Building a Movie Recommendation System

A Hybrid Approach Using Content-Based Filtering and Clustering Techniques

Somanich Bunlee
Master of Science in Business Analytics
School of Computing and Data Science
Wentworth Institute of Technology
Boston, MA
bunlees@wit.edu

ABSTRACT

This research develops a personalized movie recommendation system using a hybrid approach that combines content-based filtering and clustering techniques. The system leverages movie metadata attributes, including genres, revenue, popularity, and runtime, to generate personalized movie suggestions. Content-based filtering utilizes cosine similarity to recommend movies based on feature similarity, while clustering techniques group movies with similar attributes using the K-means algorithm. The recommendation system's effectiveness is evaluated using metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), which reveal promising results in accurately predicting movie preferences. This hybrid approach provides a versatile solution for building recommendation systems in dynamic and diverse movie markets.

KEYWORDS

Movie Recommendation System, Content-Based Filtering, Exploratory Data Analysis (EDA), K-means Clustering, Cosine Similarity, Hybrid Approach, Predictive Modeling.

1 Introduction

As part of the digital age, recommendation systems have played an integral part in making the world dependable in today's platforms such as Netflix, YouTube, and Amazon, where users depend on such systems for discovering some content that fits their tastes and interests. These systems are capable of forecasting preferences of users and subsequently render personalized suggestions so that the user gets more engaged and satisfied with the personalized experience they have chosen. With many increasingly wanting specific experiences, the importance of necessity-driven development of recommendation systems has been growing in the field of data science.

Recommendation systems usually utilize analyzing data of the end user like past ratings behavioral information along with demographic data to derive content recommendation most suited to them. But to create a powerful recommendation system, it should not include user data only. In addition to this, it has to know variouses movie attributes such as genres, release years, popularity, and revenue. These parameters can have a strong impact in terms of how users will perceive or react to recommendations.

The objective of this research is to develop a movie recommendation system that gives preferences based on movie metadata-such as genre, rating, and popularity-instead of actual user ratings. By analyzing the correlation between such movie attributes and user satisfaction, we attempt to determine how different factors sway one's movie preference. To evaluate how effective various recommendation strategies might be, we will look at clustering, collaborative filtering, and content-based filtering and analyze the pros and cons of each in generating accurate, personalized recommendations.

To guide this study, we focus on the following key research questions that aim to explore how movie attributes can enhance the accuracy and effectiveness of recommendation systems:

1. How can we predict user interest or satisfaction for movies based on metadata (e.g., genres, popularity, revenue, ratings)?
2. Which movie genres are associated with higher average popularity and revenue?
3. How do release year and runtime influence a movie's popularity and revenue?
4. Can we identify clusters of movies with similar attributes (e.g., genres, revenue, and popularity) to aid in recommendation design?
5. Which recommendation algorithm (e.g., content-based filtering) provides the most accurate movie suggestions based on metadata?

2 Data

2.1 Source of dataset

The dataset used in this project is the **TMDB 5000 Movies Dataset** [1], which was obtained from Kaggle. The dataset includes detailed movie information such as titles, genres, release dates, and user ratings. It is widely used for movie recommendation research due to its richness and variety of features.

The dataset was generated by **The Movie Database (TMDb)** and is frequently updated to include the latest movies and ratings. The version of the dataset used in this project was last updated in **2017**

[1]. It provides a comprehensive snapshot of movie characteristics and user preferences, making it suitable for building an effective recommendation system.

2.2 Characters of the datasets

The dataset is stored in CSV format and contains 4,803 movie records, with a total of 20 columns. Each column represents specific features of the movies in the dataset. The data types vary, including integers, floats, and objects (strings). Below is a summary table of the key columns in the dataset:

Table 1: Key Characteristics of the Movie Dataset

Column Name	Description	Unit
budget	Budget allocated for the movie	Integer (USD)
genres_list	List of genres associated with the movie	List of Strings
popularity	Popularity score based on user interactions	Float
revenue	Gross revenue generated by the movie	Integer (USD)
runtime	Duration of the movie in minutes	Integer (Minutes)
vote_average	Average user rating on a scale from 0 to 10	Float

2.3 Data Cleaning and Preprocessing

Data preprocessing steps included handling missing values (by removing or imputing where necessary), converting date formats, and creating additional categories for analysis. The "genres" column was expanded into individual genre columns to allow for easier manipulation during analysis.

