PMLDL. Recommender Report

Introduction

In the quest to enhance user experience through personalized content, recommendation systems stand at the forefront of technology. Leveraging the MovieLens dataset, this report delves into three distinctive models:

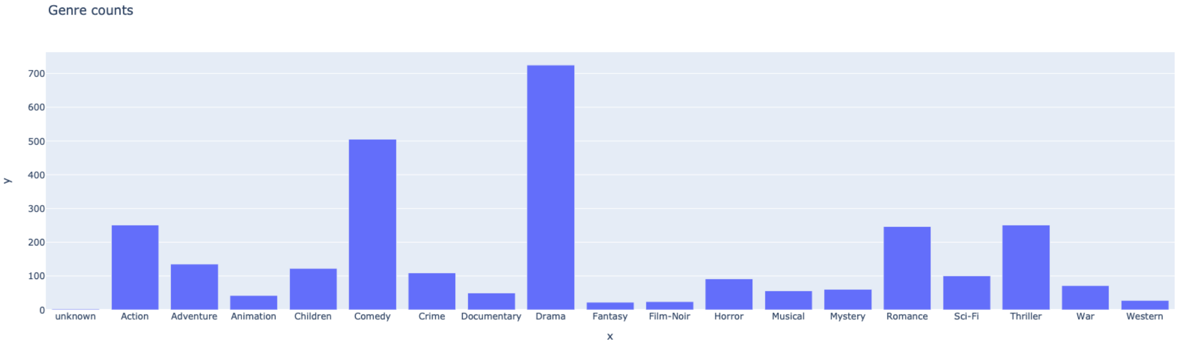
- **Collaborative Filtering**: Harnessing user-item interactions to generate recommendations.
- **Hybrid Model**: A synergistic combination of collaborative filtering and content-based methodologies.
- **Gradient Boosting Model**: Employing CatBoost to integrate a multitude of user and movie features.

The efficacy of these models is measured against their ability to predict user preferences with high accuracy.

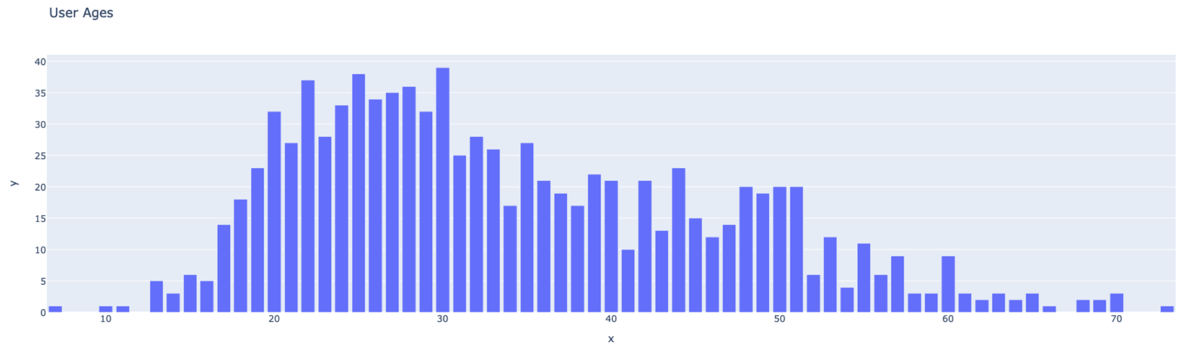
Data Analysis

An exploration of the MovieLens dataset reveals:

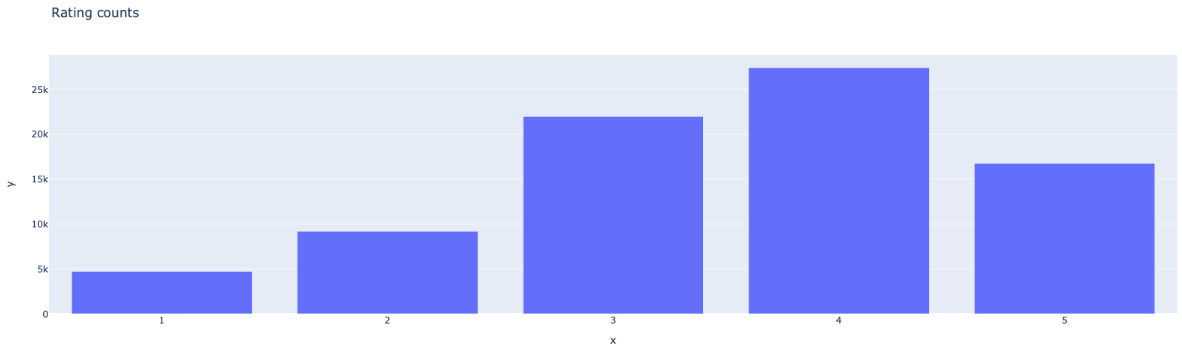
- **Genre Distribution**: A histogram showcasing the prevalence of genres across movies.



