New User Movie Recommendations

Capstone Project

Catherine Hocknell

BrainStation Data Science Diploma

New User Movie Recommendations

Catherine Hocknell

Business Case

Predictive models are becoming more and more prevalent within the retail industries, particularly within streaming services, with 35% of *Amazon.coms* revenue coming from recommendations, and **75% of what customers watch** on *Netflix* coming from their recommendation system [1]. Netflix executives have stated that the company saves up to **$1 billion each year** due to their recommendation system[2], therefore the need to come up with accurate predictions is essential. However, due to the nature of recommender systems being based on historical interests of the user in question, it is particularly difficult to successfully recommend an item to a new user to a service (a phenomenon known as *Cold Start*).

This project investigates three possible movie recommender systems, based on both the movie itself and historical reviews of other users, with the aim being to answer the following question:

"How might we use machine learning techniques to offer relevant movie recommendations for a new user to a streaming service?"

The following report will summarise the generation of these recommender systems (**Content-Based**, **User-Based** and **Collaborative**), and also provide a **Sentiment Analysis** of previous movie reviews.

Data Source

Rotten Tomatoes is a well-known movie review website, used by both professional film critics and regular audience members to review movies and provide a rating of either **'Rotten'** (bad) or **'Fresh'** (good). An extraction of over **1 million reviews** and the information for over **17,700 movies** from this website is available via two separate datasets on Kaggle[3], which between them contains specific details about the movies and their overall ratings, as well as the text reviews and scores provided by **over 11,000 reviewers**. This website scrape, actioned in October 2020, was generated for the purpose of investigating reviews made by professional film critics, where both datasets have a common column of the *Rotten Tomatoes* movie URL link so can be combined together. The table below indicates the potential target variables alongside a selection of the independent variables within each dataset.

	Target Variables	Independent Variables
Reviews	Review Type (Binary) Review Score (Continuous)	Movie URL, Critic Name, Review Type, Review Score, Review Content
Movies	List of Keywords (Feature Engineered)	Movie URL, Movie Title, Movie Description, Genre, Actors, Rating

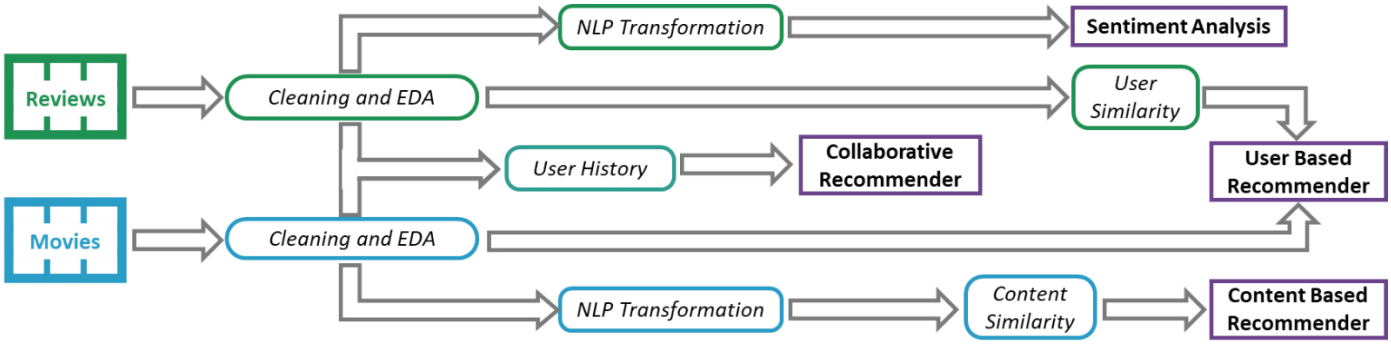
Data Wrangling

The datasets detailed above are assessed both individually and together in order to generate recommender systems and a sentiment analysis for this project. The flow of information required for each of the outputs is shown in the flow-chart on the next page, where the Sentiment Analysis and Content Based Recommender are based solely on the *Reviews* and *Movies* datasets respectively, whereas the User Based and Collaborative Recommenders utilise a combination of both datasets.