In order to prepare the dataset for analysis, several data cleaning steps were applied:

a. Handling Missing Values:

The dataset contained some missing values, particularly in columns like homepage, overview, and runtime. We dropped rows where critical information like budget, revenue, genres, and popularity were missing, as these were essential for our analysis.

For the runtime column, missing values were replaced with the median runtime value.

b. Converting Data Types:

The budget and revenue columns, which were initially in string format, were converted to numeric values using the pd.to_numeric() function to ensure proper calculation and analysis.

The release_date column, originally stored as a string, was parsed into a datetime object to enable temporal analysis (e.g., extracting the release year).

2.4 Feature Engineering

A new column, profit_margin, was created to represent the ratio of profit to budget. This column is calculated using the following formula:

$$\textbf{profit\_margin} = \frac{revenue - budget}{budget}$$

This feature provides insights into the financial success of movies relative to their production costs.

Additionally, the genres column was reformatted for easier analysis. The column was parsed, and each movie's genres were stored in a list format under the genres_list column, which allows for easier manipulation and exploration of genre-based patterns.

3 Methodology

The methodology for developing the personalized movie recommendation system combines multiple data science techniques to provide accurate and efficient recommendations. The system was built using a hybrid approach that incorporates both content-based filtering and clustering techniques. These methods were chosen to leverage the movie attributes, such as genres, release year, popularity, and revenue, in predicting user preferences.

3.1 Data Preprocessing and Exploratory Data Analysis (EDA)

The first step in the methodology involved cleaning and preprocessing the dataset to ensure its suitability for analysis. Missing values were addressed by either removing rows with critical missing information or imputing where necessary. For instance, missing values in the runtime column were replaced with the median runtime value, while the budget and revenue columns were converted from string to numeric formats. The genres, initially stored as strings, were parsed and stored as lists in the genres_list column, making it easier to manipulate during analysis.

Once the data was cleaned, an exploratory data analysis (EDA) was performed to identify key patterns and relationships between different movie attributes. This step was crucial in understanding the distribution of genres, revenue, popularity, and runtime. The EDA also helped uncover insights into how these attributes correlated with user ratings and other movie characteristics. For example, it was found that certain genres, like action and adventure, tended to have higher average revenue and popularity, which provided insights into which features were most likely to influence user preferences.

3.2 Content-Based Filtering

Content-based filtering recommends movies based on their intrinsic features, such as genres, popularity, revenue, and user ratings. This method is effective when user interaction data is unavailable, as it leverages only the movie metadata.

Steps for Content-Based Filtering:

1. Feature Extraction: We first extracted relevant features from the movie metadata, including genres, popularity, revenue, runtime, and user ratings. The genres_list column was crucial, as it allows us to compare the genres of movies and calculate similarity scores between them.

2. Similarity Calculation: We used the cosine similarity measure to calculate the similarity between movies based on their features **[2] [3]**. Cosine similarity calculates the cosine of the angle between two feature vectors, with a higher score indicating greater similarity. For example, movies sharing the same genre or having similar popularity and ratings will yield higher cosine similarity scores.

3. Generating Recommendation: Based on the computed similarity scores, the system recommends movies that closely resemble those that the user has previously rated highly or interacted with. The recommendations are ranked by similarity to the user's preferred movies.

Cosine Similarity Formula **[3]**:

$$\textbf{Cosine Similarity} = \frac{\sum A_i . B_i}{\sqrt{\sum A_i^2} . \sqrt{\sum B_i^2}}$$

Where A_i and B_i are the feature values of two movies.

