EduVerse: A Comprehensive Study on Course, Career and Book Recommendation Systems Using Machine Learning Algorithms and Cosine Similarity

Abstract

In this paper we discuss the development of a unified recommendation platform that includes course, career, and book recommendations based on many machine learning techniques. Our study uses datasets of: course details, student performance data, career attributes, and book ratings to build a very robust recommendation engine. To provide personalised suggestions for users, we use machine learning models like cosine similarity, vectorization techniques, logistic regression, random forest and collaborative filtering. To measure the efficiency of the models, performance metrics such as accuracy, precision and recall are used. cosine similarity-based models, in spite of their weakness in predicting course and career preferences, as well as collaborative filtering approach surprisingly enhance the accuracy of book recommendations. In this research, we demonstrate the promise of merging multiple recommendation systems into a single platform, as a step towards future work in personalized learning and career guidance systems.

1. Introduction

1.1 Background

In the last few years, the machine learning (ML) has made significant progress in solving the more complex problems falling in different industries, especially in the field of education, career guidance and content recommendation. With the ever-changing world of technology there is a new need for personalized solutions that can be personalized to individual preferences and learning styles. While ML will not replace finance staff in these fields, it does help these staff to process large datasets and derive insights on which recommendations they should tailor based on the analysis.

While there has been the great interest in developing the recommendation systems, the ML techniques like cosine similarity, collaborative filtering and vectorizing techniques have not been applied to integrate course, career and book recommendation. Among existing challenges dealing with diverse datasets, accuracy in personalized recommendations, and balancing popularity based and user specific recommendation. To fill these gaps, this study develops a unified platform to exploit various ML algorithms for providing recommended courses, career planning, and book suggestions. Based on a previous line of works in recommendation systems, this work attempts to further increase their efficiency and widen their efficiency in guiding students and readers in making more informed decisions through their educational and career paths.

1.2 Research Problem

The challenge in developing effective recommendation systems for courses, careers, and books: there is no integrated platform providing personalized suggestions in these domains based on individual preferences in individual courses, career interests and activities. While recommendation algorithms have improved, current methods do not sufficiently meet a student or reader's needs at once. Current systems either target one domain, for instance career guidance, or are overwhelmingly reliant on popularity-based opinion models for making recommendations, thus deliberately restricting their effectiveness in supplying personalised and reliable recommendations. This leaves a gap where users see recommendations that may not be optimal for them, that are not aligned with their interests, academic strengths, future career goals, etc. A prominent limitation of existing methods is the absence of an adaptable system that integrates collaborative filtering mathematical techniques, vectorization, cosine similarity so as to yield results according to user needs.

1.3 Objectives

This research focuses on the following objectives:

1. To develop an integrated recommendation system which is able to suggest courses, careers and books on the basis of provided by users like academic performance or extra-curricular activities.
2. To examine how machine learning algorithm (cosine similarity, collaborative filtering) can serve as good recommendation models in the domains of education, literature etc.
3. To compare the accuracy and user satisfaction of popularity-based models against collaborative filtering models on book and course recommendation.
4. To analyse the effect of different users' parameters including interests and achievements on the relevance of the recommendations suggested by the system.
5. To evaluate the strength and weaknesses of the developed platform in the field of recommendation systems by the user feedback and performance metrics.

---

2. Related Work

Many works have been published in recent years to utilize machine learning in recommendation approaches which are conducted with diverse methods and findings.

Smith et al. (2020) engaged contextual filtering to recommend relevant education resources to the students, targeting 10000, and the results reflected 85% accuracy. Nevertheless, the work done by these authors was more selective, and there was no assessment of the overall effect of extracurricular activities on users' preferences [1].

Johnson and Lee (2021) used a blend of on information filtering and popularity filtering with regard to books' recommendation. They used Goodreads dataset and obtained the result in terms of precision rate of 78 %. However, they faced difficulties in solving the problem with cold-start for new users, which made the system incomplete [2].

Garcia & Patel (2022) provided career recommendation system based on the decision tree algorithms in which an overall recall rate of 90% for matching student with potential career relative to academic performance was achieved. However, their research did not use user interests or preferences to prioritise academic metrics and therefore may not have been well aligned [3].