[1] **How retailers can keep up with customers**, *McKinsey* (https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers)

[2] **Use Cases of Recommendation Systems in Business – Current Applications and Methods**, *emerj* (https://emerj.com/ai-sector-overviews/use-cases-recommendation-systems/#:~:text=2%20%E2%80%93%20Netflix,giving%20them%20a%20good%20ROI.)

[3] **Rotten Tomatoes movies and critic reviews dataset**, *Kaggle* (https://www.kaggle.com/datasets/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset)



Each output shown above is heavily reliant on the data within the task-flow, therefore the quality of the datasets is imperative in order to produce the optimal end product. Various data cleaning and exploratory data analysis (EDA) steps have been carried to ensure this, as highlighted below.

General

1. Align the **Movie URLs** from both datasets so they can be joined.
2. Feature engineer a **numeric score** for each review.

Sentiment Analysis

1. Remove all reviews that are not in **English**.
2. Pre-process review **text data**, following the tasks defined on the right.

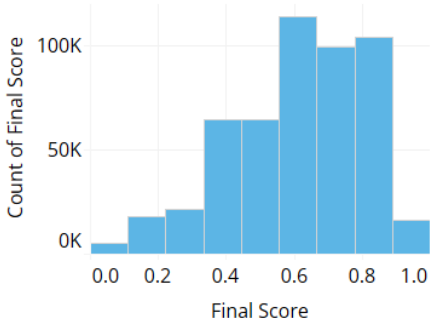
Recommender Systems

1. Generate a list of **keywords** defining each movie (Content Based).
2. Remove movies and reviews with very **low number of ratings** (User & Collaborative).
3. Evaluate historical review scores to find users with **specific genre interests** (Collaborative).

Text Pre-Processing

- Expand Contractions
- Covert to Lowercase
- Remove Punctuation
- Remove Stop-words
- Stem the words

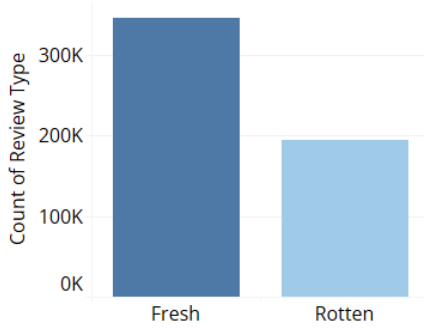
Distribution of Final Score



Following the steps carried out above, the distributions of the numeric target variables are shown here, with the database now containing approximately:

- **550,000 reviews**
- **10,000 movies**
- **2,500 reviewers**

Distribution of Review Type



Sentiment Analysis

Sentiment analysis is carried out by evaluating the text within a given review and predicting the rating associated with it, in this case either *Rotten* or *Fresh*. First, a score is given to each relevant word in the review (known as **Vectorising**), then a selection of **Classification Models** are run to find the model with the best accuracy when predicting the available test data. In this case, the **TF-IDF** vectorizer is applied on the **unbalanced data** and evaluated for **four different models**, where the accuracy metrics for each is shown below.

	Logistic Regression
AUC Score	0.869
F1 Score	0.850
Precision	0.820
Recall	0.880
Test Accuracy	0.798

	Decision Tree
AUC Score	0.674
F1 Score	0.790
Precision	0.710
Recall	0.890
Test Accuracy	0.691

	SVM
AUC Score	0.868
F1 Score	0.850
Precision	0.820
Recall	0.880
Test Accuracy	0.797

	XGBoost
AUC Score	0.860
F1 Score	0.850
Precision	0.810
Recall	0.890
Test Accuracy	0.792

The **Logistic Regression** model results in the highest overall accuracy, specifically the **AUC Score**, so this model will be used to make predictions on future reviews.