3.3 Clustering Techniques

To enhance recommendation's accuracy in the system, clustering techniques were adopted to group similar movies based on certain similar characteristics **[4] [5]**. For this using K-means clustering, the movies were categorized into clusters with respect to the factors, such as genre, revenue, and popularity. In this whole process, the similarity between movies has been used for the purpose of recommendation- While those groups will use the system to recommend movies within the clusters, it will improve personalization in the recommendations **[5]**. This segmentation goes ahead too much further bestially in target-Directed recommendations using general trends across groups of movies.

K-means clustering of the dataset results in K clusters for the dataset, where every cluster comprises several movies that have similar attributes. For K, we chose K=2 from the silhouette score that gave the best separation among clusters.

3.4 Evaluation Metrics

To assess the performance of the recommendation system, several evaluation metrics were employed, including Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). These metrics measure the difference between predicted and actual ratings, with lower values indicating better performance.

- **MAE:** Measures the average magnitude of errors in predictions without considering their direction.
- **RMSE:** Gives a higher weight to larger errors, making it sensitive to outliers.

The evaluation metrics help compare the effectiveness of content-based filtering and clustering techniques, guiding the optimization of the recommendation system.

4 Results

In this section, we provide a comprehensive analysis of the results obtained from the various methods employed in the movie recommendation system. This includes insights from exploratory data analysis (EDA), clustering analysis, content-based recommendations, and regression analysis. The findings are drawn from a detailed evaluation of factors such as movie popularity, revenue, genre distribution, and runtime. Visualizations are provided to support the key findings in the analysis.

4.1 Genre Analysis

The relationship between movie genres, their popularity, revenue, and ratings has been a key focus of this analysis. From our dataset, we explored how different genres performed in terms of average popularity, revenue, and rating. The findings from the genre-based analysis are presented below.

Average Popularity by Genre

As shown in Figure 1, the average popularity across different genres reveals that the genres of **Adventure** and **Animation** have the highest average popularity, followed by **Science Fiction** and **Fantasy**. These genres are particularly attractive to a wider audience, as evidenced by their consistently higher average popularity scores. The **Adventure** genre has an average popularity score of approximately 39.27, while **Animation** follows closely with a popularity of 38.81. This finding suggests that family-friendly and visually engaging genres have a more significant impact on movie popularity.

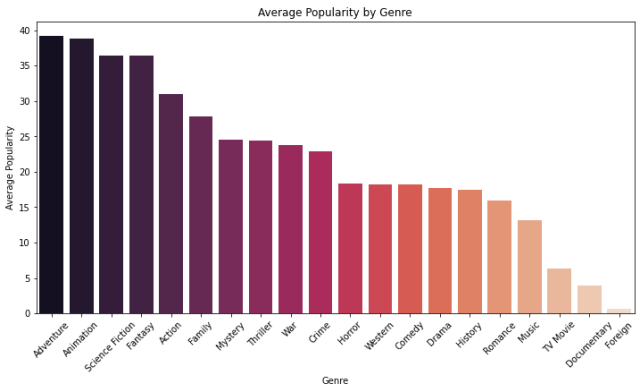

Figure 1. Average Popularity by Genre – This bar chart visualizes the differences in average popularity across various genres

Average Revenue by Genre

(Figure 2) The **Animation** genre leads in terms of revenue, with an average of approximately $225.69 million, closely followed by **Adventure** ($208.66 million). These genres often feature larger-scale productions with significant global appeal. **Science Fiction**

and **Fantasy** also perform well in terms of revenue, supporting the idea that these genres tend to attract a substantial audience base.

