Building a Personalized Movie Recommendation System∗

**Subtitle**

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

ABSTRACT

Provide a short one paragraph abstract for your report..

KEYWORDS

***Recommendation System, Collaborative Filtering, Content-Based Filtering, Data Analysis, Data Visualization***

1 Introduction

In the current digital era, personalized recommendation algorithms are now a fundamental component of the user experience on websites like YouTube, Netflix, and Amazon. By assisting users in finding material that suits their interests, these systems increase user pleasure and engagement.

The goal of this project is to create a customized movie recommendation system that predicts movie preferences by using user ratings and demographic information. Investigating the effects of demographic traits (e.g., age and gender), movie information (e.g., genres, release years), and previous user ratings on user satisfaction is the aim of this study. Utilizing sophisticated analytical methods like clustering, collaborative filtering, regression analysis, and hybrid recommendation algorithms, the research seeks to assess and improve movie prediction accuracy.

This study is particularly significant in understanding user behavior across diverse demographic groups, exploring the relationships between genres and user preferences, and ultimately creating an effective recommendation system. The insights can inform improvements in content recommendation strategies, ensuring a more tailored user experience for audiences with varying tastes.

Research Questions:

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?

Provide an introduction of your topics. Make sure you include the following part. What’s your topic? Why is it important or interesting? What’s the current research/results in this area. Include necessary citation.

Example format: xxxx.

2 Data

2.1 Source of dataset

Where did you download it? Is it a credible source? When were the datasets generated? How were the datasets generated by the creator? If you create the datasets, how did you generate it?

Example: xxxx

The dataset used in this project is the **TMDB 5000 Movies Dataset**, 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**. 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

What’s the format and size of the datasets? What parameters/columns/rows/character and their units are included in this dataset. Use a table to explain this is recommended. Did you clean the data or convert any unit in the dataset? If so, what’s the formula/rule did you apply? Did you combine any datasets? If so, how do you combine them? Did you create any new category for analysis in the datasets? If so, what and how do you create?

The dataset is stored in **CSV format** and contains **5000+ movie records**. The key columns in the dataset include:

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Unit** |
| movie\_id | Unique identifier for each movie | Integer |
| title | Movie title | String |
| genres | Genres of the movie | String (list) |
| release\_year | Year the movie was released | Integer (year) |
| runtime | Duration of the movie in minutes | Integer (mins) |
| rating | Average rating based on user feedback | Float (0-10) |
| vote\_count | Number of votes received | Integer |

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.

3 Methodology

The methodology of this study involves the development and implementation of a hybrid movie recommendation system. This section discusses the algorithms used, collaborative filtering, content-based filtering, and clustering, and explains how they were integrated to create a hybrid model. Each method was chosen for its unique strengths, and together, they address the limitations of standalone approaches.

In this part, you should give an introduction of the methods/model. First, what’s the method/model. What’s the assumption of this method/model. What’s the advantage/disadvantage of this method/model. Why did you choose it. What Python module or function do you apply to apply this method/model. Any optional input/extra work did you adjust to make the results better. If you have multiple methods, feel free to use subsection 3.1, 3.2, 3.3, … to separate them.

3.1 Collaborative Filtering

Collaborative filtering is a widely used method for recommendation systems, relying on past behavior (e.g., user ratings) to predict what a user might like. In this approach, users who share similar tastes are grouped together to suggest movies based on the preferences of like-minded users.

3.2 Heading Level 2

…

Example format: The updated template, user manuals, samples, and required fonts, all are available at the URL <https://www.acm.org/publications/proceedings-template>. It contains said information for all three versions of MS Word (Windows and 2 versions of Mac). There are also separate links to the user guide, which can be referred to by the user. This URL also contains some useful video links, which describe how to add the template, structure the paper, and generate the layout, in different clips. **Display Formula with Number**

 (1)

**Continuation part of Paragraph Text** The user must style this paragraph in **ParaContinue** style, which follows immediately after the **DisplayFormula** (numbered equation). The **DisplayFormula** style is applied only in case of a numbered equation. A numbered equation always has a number to its right. Insert paragraph text here. **Display Formula without Number**



The **DisplayFormulaUnnum** style is applied only in case of an unnumbered equation. An unnumbered display equation never contains an equation number to its right, and this unique property distinguishes it from a numbered equation.



Figure 1: Figure Caption and Image above the caption [In draft mode, Image will not appear on the screen]

**Theorem/Proof/Lemma.** Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement.

....Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract.

4 Results

In this part, you need to select a reasonable way to deliver the result of your topic. For example, equation or numerical results, or visualization of your result. You also need to provide a clear explanation of all results and how to understand the results. If there exist any unexpected results, please explain why or possible cause of this special result. You can use subsection 4.1, 4.2, … to separate your results.

4.1 Heading Level 2

Example format: In the below paragraph, it is explained how alt-txt value is placed in **MS Word 2010**. To add alternative text to a picture in Word 2010, follow these steps:

1. In a Word 2010 document, insert a picture.
2. Right click on the inserted picture and select the **Format Picture** option.
3. Select the **Alt Txt** option from the left-side panel options.
4. In the "Title:" and "Description:" text boxes, type the text you want to represent the picture, and then click "Close".

Below are steps to place alt-txt value in **MS Word 2013/2016**. To add alternative text to a picture in Word 2013/2016, follow these steps:

1. In a Word 2013/2016 document, insert a picture.
2. Right click on the inserted picture and select the **Format Picture** option.
3. In the settings at the right side of the window, click on the "Layout & Properties" icon (3rd option).
4. Expand **Alt Txt** option.
5. In the "Title:" and "Description:" text boxes, type the text you want to represent the picture, and then click "Close".

*1.1.1 Heading Level 3.* Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

*1.1.1.1 Heading Level 4.*Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

5 Discussion

Every method/project has its shortage or weakness. Please discuss the unsatisfied results in your project. And discuss the feasible suggestions of future work to revise/improve your result.

6 Conclusion

In this part, you should summarize your project. What important results did you find for your topic and what’s the effect of this result on the real-world?

ACKNOWLEDGMENTS

Insert paragraph text here. Insert paragraph text here.

REFERENCES

Use the following ACM Reference format for your citation

FirstName Surname, FirstName Surname and FirstName Surname. 2018. Insert Your Title Here: Insert Subtitle Here. In *Proceedings of ACM Woodstock conference (WOODSTOCK’18). ACM, New York, NY, USA, 2 pages.* https://doi.org/10.1145/1234567890

[1] Patricia S. Abril and Robert Plant, 2007. The patent holder's dilemma: Buy, sell, or troll? *Commun. ACM* 50, 1 (Jan, 2007), 36-44. DOI: <https://doi.org/>10.1145/1188913.1188915.

[2] Sten Andler. 1979. Predicate path expressions. In *Proceedings of the 6th. ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages (POPL '79)*. ACM Press, New York, NY, 226-236. DOI:https://doi.org/10.1145/567752.567774

[3] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. DOI:https://doi.org/10.1007/3-540-09237-4.

[4] David Kosiur. 2001. *Understanding Policy-Based Networking* (2nd. ed.). Wiley, New York, NY..

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00