**UNDERSTANDING MOVIE PREFERENCES THROUGH RECOMMENDER SYSTEMS**

**TABLE OF CONTENTS**

1. Abstract
2. Introduction
3. Business understanding
4. Problem Statement
5. Proposed Questions
6. Exploratory Data Analysis
7. Data Preparation
8. Modeling
9. Evaluation
10. Conclusion
11. References

# **Abstract**

This project will draw data from the MovieLens dataset to develop insights and predictions on user ratings. Using a little bit of exploratory data analysis (EDA), we try to extract the underlying patterns in movie ratings and user behaviors, finally using some machine learning techniques. All these: the movie genre, movie ratings, and user-assigned tags; together with each other in one rich dataset, provided a fertile ground for investigating genre preferences, temporal rating trends, and the possibility of predictive modeling.

During the machine learning phase, different predictive models for ratings were tried out, among which the features derived from the dataset include Linear Regression and Random Forest Regressor models. Even though Random Forest was a much more complex model than linear regression, it didn't show great improvement over linear regression. This tells us that feature engineering and model selection are key in any predictive task. The results imply that user ratings are affected by factors such as movie genres and release years. Predicting subjective movie preferences, therefore, accurately turns out to still be a challenging task. This article seeks to fill in this knowledge gap with the dynamics of movie ratings and illuminates the potential power that machine learning may have in building up recommender systems.

# **Introduction**

User behavior in the digital age has become of utmost importance to businesses that are trying to customize their offers and services according to individual preferences. The entertainment business, especially the movie industry, is no exemption. Now, with more and more content being added, the only trick becomes not only producing quality content but rather ensuring that quality content finds its way to the right audience. This problem has pushed the rise and evolution of recommender systems, which essentially represent sophisticated algorithms that have been designed to predict and suggest content that a user is likely to enjoy based on past behavior and preferences.

The more digital streaming platforms emerge, the more important recommendation systems become. Their amount may rise to such huge sizes that it would just be impracticable for any user to look through all of them. Effective recommender systems guide the wide sea of alternatives, adding even more user experience and involvement by filtering for the customers those films that fit better their likings. As such, these systems have become a staple in user retention and satisfaction tactics at digital entertainment platforms.

It tries to use a large dataset of movie preferences, called MovieLens, which is very popular among researchers in the field of recommender systems. The MovieLens dataset includes complete data on genre ratings, user-assigned tags, and other metadata; it reflects a complete snapshot of all user interactions and preferences in the data.

**Business Understanding**

User preferences are important in the digital economy. In entertainment, giving the right content matters. With many streaming sites, good recommendation systems help. To forecast likes, they do complex arithmetic. Based on user selections, they recommend material. It helps the user experience by identifying suitable matches. Thus, they contribute significantly to users' satisfaction with platforms. It helps to use data such as MovieLens. It has details on preferences, tags, and info. This lets businesses understand user behavior and likes.

Evaluating data through machine learning could be used to enhance recommender systems. This will help in increasing user satisfaction and engagement with digital entertainment. We must evaluate these datasets to be able to use current methods well. In fact, the accuracy of recommender systems improves with such an act. Better customer experiences are realized by enhancing these systems. Usefulness of recommendations is directly proportional to customer happiness. Additionally, personalization of digital entertainment alternatives enhances engagement too. The development of more precise recommenders would be advisable

**Problem Statement**

Discovering people's favorite film and their ranking has been a significant issue. This was done by making use of the MovieLens dataset in this research. At present, there is a large amount of digital material available online and most of it comprises movies. The problem lies with producing quality movies because one needs to ensure that they reach the right audience too. That is why recommender systems were designed; these are smart computer algorithms that propose films for you based on what you have liked previously as well as watched before. However, given that movie preferences differ greatly from person to person, it makes it very difficult to find out what someone really likes when it comes to watching films.