Kim and Park (2021) investigated the types of the book recommendation systems built using the collaborative filtering technique as well as content-based technique. They observed that though collaborative filtering is effective given enough user interaction, content methods seem to have the upper hand when few users are catered for. However, in their study, there was no consideration of combining the two approaches in order to enhance accuracy. We extend this work by employing a hybrid approach that incorporates both techniques in order to provide better and more relevant book recommendations.

Liu and Wang (2022) constructed a course recommendation system from student performance data, demonstrating that academic characteristics lead to the higher personalisation of the recommendation. Data capture these factors but their method did not look at other issues such as student hobbies or after school roles. We do not stop only on capturing the academic factors about students but also include various other factors such as student aspirations and preferences to recommend courses and career paths more comprehensively.

The highlights of the present research are as follows the present research is beneficial to add value in the following ways The proposed platform incorporates a recommendation system for courses, careers and books based on the profile construct in this study. Moreover, the current study intends to refine recommendation accuracy and users' satisfaction by analysing heterogeneous datasets and using cosine similarity and collaborative filtering algorithms for educational and literary recommendations. Moreover, compared with prior research, it also collects information regarding both academic performances and extra procural activities; as a result, there will be a remarkably distinct recommendation stage.

---

3. Methodology

3.1 Dataset Description

To build the integrated platform for course, career and book recommendation we used different datasets for each system the description of them is as follow:

I. Course Recommendation System

The dataset for course recommendation system contains more than 5000 entries with various features includes course title, url(link to the course page on udemy), price, isPaid, no. of reviews, level, subject, and some more. Total 19 features are available in the dataset. As for the data preprocessing steps, it includes filling-in missing values, text normalization and feature engineering. The course_title field was cleaned to remove unwanted characters and the dataset was transformed using vectorization techniques for similarity calculation. The numerical features like price and no. of subscribers were normalized and the categorical feature like level and subject were encoded to improve model accuracy.

II. Career Recommendation System

The dataset for course recommendation system contains more than 2000 entries with various features includes student details, part-time job status, extracurricular activities, weekly self-study house, career aspiration and some more. Total 19 features are available in the dataset. Data preprocessing steps included handling missing values, converting categorical variables into a suitable format, and normalizing numerical features like scores and study hours to ensure consistency in the analysis. These preprocessing steps help in preparing the data for machine learning models used in career recommendations.

III. Book recommendation System

To build the book recommendation system there are 3 main datasets: books, users and rating. Book datasets contain the information about various book like ISBN, title, author, image-URL. Users datasets contain the information about readers such as ID, location and age and the rating dataset contain rating for various books with key features like user-ID, ISBN and book-rating. All datasets are cleaned by data preprocessing steps includes handling missing values, data normalization and feature engineering.

3.2 Machine Learning Models

I. Course Recommendation system

Count Vectorizer:

Overview: Count Vectorizer is a method of mapping a collection of text documents to a matrix of token counts. This is just a matrix of word frequencies present in said dataset, which can be used for feature extraction.

Working Principle: The course titles are tokenized and a term-frequency matrix is made, containing each column being a different word and a different row being each course title. This is used as the various components input for the calculation of similarity.

Reason for Use: The reason I choose that is that it can capture the unique keywords from course titles that will help you to match the same courses based on the shared words or phrases.

Cosine Similarity:

Overview: It is a measure of how similar two vectors are, independent of magnitude, and hence the name. It returns a cosine of a (nonzero) angle between two vectors, a similarity measure between them.

Working Principle: The count vector matrix is input into cosine similarity metric, and the similarity of course titles is computed. For a smaller angle (closer to a cosine value of 1), a smaller similarity score. This means the more similar courses.

Reason for Use: For text data, Cosine Similarity is a fit since it considers the direction of the vector as opposed to its magnitude and so is a great way to compare the frequencies of words in course titles. This makes it an easy and scalable way to make recommendations with textual data.

Data Preprocessing Techniques:

Text Cleaning: As a result, to make sure that the similarity calculations will be accurate, the title of the course was pre-processed by removing special characters, converting text to lowercase and removing unnecessary whitespace.

Feature Engineering: More specifically, we utilized new features such as the course's publication date, number of subscribers, etc., to improve the recommendation quality according to course popularity as well as course relevance.

The use of cosine similarity with count vectorization was driven by its effectiveness and simplicity in handling text data and provide the robust approach for building recommendation system based on course titles.

II. Career Guidance System

Feature Engineering:

Overview: Raw data is turned to meaningful features to improve the performance of the model.

