What can we learn from user-movie ratings?

Niloofar Zarif, Lucie Polakova, Deepansha Chhabra

niloofar.zarf@gmail.com, lucie.polakova.18@ucl.ac.uk, deepanshachhabra21@gmail.com

Abstract

There is an abundance of available data in many domains that have to do anything with social media or online services. This may cause frustration for computer scientists and engineers as large bodies of data are not comprehensible to human-being. Systems have been and will be designed to take advantage of large datasets but without knowing enough about the existing patterns in the data, the designed systems may not be perfect. In this project, we are trying to tackle this problem in the area of user and movie interactions and extract useful knowledge from the publicly available datasets. The results of this project will help engineers to know more about user preference and design better movie recommendation systems which will yield higher user satisfaction.

Introduction

In the past decade, large companies such as Netflix, Amazon, Apple and Disney have invested a large amount of resources and effort into their streaming services[3]. In 2020 the number of subscriptions to online movie streaming services surpassed 1.1 billion and consumers spent around 80$ billion on streaming content while this number was only 12$ billion for movie theatres[1]. These numbers and records show how important it is to design movie recommendation systems that serve the users as effectively as possible. Designing such systems without knowing user preferences would be a shot in the dark. Different users certainly have different tastes and preferences but there will be patterns in the data that can point towards the most common preference criteria as well as introducing the temporal or regional user preferences.

Movie streaming companies are well aware of the complicated nature of the recommendation task. Therefore, unlike many other areas, large datasets have been made publicly available by companies who own them. These datasets are the results of companies logging interactions within their systems. For example, Netflix and Amazon Prime have been very open in releasing data and updating it every year. This gives the wider public opportunity to learn about how users act on these platforms.

As a prominent movie streaming service, Netflix has more than 209 million users as of July 2021[2]. In 2006 Netflix started a competition called the Netflix prize. The competition was open to anyone around the globe and the purpose was for the participants to design the best collaborative filtering algorithm to predict user ratings for movies. Back then collaborative filtering algorithms were the dominant solution for recommender systems. These days collaborative filtering has been dragged aside by deep learning models. Yet what the Netflix prize competition meant for the public was that a large amount of data was

published by Netflix. The culture of publishing data has continued and to this day we still can get our hands on large datasets of user-movie interactions from Netflix.

For this project we will use Netflix data to extract knowledge and discover patterns which can be found within the dataset. Assuming the rating users give to each movie is the dependent variable, we are going to see how the other independent variables including genre, duration and country of origin affect the rating. We will seek meaningful patterns in the large body of data we have and will try to use adequate visualization techniques to present them.

We are aware of how effective a nicely designed visualization can be and therefore we will seek after the suitable visualization idiom for each case. We also know the dataset we are dealing with is huge and this will make knowledge extraction and visualization more complicated. But we are ready for the challenge as we believe it can be overcome by designing creative solutions. We hope for the results of this work to be used by people trying to design and build recommender systems and make them able to come up with more effective solutions. Another group of users of this project can be the managers of the movies industry who are trying to know more about users' taste to use that knowledge while making business decisions.

The topic was chosen based on our common interest in recommendation algorithms and the motivation to enhance our understanding of how movie data can be analysed.

Related Work

There are various actual movie recommendation solutions using a range of algorithms, including Content Based Filtering, Collaborative Filtering, Hybrid Approach and Deep Learning Based Methods [4]. In terms of visualisation, there are projects and papers focusing on profitability and how it relates to different film properties [6]. There are a few solutions where data have been analyzed with the motive of making a contribution towards building a global brand and making a sustainable long term plan in terms of production of movies. The relationship between different parameters of the dataset have been shown by scatter plots, bar plots, kernel density estimate plots, histograms box plots, linear model plots, heatmaps etc., using programming tools like Python Libraries and simple and efficient tools like Tableau [5]. Based on few currently available analytical solutions, it has been observed that a particular genre movie may give the highest return per investment but is rated low and hence, in turn does not bring in high revenue. These analyses have proven to be effective tools assisting the movie production teams to get an insight of the audience's interests. Thus, our idea of identifying the correlations between user rating and other movie properties aims to fill in the gaps in the available research and provide an alternative tool when the aim is not pure financial profit.