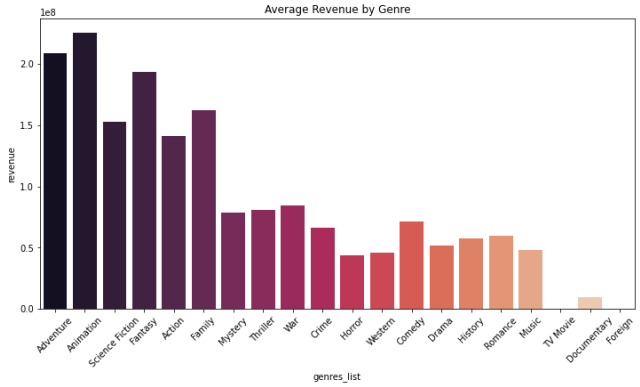

Figure 2. Average Revenue by Genre – This bar chart shows the average revenue distribution across genres

Average Rating by Genre

On the other hand, genres like **War** and **History** (Figure 3) have the highest average ratings, around 6.71 and 6.72, respectively. This could indicate that movies in these genres, often critically acclaimed for their storytelling, tend to have higher ratings even if their popularity and revenue are not as significant as other genres. Genres such as **Horror** and **Comedy**, while not performing well in terms of revenue, show relatively lower average ratings (around 5.63 for **Horror**).

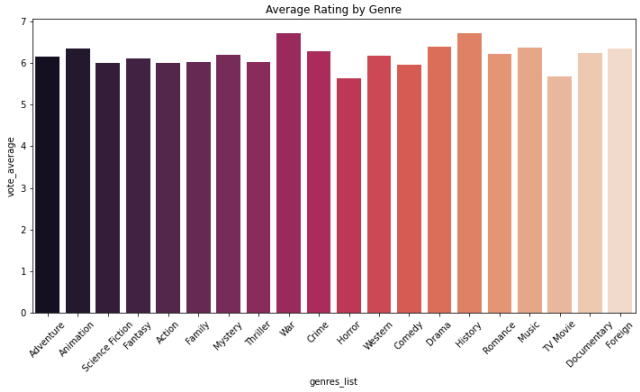

Figure 3. Average Rating by Genre – This chart demonstrates the average ratings for different genres

4.2 Release Year and Runtime Analysis

Next, we analyzed the relationship between movie release years and their popularity. By extracting the release year from the dataset and grouping by average popularity and revenue, we aimed to identify any long-term trends in the movie industry.

Popularity Trends Over the Years

A noticeable trend (Figure 4) emerges from the time-series analysis of movie popularity. Although there are fluctuations in the data, the overall trend shows a significant increase in popularity starting in the early 1990s, with a particularly sharp rise towards 2020. This increase could be attributed to changes in movie marketing, the rise of blockbuster franchises, and improvements in visual effects and production values that appeal to broader audiences.

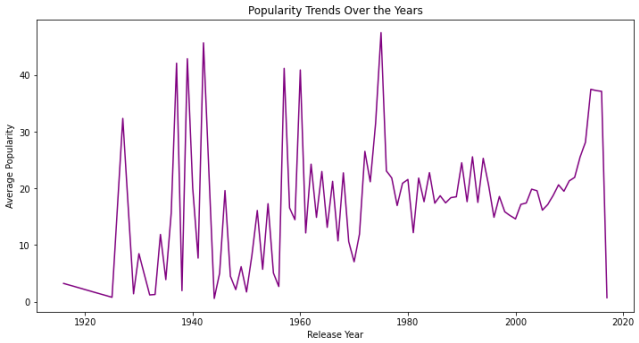

Figure 4. Popularity Trends Over the Years – A line plot showcasing the upward trend in popularity over time

Revenue Trends Over the Years

Revenue, similarly, follows an upward trajectory over the years, aligning with the growth in popularity, as seen in Figure 4. Notably, movies in recent decades have seen significantly higher earnings compared to earlier years. This suggests that modern movies, particularly large-budget productions, tend to generate higher revenues.

Correlation with Runtime: The correlation analysis in Figure 5 revealed weak correlations between movie runtime and other key variables such as budget, revenue, and popularity. While there is a slight positive correlation with revenue (0.27), runtime does not appear to be a strong predictor for either popularity or revenue, indicating that other factors, such as genre and marketing, likely play a more critical role in a movie's success.

