DSCI 599: Data Science for Business, Economics, and Society
Group Project Final Report

Trends in Online Courses: A Data-Driven Analysis

Yuheng Chen, Haoyue Xu, Jingyue Zhang

0. Team Members

Yuheng Chen, Haoyue Xu, Jingyue Zhang

1. Introduction

This project report analyzes online course data from Coursera, covering the years 2021 to 2024, to identify trends in course design and learner engagement. Utilizing machine learning techniques like clustering and regression models, we explored how course characteristics correlate with learner outcomes. Our methodology involved data collection from Kaggle, rigorous preprocessing, and analytical modeling. We used clustering to detect patterns in courses based on skills, and SHAP analysis in regression modeling to understand factors affecting course popularity. Additionally, we developed a personalized course recommendation system using Python and Streamlit, based on user preferences and our analytical insights. This report details our methods, findings, and the practical applications of our work, aiming to improve online education by linking data-driven insights to user-friendly applications.

2. Methodology

2.1 Dataset

Data was collected from Kaggle for years 2021, 2023 and 2024. Each dataset was obtained from online educational platforms Coursera, containing detailed listings of courses along with descriptions, skills, and ratings. Courses not specified in English were filtered out to maintain consistency in language for analysis.

2.2 Data Preprocessing

The preprocessing steps involved cleaning data, handling missing values, and normalizing entries for uniformity. For clustering, text data from listed skills underwent tokenization, lemmatization, and the removal of stopwords, including terms not relevant to content analysis such as "and". For the regression model and SHAP part, since the dataset consisted of features including course difficulty, type, duration, and the offering organization. We preprocessed the data by: Encoding categorical variables using one-hot encoding. Handling missing values through imputation. Normalizing the numerical features to ensure they were on the same scale.

2.3 Clustering

Features were extracted using TF-IDF vectorization, we converted text data from course skills into a numerical format that could be effectively used for clustering algorithms. Truncated SVD was employed to reduce the high-dimensional TF-IDF vectors to a more manageable size while retaining the most informative aspects of the data. KMeans clustering was utilized to group courses into clusters based on skill similarity. The optimal number of clusters was determined using the silhouette score method and also ensure meaningful segmentation.

2.4 CatBoostRegressor and SHAP

We employed the CatBoostRegressor machine learning model due to its robust handling of categorical features and implemented SHAP (SHapley Additive exPlanations) for interpretability. The model performance was evaluated using the Root Mean Square Error (RMSE), with scores indicating a high predictive accuracy.

After obtaining the model results, we employed SHAP (SHapley Additive exPlanations) to interpret the CatBoostRegressor model.

SHAP values provide insights into how each feature contributes to each prediction in the context of the model. It decomposes a prediction into the sum of effects from each feature, allowing us to understand the impact of a single feature while considering the interaction with other features.

3. Results with Analysis

As we transition to the results section of our analysis, our aim is to illuminate the overarching trends within online courses, particularly emphasizing the distinct patterns emerging from technical and humanities disciplines. Our combined methodologies using K-Means clustering and CatBoostRegressor with SHAP explanations have enabled us to explore these fields in depth, identifying unique preferences and behaviors among learners engaging with these two broad categories of subjects. The forthcoming results will delineate how user engagement and course offerings differ significantly across technical and humanities courses, offering insights into the divergent dynamics that characterize these educational paths. This analysis is crucial for understanding the current educational landscape and for guiding the development of future online courses that meet the varied needs of learners interested in both technological skill acquisition and humanities-based critical thinking.

3.1 Clustering

The plots illustrating the clustering results and tables listing the top 5 terms for each cluster are included in the Appendix A.

In 2021, the silhouette score of 0.524 suggested moderate separation between clusters. The curriculum includes a broad spectrum of academic disciplines and skills, including life sciences, education, arts, and foundational computer science. This diversity indicated a wide-ranging approach to course topics, catering to a varied set of academic interests and basic skill development. Progressing to 2023, the silhouette score is 0.618, indicating enhanced separation between clusters. Courses during this period began to exhibit a more defined focus on professional and technical skills, such as cloud computing, SQL, data analysis, and machine learning. This shift underscores a trend towards specialization, aligning course offerings more closely with practical and professional skill requirements that respond to changing job market needs. By 2024, the silhouette score further increased to 0.778, clearly demonstrating the distinct nature of each cluster. The analysis of cluster terms from this year revealed a strong emphasis on advanced technological skills, including software development, project management, and data science, particularly using TensorFlow. This marked progression reflects the industry's increasing demand for high-level technical expertise and the rapid pace of technological advancements impacting educational content. From a comparative perspective, there was a notable shift over the years from general academic topics to specific, technology-oriented skills, mirroring the growing necessity for digital and technical competencies in the contemporary job market.

Despite the technical focus, an enduring emphasis on management and leadership skills was evident, highlighting the sustained demand for these competencies in professional development. This suggests that while technical proficiency is increasingly prioritized, complementary skills in leadership and management remain invaluable.