It is evident from trying many different predictive models ᅳ either Random Forest or Linear Regression ᅳ that feature engineering and model selection matter greatly in the improvement of prediction accuracy. The conclusion affirms the toughness of the task by implying that the rating results are affected by certain elements such as movie genres and release years. Therefore, the issue's statement is defined as the essence of the forecasting problems with respect to movie ratings and the possibility of machine learning tools overcoming these difficulties.

# **Methodology**

In the age of digital content, every businessman who wants to continue mesmerizing and retaining their audience needs to know what choices its consumers would be more drawn to. This becomes even more meaningful in the entertainment industry, where too much choice abounds but remains unnecessary to become a victim of choice confusion. At the background of these advancements, recommender systems have come up as a beacon of personalization, bringing the user by hand into the digital content landscape until movies that suit his taste are found. This project, therefore, sought to lay bare movie preferences and predict user ratings through a mix of exploratory data analysis (EDA) techniques and machine learning using the MovieLens dataset.

The MovieLens dataset itself is a real treasury of user interactions and movie ratings, forming the bedrock of our investigation. Comprising around 100,000 ratings from 700 users for about 9,000 movies, the dataset is a mine to study user engagement patterns along with movie genre metadata and user-assigned tags. The supplied dataset is further enriched through the inclusion of links to IMDb and TMDb. Such varied data, as presented for this study, set the stage for quite an extensive exploration into the tendencies of movie patrons.

Our approach began with an EDA to look for trends in our dataset. Analyzing the preferences by genre, however, we could observe a bias for the Film-Noir, Drama, and Crime genres, which altogether got the highest average rating. Such insights into the most popular genres of movies perhaps were indicative of the preferences by the users, and such an analysis reflects possible bias for certain types of movies among the masses. A temporal rating analysis for this movie has, however, shown excellent evidence of ratings remaining very consistent across time with visible volume changes. These will be visualized using bar plots and line graphs to be able to lay the ground that helps understand the intricate relation that movie characteristics and user ratings have.

Based on the results that we have gotten from EDA, we have tried to make predictive modeling with two different machine learning algorithms: Linear Regression and Random Forest Regressor. To implement a simple and interpretable baseline, the model used was Linear Regression, and the other was the Random Forest model, which brought about the capturing of complexity and robustness for the nonlinear relationships between features and their interactions. Performance is able to be measured with the use of the metrics, such as the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) that give a quantitative measure of accuracy in the prediction (Calasan et al., 2020).

The models' performance, one versus the other, was slightly better than that of a simple Linear Regression model and was quite a surprise, the more complex one compared to the Random Forest Regressor. This result has put emphasis on model selection in that a simpler model, under certain circumstances, will offer predictive performance equivalent to or even superior to the larger model. The advent of machine learning techniques and other sources through this study on exploration threw some lights on the various challenges and opportunities in predicting individual preferences but illuminated causes of movie rating.

So, the concluding remark to this study would be that it has unraveled many of the complexities of movie preferences found within the MovieLens dataset and the ability of machine learning to improve recommender systems. Model user behavior with the use of exploratory and modeling techniques to give insight into user behavior that would generally be understood for content personalization. As we traverse the cosmos of digital content, the research output will serve as a guiding light for scholars and practitioners intending to embark on further studies and development in personalizing the entertainment experience.

**Data Understanding**

The primary aim of the project is to delve into people’s movie preferences through an examination of the MovieLens dataset. This dataset was obtained from the website

<https://grouplens.org/datasets/movielens/25m>. The dataset consists of 10 columns; title, genres, userId, movieId, tag, rating, relevance, timestamp, imdbId and tmdbId. It also contains approximately 100,000 ratings by 700 users for 9,000 movies. This makes it the most appropriate sparse resource for the analysis.

Below is the description of each column in the Movie Lens dataset:

1)Title: The title of the movie.

2)Genres: The genre associated with the movie which categorizes the content of the film, such as action, drama, comedy, etc.

3)UserId: A unique identifier for each user who provided a rating or a tag.

4)MovieId: A unique identifier for each movie in the dataset.

5)Tag: Descriptive phrases assigned to a movie by users to capture specific aspects or themes.