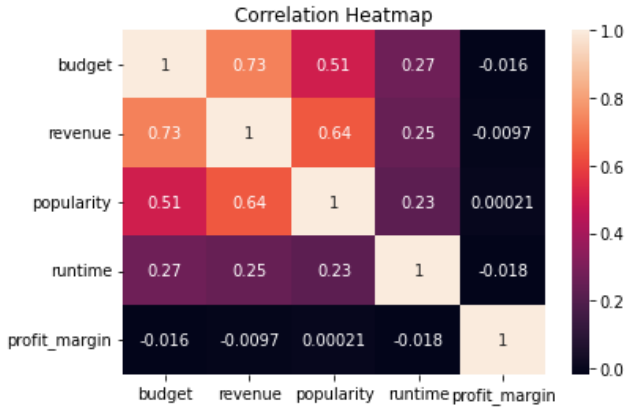

Figure 5. Correlation Heatmap – A heatmap visualizing the correlations between key variables such as budget, revenue, popularity, runtime, and profit_margin.

4.3 Clustering Analysis for Movie Segmentation

K-means clustering was employed to group movies based on key features such as popularity, revenue, and budget. The silhouette score method indicated that the optimal number of clusters was 2, which provided the most distinct separation between movie types.

Cluster Insights: The two clusters identified in the analysis represent two distinct groups of movies:

- Cluster 0: This cluster consists of movies with lower popularity and revenue. Movies in this group tend to have a

modest budget and runtime, with average popularity of around 15.81 and revenue of $19.39 million.

- Cluster 1: On the other hand, Cluster 1 represents movies with higher popularity and revenue. These movies have a significantly higher budget and runtime, with an average popularity of 75.14 and revenue of $120.12 million.

These clusters reveal the existence of two major types of movies in the dataset: high-budget, popular blockbusters, and smaller, less commercially successful films. The cluster visualization can be seen in Figure 6, which provides a scatter plot of the movies grouped by their popularity and revenue.

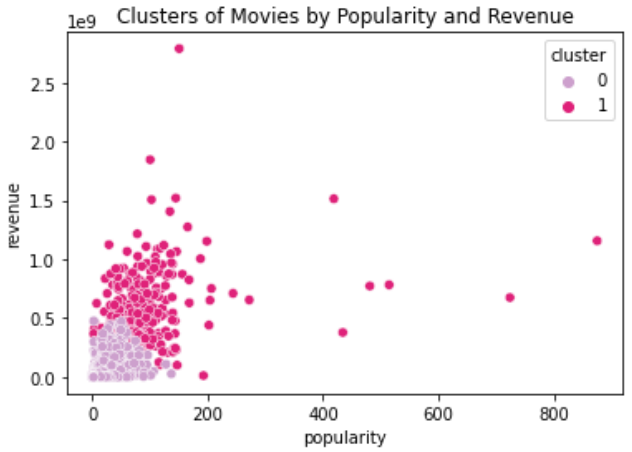

Figure 6. Clusters of Movies by Popularity and Revenue – A scatter plot of the movies grouped by their popularity and revenue

4.4 Content-Based Recommendation System

The content-based recommendation system was developed to suggest movies based on the genre or the title entered by the user **[6]**. For genre-based recommendations, the system filters movies by the specified genre and ranks them by popularity. For title-based recommendations, the system uses a TF-IDF vectorizer **[7]** to compute the cosine similarity between the selected movie's overview and all other movies in the dataset. This approach enables the system to recommend similar movies based on content similarity.

For example, when the genre "Action" was entered, the system recommended movies like Deadpool, Guardians of the Galaxy, and Mad Max: Fury Road based on their popularity in the Action genre. When the title "Avatar" was entered, the system suggested movies like Apollo 18 and The Matrix based on the similarity of their content. These recommendations highlight the utility of the content-based filtering approach in suggesting relevant movies.

