MovieLens - Genre Analysis

Christian R Chasteau

Dataset

Which dataset did you use of the following:

- MovieLens Dataset - from week four tutorial

Motivation

In the Movie business, hundreds of millions of dollars invested in a movies, with no idea of the potential returns or box office bomb[2].

But Sequels; how many *Pirates of the Caribbean, Star Wars, Star Trek, Fast and Furious* I won't mention *Harry Potter.* Which genres were these? Multi genre perhaps!

Did someone analyse the data and figured that maybe it's a good idea to make some more? The more insight you have the better[6]. Not that anyone would say that they didn't make that movie because a data scientist said it will flop. Nope, not going to happen; will you get to hear about it, if it did?

Who else thought it's a good idea to analyse data in this way. Neflix[5], for their recommender system.

The dataset used is created by the MovieLens[4] recommender system by its Users using Collaborative Filtering[4] to recommend a movie and allowing its Users to rate movies. Movies in this dataset are rated and the genre is also rated indirectly as part of the movie rating. So, using this dataset, can some insight be gained into how different genres behave when compared with each other.

Research Question(s)

One step in exploring the abstract [6] from the motivation would be:

1. How does most highly rated genre compare with sci-fi genre on a year by year count for movies released across a given years range.
2. Are there any limitations using this specific dataset to gain greater insight?

Findings

Ratings dataset

- a one-to-one relationship between movie and rating

Movies dataset - Genre

- a one-to-many relationship between movie and genres
    - Example of many genres : `Crime|Drama|Horror|Mystery|Thriller`

Genre splitting itself is ok, the key is when you merge this with the ratings data, now a deficient arises - Genres are not directly rated. Unfortunately, as it's the movie that is rated, this implies that the rating should apply equally for all genres for that movie. ***This is an assumption.***

Findings

Movies dataset - other attributes

- Earliest Sci-Fi movies dates to 1898; a few movies made during the early years '***Early years Sci-Fi Movies***' below, means that counts will need to be for the same year for all genres to make for a realistic comparison plot
- The final year's data must be for a complete 12 months, not anything else as seen in the '***Final year Movies***' below

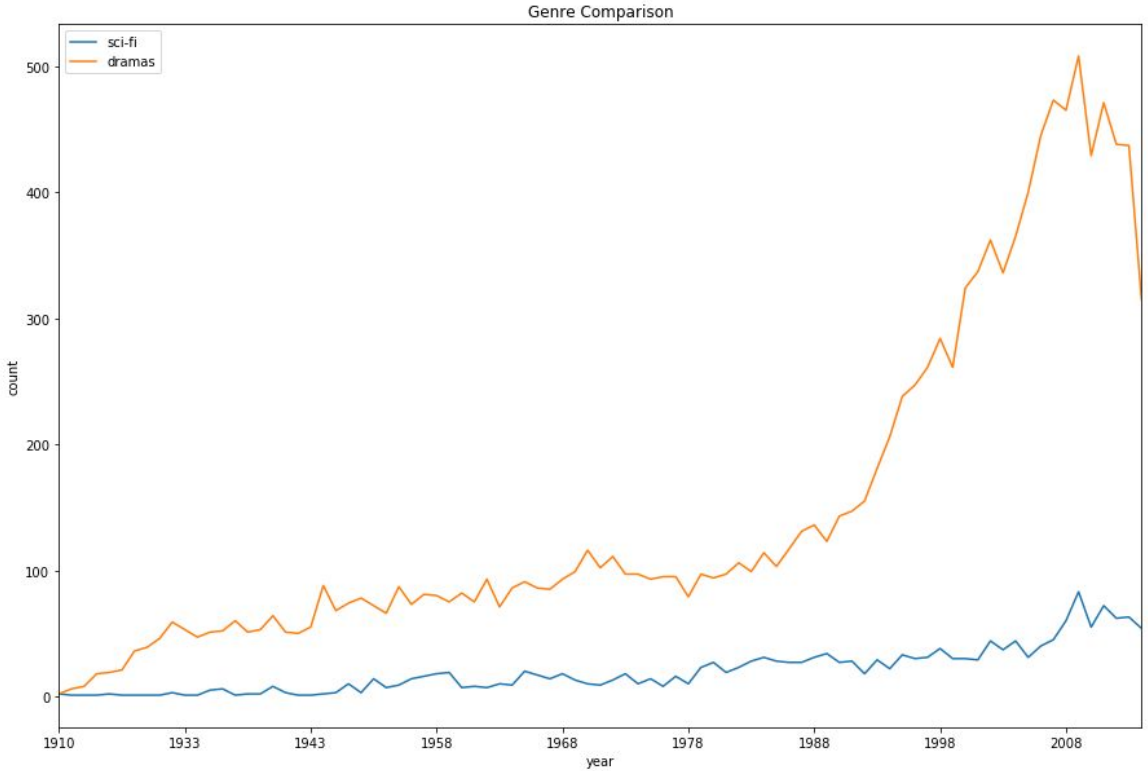
	year	count
0	1898	1
1	1902	1
2	1910	2
3	1916	1

Early years Sci-Fi Movies

	year	sci-fi	dramas
83	2011	72	471
84	2012	62	438
85	2013	63	437
86	2014	54	315
87	2015	12	46

Final years Movies

Findings - Visualisation



Findings - Conclusions

1) In the section ‘Visualisation’ I have shown that using this dataset, a plot between the highest rated genre and Sci-fi is possible over a defined year-on-year taking into account the years for movie release.

2) Genre splitting mentioned above; showed that the MovieLens dataset does not have enough detailed information on the rating for each specific genre for that movie. For a more accurate Genre Analysis, especially when investigating comparisons for a business case, it would be preferable if the movie’s associated genres are rated individually instead of being lumped together with the movie rating being used. This is the greatest limitation of this dataset when trying to compare genres.

Therefore, the goal of the questions was achieved.

Acknowledgements

As part of this mini project, background reading for the motivation section, I’d like to mention

- master’s thesis - ‘How Do Movie Producers Identify the Genre Shifting Trend?’ by Author: Xinri Fu, Xiaoyue Yao ; Jönköping May, 2010

No feedback was given from anyone else

References

[1] This work is based upon the tutorial from week four

[2] Wikipedia, List of box office bombs, (30 August 2017) retrieved 31 August 2017, from https://en.wikipedia.org/wiki/List_of_box_office_bombs

[3] Wikipedia, Collaborative filtering, (28 August 2017) retrieved 31 August 2017, from https://en.wikipedia.org/wiki/Collaborative_filtering

[4] Wikipedia, MovieLens, (4 February 2017) retrieved 31 August 2017, from https://en.wikipedia.org/wiki/MovieLens

[5] Wikipedia, Netflix Prize, (23 August 2017) retrieved 31 August 2017, from https://en.wikipedia.org/wiki/Netflix_Prize

[6] Xinri Fu, Xiaoyue Yao (2010-05-19). How Do Movie Producers Identify the Genre Shifting Trend? (Master's thesis) Jönköping University, from http://www.diva-portal.org/smash/get/diva2:352538/FULLTEXT01.pdf