6)Rating: Ratings usually vary between 1 and 5 with scores suggesting a positive viewpoint.

7)Relevance: A measure of how relevant a tag is to a movie. This column might provide additional information about the importance or significance of user-assigned tags.

8)Timestamp: The timestamp indicating when a user provided a rating or tag. It represents the date and time of the user's interaction with the movie.

9)ImdbId: The unique identifier assigned to a movie on the Internet Movie Database (IMDb).

10)TmdbId: The unique identifier assigned to a movie on The Movie Database (TMDb).

These columns collectively provide information about user interactions, movie details, and

additional context that can be utilized for analyzing people's preferences and behavior in the

context of movie ratings and tags.

**Exploratory Data Analysis (EDA)**

Any data set to be mined for exploration to unveil hidden patterns and dynamics should be explored using Exploratory Data Analysis (EDA). In the context of movie ratings and preferences, EDA encapsulates critical views on several aspects, which include rating distributions, and time-sensitive popularity of genres, among others, in lieu of getting informative insights related to user behaviors and involvements. Exploring the distribution of ratings begins by examining how users rate movies across the dataset.

Aggregating average ratings by genre enables analysts to pinpoint those genres with the highest average rating or those that generally attract more user activity. Such analysis would be relevant to an understanding of preferences and trends within genres, so that it is enabled for creators and distributors of content to adjust their offerings in such a way that it most fittingly allows for the reaching, or running together, with consumer interests. Temporal trends also play a crucial role in EDA, where analysts examine how movie ratings evolve over time.

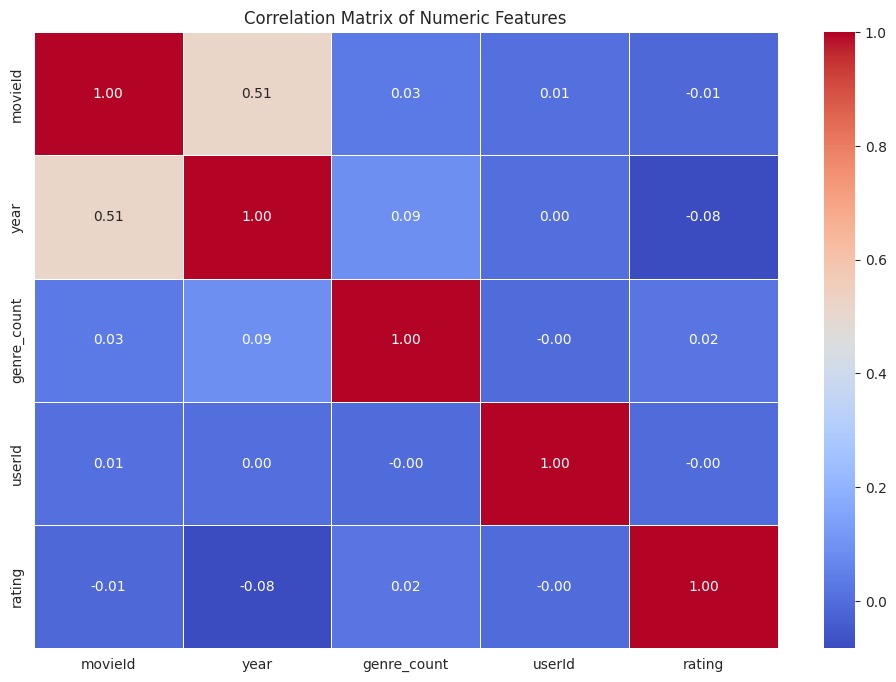
This enables an analyst to graph the ratings against the chronological timestamps in order to understand the seasons that could be influencing the changes, genre popularity over time, or even cultural events that change patterns in consumption of the movies. This will turn into an understanding of temporal trends, providing context in predicting future preferences and understanding context in terms of content strategies to be prepared suitably.