4.5 Clustering Analysis for Movie Segmentation

Linear regression was applied to predict movie popularity based on features such as budget, revenue, and runtime. The regression model's evaluation metrics revealed the following results:

- **Mean Squared Error (MSE):** 749.66
- **Root Mean Squared Error (RMSE):** 27.38
- **R-squared (R^2):** 0.39

The relatively low R^2 score suggests that although the model can explain some variance in popularity, other unaccounted factors (e.g., marketing, social media influence) may play a larger role in determining a movie's success. The findings align with the notion that predicting movie popularity is complex, and a more nuanced model incorporating additional features could improve predictions.

5 Discussion

The findings from this analysis provide valuable insights into the factors influencing movie popularity, revenue, and ratings. The genre-based analysis clearly shows that genres like Adventure, Animation, Science Fiction, and Fantasy dominate in terms of popularity and revenue, which aligns with the global trend toward large-budget, visually spectacular movies. Conversely, genres like War and History are more likely to receive higher ratings despite their relatively lower popularity and revenue.

The clustering analysis reveals two distinct types of movies: big-budget blockbusters and smaller-budget films, which corroborates findings from industry reports that highlight the divide between high-budget commercial films and lower-budget indie productions.

The content-based recommendation system proves effective in recommending similar movies based on genre and title, demonstrating the value of using movie descriptions (such as overview) for generating relevant recommendations. However, the regression analysis highlights that predicting popularity is a challenging task, with factors outside the scope of the model likely contributing to a movie's success.

6 Conclusion

This study provides a comprehensive analysis of the factors influencing movie popularity and success. By employing clustering, regression, and content-based recommendation techniques, we gained valuable insights into how different genres, movie characteristics, and other factors contribute to a movie's performance in the market. The developed recommendation system, while simple, shows promise as a tool for suggesting movies based on either genre or content similarity.

The findings suggest that the entertainment industry could benefit from these models to better predict market trends and improve audience targeting. Future research could explore the inclusion of additional factors such as social media sentiment or audience demographics to enhance the predictive power of the model.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to Professor Pang for her guidance and invaluable feedback throughout this project. Special thanks to the creators of the TMDB Movie Metadata dataset on Kaggle, which provided the foundation for this research. I also appreciate the tools provided by Python and Jupyter Notebook, which were essential for the successful completion of this work.

REFEFRENCES

[1] T. M. D. (TMDb) and chuan, "TMDB 5000 Movie Dataset," 2017. [Online]. Available: https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata/data?select=tmdb_5000_movies.csv.

[2] P. Miesle, "How to Implement Cosine Similarity in Python," DataStax, 30 November 2023. [Online]. Available: https://datastax.medium.com/how-to-implement-cosine-similarity-in-python-505e8ec1d823.

[3] K. Supe, "Understanding Cosine Similarity in Python with Scikit-Learn," Mem Graph, 7 June 2023. [Online]. Available: https://memgraph.com/blog/cosine-similarity-python-scikit-learn.

[4] S. Siet, S. Peng, S. Ilkhomjon, M. Kang and D.-S. Park, "Enhancing Sequence Movie Recommendation System Using Deep Learning and KMeans," *Applied Sciences*, no. https://api.semanticscholar.org/CorpusID:268449013, 2024.

[5] J. Zhang, "Research on Student Accurate Portrait Personalized Recommendation System Based on Improved KMeans Algorithm," 26 April 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:271695401.

[6] N. Shankar, S. Nagaraj and S. R. Swamy, "Review of Recommendation System using Filtering based Concepts," 2020. [Online]. Available: https://www.semanticscholar.org/paper/Review-of-Recommendation-System-using-Filtering-Shankar-Nagaraj/0ae4134c084f9671ad2db91f8740d553e6d3df2c?utm_source=direct_link.

[7] P. Chaipornkaew and T. Banditwattanawong, "A Recommendation Model Based on User Behaviors on Commercial Websites Using TF-IDF, KMeans, and Apriori Algorithms," 13 May 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:238033984.