Movie Analysis: Predictions and Recommendations Project Proposal

Dharini Baskaran*
Vignesh Kumar Karthikeyan*
Sabina Adhikari*
Dharini.Baskaran@colorado.edu
VigneshKumar.Karthikeyan@colorado.edu
sabina.adhikari@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

1 INTRODUCTION AND MOTIVATION

With the abundance of data collection, data mining methods and machine learning have been handy to make predictions and recommendations, which filter out useful information relevant for users and various stakeholders. Recommendation system is a type of content filtering system, which suggests items related to search items and history of the items. Recommendation system sorts out relevant information, in the presence of overwhelming data these days. Similarly, prediction methods help to make forecasts based on historical data.

Both prediction and recommendation methods have applications in many industries like e-commerce, entertainment and banking. Digital platforms like Netflix, Prime Video, Spotify, online advertisements, shopping or job recommendation sites follow recommendation systems to make suggestions. This recommendation system could be useful for viewers to understand the choices being offered by different companies. On the other hand, recommendation methods are significant to companies to promote carefully catered products to its customers based on their previous usage and preferences. Prediction methods are appropriate and effective for any financial companies, health companies and entertainment businesses to forecast earnings or sales or number of customers. In addition, prediction method could be instrumental to identify factors or attributes needed to be modified to achieve certain goals or to support various business decision-making activities.

Our group wants to understand the underlying mechanisms of these recommendation and prediction methods. As our group loves movies and there is vast dataset on movies, we will perform data analysis on a movie dataset from Grouplens. Movie industry is a significant contributor of the global economy. This data analysis will be pertinent to make predictions on quantities like movie ratings and revenues. We will shed light on whether factors like genre, actors, ratings, date of movie release, and budget affect revenue and movie ratings. Predictions of revenues and movie ratings could be instrumental for companies, producers and investors to understand the viewers' movie choices and to choose actors or genres or release dates for successful movies. On the other hand, as our group does not have much prior experience on recommendation systems, this project will be an opportunity to learn about recommendation algorithms and understand the strategies employed by digital platforms.

*All three authors contributed equally to this work.

2 LITERATURE REVIEW

There have been many prior works done on both recommendation systems and prediction methods. Recommendation systems were first introduced in the 1990s. The concept of **Collaborative Filtering** was introduced in 1992 which was experimentally applied to personal emails and information filtering [3]. Personal recommendations are present everywhere, leading to growing interest in exploration of different recommendation systems and their effectiveness.

In [1], the authors have realised the bias and unfairness in existing recommendation systems, They focused on their paper to find the anomaly's origin. And this is achieved by **Soft Matrix Factorization (SoftMF)** on MovieLens dataset, which tries to balance the predictions of different types of users to reduce the present inequality.

In [4], the authors review various recommendation systems like **collaborative filtering, content-based filtering, context-based filtering** and **hybrid filtering**. The authors also present various machine learning algorithms like K-Means Clustering and Principal Component Analysis and measure the model accuracy.

In [6], the authors use the collaborative filtering with three different user similarity measures: **Cosine similarity, Correlation based similiarity** and **Euclidian similarity** to predict ratings of various movies.

In [9], the authors introduce the novel **k-clique** method on social networks to improve the efficiency of collaborative filtering.

In [8], the authors have discussed various existing methods of recommendations system in current practice. The paper discusses on improving the recommendation system's performance and agility through collaborative filtering method. To achieve this, **K-means, Content-Based recommendation** and **SVD** methods were deployed. They calculated mean and cross validation metrics to evaluate and show that their approach indeed results in increased performance.

There have been ample number of studies done on prediction methods too. In [7], the authors conduct performance of seven different machine learning methods to predict profit value of movies and conclude that **Multilayer Perceptron Neural Network** gives the best output.

In [5], the authors propose the **Support Vector Method (SVM)**-based machine learning method to use economic factors to predict movie box-revenues of China and the US. They also compare the

SVM method with random forest based and neural network based machine learning method.

The paper [2], tries to find whether multi-model or single-model prediction system yields better results. This is tackled because the revenue prediction on box office have always shown conflicting results as different data is used. The main sources are either movie reviews or metadata. This experiment proves that using metadata alone, we can predict the box office revenue. This is done utilizing **EM(Expectation Maximization)** algorithm.

Hence, both recommendation methods and prediction methods are widely used to analyse the movie datasets and many other applications. Recently many works on comparison of different methods with various modifications are being explored to deal with issues like size and sparsity of datasets, and efficiency of different algorithms.

3 PROPOSED STUDY

For our study, we will use two different datasets. First one is the small MovieLens dataset with 100000 ratings and 3600 tag applications applied to 9000 movies by 600 users between 1996 and 2018, available on https://grouplens.org/datasets/movielens/. Second we will use *The Movie Dataset* available on https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset.

With the MovieLens dataset, we will first preprocess the dataset to remove any redundancies, identify missing attributes, and maintain consistency. We will then perform statistical analysis to find any correlations or associations between variables like tags, ratings and timestamps. We will play with visualizations to present results like ratings distribution by genre, tags or correlation between release dates and ratings. This dataset will also be useful to see the movie trend throughout these years. We will then use item-based collaborative filtering with k nearest neighbors algorithm with cosine similarity metric (*cosine similarity between A and B is given by* $S(A, B) := \cos\theta = \frac{\mathbf{AB}}{\|A\|\|B\|}$) to develop a recommendation system.