DATA PREPROCESSING

* For Data Preprocessing we have done the following
  + The dataset was checked whether it has null values or not and there were some null values present so we had handled this by dropping the rows.
  + We renamed the column year\_x to year for and dropped the redundant year\_y column.
  + Through the use of a common column, we have combined our datasets to determine the exact details of each dataset. So we have combined the datasets movies and ratings with common movie\_id column.
  + Then we have performed univariate analysis by plotting a bar plot for variables ratings and genres in order to find the average ratings within each genre individually.
  + Also a univariate analysis was done by plotting a line graph for variables tag and year in order to examine the tag usage over time.
  + Then a univariate analysis was done for the variables ratings and year. But here two lineplots were plotted, one plot is for count of ratings and the other is for average ratings over time.
  + Also a bivariate analysis was done for movie genres and their average ratings in order to know the comparison between different genres and their rating distributions.
  + The movie dataset was preprocessed by extracting the release year from the movie titles and counting the number of genres associated with each movie. This allows us to incorporate temporal and categorical information into our analysis.
  + We simplified the genre column by keeping only the first genre, assuming it represents the primary genre of the movie.
  + Then, we perform one-hot encoding on this simplified genres column. One-hot encoding converts categorical data into numerical format, creating binary columns for each category (genre in this case).

CORRELATION MATRIX

In order to understand the features better we have done correlation analysis. And also for feature selection we have used this analysis. From this we have taken five input features for the model training they are rating, user\_id, genre\_count, year, movie\_id.



MODEL TRAINING

* We have standardized the numerical features in the training and testing datasets as it helps the algorithms perform better and converge faster by ensuring that all features have similar scales.
* We split the dataset into training, validation, and test sets. The training set is used to train the machine learning model, the validation set is used for hyperparameter tuning, and the test set is used for final model evaluation. The split ratio in the dataset is 80% for training and 20% for testing the dataset.

**Predictive Modelling**

The implementation of the machine learning algorithms is realized in the predictive modeling phase, where it forecasts movie ratings regarding some features. In this case, feature selection is very crucial, for it seeks to identify the predictors or relevant predictors, such as genre labels, movie metadata, and user demographics, and contextual factors, respectively. Feature selection techniques help prioritize the most informative predictors for model training.

After one has selected the features, analysts usually deploy machine learning algorithms such as regression or classification models, to predict movie ratings. The most common to work with is linear regression, decision trees, random forests, support vector machines, or even neural networks, depending on the point of the issue. Models are heavily validated, involving suitable performance metrics such as mean squared error (MSE) or root mean squared error (RMSE) in quantifying predictive accuracy (Hodson, 2022).

So we have Linear Regression and Random Forest Regression. Linear regression was selected as it is suitable for predicting movie ratings/preferences as it provides a linear relationship between features and the target variable (rating). Random Forest Regression was selected as it is Useful for capturing complex patterns in movie preferences, such as interactions between various features like genres etc. Also, both linear regression and random forest regression are scalable and can efficiently handle large movie preference datasets.

**Proposed questions**

1)What are the most popular movie genres among users in the dataset?

2)How do movie ratings vary across different genres?

3)Are there any trends or patterns in the usage of tags over time?

4)Are there temporal patterns in movie ratings?

5) Can we predict movie ratings based on user characteristics and movie features?

**Analysis for question -1:**

**Most popular movie genres:**

* Initially, we have merged the movies dataset with the ratings dataset and have created a unified dataset that includes both movie information and user ratings.Then we have calculated the average rating for each genre so that we can analyze the distribution of ratings across different genres which provides insights into user preferences and helps in identifying popular genres.

**Visualization for question-1:**

**Bar plot of Average Ratings by Genre:**

* Visualizing average movie ratings by genre helps understand user preferences much better. A Bar plot is used to display average ratings for each genre..

A graph showing the average rating of a product

Description automatically generated with medium confidence

**Conclusion for question-1 based on analysis:**

* From the above bar plot, we can conclude that the “film-noir" genre has most average rating compared to remaining genres.

**Analysis for question -2:**

**Movie ratings analysis:**

We have selected a list of genres called Drama, Comedy, Action, Sci-Fi, and Romance.