The transition from traditional academic skills to cutting-edge technical skills suggests that online educational platforms are swiftly adapting to meet market demands and technological evolutions. Additionally, blending technical training with management and communication skills could provide a holistic educational approach. Regular analysis of these trends is advised to continuously adapt to and anticipate future educational demands.

3.2 SHAP RESULT

The CatBoostRegressor was chosen for its performance with categorical data and its gradient boosting capabilities. The split is done with a 50% test size, a random seed of 0, and stratification based on the selected column. And we initialize the training and testing data pools using CatBoost's Pool class to create a training data pool and a testing data pool with categorical features.The model yielded an RMSE score of 0.181 points for the training set and

0.153 points for the test set, indicating the model's effective learning and the model is highly predictive and generalizes well. We will display the visualizations in **Appendix B**.

We initialized the SHAP values for visualizations (see **Appendix B - Figure 1**) and created a TreeExplainer object for the model. After calculating the SHAP values for the test data using the TreeExplainer, we generated a summary plot to visualize the impact of features on model predictions. The prominence of programming skills, particularly in Python, underscores the high demand for courses that are technical in nature. This trend reflects the real-world market demand for programming skills and suggests that there could be benefits from a greater focus on or expansion of technical offerings. Features like **Duration** and **Type** indicate that the length and format of courses are significant factors for students. Courses with an optimal duration and type that match learners' preferences are more likely to succeed, indicating a trend towards flexible, yet comprehensive learning experiences. Skills related to **Data Analysis**, **Machine Learning**, and **Cloud Computing** feature prominently, suggesting a trend towards courses that provide practical and immediately applicable skills that are relevant in today's job market.While technical skills dominate, the presence of features like **Communication skills** and **Leadership** and **Management skills** indicates a balanced interest and the importance of soft skills for professional development.

We also created a bar chart with error bars to visualize SHAP values**(**see **Appendix B - Figure 2)**. The SHAP value plot indicates a strong positive correlation between technical skills and course ratings. Features such as **Skills_Python_Programming**, **Skills_Data_Analysis**, and **Skills_Machine_Learning** exhibit lengthy red bars, signifying their potent influence in elevating course ratings. This positive association suggests that courses which impart technical skills are in high demand and are perceived as providing valuable career-advancing knowledge, thus receiving higher ratings from learners. Conversely, humanities-related skills such as **Skills_Communication** and **Skills_Leadership_and_Management** show shorter bars on the SHAP plot, reflecting a more subdued influence on course ratings. While these skills are essential for a well-rounded skill set, their impact on course ratings is less pronounced, which may reflect a current learner preference towards more technical, marketable skills in the online education sphere.

4. Application

After performing a cluster analysis on our dataset, we gained valuable insights into the different categories of courses offered by Coursera. These clusters reveal the skills included in course content, allowing us to better understand the online learning landscape and trends. Furthermore, our SHAP analysis shed light on the most influential features driving course popularity and user engagement. By predicting course ratings on Coursera, we can uncover valuable trends that not only guide course creators but also reveal broader patterns in online learning preferences.

Armed with these insights, we recognized an opportunity to leverage our findings to assist Coursera in making data-driven decisions. We envisioned a course recommendation system that could suggest relevant courses to users based on their input requirements, taking into account the trends and patterns uncovered through our clustering and SHAP analysis. By doing so, we aimed to not only improve the user experience by providing personalized recommendations but also to showcase how our analytical work could be translated into a practical application.

To bring this idea to life, we implemented the application using Python and the Streamlit library for creating the user interface. The application allows users to input their desired skills, preferred course type and difficulty level. It then processes the user's input and matches it against a dataset of courses to provide personalized recommendations.

4.1 Implementation

The implementation of our project involved several critical steps, beginning with data preparation. We utilized the "coursera_course_2024.csv" dataset, which includes comprehensive details on various courses, such as titles, descriptions, skills covered, difficulty levels, and other pertinent information. Following this, we focused on skill extraction by identifying all unique skills within the dataset and calculating their frequencies. We then prioritized the top skills that represented 90% of the total occurrences, allowing us to concentrate on the skills that are most relevant and in demand.

Next, the application engages users by prompting them to enter their desired skills, preferred course format, and difficulty level. To facilitate user input, menus were created for the course format and difficulty level, utilizing the unique values identified in the dataset. We then employed OpenAI's GPT-3.5-turbo model to perform keyword matching. This model extracts relevant keywords from the user's input and matches them against the available skills, course types, and difficulty levels in the dataset. It returns a dictionary containing the matched keywords, which are essential for the next step.

Course filtering is then executed based on the matched keywords. We constructed a query to filter the dataset effectively, retrieving courses that align with the user's specific skills, type, and difficulty. The query conditions are dynamically generated based on user input to ensure precision in the course selection process. Once the relevant courses are identified, they are sorted by review count in descending order to prioritize the most popular ones.

