Project and Data Management Plan

Predicting Movie Ratings and Analyzing Feature Importance Using Tags and Genres

Project Overview

- The project focuses on predicting movie ratings based on user-provided tags and movie genres using the MovieLens 32M dataset.
- The research will use machine learning models to predict movie ratings with user-provided tags and movie genres.
- Using explainable AI tools like SHAP, the project will identify the most impactful tags and genres on user ratings.
- The research will then investigate the current trends in user ratings to understand changing viewer preferences.

Project Overview

- Research will assist content providers and movie platforms in identifying genres and themes associated with higher or lower ratings.
- Findings will enhance recommendation systems' accuracy, leading to personalized movie suggestions and better content prioritization.
- Offers insight into evolving user preferences in data-driven decision-making.

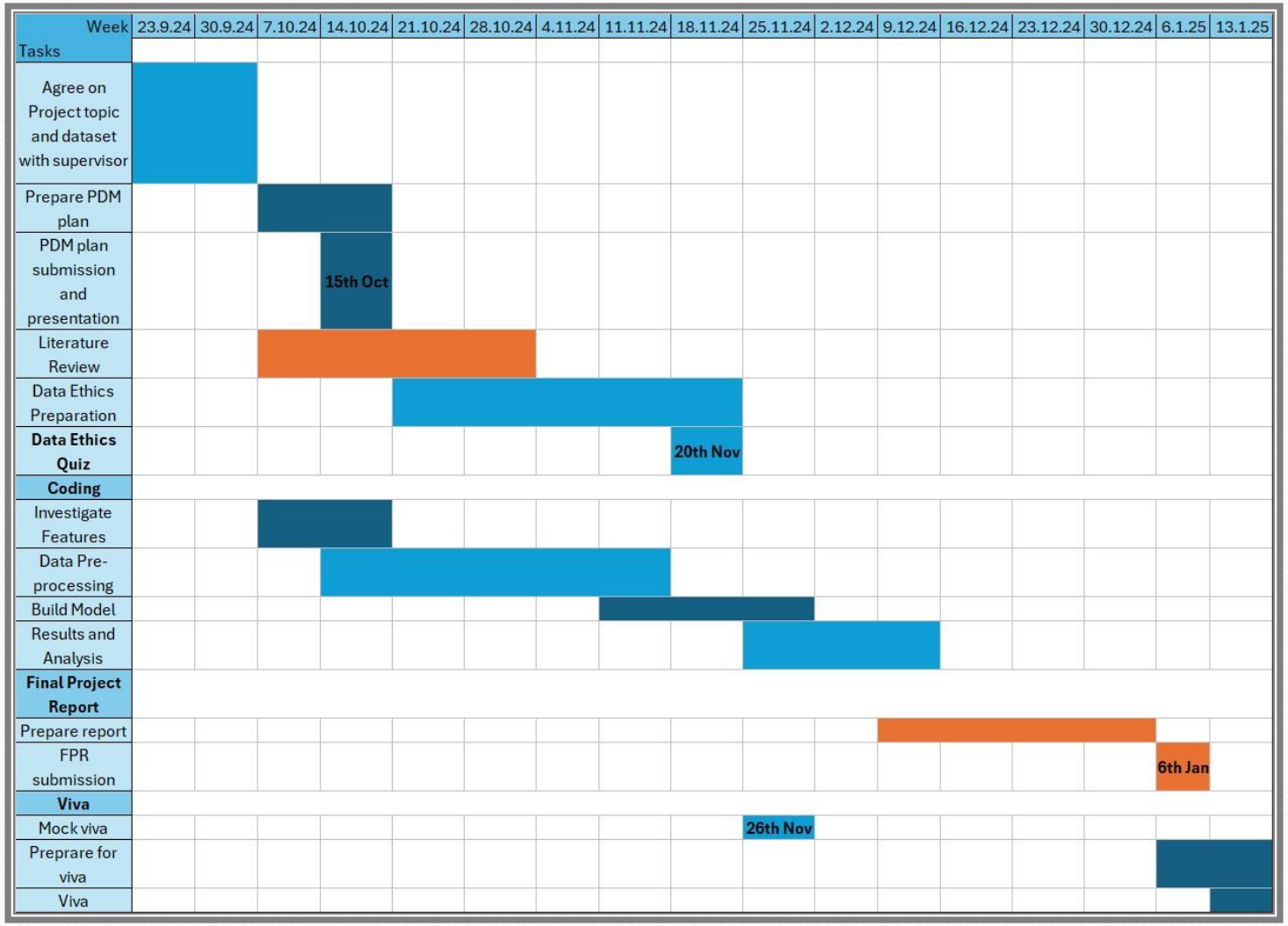
Research Question

How can user-generated tags and movie genres be used to predict movie ratings, and which are the tags or genres influencing these predictions in the current trend landscape?

Project Objectives

1. Predict Movie Ratings using tags, genres.
2. Analyze Feature Importance to find key tags and genres.
3. Investigate Trends in ratings and tagging patterns.
4. Provide Actionable Insights for content curation.

Project Plan



Dataset Overview

- The MovieLens 32M dataset contains 32,000,204 ratings, 2,000,072 tag applications, and 87,585 movies rated by 200,948 users.
- It was collected from 200,948 users between 1995 and 2023.
- Each user has rated at least 20 movies.
- UTF-8 encoded, comma-separated value (CSV) files.
- Dataset size: 239 MB
- The dataset is publicly available from GroupLens website at: https://grouplens.org/datasets/movielens/32m/.

Dataset Overview

- Organized into four files: ratings.csv, tags.csv, movies.csv, and links.csv.
- ratings.csv contains 5-star movie ratings.
- tags.csv contains user-generated movie tags.
- movies.csv contains movie titles and genres.
- links.csv contains IMDb and TMDb IDs of movies to these external websites.

Document Control, Storage and Security

- Project is stored at https://github.com/aaronmj7/MSc_project.
- File names follow a structured convention for readability and easy tracking.
- A ReadMe file will be provided to help other coders who may want to use the code.
- Dataset is available at https://grouplens.org/datasets/movielens/32m/ and stored in my OneDrive.
- Code is automatically saved to Google Drive and will be uploaded weekly to GitHub and OneDrive.

Ethical Considerations

- Each user is represented by an ID, and no other information is provided.
- Data is anonymized, avoiding individual identification. Hence, it does not come under the General Data Protection Regulation (GDPR).
- The research follows the University of Hertfordshire's ethical policies.
- The data is ethically sourced and does not require further authorization for academic use.

References

- Boudiba, T. and Dkaki, T. (2022). Exploring Contextualized Tag-based Embeddings for Neural Collaborative Filtering. Proceedings of the 14th International Conference on Agents and Artificial Intelligence. Available at: https://api.semanticscholar.org/CorpusID:246922630 (Accessed: 14 October 2024).
- Harper, F. M. and Konstan, J. A. (2015), The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS). Available at: https://doi.org/10.1145/2827872 (Accessed: 14 October 2024).
- Hassan, H. A. M., Sansonetti, G., Gasparetti, F., and Micarelli, A. (2018). Semantic-based tag recommendation in scientific bookmarking systems. Proceedings of the 12th ACM Conference on Recommender Systems (RecSys '18). 497, pp.465–469. Available at: https://doi-org.ezproxy.herts.ac.uk/10.1145/3240323.3240409 (Accessed: 14 October 2024).