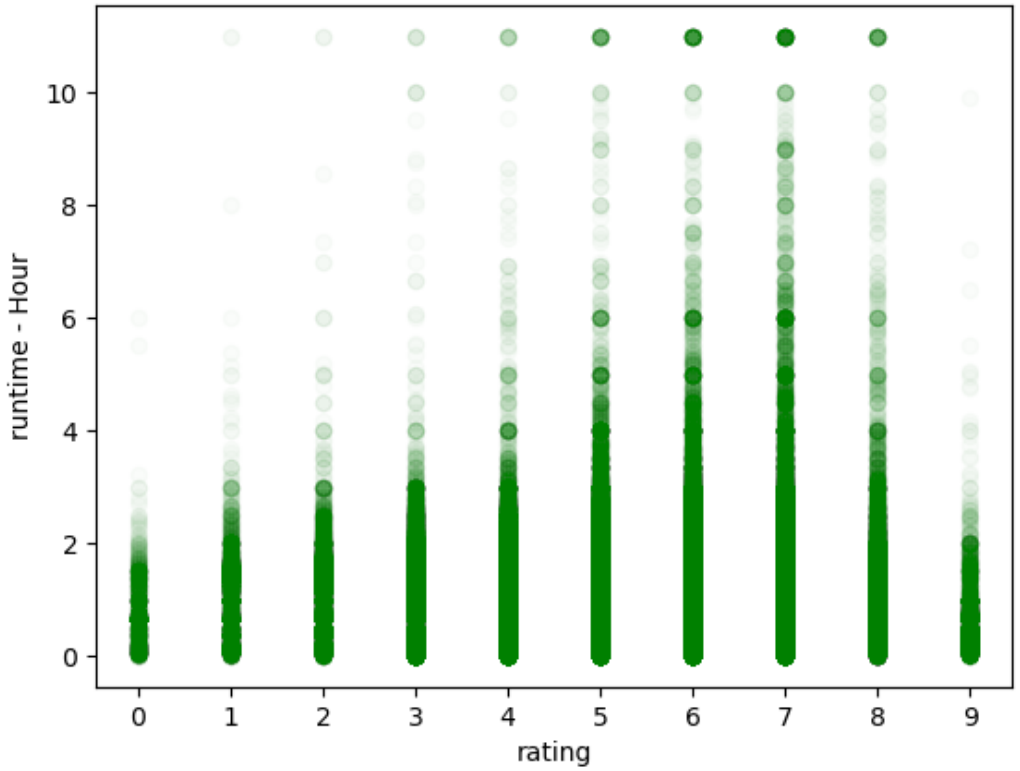


GENRES VS RATING

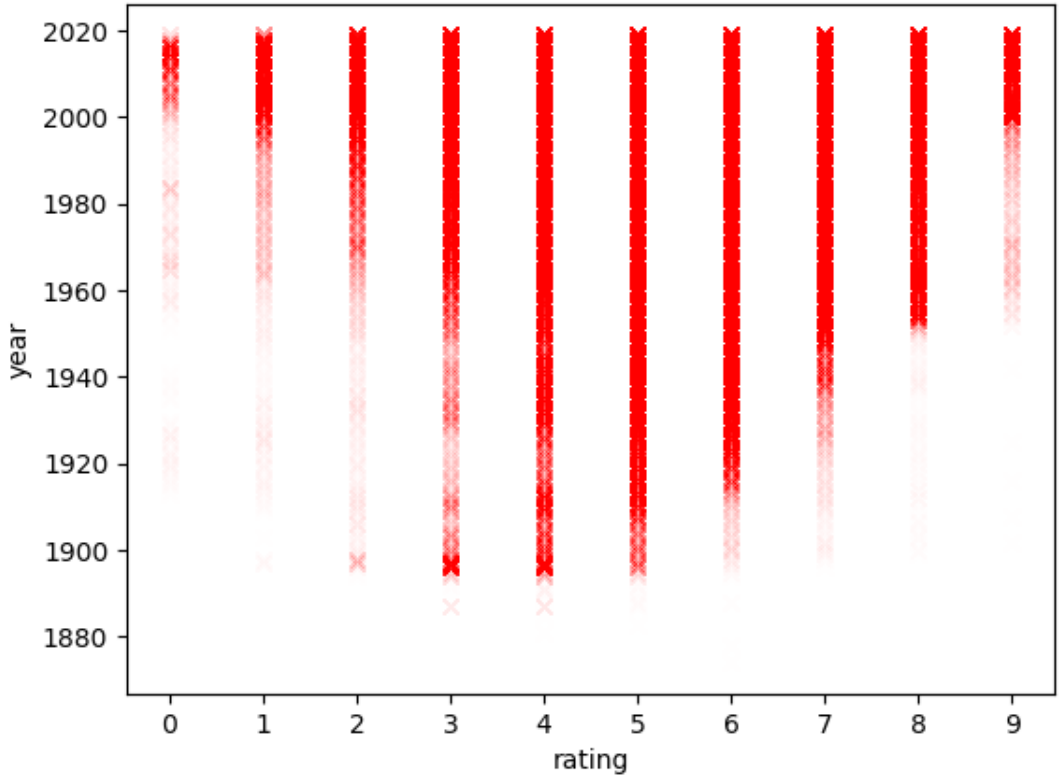


Initial Results

RUN-TIME VS RATING



YEAR VS RATING



Challenges

- The dataset has around **one million rows** (movies).
- **Unbalanced** labels.
- Pre-processing is hard for some features like Directors. (we have thousands of unique values).
- There may or may not be a relation between our features and our prediction labels (ratings).
- After preprocessing, we have **100 features** which may cause problems related to the Curse of Dimensionality.