Enhancing Movie Recommendation Systems: A Comprehensive Analysis of IMDB Data

Tsung-Hsiang Ma, Guan-Yu Shih
University of California San Diego

Abstract

This thesis explores movie recommendation systems using IMDb's dataset. Focusing on ”would watch” prediction and rating estimation, diverse methodologies, including Jaccard similarity, Bayesian Personalized Ranking, Logistic model, and neural network, are applied. These analyses deepen our understanding of user behavior and contribute to personalized recommendation systems. The study also evaluates simple and regular latent-factor models for rating prediction, aiming to enhance accuracy. Subsequent chapters provide a nuanced analysis of each methodology's strengths and limitations, contributing valuable insights to advance our understanding of movie recommendation systems within the dynamic landscape of digital entertainment.

1 Introduction

In the dynamic world of digital entertainment, IMDb stands as a globally recognized beacon for movie enthusiasts, offering comprehensive reviews and ratings. This thesis explores the intricate domain of movie recommendation systems, utilizing IMDb's vast dataset to address two key tasks: ”would watch” prediction and rating estimation.

The ”would watch” task employs diverse methodologies, including Collaborative filtering, Bayesian Personalized Ranking, Logistic model, and Neural Network, to predict user interest in a particular movie. This predictive analysis enhances our understanding of user behavior, facilitating the development of personalized recommendation systems.

Additionally, the thesis delves into rating prediction, focusing on both simple and regular latent-factor models. By evaluating the efficacy of these models, we aim to refine the accuracy of predicting user ratings, ultimately elevating the user experience on IMDb.

Subsequent chapters provide a detailed examination of each methodology, presenting a nuanced analysis of their strengths and limitations. This research aims to contribute valuable insights, advancing our understanding of movie recommendation systems and predicting user behavior in the ever-evolving landscape of digital entertainment.

2 Literature Review

The IMDb dataset, compiled through the extraction of data from the IMDb website, encompasses a wealth of information, including movie titles, genres, release years, user ratings, and more. Widely utilized across various research domains, this dataset has become a cornerstone for tasks ranging from sentiment analysis to the development of recommendation systems.

In the realm of sentiment analysis, as illuminated by Maas et al. (2011)[0], understanding user sentiments within movie reviews takes center stage. Although the primary focus may not explicitly be on predicting movie ratings, acknowledging the sentiments expressed by users becomes paramount. This awareness contributes significantly to the broader understanding of user attitudes, thereby influencing the accuracy of predictive models within the context of movie recommendation systems.

Matrix factorization techniques, introduced by Koren et al. (2009)[0], lay the groundwork for collaborative filtering within recommendation systems. By decomposing the user-movie rating matrix, collaborative filtering emerges as a powerful tool for predicting user ratings. This method, pivotal in recommendation systems, facilitates the identification of similar users and the subsequent recommendation of movies based on their preferences. The collaborative filtering technique exemplified in this work becomes indispensable for enhancing the precision of movie rating predictions.

In a broader context, Adomavicius and Tuzhilin (2005)[0] contribute to the literature by presenting a comprehensive survey on user behavior modeling for recommender systems. Although not specific to IMDb, this survey establishes the foundational understanding of how user preferences and behaviors can be effectively modeled. This knowledge proves essential for accurately predicting movie ratings and enhancing the overall performance of recommendation systems, providing valuable insights that contribute to the development of robust predictive models.

Beyond IMDb, other datasets such as the Netflix Prize dataset, widely used since its inception, feature a vast array of ratings from Netflix users. Similarly, the MovieLens Datasets, encompassing user demographics, movie genres, and ratings, have found extensive use in prediction tasks and recommendation systems, mirroring the diverse applications of the IMDb dataset in the cinematic analytics landscape.

3 Data Analysis

This project leverages the extensive IMDb dataset [0], encompassing nearly 1 million unique movie reviews across 1,150 IMDb movies, spanning 17 genres. Genres include Action, Adventure, Animation, Biography, Comedy, Crime, Drama, Fantasy, History, Horror, Music, Mystery, Romance, Sci-Fi, Sport, Thriller, and War. The dataset further provides movie metadata, encompassing release date, runtime, IMDb rating, movie rating (e.g., PG-13, R), IMDb raters, and reviews per movie.