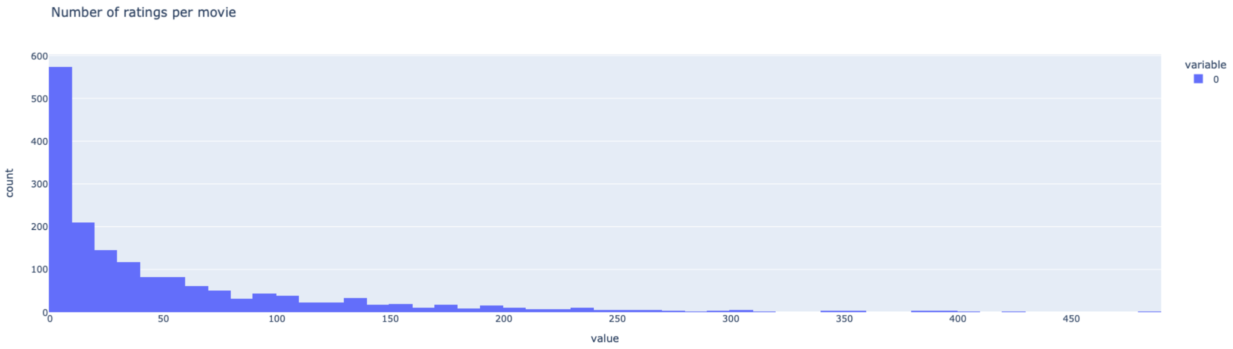
- **User Age Distribution**: Highlighting the demographic's inclination towards younger age groups.

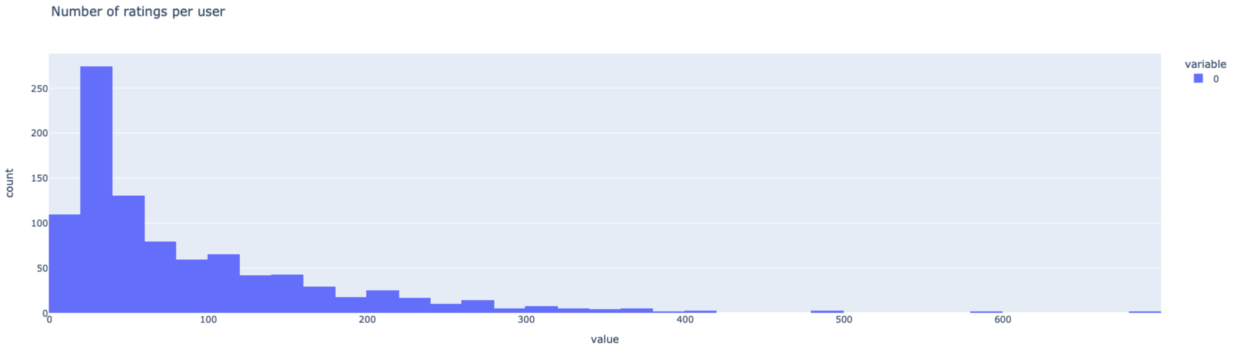


- **Rating Distribution**: Indicating users' propensity to award higher ratings, with a significant skew towards the 4-star rating.



Notably, the number of ratings per user and per movie exhibits a **heavy-tailed distribution**, signifying a concentration of ratings among a limited subset of users and movies. This phenomenon poses a challenge, particularly for movies and users with sparse ratings, hinting at an *almost-cold start* problem in recommendation systems.





Model Implementation

The models are crafted as follows:

- **Collaborative Filtering with SVD**: It decomposes the user-item matrix to unearth latent factors that drive user preferences.
- **Hybrid Approach**: It amalgamates the score predictions of collaborative filtering with a content-based model that analyzes movie genres.
- **Gradient Boosting with CatBoost**: This model synthesizes all available features into a unified DataFrame, treating user and movie identifiers as categorical variables in the prediction process.

Model Advantages and Disadvantages

A critical analysis uncovers:

- **Collaborative Filtering**: While adept at discerning hidden user patterns, it struggles with new items due to the cold start problem.
- **Hybrid Model**: By combining two methodologies, it aims for enhanced accuracy but risks performance if the content-based segment underperforms.
- **Gradient Boosting**: Although capable of modeling complex interactions, its success is contingent upon the predictive power of the underlying features.

Training Process

The models undergo meticulous training, with hyperparameters fine-tuned for optimal performance. The training phase is cautious in splitting the dataset to affirm the model's ability to generalize beyond the training data.

Evaluation

The models are assessed using metrics tailored for recommendation systems:

- **Hit Rate**: Evaluates whether the recommended set contains items of interest.
- **ARHR**: Weighs the ranking of successful recommendations.
- **NDCG**: Provides a normalized score reflecting the quality of the recommendation list ordering.

These are juxtaposed with a random rating predictor's performance. The metrics yield the following scores:

- **Collaborative Filtering**: (0.53, 0.34, 0.065)
- **Hybrid**: (0.44, 0.31, 0.05)
- **Boosting**: (0.51, 0.33, 0.06)
- **Random**: (0.32, 0.36, 0.03)

Results

All models demonstrate a marked improvement over the random baseline. The suboptimal performance is attributed to the skewed distribution of ratings. Specifically, the hybrid model's underwhelming results may stem from the content-based component's insufficient accuracy, which, when averaged with collaborative filtering, dilutes the overall efficacy. The gradient boosting model's inability to exceed collaborative filtering's performance suggests that the amalgamated features lack the depth of information necessary to enhance predictions significantly.