Bilkent CS 464 Machine Learning
Project Proposal
Group 15

Project Title: Book Recommendation System

Team Members:
Ferhat Korkmaz 21901940, Ömer Oktay Gültekin 21901413, Utku Kurtulmuş 21903025,
Dilay Yigit 21602059, Muhammet Oğuzhan Gültekin 21801616

Dataset Description: We plan to use a dataset we found on kaggle.com [1]. It contains 3 million book reviews for 212404 unique books and users who give these reviews for each book. The dataset we plan to use contains 2 files. The initial document provides an overview of a file that contains evaluations, encompassing feedback. This dataset is a component of the Amazon review dataset, which includes product reviews and associated information from Amazon. It encompasses a staggering 142.8 million reviews, dating from May 1996 to July 2014 [1]. The second file, referred to as the "Books Details file", provides comprehensive information about each book. This file is created by utilizing the Google Books API to obtain detailed data about the books such as description of the book, rating counts and genre. We filtered some review data from Books_rating.csv. We found out that there are 7209 users that have reviewed at least 20 books.

Problem Definition: The primary objective of this project is to develop a personalized book recommendation system that predicts users' preferences based on their historical reading behavior and interests. By leveraging machine learning techniques, we aim to create a system that accurately predicts how much a user would enjoy a specific book, enhancing the overall reading experience and encouraging exploration of diverse literary genres

The Planned Milestone: In order for us to work on the dataset more efficiently, firstly, we will eliminate the users that have not reviewed less than 20 books. Also, we will eliminate the fields price, profileName, review/helpful, review/time, review/summary, review/text, image, etc.
Our strategy involves a hybrid approach that combines collaborative filtering and content-based methods to enhance the accuracy of our predictive ratings [2]. By leveraging both user rating data and book features, we aim to provide more precise and tailored recommendations. To facilitate this, we intend to employ Cosine Similarity to quantify the similarity levels between user ratings, subsequently integrating this information with the book features to capture both user and content similarities effectively.
To streamline the process, we will consider using the similarity levels as features and apply Principal Component Analysis (PCA) to reduce feature dimensions, ensuring efficient data handling. While we initially plan to adopt this hybrid framework, we remain open to exploring a content-based approach if it demonstrates superior efficacy during the project's progression.
For the predictive modeling phase, we have selected Linear Regression, Decision Tree Regression, and Random Forest Regression as they are well-suited for variable prediction tasks [3]. Leveraging the rating as the target variable and incorporating other relevant features as input variables, we aim to meticulously train these models. Our focus will involve thorough analysis and potential hyperparameter tuning to optimize the performance of each model. Notably, the inclusion of the Random Forest Regression model will enable us to harness the benefits of an ensemble learning method, allowing for improved prediction accuracy and robustness, particularly in handling complex, non-linear relationships within the data [3]. This comprehensive approach is anticipated to yield more precise and reliable predictions, thereby enhancing the overall effectiveness of our book recommendation system.
In addition to these traditional machine learning models, we have opted to incorporate Multilayer Perceptrons (MLPs) as our chosen neural network architecture. This decision is based on the suitability of MLPs for regression prediction problems, particularly when dealing with tabular data, as is the case with our dataset. To assess the performance of our selected approaches, we plan to employ the Mean Absolute Error (MAE) metric, ensuring a comprehensive evaluation process that validates the efficacy of our hybrid methodology

References:

[1] “Amazon Books Reviews”, Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews/data

[2] “How to Build a Recommendation System”, Analytics Vidhya, 2022. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/build-book-recommendation-system-unsupervised-learning-project/

[3] “7 of the Most Used Regression Algorithms and How to Choose the Right One”, 2021. [Online]. Available: https://towardsdatascience.com/7-of-the-most-commonly-used-regression-algorithms-and-how-to-choose-the-right-one-fc3c8890f9e3