To optimize feature selection for calculations, we conducted a thorough analysis. This involved examining average genre ratings, an-

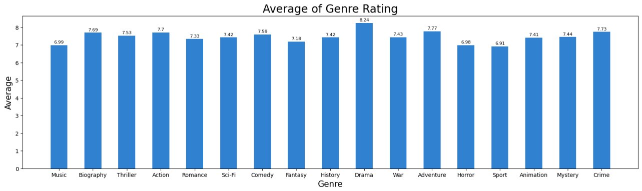

Figure 1: *Average rating of genres*

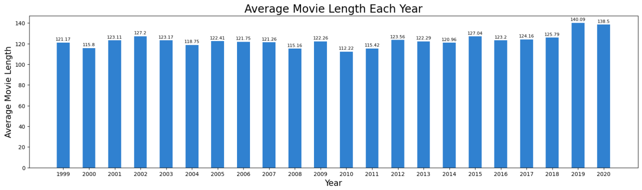

Figure 2: *Length of movie each year*

nual movie lengths, annual ratings, the number of movies reviewed by users, and users' preferences (BetaU distribution). This meticulous approach ensures the relevance and efficacy of chosen features in our analytical processes.

Analysis of average genre ratings in Figure 1 reveals noteworthy distinctions among genres. Sport emerges with the lowest rating, registering at 6.91, while drama claims the pinnacle with an impressive 8.24. This variance underscores inherent differences in audience reception across genres, with certain genres consistently garnering higher ratings while others tend towards lower evaluations. These findings substantiate the establishment of genre-specific baseline ratings, providing valuable insights into the comparative reception of different genres. Such nuanced considerations contribute to a more informed understanding of audience preferences and genre-specific expectations within the diverse landscape of cinematic appraisal.

Examining the annual trends in movie lengths depicted in Figure 2 reveals a discernible uptrend in recent years, suggesting a potential inclination among viewers for longer cinematic experiences. Moreover, juxtaposing this data with annual ratings in Figure 3 indicates a positive correlation between movie length and rating, underscoring the notion that audiences tend to appreciate longer films. However, the anomaly observed in 2020, marked by a significant drop in ratings, implies the influence of external factors on audience perceptions. This anomaly prompts further investigation into the unique circumstances of 2020 that may have impacted viewer assessments, thus enriching our understanding of the nuanced dynamics shaping film ratings.

The analysis of movies reviewed for each user, as depicted in Figure 4, unveils a distinctive user distribution pattern concerning the number of movies reviewed on IMDb. The accompanying bar graph sheds light on the prevalence of user behavior in terms of reviewing movies.

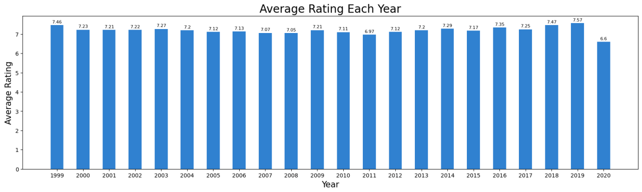

Figure 3: *Annual ratings*

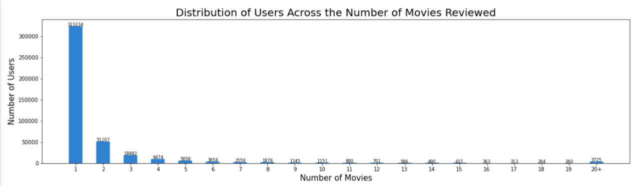

Figure 4: *Distribution of Users Across the Number of Movies Reviewed*

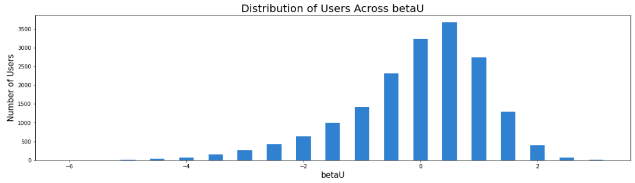

Figure 5: *Distribution of Users Across β_u*

Remarkably, a substantial majority of users, constituting nearly 80 percent, have only reviewed a single movie on IMDb. Conversely, a relatively smaller proportion of users have reviewed more than two movies. These findings underscore the potential challenges in building a recommendation system or predicting user ratings when considering the entire dataset. The prevalence of one-time reviewers poses a significant hurdle, as predicting the preferences of users who have reviewed only a single movie can lead to substantial errors. This becomes particularly pronounced when attempting to predict the preferences of users who are yet to be encountered in the dataset.

In light of these insights, it becomes evident that pre-processing steps are essential before constructing our recommendation system or engaging in user rating prediction. Addressing the unique characteristics of users who have reviewed only one movie is crucial for mitigating potential errors and enhancing the overall accuracy of the model.

The examination of user preferences involves a detailed analysis of the distribution of users across β_u, illustrated in Figure 5. β_u, representing the user-specific bias, serves as a crucial indicator, revealing each user's inclination to rate movies above or below the global average. This metric provides valuable insights into the distribution of individual users' rating preferences.