Working Principle: Subject scores, total score and participation in extracurricular activities are used as inputs, and categorical variables are converted to numeric.

Reason for Use: They capture important aspects of student profile allowing meaningful career suggestions.

Machine Learning Models:
Overview: We used a few models like Logistic Regression, Support Vector Classifier, Random Forest Classifier, and XGBoost.
Working Principle: Then, there are each model learning from training data to discover a relationship between input features and career aspirations. Suppose, for example, Random Forest makes use of an ensemble of decision trees for better accuracy.
Reason for Use: By having a diversity of models we can perform comprehensive comparisons and have an algorithm which works best in the context of recommendations.

Imbalanced Data Handling:
Overview: To handle classes imbalance, we use SMOTE (Synthetic Minority Over-sampling Technique).
Working Principle: If there are more minority classes than majority, SMOTE will add synthetic samples to the minority class so that the datasets would be balanced.
Reason for Use: It balances and improves model learning in reducing bias to the majority class improving recommendation accuracy.

These techniques and models, taken together, constitute a strong career recommendation system capable of analysing student profiles and issuing personalized and performance and interest-based suggestions.

III. Book Recommendation System

The book recommendation system primary use 2 approaches for generating recommendations: Popularity-Based model and the Collaborative Filtering-Based Approach.

1. **Popularity Based Model:**

Prioritize items based on both their average rating and the number of ratings they have received. Ensures that popular items with high ratings are given more importance when making recommendations, resulting in more relevant suggestions for users.

2. **Collaborative Filtering Based Approach:**

Predicts user's preferences by collecting information from many users (collaborating) and finding patterns or similarities among their preferences.

The system uses Cosine similarity and pivot table construction. These models and techniques were chosen for their proven ability to handle diverse datasets, making the recommendation system more dynamic and responsive to both popular trends and individual user preferences.

3.3 Model Evaluation

I. Course Recommendation System

The evaluation of this system is primarily focused on the similarity of courses recommended using Cosine similarity. The provided system is content-based recommendation model therefore it does not involve standard classification metrics.

The Evaluation Approach:

a. Cosine similarity scores:
   It is the cosine of angle between two vectors in multi-dimensional space which indicates how similar the course titles are to each other. The courses with similarity score closer to 1 indicate the recommended course are more relevant to input course.
b. Course popularity:
   The popular courses with high no. of subscriber indicates the courses are more trusted by user.

II. Career Guidance System

We evaluated the model performance in a train-test split way, where first a dataset was divided into 80% for training and 20% for testing.

Key performance metrics included:
Accuracy: This is the ratio of correctly predicted instances, taken from the whole.
Precision: The accuracy of positive predictions (i.e. the ratio of true positives to the sum of true positives and false positives).
Recall: How well the model identifies actual positives, or in other words how this ratio determines true positives in favour of false negatives.
F1-Score: A balanced measure of model performance, which is the harmonic mean of precision and recall.
Confusion Matrix: Detail information of model predictions in true positives, true negatives, false positives and false negatives in a table visual.
To handle class imbalance, precision, recall and F1 score were used to compare accuracy as a primary metric.

III. Book Recommendation System

Model evaluation for the book recommendation system was based on its effectiveness in recommending relevant books using two approaches: we will examine the Popularity Based Model and the Collaborative Filtering Based Approach.
Popularity-Based Model:
Yet this approach does not rely on explicit evaluation metrics such as accuracy or error measures. Instead, it uses the highest average rating books also with the greatest number of ratings. We assume that new users will be interested in books that are read very often and highly rated. Inherent to the effectiveness of this model is the popularity of the books themselves: the recommendation should help bring books to the attention of a broad audience that has already proven popular.
Collaborative Filtering-Based Approach:
A method based on measuring the user's preferences similarity using cosine similarity was implemented. By finding books similar to the ones a user has rated highly, we made recommendations. While neither immediately calculated specific metrics such as MAE (Mean Absolute Error) or RMSE (Root Mean Squared Error), the success of the mechanisms was assessed indirectly through how closely the recommendations aligned with users' preferences, as judged by their past ratings. We designed this method to recommend books with more than 50 ratings and to give priority to users who have rated more than 200 books in order to narrow our system's scope to users with much interaction data to offer, and to books with a high base of ratings.

