Project and Data Management Plan

Aaron Modiyil Joseph
22018497

Predicting Movie Ratings and Analyzing Feature Importance Using Tags and Genres: Insights from the MovieLens 32M Dataset

Project Overview

The project focuses on predicting movie ratings based on user-provided tags and movie genres using the MovieLens 32M dataset. The dataset contains user-generated tags, ratings, movie titles, and genres. User-generated tags provide descriptive information about movies, whereas genres categorize them into broader themes. The goal of the project is to explore the relationship between these features and movie ratings to uncover trends in user preferences.

Using machine learning models, the project will identify which tags and genres have the greatest impact on user ratings. The analysis will also investigate current trends in user ratings, providing insights into changing viewer preferences over time. This research will assist content providers and movie platforms in determining which genres and themes are consistently associated with higher or lower ratings, as well as providing actionable recommendations for content curation. Furthermore, by improving the understanding of user behaviour, the findings will enhance the accuracy of recommendation systems, leading to more personalized movie suggestions and better content prioritization.

As data-driven decision-making becomes more prevalent, this project offers an opportunity to better understand how tags and genres influence ratings, ultimately providing insights into evolving user preferences.

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:** Develop a machine learning model that predicts movie ratings based on user-provided tags, genres, and other relevant features.
2. **Feature Importance Analysis:** Identify the most important tags and genres that influence movie ratings using feature importance techniques (e.g., SHAP values)
3. **Trend Analysis:** Investigate how the importance of tags and genres has evolved and how these changes reflect current user trends in movie preferences.
4. **Actionable Insights:** Provide recommendations on which genres or themes are currently trending based on the ratings and tagging patterns, offering insights to movie platforms on how to tailor their content for better engagement and satisfaction.

References

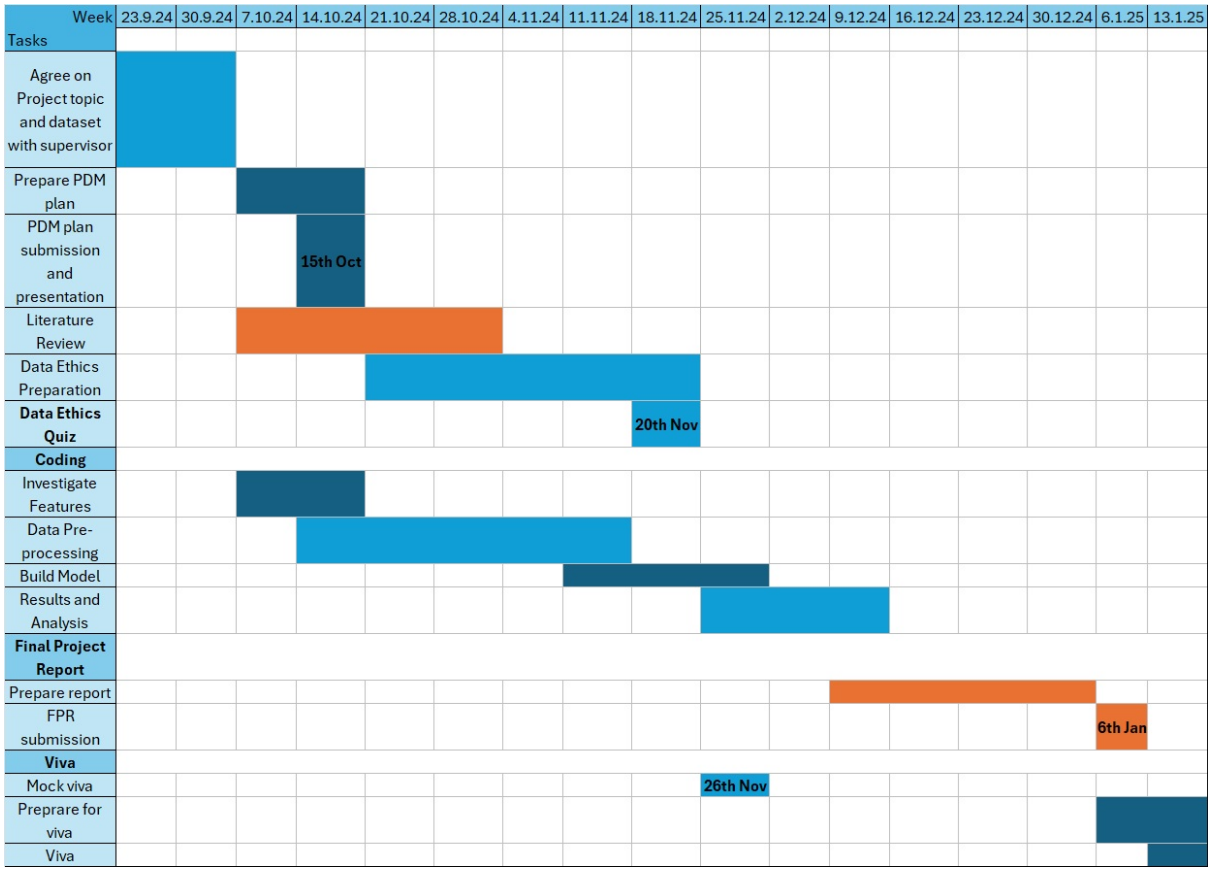
Boudiba, T. and Dkaki, T. (2022). Exploring Contextualized Tag-based Embeddings for Neural Collaborative Filtering. Proceedings of the 14th International Conference on Agents and Artificial Intelligence.