Data and Task Abstraction

Domain

As mentioned, big data analysis is gaining more popularity in the media industry as well and some of the major providers actively seek public involvement in optimising the existing algorithms on working with

this data. This project therefore aims to provide a clear overview of the different patterns in the publicly available data by learning more about the user preferences. The target users are independent computer scientists and engineers that do not have the time and resources to conduct robust customer research and big data analysis, but want to contribute to the optimization of the recommendation algorithms and design new and better solutions. Managers in the movie industry could use the results to make business-related decisions with the potential of profit maximisation.

Task

The main task of the visualisation is to discover patterns that can be found within the datasets and clearly show the different relationships between user rating and other film properties (genre, country of origin, year, duration, original language etc.). As a result, the properties that have the highest effect on user rating will be identified.

Data

Data for this project are sourced from Kaggle and the final dataset will be obtained by linking multiple related datasets to gain a wide variety of properties for each attribute. In total, we will be working with a dataset of 45,000 items (movies) and the average rating will be obtained from a dataset containing 26 million items (ratings). The attributes we intend to use in our analysis are both categorical (year, country of origin, genre, original language) and ordered (rating, runtime, number of votes). Specifically, the levels of categorical attributes are as follows:

- Year: 61
- Country of origin: 53
- Genre: 15
- Original language: 26

The ranges for ordered attributes are:

- Rating : 1-5
- Runtime: 50-153
- Number of votes: 1-563,9

Solution

The final solution of this project will consist of a dashboard with multiple static as well as interactive visualisations, each addressing a different property and how it relates to user rating. Potential idioms and design choices include:

- Bar graph - to show correlation coefficients between user rating and different properties
- Heat map - clustered rating and the different properties to show an overview of the relationships
- Line graph - to show change of rating over time for different genres/other properties and identify the latest trends
- Box/jitter plot - to show the distribution of user ratings based on years/genres/etc.

Additional channels will likely be used to enhance the user experience and interaction with the visualisations. Given that a lot of our data is categorical, color and its properties (e.g. hue, luminance) might be used to distinguish between the different categories.

Potential features also include the ability to manipulate view of the visualisation such as allowing a temporary selection of items and properties that will be displayed and included in the dashboard.

Implementation

Datasets in a CSV format will be loaded into a pandas data frame in Python and will be cleaned and linked together based on the movie names. The combined datasets will be then further processed in Tableau Prep to ensure they are ready for analysis and visualisation. All visualisations will be designed using Tableau Desktop.

Scenarios

- Indie movie theatre manager is looking for what movies he should stream during the upcoming season in order to please their demanding and critical clientele. By interacting with our visualisation dashboard, he can clearly see the genre that has been trending in the past couple of years, which countries might be generally producing good movies and what runtime might most likely lead to satisfied customers.
- Independent computer engineer decided to work on a new machine learning recommendation system in order to contribute to an open call of a movie streaming company. The large amount of data available means that he has enough input, however, knowing what properties he should prioritize would remarkably shorten the development period. By interacting with our solution and selecting the different properties to display, he can clearly see the most significant starting points for his project.