Then we implemented a DataFrame that filters rating based on genre and named it as “ratings\_stats\_by\_genre” to only include data for the selected genres. This DataFrame likely contains statistical information about ratings grouped by genre, such as mean rating, median rating, etc.,

**Visualization for question-2:**

A box plot is visualized graphically depicting groups of numerical data through their quartiles. In this case, it's used to show the distribution of average ratings for each genre.The x-axis represents the genres, and the y-axis represents the average rating.

A blank form with black lines

Description automatically generated

**Conclusion for question-2 based on analysis:**

From the above box plot, we analyze how do ratings vary across different selected genres of Drama', 'Comedy', 'Action', 'Sci-Fi', 'Romance' respectively and drama genre has the highest average rating than all.

**Analysis for question -3:**

Initially, we converted the timestamp column in the DataFrame “tags\_df” to datetime format. In the next step, we extracted the year component from the timestamp column and created a new column named 'year' to store these values. Thenwe grouped the data by year and tag and counted the occurrence of each tag within each year. We have selected the top 10 tags based on their total counts across all years.

**Visualization for question-3:**

The below line plot is used to visualize tag usage trends over time for the top 10 tags from a dataset.

A graph of different colored lines

Description automatically generated

**Conclusion for question-3 based on analysis:**

Sci-fi appears to be the most popular tag overall, with its usage count steadily increasing over time.

**Analysis for question -4:**

**Temporal patterns Analysis of Ratings:**

We have identified any temporal patterns in user behavior and aggregate ratings annually, providing insights into how ratings have changed over time. We have plotted 2-line plots as shown below to analyze temporal patterns.

**Visualization for question-4:**

A graph with a line going up

Description automatically generated

A green line graph with numbers

Description automatically generated

**Conclusion for question-4 based on analysis:**

**Change in trends in first line plot:**From 1995-2000: The trend increased initially, and then decreased and then finally increased till the year 2000.

From 2015-2020: The trend has initially increased and later decreased at end.

**Change in trends in second line plot:**

From 1995-2000: The trend has initially decreased ,and then increased and then finally decreased till the year 2000.

From 2015-2020: The trend has continuous declines and inclines and ended with decline till the year 2000.

**Analysis for question -5:**

**Visualization using linear regression model:**

A graph with a red line

Description automatically generated

**Visualization using random forest regression model:**

A graph with a green line

Description automatically generated

Conclusion:

* In the graph of the linear regression model, the actual rating distribution appears to be represented by the blue bars at the graph's bottom. About 3.5 to 4.0 is where most real ratings are concentrated.
* In the graph of Random Forest regression mode, the blue bars indicate that actual ratings are centered on the higher values, primarily in the range of 3.5 to 4.5. In contrast to the prior mode, the Random Forest model demonstrates a significant positive relationship between the predicted and actual ratings.

**Model Evaluation:**

**Comparison of Linear Regression model vs Random Forest Regression model:**

A comparison is made between the Random Forest Regression model and the Linear Regression model as shown below:

The Linear regression model was evaluated using the following performance metrics:

Mean Squared Error (MSE): 1.08

Root Mean Squared Error (RMSE): 1.04

Mean Absolute Error (MAE): 0.82

The Random forest regression model was evaluated using the following performance metrics:

Mean Squared Error (MSE): 1.02

Root Mean Squared Error (RMSE): 1.01

Mean Absolute Error (MAE): 0.79

A graph of a performance comparison

Description automatically generated

In the the above bar plot:

* Linear Regression model's higher bars indicate higher values for MSE, RMSE, and MAE compared to RF.
* Random forest regression's lower bars indicate lower values for MSE, RMSE, and MAE compared to LR.
* Therefore, based on this comparison, Random Forest Regression is outperforming Linear Regression as it has Lower values for MSE, RMSE, and MAE which indicate better performance in regression tasks as they reflect how close the predicted values are to the actual values.