Upon scrutiny of the graph, a discernible pattern emerges. The distribution exhibits characteristics akin to a normal distribution, albeit with a slight imbalance on either side, attributable to varying rating ranges. Notably, smaller rating ranges attract a higher concentration of users, while larger ranges accommodate fewer users. Interestingly, a prevalent trend is observed where the majority of users tend to rate movies slightly above the average. This skew in preference offers a nuanced understanding of how users deviate from the global rating mean.

In essence, the graph provides a preliminary yet insightful examination into each user's rating preferences. The distribution across β_u serves as a valuable tool for gaining a holistic understanding of users' biases, laying the foundation for more refined and targeted approaches in recommendation system development and user rating prediction.

4 Predictive tasks

Utilizing the IMDb dataset, this study undertakes two predictive tasks: Would Watch Prediction and Rating Prediction. To ensure data integrity, users with fewer than three reviews—insufficient for meaningful evaluation—are identified as outliers and consequently excluded from the dataset. The data undergoes an 80-20 split into training and validation sets, respectively.

4.1 Would Watch Prediction

In the realm of Would Watch Prediction, accuracy serves as the model evaluation metric. Collaborative Filtering serves as the baseline model for comparative analysis. Bayesian Personalized Ranking and Logistic Model leverage user and movie similarities as features, while the Neural Network model incorporates genres, release year, rating, number of raters, and reviews as predictive features. Validation on the dataset ensures the robustness of these models in predicting user behavior within the dynamic cinematic landscape.

4.2 Rating Prediction

In the context of movie rating prediction, our model employs mean square error as the evaluation metric. This study leverages user, movie, and rating features to perform the prediction task. The baseline model adopts a simplistic approach by predicting ratings based on the relationships among the average rating of all movies, the rating associated with each movie, and the rating with respect to each user. While this baseline model yields satisfactory results, it treats users and items independently. To address this limitation and enhance predictive accuracy, we introduce Latent-factor models, incorporating the inherent relationship between users and movies. The use of Latent-factor models aims to provide a more refined solution to the prediction task. To assess the robustness of these models in predicting user ratings, the dataset is divided into training and testing sets, with validation conducted on the unseen portion of the dataset.

5 Model

5.1 Would Watch Prediction

Within the domain of Would Watch Prediction, an array of sophisticated methodologies, including Collaborative Filtering, Bayesian Personalized Ranking, Logistic Model, and Neural Network, were deployed to discern user inclination towards specific films. This comprehensive approach aims to accurately predict users' decisions regarding movie selection. By leveraging the amalgamation of these advanced techniques, this study seeks to unravel nuanced patterns in user behavior, providing a refined understanding of preferences and contributing to the enhancement of predictive models within the dynamic landscape of digital entertainment.

5.1.1 Collaborative filtering

Within the collaborative filtering model, Jaccard Similarity serves as the foundational metric, as outlined in the Equation 1. A nuanced approach is taken by incorporating the mean of the three most substantial similarities, coupled with the inherent popularity of a movie as the algorithm 1. This comprehensive methodology is employed to discern user inclination, enhancing the accuracy of predicting whether a user is likely to engage with a specific movie. The utilization of Jaccard Similarity in conjunction with both the mean of key similarities and movie popularity underscores the sophistication of this predictive model, contributing to a more nuanced and precise understanding of user behavior within the cinematic landscape.

$$Jaccard(U, V) = \frac{|U \cap V|}{|U \cup V|} \tag{1}$$

Algorithm 1 Collaborative filtering

```
procedure ISWATCH($sims, simTh, movie, popularTh$)
    $sims.sort()$
    $avg \leftarrow sims[: -3]$
    $nWatch \leftarrow len(usersPerMovie(movie))$
    if $avg > simTh$ or $nWatch > popularTh$ then
        return $True$                    ▷ The user would watch movie
    return $False$                  ▷ The user would not watch movie
```

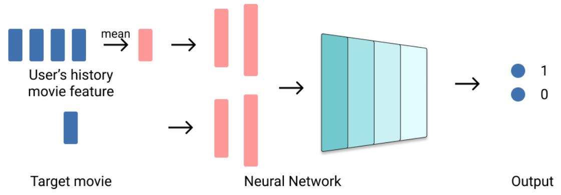

Figure 6: *Neural network*

5.1.2 Bayesian Personalized Ranking

The Bayesian Personalized Ranking is to maximize the probability of correctly predicting pairwise preferences as Equation 2

$$p(i, j) = \sigma(\gamma_u \cdot \gamma_i, \gamma_u \cdot \gamma_j) \tag{2}$$

where i and j are items, $\sigma()$ is sigmoid function,and γ_u, γ_i, and γ_j signify the latent factors for user u and item i, and user u and item j.

