MovieLens Case Study: Exploratory Data Analysis and Movie Rating Prediction

The MovieLens case study involves analyzing a dataset containing ratings of approximately 3,900 movies made by users who joined MovieLens in 2000. The objective of this project is to perform Exploratory Data Analysis (EDA) and build a model to predict movie ratings based on various features.

1. Importing and Preparing the Data

The first step is to import three datasets: Ratings.dat, Users.dat, and Movies.dat. These datasets contain information about user ratings, user demographics, and movie details, respectively. After importing the datasets, we create a new dataset called Master_Data by merging information from Ratings.dat, Users.dat, and Movies.dat using the primary keys MovieID and UserID.

2. Exploratory Data Analysis

During the EDA phase, we explore the datasets using visual representations such as graphs and tables. Some of the key insights we obtain are:

- User Age Distribution: We analyze the distribution of users' ages to understand the age groups of the MovieLens user base. This helps us identify the dominant user demographics and their movie preferences based on age.
- User rating of the movie "Toy Story": We specifically examine the ratings for the movie "Toy Story" to determine its popularity and reception among users.
- Top 25 movies by viewership rating: We identify the top 25 movies based on the number of ratings they received, indicating their popularity and viewership.
- Ratings for movies reviewed by a particular user (UserID = 2696): We extract and present the ratings given by user 2696 for various movies, helping us understand their preferences and interests.

3. Feature Engineering

In this phase, we use the genres column in the Movies dataset to create one-hot encoded columns for each genre category. This transformation allows us to analyze the impact of movie genres on user ratings. For example, we can examine whether specific genres receive higher or lower ratings compared to others.

4. Model Building and Prediction

To predict movie ratings, we use the features from the Master_Data dataset, including age, gender, occupation, and one-hot encoded genre categories. We then build an appropriate predictive model for movie ratings. For this case study, we used LGBM and XGBoost algorithms.

Results and Conclusion

After completing the analysis and building the models, we obtain the following results:

- LGBM accuracy score: 36.32%

- XGBoost accuracy score: Continuous Error

The LGBM model achieves an accuracy of 36.32%, indicating its ability to predict movie ratings to some extent. However, the XGBoost model yields an error, suggesting it may require further tuning or feature selection to perform effectively.

In conclusion, the MovieLens case study demonstrates the power of Exploratory Data Analysis in understanding user preferences and patterns in movie ratings. By leveraging demographic information and movie genres, we can build models to predict movie ratings. However, achieving higher accuracy might require more sophisticated algorithms or additional features.

Project Takeaways

This project provides valuable insights into user behavior and movie preferences. The results indicate that demographic factors, such as age and occupation, may influence movie ratings. Additionally, certain movie genres may be more appealing to specific user groups, impacting their ratings.

However, there are some limitations to consider. The dataset might not be fully representative of all MovieLens users, as it includes only users who provided demographic information voluntarily. Also, the movie genre information entered by hand may introduce errors and inconsistencies.

For future work, more advanced machine learning algorithms and feature engineering techniques can be explored to improve prediction accuracy. Gathering additional data, such as user reviews and external movie ratings, could further enhance the models' performance.

In conclusion, the MovieLens case study provides a valuable learning experience in data analysis, feature engineering, and predictive modeling. Understanding user preferences and building accurate movie rating prediction models have practical applications in the entertainment industry, such as personalized movie recommendations and targeted marketing campaigns.