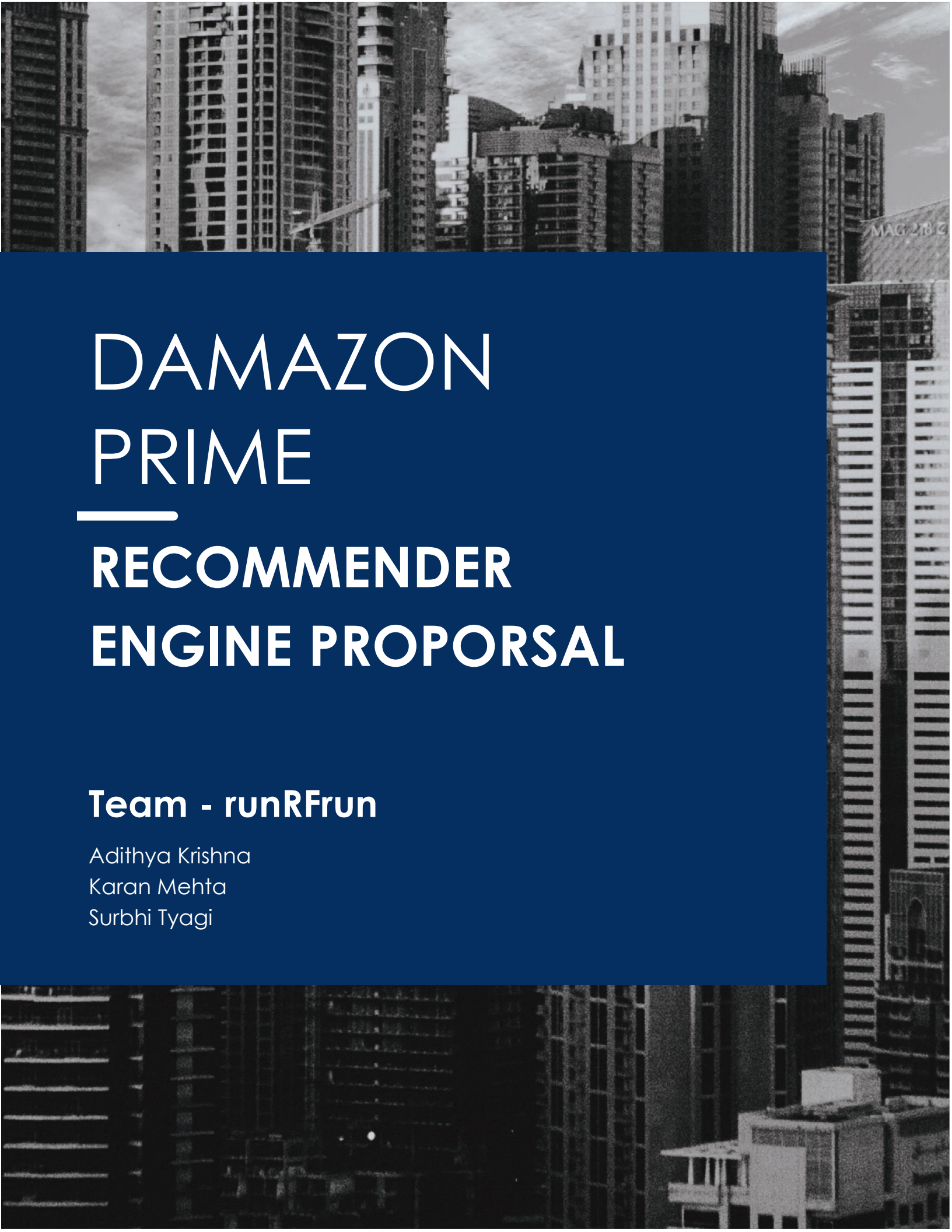


DAMAZON PRIME

RECOMMENDER ENGINE PROPORSAL

Team - runRFrun

Adithya Krishna
Karan Mehta
Surbhi Tyagi

DAMazon Prime

Recommender Engine Proposal

Business Objective

As a business looking to grow its user base over the coming quarters, helping users discover movies they will enjoy is integral to DAMazon Prime achieving its objectives. The Content Management team has commissioned the RRFR Data Science team to build a recommendation engine to provide accurate predictions of movies current users would enjoy. The benefits of implementing a recommendation engine for DAMazon Prime include potential increases in user retention rates by providing a service which is relevant for them and building habits around viewing from the DAMazon Prime platform.