5.1.3 Logistic model

Maximize probability when label is positive and minimized when label is negative

$$\sigma(z) = \frac{1}{1 + e^{-z}} z = X_i \cdot \theta \tag{3}$$

where $\sigma()$ is sigmoid function, X_i is features of item i, and θ is parameters used to maximize probability.

5.1.4 Neural Network

Within the Neural Network model shown in the figure 6, a meticulously chosen set of features, including one-hot encoded genres, year of release, rating, number of raters, and number of reviews, contribute to the model's predictive capabilities. The neural network architecture incorporates two inputs: the mean of the feature vector in a user's movie history and the feature vector of the target movie. Each input undergoes individual network processing before converging into a larger network. The ultimate objective is to predict user engagement with the target movie, showcasing a sophisticated and comprehensive approach that integrates diverse features within the neural network framework for enhanced predictive accuracy.

5.2 Rating Prediction

5.2.1 Baseline Model

The foundation of our recommendation system is established through a baseline model, as inspired by the approach employed in the Netflix Prize competition. The model is formulated as

$$f(u, i) = \alpha + \beta_u + \beta_i \tag{4}$$

In this expression:

- α denotes the global average rating for all movies, serving as a fundamental baseline.
- β_u captures the user-specific bias, signifying the inclination of a user to rate movies above or below the global average.
- β_i accounts for the item-specific bias, indicating whether an item tends to receive ratings higher or lower than the global average.

This model provides an initial framework, assuming independence between users and items, and offers a baseline for subsequent enhancements.

The optimization of the baseline model is conducted through manual gradient descent, with a focus on minimizing the mean square error of the predicted results. The regularization factor (λ) is strategically selected to fine-tune the model. This approach allows us to iteratively adjust the parameters to enhance the accuracy of the baseline predictions.

5.2.2 Latent Factor Model

Building upon the baseline model, we introduce a more sophisticated latent factor model to better capture intricate user-item interactions. The augmented model is expressed as

$$f(u, i) = \alpha + \beta_u + \beta_i + \gamma_u \cdot \gamma_i \tag{5}$$

Here, γ_u and γ_i signify the latent factors for user u and item i.

This extension allows for the discovery of nuanced relationships by learning latent factors that encapsulate underlying patterns in user-item interactions. As we delve into the latent factor model, we aim to enhance the predictive capabilities of our recommendation system, surpassing the limitations of the baseline model and delivering more accurate and personalized suggestions to users.

For the implementation of the latent factor model, we leverage the Surprise library to streamline the process. However, default parameters fail to yield satisfactory results, and in fact, the predictions often fall short of the baseline model's performance. To address this, an exhaustive exploration of parameters and a rigorous validation process are undertaken.

Through systematic parameter tuning, we seek to minimize the mean square error of the predicted ratings. While the majority of parameters exhibit minimal impact on the prediction quality, certain parameters prove instrumental in achieving this goal. Key optimizations include an increase in the number of epochs and factors, meticulous search for optimal regularization factors for each term, and the integration of cross-validation techniques. These nuanced parameter adjustments collectively contribute to a substantial enhancement in the accuracy and effectiveness of the latent factor model, ultimately yielding results that surpass those of the baseline model.

6 Result and discussion

6.1 Would Watch Prediction

The Collaborative Filtering, Bayesian Personalized Ranking, Logistic Model, and Neural Network achieved respective accuracies of 0.67, 0.62, 0.70, and 0.78. The notable superiority of the Neural Network stems from its ability to incorporate rich movie features. Leveraging a user's movie history, which encapsulates genre preferences, years of favorited movies, and key popularity indicators like ratings, raters, and reviews, the Neural Network excels in comparing this information with the target movie's features.

Table 1: *Mean square error (MSE) for differnt selection of λ*

λ	**MSE**
0.1	3.9824
1	3.7225
2	3.6953
2.2	**3.6896**
2.5	3.7015
3	3.7271

Table 2: *Mean square error (MSE) for differnt selection of regularization factor (REG)*

REG	**MSE**
0.02 (default)	4.1236
0.1	3.6914
0.15	**3.5967**
0.2	3.6826
0.3	3.7125
1	4.2167

In contrast, Collaborative Filtering, Bayesian Personalized Ranking, and Logistic Model exclusively focus on user-movie relationships without considering movie features. Logistic Model, however, outperforms the former two by autonomously learning parameters from provided features. This underscores the importance of integrating movie features for enhanced predictive accuracy.