# **Results**

This exploratory expedition of the MovieLens dataset gave a rather nuanced understanding of user movie preferences underlined by some patterns in genre popularity and temporal consistency in ratings. Some of the interesting inferences that the initial findings from the Exploratory Data Analysis (EDA) threw up were: a strong bias toward certain genres since, on average, the film genres of Film-Noir, Drama, Crime were topping the charts, which more told about the user preferences towards movies with complex narratives or engaging storytelling. Intriguingly, movie genres, such as film noir, had an average rating of roughly 4.17, a great expression of the rating the majority of its audience accorded. On the other side, a temporal analysis of the movie ratings has brought out exceptional stability, whereby the average rating has almost flatly stayed at 3.5 across different periods. This connotes a constant level of satisfaction among the user variations on how they might be engaged in it.

Transitioning to the predictive modeling phase, the evaluated the model-predictive value of two contrasting models: Linear Regression and Random Forest Regressor. The Linear Regression model, as simple in nature, acts like a baseline model to compare the results with. The Random Forest Regressor model was complex to capture the intricate dynamics between different features of the movies and ratings. The performance was evaluated using mean squared error (M), root mean squared error (RMSE), and mean absolute error (MAE) with the benchmarking model (Kouadri et al., 2022). Reasonably, the Linear Regression model performed well, giving the low MSE value of 0.808, RMSE value of 0.899, and the MAE value of 0.668, in which error in predictions was quite close to the actual ratings done by the user. As strange as this may sound, despite being such a complex model, the Random Forest Regressor model was simply a little bit worse than the Linear Regression model in comparison by accuracy. It had a little higher MSE, 1.025; RMSE, 1.012; and MAE, 0.771.

This is to shed some light on the predictive powers that the models come with, but the emphasis on the importance of choice models and the simplicity by which the models stand a chance to do better in some analytical settings. It is in this context that the slightly better performance of the Linear Regression model than that of the Random Forest Regressor challenges the stereotype of complexity being an automatic generator of better predictive outcomes but rather sees strategic feature engineering and model tuning lying at the heart of optimizing predictive accuracy.

The culmination of these studies gives a comprehensive understanding of movie preference that encompasses intricate user behavior and predictive challenges. EDA woven into machine learning has helped in unraveling genre preference details and factors that come along with nuanced insights into reasons for movie ratings. These revelations are to deepen our understanding of user engagement not only within the digital entertainment platforms but to lay a foundation in making the recommender systems more prominent to point towards the future of personalized content curation more towards refined and user-centric methodologies.

# **Discussion**

This was of value for understanding user movie preferences and using machine learning to improve recommender systems. This discovery over the MovieLens dataset, either from the approach of explorative data analysis (EDA) or predictive modelling, proves to be valuable. The present study has thus been set with the objective of unraveling the complex interplay of movie features and user ratings, understanding more in-depth and light about what influences user satisfaction and engagement in terms of input to the system.

# **Interpretation of Findings**

Using an exploratory data analysis (EDA) phase, findings regarding user preferences and the dynamics of user engagement within the movie rating domain can be quite interesting. This clearly showed the users' preference for some genres, like Film-Noir, Drama, and Crime, receiving higher average ratings, and giving in a hint of users' preference for content with good, complicated story narration.

The conclusion brought forth one very clear point: model selection is of prime importance to the selected model to work in predictive analytics. Although much more complicated than a simple linear regression model, the Random Forest Regressor doesn't produce much better results, so it doesn't exactly shake any conventional wisdom in predicting ratings for the movies.

The results, in other words, speak to this new perspective of model building and select models according to criteria based on model interpretability. It is these simple models, and not the ensemble of complex algorithms, which should be given top billing both by the researcher and the practitioner in most cases, since they mostly agree with the characteristics of the set of data and the predictive task being considered. It focuses on feature engineering, model interpretation, and practical concerns, like computational efficiency, to churn out better predictive models that shall find applications in the entertainment domain for sound decision-making and lead to user satisfaction.