Data Understanding

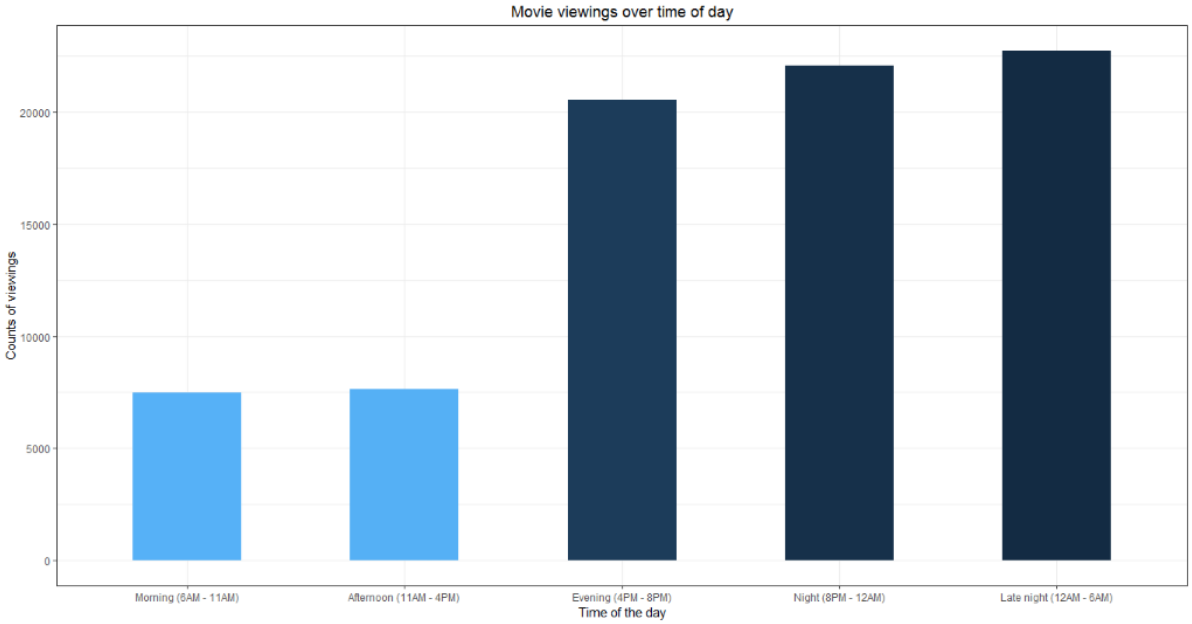
The data used for the analysis and modelling for the recommendation engine included information relating to user demographics (age, gender, occupation), movie features (genre, movie length, release date), and the user rating for every movie viewed by them. The following assumptions were defined prior to the analysis:

- Genre tags are correctly labelled by producers
- User information is reliable and accurate
- Ratings seen are absolute values and not personalised based on user/item features
- IMDb scrape data does not require any cross verification

The dataset used contains over 1,600 movies, reviewed by around 950 viewers, rated between the period of September 1997 and April 1998. Initial analysis of the data provided some key insights into viewing habits, rating habits, and content preferences.