The results underscore that augmenting models with movie-specific features significantly elevates performance. The Neural Network's proficiency in discerning similarities between target movie features and a user's historical preferences highlights the value of informative features. Consequently, providing detailed movie information becomes pivotal in enhancing model capabilities and accurately predicting user engagement.

6.2 Rating Prediction

For the baseline model, a critical aspect of our optimization strategy involved the careful selection of the regularization factor (λ). Table 1 outlines the mean square error across various λ values. Following this iterative process, we identified that setting λ = 2.2 yielded the most favorable outcome, resulting in a minimal mean square error of 3.6896.

Transitioning to the latent factor model, initial experiments with default parameters proved suboptimal, with a mean square error reaching 4.1259, worse than baseline model's performance. Undeterred by this setback, a meticulous parameter tuning process was initiated to uncover the latent potential of the model.

Through systematic exploration, several key insights into parameter optimization were revealed. Notably, augmenting the number of epochs and factors, precision-tuning regularization factors for each term, and the strategic integration of cross-validation methodologies emerged as pivotal strategies. By selecting a regularization factor of 0.15, we achieved our most favorable result, with a mean square error reduced to 3.5967.

Table 2 provides a comprehensive overview of the mean square error corresponding to different parameter selections, showcasing the impact of each adjustment on the predictive accuracy of our latent factor model.

These results underscore the iterative nature of model refinement and the significance of parameter tuning in achieving optimal predictive performance. The transition from the baseline to the latent factor model not only addressed initial performance discrepancies but also substantially improved the accuracy of our user rating prediction model.

7 Conclusion

7.1 Would Watch Prediction

In conclusion, the comparative analysis of Collaborative Filtering, Bayesian Personalized Ranking, Logistic Model, and Neural Network reveals distinct performance disparities, with the Neural Network exhibiting notable superiority due to its adept incorporation of comprehensive movie features. The Neural Network leverages a user's movie history, encompassing genre preferences, historical film choices, and key popularity indicators, enabling it to excel in aligning this information with the features of the target movie.

Conversely, Collaborative Filtering, Bayesian Personalized Ranking, and Logistic Model, focusing solely on user-movie relationships, fall short in considering critical movie-specific features. Notably, the Logistic Model's superior performance suggests the efficacy of autonomously learning parameters from provided features.

This study emphasizes the pivotal role of integrating movie features for augmented predictive accuracy. The Neural Network's proficiency in recognizing similarities between target movie features and a user's historical preferences underscores the significance of informative features. In light of these findings, the provision of detailed movie information emerges as a crucial factor in enhancing model capabilities, ultimately facilitating more precise predictions of user engagement.

7.2 Rating Prediction

In this study, we embarked on the task of building and optimizing a recommendation system, drawing inspiration from the Netflix Prize competition. The journey began with the formulation of a baseline model, where manual gradient descent was employed to fine-tune parameters and minimize the mean square error of predicted ratings. This approach allowed us to establish a fundamental understanding of user-item interactions, paving the way for further refinement.

As we transitioned to the latent factor model, implemented using the Surprise library, we encountered initial challenges with default parameters. Surprisingly, the baseline model outperformed the initial latent factor model predictions. Undeterred, we delved into a systematic exploration of parameters, recognizing that default configurations were insufficient for our specific dataset.

Through a comprehensive parameter tuning process, we unearthed crucial insights into optimizing the latent factor model. Significantly, increasing the number of epochs and factors, identifying optimal regularization factors for each term, and incorporating cross-validation methodologies emerged as key contributors to improved predictions. While many parameters exhibited marginal impact, these strategic adjustments led to substantial progress in predicting user ratings.

In conclusion, our study demonstrates the iterative nature of model development and the importance of nuanced parameter tuning. By combining manual optimization for the baseline model and strategic parameter adjustments for the latent factor model, we not only overcame initial setbacks but also significantly improved the predictive capabilities of our recommendation system. The incorporation of these insights contributes to a deeper understanding of user-item interactions and enhances the practical application of recommendation systems in real-world scenarios.

References

Aditya Pal, Abhilash Barigidad, and Abhijit Mustafi. Imdb movie reviews dataset. *IEEE Dataport*, 2020.

Adomavicius, G., Tuzhilin, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering, 17(6), 734–749.*, 2005.

Koren, Y., Bell, R., Volinsky, C. Matrix factorization techniques for recommender systems. *Computer, 42(8), 30–37.*, 2009.

Maas, A. L., et al. Learning word vectors for sentiment analysis. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, 2011.