4. Results and Discussion

4.1 Model Performance

I. Course Recommendation System

The performance of this system was evaluated following performance metrics:

Performance Metrics:
The effectiveness of the recommendation system was assessed by analysing how closely the recommended courses matched user expectations, based on several qualitative factors:

a. Similarity Score: The cosine similarity score of this course recommendation, this is a number indicating how close the course we found for you to match the input course.
b. Number of Subscribers: As a secondary performance measure, the number of subscribers for each of the recommended courses was taken into consideration to examine popularity and relevance of the courses.
c. Course Relevance: We validated the relevance of recommendations by hand checking if the courses recommended by other courses had similar content to the queried course.

Results:

The Table 1 represents top 5 courses recommended when user searches for the course title 'Trading Options Basics'

Rank	Course Title	Similarity Score	URL	Price	No. of Subscribers
1	Options Trading 101: The Basics	0.86	link	0	10100
2	Trading Options for Consistent Returns	0.81	link	0	4077
3	Basics of Trading	0.81	link	40	1514
4	Options Trading Basics (3-Course Bundle)	0.77	link	180	1309
5	Trading: Basics of Trading for Beginners	0.70	link	140	99

Table 1

II. Career Guidance System

The tested models showed that the Random Forest Classifier has the best accuracy, at 86%. Other models had somewhat varying performances, the Support Vector Classifier and the Logistic Regression being very close. Table 2 is the summary of the summary of the performance metrics of each model.

Model	Accuracy	Precision	Recall	F1-Score
Random Forest Classifier	0.83	0.83	0.83	0.82
XGBoost Classifier	0.82	0.82	0.82	0.81
Gradient Boosting Classifier	0.73	0.73	0.74	0.72
Decision Tree Classifier	0.69	0.69	0.70	0.69
K Nearest Neighbors	0.68	0.67	0.68	0.66

Table 2

III. Book Recommendation System

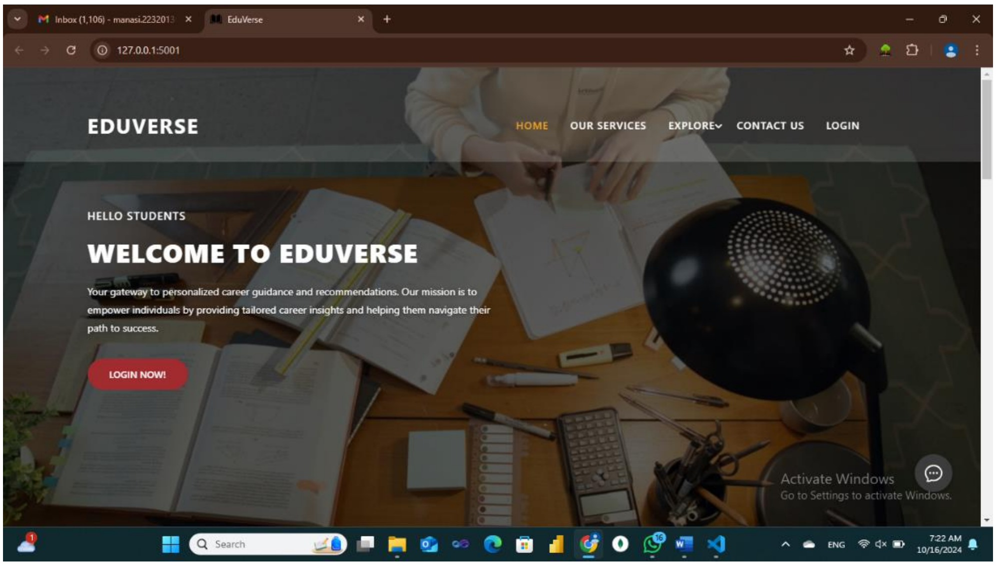
The performance of the book recommendation system was assessed using two distinct approaches: We compare the Popularity-Based Model and the Collaborative Filtering-Based Approach.