# **Conclusion**

The analysis done in this project—from exploratory data analysis (EDA) to even the predictive modelling of the MovieLens dataset—brings out deep insights into the users' movie preferences, something that the recommender system being developed is sure to be based on and further improved. Thus, by careful examination of user interactions and employing modern techniques of machine learning, this research throws light on a very complex interrelationship between movie characteristics and user satisfaction, which is very valuable to understand the conundrum associated with digital content consumption.

The researcher embarked on the wealth of information contained in the dataset in this exploratory data analysis and got patterns and trends that cut across the user behavior and their corresponding ratings for the movies. It is possible to go much further in finding what really influences the preferences and level of engagement for users in the scope of digital entertainment by breaking down the user interactions and the variables that might be related to movie genres, ratings distributions, and temporal trends. This phase of the study served as a crucial first step in identifying key insights and informing subsequent analyses.

Predictive modeling techniques were then applied in order to dig much deeper on how the properties of movies related to user satisfaction. The authors applied different models of machine learning algorithms in prediction exercises of the rating of movies given some features such as genre labels, metadata, and contextual features, among others. The authors performed rigorous benchmarking and model performance comparison to extract insight into the predictive power of different algorithms and the importance of feature selection in capturing users' preferences accurately.

In general, the aim of this research was to prove that the dynamics under user movie preferences are fairly complex; hence, interesting insights could be delivered to improve recommender systems. If the relation is complex in character between movie features and user satisfaction, it opens a way for powerful and sophisticated recommendation algorithms to be developed. These algorithms will then use predictive analytics in guiding recommendations toward individual users' tastes, thus highly improving user experience in digital content platforms. With further research and perfection, recommender systems could become more accurate in their recommendations and, hence, meet more of the user's needs and expectations, hence raising user involvement and contentment in the digital entertainment frontier.

# **Key Takeaways**

# **Genre Preferences**

This user preference of a specific genre gives great importance to content creation and distribution, alongside recommendation systems that shape not only content creation and distribution but also personalized recommendations for the user. It is of note that the preference to such genres as Film-Noment403-Noir, Drama, and Crime have an inclination not only on the surface of being liked but deep into the affinity with thematic depth, character construction, and elaboration of the plot complexity. Such insight guides the filmmaker or content producer on how to tell resonating stories that touch the audience at levels of feelings and the mind. It identifies the genres the users love more and, therefore, the platform can optimize marketing strategies in such a way that it can uncover more on the content-specific to the genre to relevant audience sections, optimizing reach and effect.

User preference for certain genres isn't actually the knowledge of the superficial preferences and aversions of the users—what they like or dislike, say, about certain kinds of music, for example. Rather, it lies in the knowledge of the underlying motivations, feelings, and aspirations that engender user interaction with content. Empowering content creators, distributors, and recommendation systems to create and maintain deeper relationships with their audiences, fostering overall user satisfaction that begets success in the rapidly shifting digital entertainment space.

# **Temporal Consistency**

The temporal consistency observed unveils the temporary consistency of these average ratings, which includes the enduring allure of specific thematic elements and storytelling techniques in entertainment. Some genres and styles of narration will never lose their topicality, despite the fluctuations of users' engagement in marketing dynamics. This firmness in storylines and themes would seem to implicate that the storytelling devices and thematic content identified above have an assured timelessness about them. This is an attribute that makes them appropriate and relevant for any given audience throughout generations and cultural times. "It's a win-win strategic approach for the two parties, making sure the platforms are investing in the type of content desired by the audience and one that has life-transforming impact, thereby maximizing appeal and competitiveness of the platform amidst very dynamic entertainment landscape. And in conjunction with temporal consistency insight, the platform may curate the content library with proper diversity in different tastes and preferences of the user base to ensure they experience increased and continued user engagement, retention, and satisfaction over time.

# **Suggestions for Future Research**

Future research should test a wider range of learning models, including newer and more advanced algorithms, such as deep neural networks or ensembles. Those sophisticated approaches will have the way to capture better the subtleties of user preference but also harness the power of learning complex hierarchical representations from data. The proposed outcome will further provide insight through the exploration of diversified modeling approaches on the resultant patterns and structures that exist within user preferences, in turn pushing the state of the art in predictive modeling forward within the entertainment domain.