Recommender Systems

Three different recommender systems have been generated for this project task, in order to provide as much personalised information as possible to the new user. These are defined in more detail below.

Content-Based

- Keywords list generated from description, genre, directors, actors and rating.
- Matrix of keywords generated using TF-IDF vectorizer.
- Similarity between keywords evaluated.
- **Output**: Top ten most similar movies to given input movie.

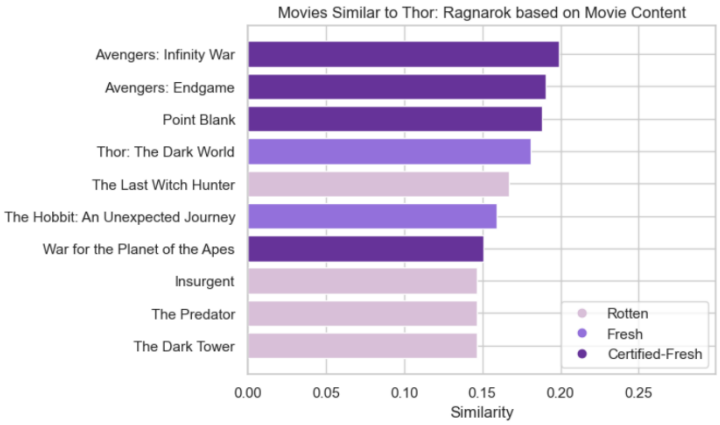
User-Based

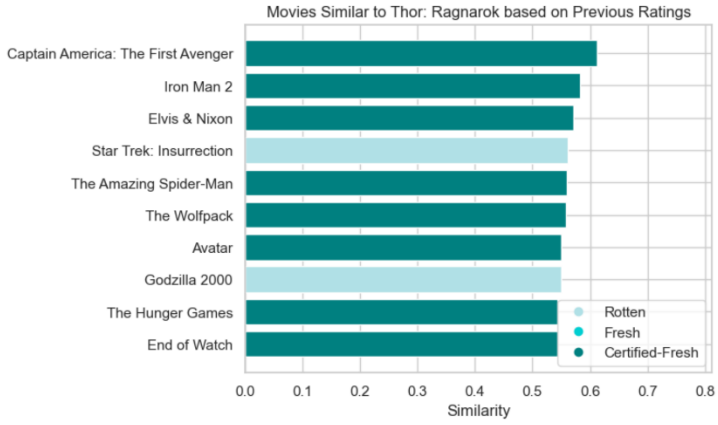
- Matrix generated from previous users and their ratings for each movie.
- Similarity between user ratings is evaluated for each movie.
- **Output**: Top ten most similarly rated movies to given input movie.

Collaborative

- Generated User and Movie matrices to calculate ratings using FunkSVD.
- Find similar user based on historic critic ratings.
- Ratings calculated for each movie for the similar user.
- **Output**: Top ten movies estimated for given user.

Both the Content-Based and User-Based recommender systems take a movie title input, and output the top ten most similar movies; the first in terms of content, the second in terms of user ratings. See how these compare below when each is asked to recommend a list of movies based on an input of *'Thor: Ragnarok'*.





This form of recommendation system does not require any previous information on a user for it to be implemented, with recommendations made purely on similarity of movie content and previous reviewer ratings. These are therefore **ideal to be used for a new user** to a streaming service.

As for the Collaborative recommender system, this requires a **User ID input** for a previous critic with similar interests to the new user, for which a list of options have been generated as part of the *Data Wrangling* phase. New users will be able to select an ID that is most similar to their interests and use this as an input to the recommender system.