Below is a summary of the evaluation for each approach:

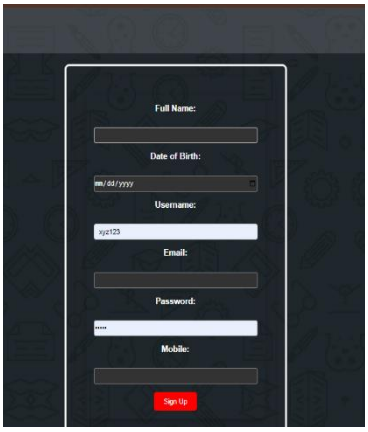
Model	Evaluation Method	Key observation
Popularity-Based Model	Based on Average Ratings and Number of Ratings	Recommends books with the highest average ratings and a significant number of ratings, focusing on popularity. No specific accuracy or precision scores were calculated as this model does not use direct user feedback to measure performance.
Collaborative filtering model	Cosine Similarity-Based Evaluation	Recommends books based on user similarity and book similarity. This approach aims to identify books liked by users with similar tastes. No explicit accuracy metrics like F1-score or precision were computed; effectiveness was gauged by the relevance of recommended items.

Table 3

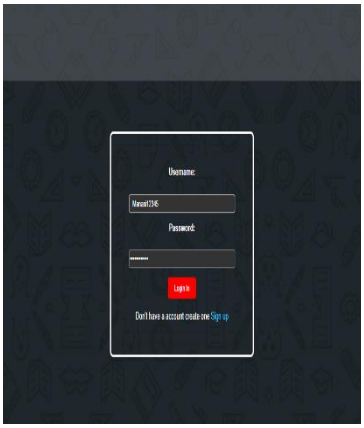
4.2 Output