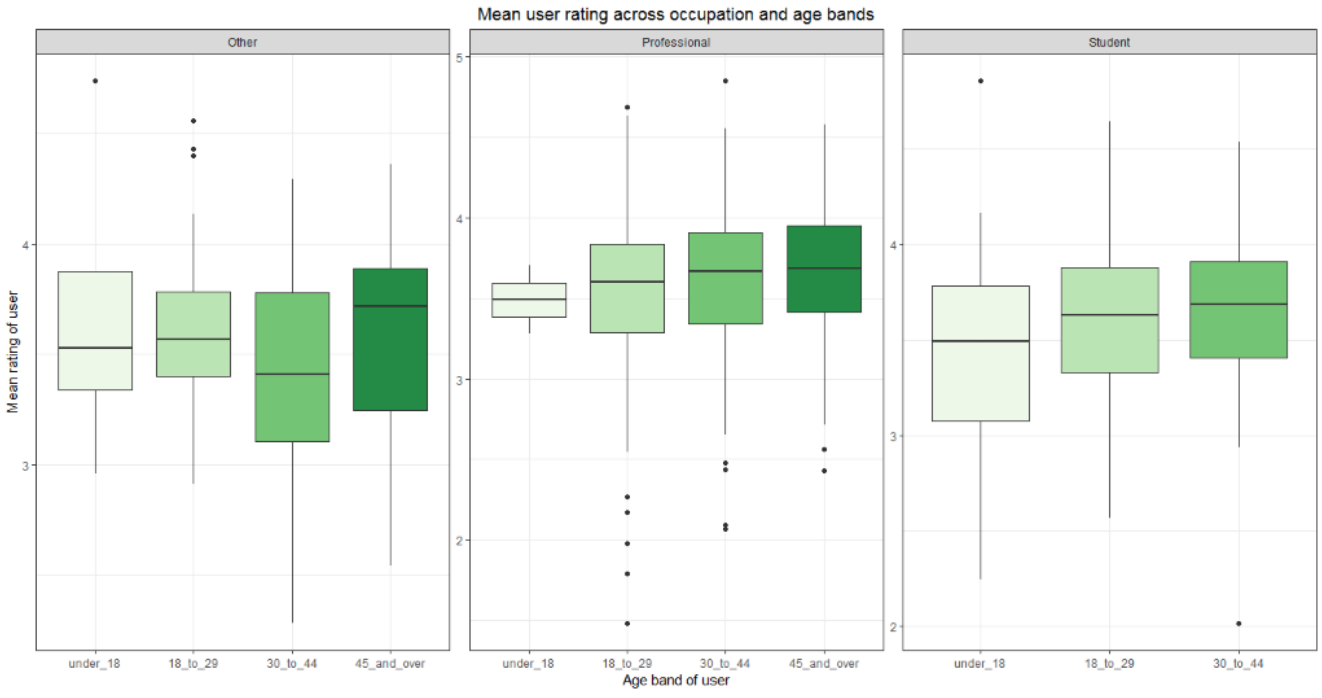
VIEWING HABITS

Current viewing habits indicate that a majority of DAMazon Prime viewers are accessing movies after 4pm during the day. The following graphs highlights the low viewership in the early parts of the day and midday:



RATING HABITS

Across the age groups, older viewers seem to be less likely to watch movies on DAMazon Prime but are however likely to rate those movies highly. The graph below highlights the general trend of increasing mean ratings as age increases across the three broad occupational categories of Student, Professional, and Other (i.e., Homemaker, Retired, Unknown, and None)



CONTENT PREFERENCES

With regards to content, users were more likely to watch War and Sci-Fi genre movies, while Documentaries and Drama were least viewed. However, the genres which were rated the highest include Film Noir and War, while those with the lowest ratings include Fantasy and Horror (see Appendix – Table 1).

The DAMazon Prime data highlighted some issues for consideration regarding the collection of data for future enhancements. These anomalies included the following:

- Item ID duplicates – 18 movies had duplicated IDs which may cause issues with accuracy of predictions to users.
- Rating Timestamp - Over 100 cases of movie rating timestamps prior to the release date of those movies which can impact the ability to use this variable for modelling purposes.

Data Preparation and Feature Engineering

Based on the initial analysis, variables such as user zip code (too many values given the user count), video release date (all NULLs) and rating timestamp (as noted above) were not considered for the modelling exercise.
For the modelling process, the following features were engineered based on research and insights from preliminary analysis:

1. user_mean_rating: mean of each user rating across all movies they have viewed;
2. user_genre_mean_rating: mean of how certain age bands, gender have rated genres (except 'Unknown' and 'Western' genres which were dropped due to low prevalence in the test set);
3. userIndexMeanRating: mean rating of how users of a certain age, gender, occupation have rated items at a certain hour in the day;
4. itemIndexMeanRating: mean rating of items of similar ratings, item length, and mature rating.

All four of the above features were used in the final model for predicting the rating, in addition to item_mean_rating.

Modelling & Evaluation

To ensure the most accurate predictions for user preference were suggested, several modelling techniques were considered for the recommendation engine. This included:

- Collaborative Filtering (both user-based and item-based);
- Linear Regression;

- Matrix Factorisation; and
- Tree based models such as Random Forests, Gradient Boosting Machine (GBM), and XGBoost

A key business measure of success is to help users discover new movies they will enjoy. The model which is likely to achieve this is the one with the lowest prediction error (Root Mean Square Error – RMSE), and hence models were evaluated based on this criteria. The table below summarises the error scores for the various models.

Modelling Technique	Test / Train RMSE Score	Kaggle RMSE Score
User Based Collaborative Filtering	1.395126	1.53334
Item Based Collaborative Filtering	2.373015	N/A (not submitted)
Linear Regression	1.015678	N/A (not submitted)
Matrix Factorisation	1.0159873	1.19350
Random Forest	0.9969543	0.97422
XGBoost	0.9136629	0.94229
GBM	0.9071594	0.93172

While Collaborative Filtering methods have been popularised by numerous players in the recommendation engine market, the current DAMazon Prime subscriber count limits its ability to predict accurately. Additionally, Collaborative Filtering methods also pose limitations in scenarios of new movies and new users being introduced to the system. Finally, these and linear regression methods can also reduce the generalisability of the predictions beyond this dataset.