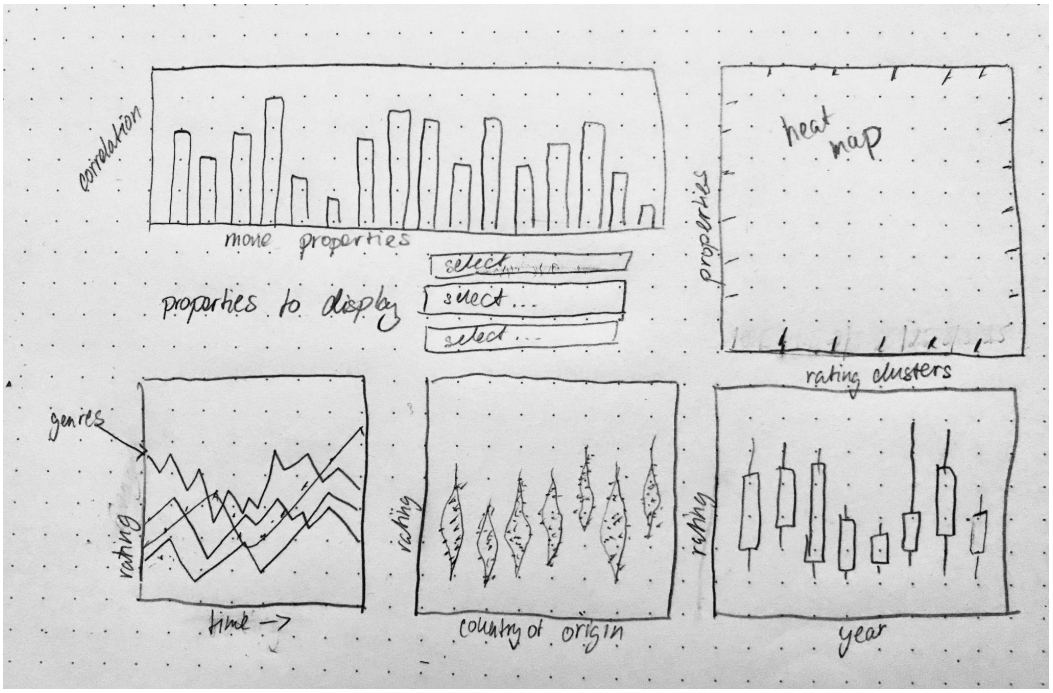

Figure 1: Low-fidelity prototype of final dashboard

Milestones

October 13: Preliminary Work

- (4 hours, All) Prepare project pitches, exploration and discussion on data source
- (5 hours, All) Meetings to discuss project direction

October 21: Project proposal

- (5 hours, all) Our first step is choosing the best dataset we can find and clean it such that it will be ready for our further use
- (15 hours, All) Develop and edit project proposal

November 10: - Preliminary data exploration, data preparation and vis finalisation

- (15 hours, Niloo) Cleaning data and linking relevant datasets to create a final dataset for analysis.
- (15 hours, Niloo) aggregation of rating records based on movie name
- (15 hours, Lucie and Deepansha) Further exploration of vis options, finalising the visualisations used, sketching/designing low-fidelity prototypes
- (2 hours All) Group discussion of prototypes and design finalisation

November 16: Update - Create first version of vis software using finalized idioms

- (4 hours, Lucie) Knowledge extraction phase, preparing data in Tableau Prep, preliminary statistical analysis, seeking meaningful patterns and useful knowledge
- (8 hours, Niloo) Finalise related works section for the final paper
- (10 hours, Deepansha and Lucie) Begin data analysis and correlation overview in Tableau.
- (10 hours, All) Write update report

November 24: Visualisation design

- (15 hours, Deepansha) Design 2 idioms on movie properties
- (20 hours, Lucie) Design 4 idioms on movie properties

December 3: Finalize vis dashboard

- (20 hours, All) Final refinements and bug fixes
- (20 hours, All) Developing other features of the vis and final dashboard (multiple views together with chosen idioms, interactivity to support selection/filtering)

December 15: Final presentation

- (10 hours, Deepansha) Slides
- (3 hours, All) Rehearsal, demo

December 17: Final paper

- (20 hours, All) Documenting and reporting the results, finalisation of the paper, formatting and citation check

Discussion, Future Work, and Conclusion

TBD

Bibliography

[1] Mendelson, Scott. “Streaming Subscriptions Top One Billion: A New Normal, Or A Temporary Disruption?” *Forbes*, 2021. *Forbes*, https://www.forbes.com/sites/scottmendelson/2021/03/18/streaming-tops-one-billion-subscribers-a-new-normal-or-a-temporary-disruption/?sh=37e626c8465a. Accessed 18 10 2021.

[2] Kats, Rimma. “Netflix statistics: How many subscribers does Netflix have?” *Insider Insights*, 2021, https://www.insiderintelligence.com/insights/netflix-subscribers/. Accessed 18 10 2021.

[3] Stodart, Leah, and Ashley Keegan. “Best streaming sites for movies.” *Mashable*, 2021, https://mashable.com/roundup/best-movie-streaming-sites. Accessed 18 10 2021.

[4] Sunilkumar, Chaurasia Neha. "A review of movie recommendation system: Limitations, Survey and Challenges." ELCVIA Electronic Letters on Computer Vision and Image Analysis 19, no. 3, 2020: 18-37.

[5] Lee, Jeremy. “Exploratory Data Analysis With Movies - Towards Data Science*”*. *Medium*, 2020, September 8 https://towardsdatascience.com/exploratory-data-analysis-with-movies-3f32a4c3f2f3. Accessed 20 10 2021

[6] Panchal, Kishan and Swalin, Alvira. “MSDS 622 Final Project”. *GitHub*, 2021, August 19. https://github.com/k7p/dataviz_project. Accessed 20 10 2021