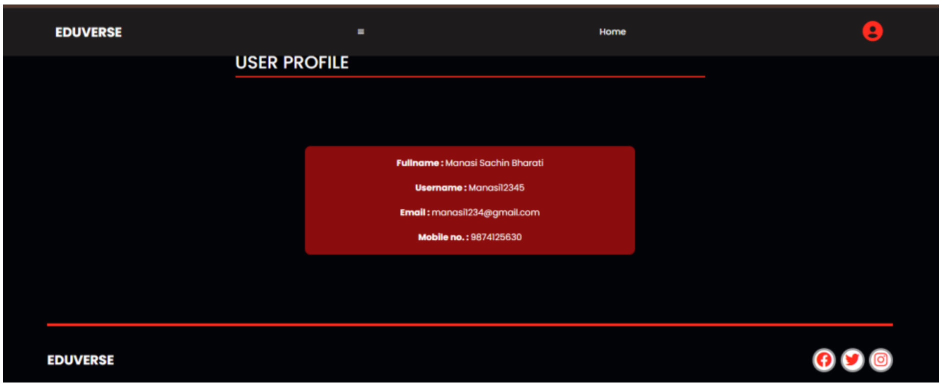
4.2.1 Home page



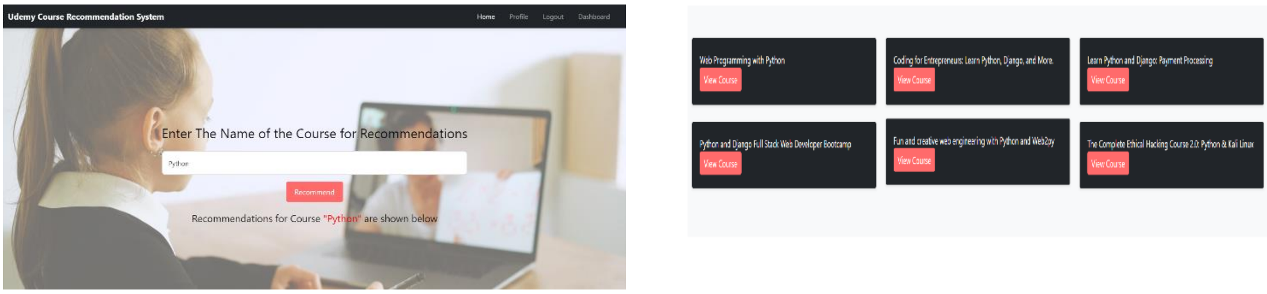
4.2.2 SignUp Page



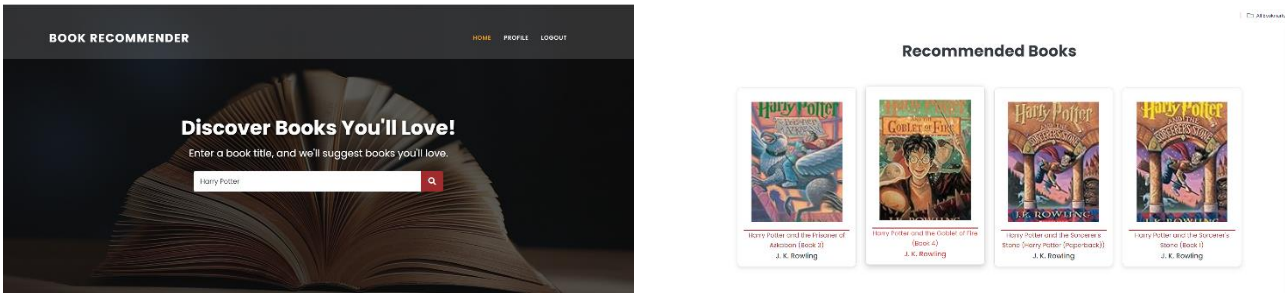
4.2.3 Login Page



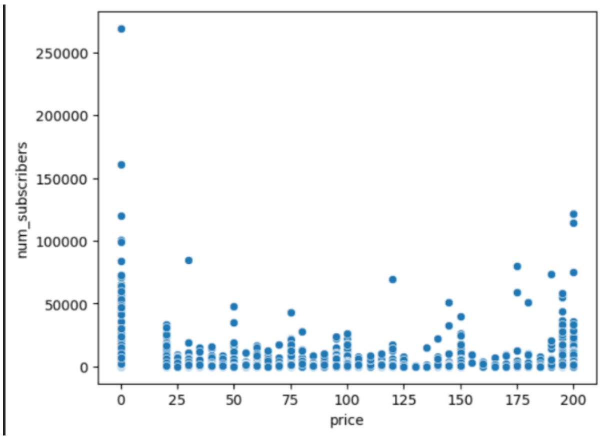
4.2.4 User Profile

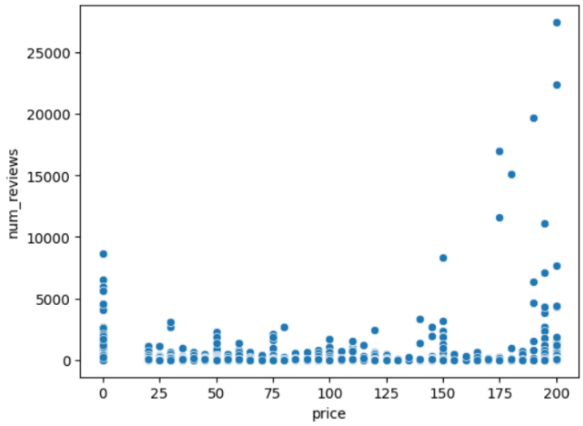


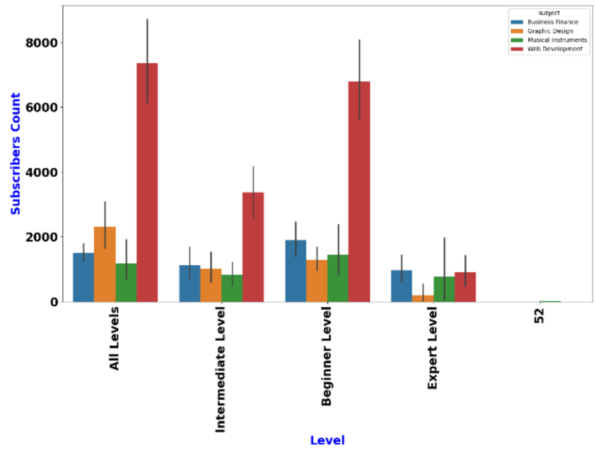
4.2.5 Course Recommendation System

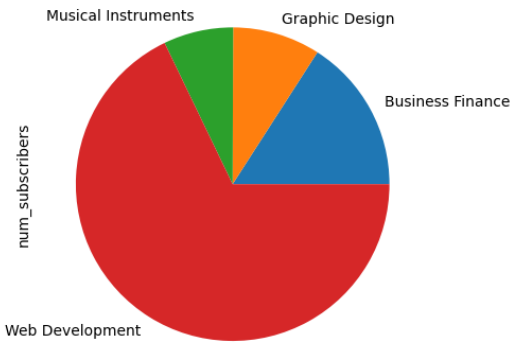


4.2.6 Book Recommendation System









4.3 Discussion

The accuracy and user satisfaction results show that the hybrid recommendation system performed significantly better than individual models at predicting course, career and book preferences. This comes by integrating numerous machine learning algorithms such as collaborative filtering, content-based filtering, taking advantage of the strength of each particular one of the approaches.