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).

Project Plan



The project begins with selecting and agreeing on a research topic and dataset with the supervisor. This is followed by the preparation of the Project and Data Management (PDM) plan, which is submitted and presented on October 15. A four-week literature study of relevant research and methodology will be undertaken parallel with preparation for a data ethics assessment, which will culminate in a quiz on November 20.

In terms of practical work, the initial phase involves exploring and analyzing the dataset's features, followed by necessary data pre-processing. After the data is prepared, appropriate models will be selected and trained to predict movie ratings. The model results will then be analyzed for key insights. The report writing for the final project will begin in early December, with submission on 6 January 2025.

Furthermore, a mock viva is scheduled for 26 November and preparation for the final viva will begin on 6 January, with the viva scheduled on 13 or 14 January 2025.

Data Management Plan

Overview and Metadata of the Dataset

The MovieLens 32M dataset contains extensive data on user interactions with movies from MovieLens, a movie recommendation service, including 32,000,204 ratings and 2,000,072 tag applications across 87,585 movies. The data were collected from 200,948 users between January 9, 1995, and October 12, 2023. Each user rated at least 20 movies, but no demographic information is provided, and users are only identified by anonymized IDs and were selected at random for inclusion. The dataset, sized at 239 MB (compressed), was collected in October 2023 and released in May 2024.

The dataset is organized into four main files: ratings.csv (containing 5-star movie ratings), tags.csv (with user-generated movie tags), movies.csv (a list of movie titles and genres), and links.csv (IMDb and TMDb IDs of movies to these external websites). The dataset is formatted as UTF-8 encoded, comma-separated value (CSV) files.

Data Collection

The MovieLens 32M dataset is available for public download on the GroupLens website at `https://grouplens.org/datasets/movielens/32m/`. GroupLens is a research lab in the Department of Computer Science and Engineering at the University of Minnesota, Twin Cities, focused on recommender systems, online communities, mobile and ubiquitous technologies, digital libraries, and local geographic information systems.

Document Control

The project is stored in my GitHub repository at `https://github.com/aaronmj7/MSc_project`, where I will be committing code regularly, with updates planned at least twice a week to ensure the project stays current and well-maintained. File names will follow a structured and descriptive convention (e.g., project_v1.ipynb, etc.) to maintain readability and facilitate easy tracking of changes. My repository is public, allowing markers to view the code and track progress throughout the development.

ReadMe File

At the end of the project, the ReadMe file will provide a concise overview of how the MovieLens dataset was used to predict movie ratings based on tags and genres. It will highlight the use of machine learning models to identify key features that influence ratings and trends in user preferences. The document will include setup instructions, dataset details, project structure, code execution guidelines, and a summary of the analysis and results to assist others in implementing the findings.

Security and Storage

The dataset is publicly available at `https://grouplens.org/datasets/movielens/32m/` and is downloaded and stored in my OneDrive. Since I'm using Google Colab, the code will be automatically saved to my Google Drive and will be uploaded to GitHub and OneDrive once a week.

Ethical Requirements

The data is anonymized and does not allow for the identification of individuals. Users are only identified by anonymized IDs; hence, it does not come under the General Data Protection Regulation (GDPR). Furthermore, the research follows the University of Hertfordshire's ethical policies. The data is ethically sourced and does not require further authorization for academic use.