Given the limitation of models mentioned above and noting that the lowest error in prediction was observed with the GBM modelling technique (Table above), this technique is proposed for the recommender engine. The GBM technique minimises error by iteratively learning from sequential models and averaging the error from the individual predictive models, thus making it the preferred technique for the solution.

From the models trialed, it was noticed that the ratings given by a user are best predicted by the average rating of the movie (*item_mean_rating*), how other users of the same demographic have rated the movie (*userIndexMeanRating*), the typical rating behaviour of the user (*user_mean_rating*), and the average rating of the user for movies of the same genre (*user_genre_mean_rating*) (see Appendix – Table 2). From a business standpoint, this may mean that users are highly influenced by how movies are rated by other users, specifically those belonging to the same demographic, and noting that users typically have strong genre preferences.

Implementation

Implementation of the proposed recommendation engine can have significant impact to DAMazon Prime's ability to provide its users with content they will enjoy. Consequently, this is likely to have an impact on the level of user retention, subscriber counts, and content management.

While the proposed recommender engine will provide users with accurate predictions, there is room for DAMazon Prime to further refine and enhance the accuracy of predictions made by considering additional data collection in relation to users and their viewing habits. It is generally accepted that with more data, accuracy of predictions improves. Furthermore, by monitoring aspects of viewing behaviour, such as click throughs based on recommendations, movie completion rates, and incorporating additional external ratings data (i.e. Rotten Tomato, Fandango, Google, Metacritic, etc.), DAMazon Prime's ability to provide the right content to its users can be improved.

Ethics

Recommender engines, while beneficial for both DAMazon Prime and its users, can pose some ethical questions for the business to consider. There are issues relating to privacy, behaviour manipulations, discrimination, misleading content, and the potential threat of identity theft (Paraschakis, 2017).

As mitigation strategies to reduce privacy and ethical concerns, DAMazon Prime should consider:

- Policy & Framework – user access to privacy policy in relation to user data storage and usage
- Transparency – garnering customer consent in the use and storage of data as outlined in the policy to ensure customer trust is fostered
- Security – strong cybersecurity measures should be developed to ensure mitigation of risks associated with data breach and associated reputational and business impact
- User Controls – user interfaces which enable users to adjust profile data by defining interests and privacy settings

To ensure users place trust within the DAMazon Prime offering, it is important to ensure that the systems are legally, algorithmically, and ethically sound.

References

Paraschakis, D. (2017). *Towards an Ethical Recommendation Framework.* 11th International Conference on Research Challenges in Information Science (RCIS), pp 211-220.

Appendix

Table 1- Content Preferences

Genre	Average Rating	Views per Movie
Film Noir	3.92	50
War	3.81	106
Documentary	3.69	12
Drama	3.68	44
Crime	3.64	60
Mystery	3.64	69
Romance	3.62	63
Animation	3.59	69
Western	3.58	55
Sci Fi	3.55	101
Musical	3.52	72
Thriller	3.51	70
Adventure	3.50	82
Action	3.48	82
Comedy	3.39	47
Childrens	3.35	47
Horror	3.29	47
Unknown	3.20	5
Fantasy	3.20	49

Movies with n genres tagged to them have been considered n times for the purpose of this table (i.e., a movie tagged as sci-fi and war has been considered twice, for calculation of the above stats for sci-fi and war genres both).

Table 2 – Variable Importance from the final model

Variable Name	Importance
Average Movie Rating - (*item_mean_rating*)	41.47524
Average Ratings by User Demographics - (*userIndex_mean_rating*)	30.19813
Average User Rating - (*user_mean_rating*)	17.46983
Average User Rating by Genre - (*user_genre_mean_rating*)	10.85681

Team Contribution Statement

The runRFrun team worked collaboratively on the AT2b project as required by the assignment brief, while focusing our efforts on everyone's capabilities.
Key responsibilities for the team were as follows:

- Surbhi Tyagi – EDA, reporting and AT2b submission
- Adithya Krishna – EDA, modelling, and research
- Karan Mehta – EDA, modelling and Kaggle submissions

In addition to the responsibilities noted above, the team worked cohesively to resolve process issues and generate ideas with all team members contributing evenly across the various tasks required for the assignment.

The working files for all Kaggle submissions and model iterations tested can be found here:

https://github.com/AdithyaKrishna7/runRFrun