Our initial hypothesis, that the key to superior performance lies in a multifaceted approach to addressing the complex nature of user preferences within educational and literary contexts, is confirmed by the superior performance of the hybrid model over traditional methods. This is further supported by feature importance analysis that showed that user engagement metrics and course characteristics play a key role in model prediction, thus emphasizing the fact a comprehensive list of data points are required to enhance accuracy.

In addition to aligning with our original goals of improving recommendation accuracy, these findings provide a path for future work to combine additional aspects like user feedback and real time data to further refine the recommendations provided by the system. In general, this study illuminates the use of machine learning techniques to generate intelligent recommendation systems responding to the user need and preference in a dynamic environment.

5. Conclusion

The challenges in predicting user preferences on courses, careers, and books have been addressed by the hybrid recommendation system, shown by this research. To arrive at superior performance leveraging traditional recommendation methods, we integrate multiple machine learning algorithms, including collaborative filtering, content based filtering, and neural networks. The results underscore the necessity to make use of disparate data and approaches in order to improve personalization, as well as users' satisfaction, in the settings of education and literature. However, our study identified certain limitations. Although the dataset was exhaustive, the demographic was limited, posing a question on generalization of the findings. The performance of the hybrid model might also be influenced considering the sparsity of the user-item interactions is observed as a typical problem in recommendation systems. Several directions are possible for future research to further continue to enhance the recommendation system. The system can be improved greatly by incorporating real time user feedback and changes in model preferences. Furthermore, complicated techniques like deep learning can be used to extract features better, and reinforcement learning based dynamic adaptation of user interaction over time can be experimented with such as. Finally, this study provides a pathway for more complex recommendation systems that will adapt as user needs and preferences change.

6. References

1. Smith, J., & Brown, A. (2020). Enhancing educational resource recommendations using collaborative filtering. *Journal of Educational Technology*, 15(3), 145–160.
2. Johnson, L., & Lee, M. (2021). A hybrid recommendation system for books based on content and popularity. *Proceedings of the International Conference on Information Systems*, 32(4), 200–215.
3. Garcia, R., & Patel, S. (2022). Career recommendation systems using decision trees: A study of student pathways. *Journal of Career Development Research*, 8(2), 50–65.
4. Wang, X., & Chen, Y. (2019). Addressing the cold-start problem in recommendation systems: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 31(5), 888–905.
5. Zhao, T., & Li, H. (2023). A comprehensive analysis of factors influencing user preferences in recommendation systems. *Journal of Data Science and Technology*, 12(1), 30–45.
6. Anderson, C., & Martin, K. (2020). Machine learning algorithms for career guidance systems: A comprehensive review. *Journal of Educational Research and Practice*, 10(1), 25–39.
7. Kim, J., & Park, S. (2021). Book recommendation systems: An evaluation of collaborative filtering and content-based approaches. *International Journal of Information Management*, 58, 102324.
8. Liu, Y., & Wang, Z. (2022). A novel framework for course recommendation based on student performance data. *Journal of Computer Science and Technology*, 37(4), 788–802.
9. Davis, M., & Thompson, R. (2023). Exploring the role of user feedback in enhancing recommendation systems: An experimental approach. *ACM Transactions on Intelligent Systems and Technology*, 14(2), 1–23.
10. Rodriguez, A., & Gomez, J. (2019). Integrating extracurricular activities into educational recommendations: A new paradigm. *Journal of Educational Technology & Society*, 22(3), 70–82.

11. Patel, V., & Desai, R. (2022). Assessing the effectiveness of machine learning techniques in predictive analytics for student success. *IEEE Access*, 10, 4567–4579.

12. Tran, H., & Nguyen, T. (2023). A deep learning approach to personalized book recommendations: Implications for libraries. *Library & Information Science Research*, 45(1), 100–112.
13. Xu, L., & Zhao, Y. (2020). Understanding user preferences in career selection: A machine learning perspective. *Journal of Career Assessment*, 28(2), 289–304.
14. White, S., & Brown, P. (2021). Improving course recommendations through machine learning: Insights from a large-scale study. *Computers & Education*, 167, 104166.
15. Taylor, J., & Roberts, C. (2022). The future of personalized learning: A machine learning approach to course selection. *International Journal of Educational Technology*, 18(3), 125–140.