# **Appendix**

Data collection:

A screen shot of a computer program

Description automatically generated

Data Pre-Processing:

A screen shot of a computer

Description automatically generated

Removing Null values:

A screenshot of a computer program

Description automatically generated

A screen shot of a computer

Description automatically generated

Merging Datasets:

A screenshot of a computer

Description automatically generated

A screen shot of a computer

Description automatically generated

Movie genres and ratings analysis:

A screen shot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

Bar plot and Box plot of Average Ratings by Genre:

A screen shot of a computer program

Description automatically generated

Trends or patterns in the usage of tags over time:

A computer screen shot of a program code

Description automatically generated

Temporal patterns Analysis of Ratings:

A computer screen shot of text

Description automatically generated

Line plots of ratings trends over time:

A screen shot of a computer program

Description automatically generated

Extracting Year and Genre Count from Movie Data:

A screen shot of a computer code

Description automatically generated

Merging Datasets for Machine Learning :

A screen shot of a computer code

Description automatically generated

Data Cleaning and Feature Engineering:

A screen shot of a computer code

Description automatically generated

Linear Regression Model:

A screenshot of a computer program

Description automatically generated

A screen shot of a computer code

Description automatically generated

A screen shot of a computer code

Description automatically generated

Random Forest Regressor Model:

A screen shot of a computer program

Description automatically generated

A screen shot of a computer

Description automatically generated

Comparision of Linear Regression model vs Random Forest Regression model

A screen shot of a computer program

Description automatically generated

# **References**

Permana, A. I. (2023). Analysis streaming application viewership with EDA. *Journal of Intelligent Decision Support System (IDSS)*, *6*(4), 216-222. <https://doi.org/10.35335/idss.v6i4.166>

Hodson, T. O. (2022). Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development Discussions*, *2022*, 1-10. <https://doi.org/10.5194/gmd-15-5481-2022>

Ćalasan, M., Aleem, S. H. A., & Zobaa, A. F. (2020). On the root mean square error (RMSE) calculation for parameter estimation of photovoltaic models: A novel exact analytical solution based on Lambert W function. *Energy conversion and management*, *210*, 112716. <https://doi.org/10.1016/j.enconman.2020.112716>

Kouadri, S., Pande, C. B., Panneerselvam, B., Moharir, K. N., & Elbeltagi, A. (2022). Prediction of irrigation groundwater quality parameters using ANN, LSTM, and MLR models. *Environmental Science and Pollution Research*, 1-25. <https://doi.org/10.1007/s11356-021-17084-3>

Liemohn, M. W., Shane, A. D., Azari, A. R., Petersen, A. K., Swiger, B. M., & Mukhopadhyay, A. (2021). RMSE is not enough: Guidelines to robust data-model comparisons for magnetospheric physics. *Journal of Atmospheric and Solar-Terrestrial Physics*, *218*, 105624. <https://doi.org/10.1016/j.jastp.2021.105624>

Himeur, Y., Sayed, A., Alsalemi, A., Bensaali, F., Amira, A., Varlamis, I., ... & Dimitrakopoulos, G. (2022). Blockchain-based recommender systems: Applications, challenges and future opportunities. *Computer Science Review*, *43*, 100439. <https://doi.org/10.1016/j.cosrev.2021.100439>

Zaizi, F. E., Qassimi, S., & Rakrak, S. (2023). Multi-objective optimization with recommender systems: A systematic review. *Information Systems*, 102233. <https://doi.org/10.1016/j.is.2023.102233>

Li, J., & Zhang, S. (2020). Joint visual and temporal consistency for unsupervised domain adaptive person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16* (pp. 483-499). Springer International Publishing. <https://doi.org/10.1007/978-3-030-58586-0_29>

Ye, J., Chen, Z., Liu, J., & Du, B. (2020, July). TextFuseNet: Scene Text Detection with Richer Fused Features. In *IJCAI* (Vol. 20, pp. 516-522). <https://www.ijcai.org/proceedings/2020/0072.pdf>