We realize that collaborative algorithm does not account for users' preferences. To address this issue, we will use the content filtering to develop a recommendation system according to users preferences.

The MovieLens dataset does not contain information about many features like budget, release dates and more, which are crucial for the prediction method. For this reason, we will use *The Movie Dataset* from *Kaggle* to predict movie revenue and ratings. As this dataset is extensive and includes many features, we may reduce the dimension by using a subset of attributes. Similar to the previous dataset, it will be interesting to perform statistical analysis to see any correlations between attributes like ratings and opening weekend grosses, or analyse movie ratings based on gender or language or budgets or genre. We think that we will be able to get some nice visualizations to represent key statistical properties of this dataset. We will then develop a multi-variable regression model to predict ratings and revenues.

4 EVALUATION

We will train our models on a subset of datasets and evaluate them on the remaning dataset. To evaluate recommendation system, we consider:

(1) Precision and Recall Method:
To calculate both precision and recall measures, we first get the count of True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) for our recommendations. Then we calculate Precision to find the proportion of correct positive identification:

$$\text{Precision} = \frac{TP}{TP + FP}.$$

We calculate Recall to find the proportion of actual positives that were correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN}.$$

Given the big data size, this method should work well for the recommendation methods.

To evaluate the prediction methods, we consider two different methods:

(1) Root Mean Squared Error (RMSE): Let y_i be the actual rating or revenue and $\hat{y_i}$ be the predicted rating or revenue. We use the formula

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y_i})}{n}},$$

where n is the total number of predictions. This method will ensure to assign different weights to outliers and non-outliers predictions. The decreasing value of RMSE implies improving performance of our model. Nevertheless, these values are not intuitive on its own. So, we will also calculate the R^2 statistic.

(2) R^2 statistic:
This measures the goodness of fit of our model. Let y_i be the actual rating or revenue, $\hat{y_i}$ be the predicted rating or revenue and $\hat{y}$ be the average of the predicted values. We use the formula

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y_i} - y_i)^2}{\sum_{i=1}^{n}(\hat{y} - y_i)^2},$$

where n is the total number of predictions. Higher value of R^2 implies good fit for our data.

5 MILESTONES

Here are the milestones we are aiming to achieve:

(1) October 10 - Understanding and preprocessing the data.
(2) October 24 - Perform statistical analysis of both datasets and evaluate if our proposed methods are suitable. Create visualizations to present statistical properties.
(3) October 31 - Project Progress Report/ Discuss with Ravi.
(4) November 16 - Develop and implement the recommendation model and regression model.
(5) November 28 - Evaluation of the methods implemented.
(6) December 7 - Paper write up and presentation completed.

REFERENCES

[1] Álvaro González, Fernando Ortega, Diego Pérez-López, and Santiago Alonso. 2022. Bias and Unfairness of Collaborative Filtering Based Recommender Systems in MovieLens Dataset. *IEEE Access* 10 (2022), 68429–68439. https://doi.org/10.1109/ACCESS.2022.3186719

[2] Guijia He and Soowon Lee. 2015. Multi-model or Single Model? A Study of Movie Box-Office Revenue Prediction. (2015), 321–325. https://doi.org/10.1109/CIT/IUCC/DASC/PICOM.2015.46

[3] Dietmar Jannach, Pearl Pu, Francesco Ricci, and Markus Zanker. 2021. Recommender Systems: Past, Present, Future. *AI Magazine* 42 (2021), 3–6. Issue 3.

[4] Sambandam Jayalakshmi, Narayanan Ganesh, Robert Cep, and Janakiraman Senthil Murugan. 2022. Movie Recommender Systems: Concepts, Methods, Challenges, and Future Directions. *Sensors* (2022). https://www.mdpi.com/1424-8220/22/13/4904

[5] Dawei Li and Zhi-Ping Liu. 2022. Predicting Box-Office Markets with Machine Learning Methods. *Entropy (Basel, Switzerland)* 24 (2022). Issue 5. https://doi.org/10.3390/e24050711

[6] Rahul Pradhan, Ashish Chandra Swami, Akash Saxena, and Vikram Rajpoot. 2021. A Study on Movie Recommendations using Collaborative Filtering. *IOP Conf. Series: Materials Science and Engineering* (2021). https://iopscience.iop.org/article/10.1088/1757-899X/1119/1/012018/pdf

[7] Nahid Quader, Md. Osman Gani, and Dipankar Chaki. 2017. Performance evaluation of seven machine learning classification techniques for movie box office success prediction. *2017 3rd International Conference on Electrical Information and Communication Technology (EICT)* (2017), 1–6. https://api.semanticscholar.org/CorpusID:25140665

[8] Mojtaba Sadeghian and Mohammad Khansari. 2018. A Recommender Systems Based on Similarity Networks: MovieLens Case Study. (2018), 705–709. https://doi.org/10.1109/ISTEL.2018.8661141

[9] Phonexay Vilakone, Doo-Soon Park, Khamphaphone Xinchang, and Fei Hao. 2018. An Efficient movie recommendation algorithm based on improved k-clique. *Human-centric Computing and Information Sciences* (2018). https://doi.org/10.1186/s13673-018-0161-6