Recommender Engines with Movies Dataset

Table of contents

01

Overview of the project

Overview

Context of project

The Author chose to tackle the movie industry, with the intention of exploring Recommender Engines that is widely adopted in movie streaming services such as Netflix, Disney+ etc

Tech Stack Used

Python, SQLAlchemy, Matplotlib/Seaborn, Tensorflow/Keras, Streamlit

Methods Adopted

Data Acquisition – requests, BeautifulSoup
Data Handling – SQLAlchemy
Data Cleaning – pandas, missingno
Data Visualization – matplotlib, altair, seaborn
Machine Learning/Modelling – Tensorflow/keras
Deployment – Streamlit

Objectives

Data Analysis

Exploiting insights from the IMDb Dataset to understand characteristics and trends of the movie industry in the past 10 years

Building a Recommender Engine

Creating a Recommender Engine to recommend movies of interest to users

02

Datasets used

Datasets

1. Initial exploration of using **IMDb dataset** + scraped values of user reviews + revenues
    - Too little data scraped + insufficient time – Data density of 0.04% (ie 99.96% missing values)
    - Initial Keras model had MSE of ~3.5
    - Dataset used for data analysis + building of simple cosine similarity engine instead (demo in streamlit)
    - Synthesized 2 columns: popularity_score* & recommendation_propensity

$$popularityscore = R * \frac{v}{v+m} + C * \frac{m}{v+m}$$

$$recommendationpropensity = popularityscore - k * \frac{age}{10}$$

v – number of votes
R – average rating
m – min votes
C – mean rating across all movies

2. Retrieved **Movie Lens 1m dataset**
    - Popular dataset for training of recommender engine models
    - Data density of 3% (ie 97% missing values)

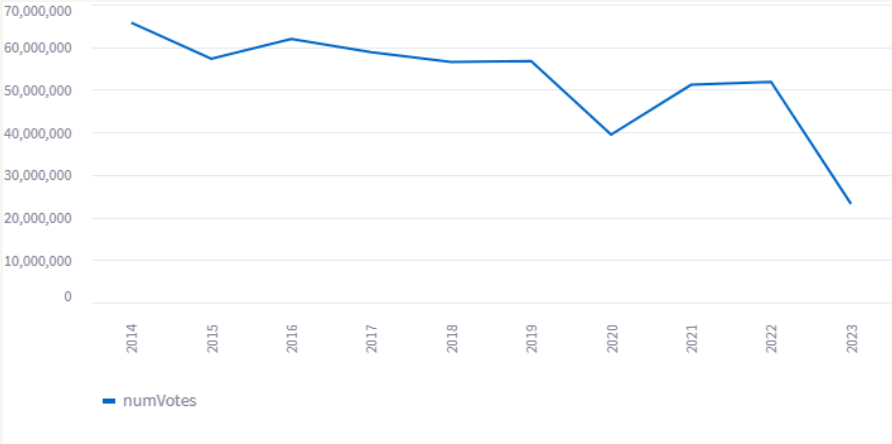
03

Key Insights

Key Insights

1. Steady Decline of total voters in IMDb

With the rise in popularity of streaming services provided by tech giants, such as Netflix, Apple and Amazon, coupled together with their platform for users to rate and review movies, this has **taken market share away from IMDb** as evidenced by the falling number of voters



Key Insights

2. The rise and fall of select genres

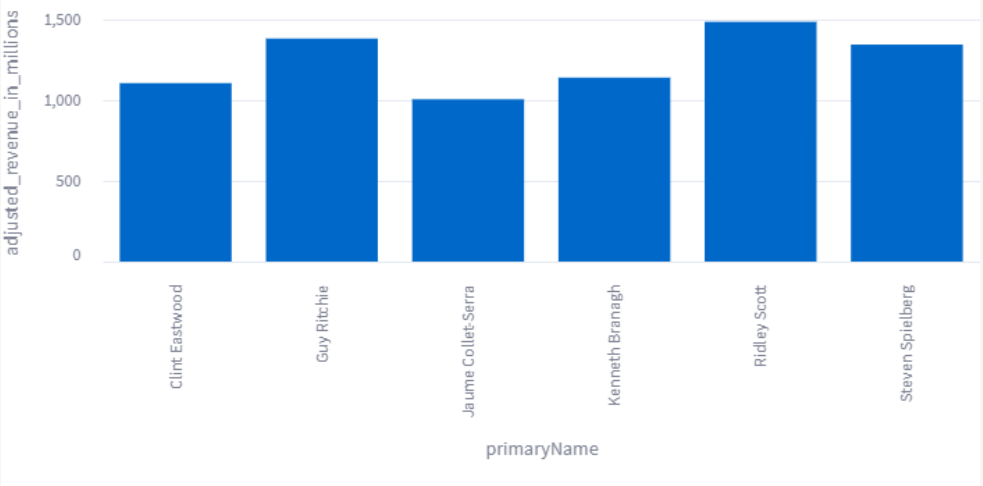
Over the years, there have been a clear increase in popularity for genres like Fantasy, Horror, Mystery and Romance.

Traditional genres such as Biography, Documentary and Musicals have somewhat tapered off, perhaps due to the increasing degree of digitalization of the world resulting in traditional preferences being eroded

Key Insights

3. Big earners in the industry

Revenue figures for films by successful directors can yield up to billions of dollars



Total revenue earned from directing films between 2013 to 2022

Key Insights

4. Revenues took a hit during covid, but some genre of movies recovered faster than others

Genres like Adventure and Animation took a significant decline in average revenue per film in 2020, but made a quick recovery in average revenue per film

Some genres (like mystery, war, biography) are still trying to play catch-up to pre-2020 figures

04

Recommender Engines

Models Used

Cosine Similarity

- Attribute based recommendations
- Unsupervised form of recommendation
- Limited to non-personalized recommendations
- Dataset used: IMDb dataset

Tensorflow/Keras

- User based collaborative filtering
- Supervised form of recommendation with MSE of ~0.94
- 2 Embedding layers with latent factors of 5, 1 Flatten layer and 4 Dense layers
- Dataset used: Movie Lens 1m dataset

05

Streamlit

06

Whats Next

Next Steps

Further Data Analysis

Adding more interactive pages to Streamlit web app

More Machine Learning

Re-training a user-based collaborative filtering model using Movie Lens 20m Dataset

Incorporating Bayesian Ranking to update recommendations based on selections

UI/UX Refinement

Repositioning some elements/pages in Streamlit web app

Refining language used and proofreading writeup

Thanks!

Do you have any questions?