An example of this in action is shown on the right, where the user *Sarah Chauncey* is considered to have the following interests:

- Likes: Romance
- Likes: Drama
- Dislikes: Horror

```
Critic Name: Sarah Chauncey
Number of Reviews: 122
Favourite Movies:  The Times of Harvey Milk, Broadcast News, Thirteen, Hotel Rwanda, Dead Poets Society, Ordinary People
Least Favourite Movies:  Taxi, The Honeymooners, My Boss's Daughter, Soul Plane, Timeline, AVP - Alien Vs. Predator
```

	Movie Title	Description	Genres	Rating	Release Date
9524	Waiting	Staffers at the restaurant Shenaniganz engage ...	Comedy	Rotten	2005-10-07
3792	Garden State	After many years away, television bit part act...	Comedy, Drama, Romance	Certified-Fresh	2004-07-28
4356	The Horse Whisperer	When teenage Grace (Scarlett Johansson) is tra...	Drama, Western, Romance	Fresh	1998-05-15
4024	The Greatest Game Ever Played	Blue-collar Francis Ouimet (Shia LaBeouf) figh...	Drama	Fresh	2005-09-30
4180	Hateship Loveship	A shy caretaker (Kristen Wiig) believes that t...	Comedy, Drama	Rotten	2014-04-11
830	Step Brothers	Brennan Huff (Will Ferrell) and Dale Doback (J...	Comedy	Rotten	2008-07-25
9323	Tyler Perry's A Fall from Grace	When a woman is indicted for murdering her hus...	Mystery & Suspense	Rotten	NaN
854	Hofshat Kaits (My Father My Lord)	The leader of a small ultra-orthodox community...	Art House & International, Drama	Fresh	2008-07-11
2275	But I'm a Cheerleader	Megan (Natasha Lyonne) considers herself a typ...	Comedy	Rotten	2000-06-23
5035	Kumaré	Lapsed Hindu Vikram Gandhi conducts an experim...	Documentary, Drama, Special Interest	Fresh	2012-06-20

Key Findings and Insights

Sentiment Analysis

- The Sentiment Analysis was able to successfully predict the type of a review to a test accuracy score of **79.8%**, meaning that the model accurately predicted the rating associated to a review on almost 80% of all test inputs.
- The words most indicative of a *Fresh* review are shown in the graphic on the right, with the opposing words for a *Rotten* review shown in a similar image below, both of which indicate words that you would expect to find in positive and negative reviews respectively.

- A number of these models were very **computationally expensive** due to the size of the dataset. This led to memory issues when optimising models, particularly for the **XGBoost**, which should be investigated further to see if the result can be improved.
- There is further potential to expand this analysis towards a **Linear Regression** model by assessing against the review score as a continuous target.

Recommender Systems

- The accuracy of a recommendation system is **difficult to quantify**, with the aim being to assess the recommended results based on the response to the output – i.e. the number of times a movie that has been recommended is then chosen to be watched. Without the ability to assess this, in the case of this project the accuracy of the recommendations is only evaluated using prior movie knowledge.
- The content-based recommender is generally good at finding related movies, however there is potential benefit in looking into each of the **keyword categories separately** and applying more weight to certain values (i.e. director or rating), as currently they seem to be highly **weighted towards actors**.
- Although largely accurate, there is a wide variety in the results coming from the user-based recommender systems, likely due to the nature of *Rotten Tomatoes* being that a large proportion of the reviews are made by **professional film critics** who write for industry publications. This inherently means they **will not have specific opinions** on genres which they like/dislike and will have rated a wide range of movies, meaning it is difficult to narrow down on any specific relationships.
- This is specifically the case for the Collaborative recommender system, where in order to utilise this method as a new user, an accurate selection of a **previous critic with similar interests** must occur in order to produce relevant recommendations. In order to harness this method of recommendation to its best ability, an improved method for finding the likes/dislikes of existing reviewers must be found, which may involve locating a new dataset of **audience reviews**, rather than critic reviews.

Next Steps

Following on from the above findings, the sentiment analysis and various recommender systems will be incorporated into an **app for new users** to be assigned a selection of recommendations based on their interests. In order to build on the user-based and collaborative recommender systems, using an **audience review dataset** is likely to improve the overall results found from these recommenders, both for user similarity and in finding an existing user with specific interests.