Finally, the top recommended courses are presented to the user. Each course listing includes detailed information such as the course title, number of enrolled students, review count, ratings, description, skills covered, difficulty level, course type, and duration. Additionally, a link to the course website is provided for users who wish to explore further. This comprehensive approach not only tailors the search to individual user preferences but also enhances the likelihood of satisfying their educational needs.

The screenshot in **Appendix C** demonstrates our recommendation system.

5. Learning Experience

Throughout our project, we encountered various challenges and limitations.

To analyze the trends in online learning during and after the COVID pandemic, we focused on data from Coursera, a leading online platform. Since Coursera doesn't publicly share their data, we spent considerable time finding suitable datasets online. These datasets were well-structured but had skills information in varying formats, making it hard to analyze directly. Initially, our clustering attempts using the KMeans algorithm gave us low silhouette scores, suggesting that our data had too much noise.

To improve our results, we decided to transform the data. We used lemmatization, a process that groups different forms of the same word together. This helps the algorithm treat "run," "running," and "ran" as the same word, for example. Additionally, we removed common words, known as stopwords, which don't add much information to the analysis. These steps helped reduce noise and significantly improved our clustering results, giving us clearer insights into online study trends.

While working with the CatBoostRegressor model, selecting appropriate vectorization parameters, such as min_df and max_features for CountVectorizer, had a significant impact on the quality of the final features and the model's performance. Improper parameter settings could result in the loss of important information or the generation of too many irrelevant features. When using RareLabelEncoder to handle rare labels, setting the n_categories and tol parameters also posed a challenge, as inappropriate thresholds might cause too many categories to be classified as "Other," obscuring potentially useful information.

In terms of model training, although CatBoost has good resistance to overfitting, it can still occur when there are numerous features or insufficient data. Therefore, careful tuning of model parameters and the use of appropriate cross-validation methods are necessary to monitor and avoid overfitting. Ultimately, the model's effectiveness heavily relies on the quality and representativeness of the input data. If the training data does not cover the prediction scenarios well, the model's generalization ability may be limited. Furthermore, any biases introduced during the data preprocessing stage directly affect the model's output and the interpretation of SHAP values, which is an ongoing concern that we need to address throughout the project's implementation.

Applying our findings to develop a user-friendly recommendation system presented its own set of challenges. Matching user input with course metadata is a complex task. We leveraged NLP

techniques, such as keyword extraction and semantic similarity, which opened up new possibilities for understanding user intent and matching it with relevant courses. We had to experiment with different approaches, including exact keyword matching and semantic similarity, to obtain accurate results. Additionally, striking a balance between recommending highly relevant courses and promoting diversity in suggestions was challenging. We had to fine-tune our recommendation algorithm to ensure a mix of popular and niche courses. Moreover, developing an interactive web application allowed us to design with a user-centric approach. By utilizing Streamlit, we created an intuitive interface and clear communication with minimal code, making our recommendation system accessible and engaging for users.

However, due to the lack of more user preference data, some common approaches in recommendation systems, such as collaborative filtering, were not feasible. To provide better recommendations to users, we need more indicators of individual user preferences, such as ratings and viewing data. Furthermore, when making future recommendations, we should prioritize content that aligns more with the trends identified through our clustering and SHAP analysis. Our main objective is to ensure that our analysis brings more possibilities for application, rather than remaining solely on paper. It should lead to the optimization of the user experience and the growth of our user base in the future.

6. Conclusions

Our study provides a comprehensive analysis of the course offerings on Coursera from 2021 to 2024. Utilizing advanced clustering and regression techniques, we were able to detect significant trends in course characteristics and learner preferences, informing strategic course development and personalization of learning pathways.

The clustering analysis revealed distinct groupings of courses that reflect the shifting focus from general academic skills to more specialized, technology-oriented skills. This transition underscores a growing demand for digital proficiency in today's job market, matched by an equally persistent need for leadership and management skills. For learners, the findings offer guidance on what skills to develop for future-proof careers. For educators and online platforms, the analysis serves as a compass for curriculum development and the strategic planning of course offerings.

From the learner preference perspective,The SHAP value analysis reveals a definitive learner preference for courses on Coursera. Students favor beginner-level, professionally certified courses that fall within a one to three-month duration. This trend indicates that learners are seeking accessible, practical skills with recognizable credentials that can be achieved within a manageable time frame. However, there's a notable disinclination towards courses that are perceived as too brief or too demanding, such as guided projects or courses labeled as intermediate or advanced in difficulty.

For Coursera, the strategic implication is to prioritize the development and promotion of courses that align with these learner preferences. The data suggests an optimal course design that

balances comprehensibility with professional applicability. In response to the negative SHAP values associated with certain course types and difficulty levels, Coursera should consider enhancing course support structures, tailoring content to meet diverse needs, and setting clear expectations to improve overall course ratings.

Appendix A

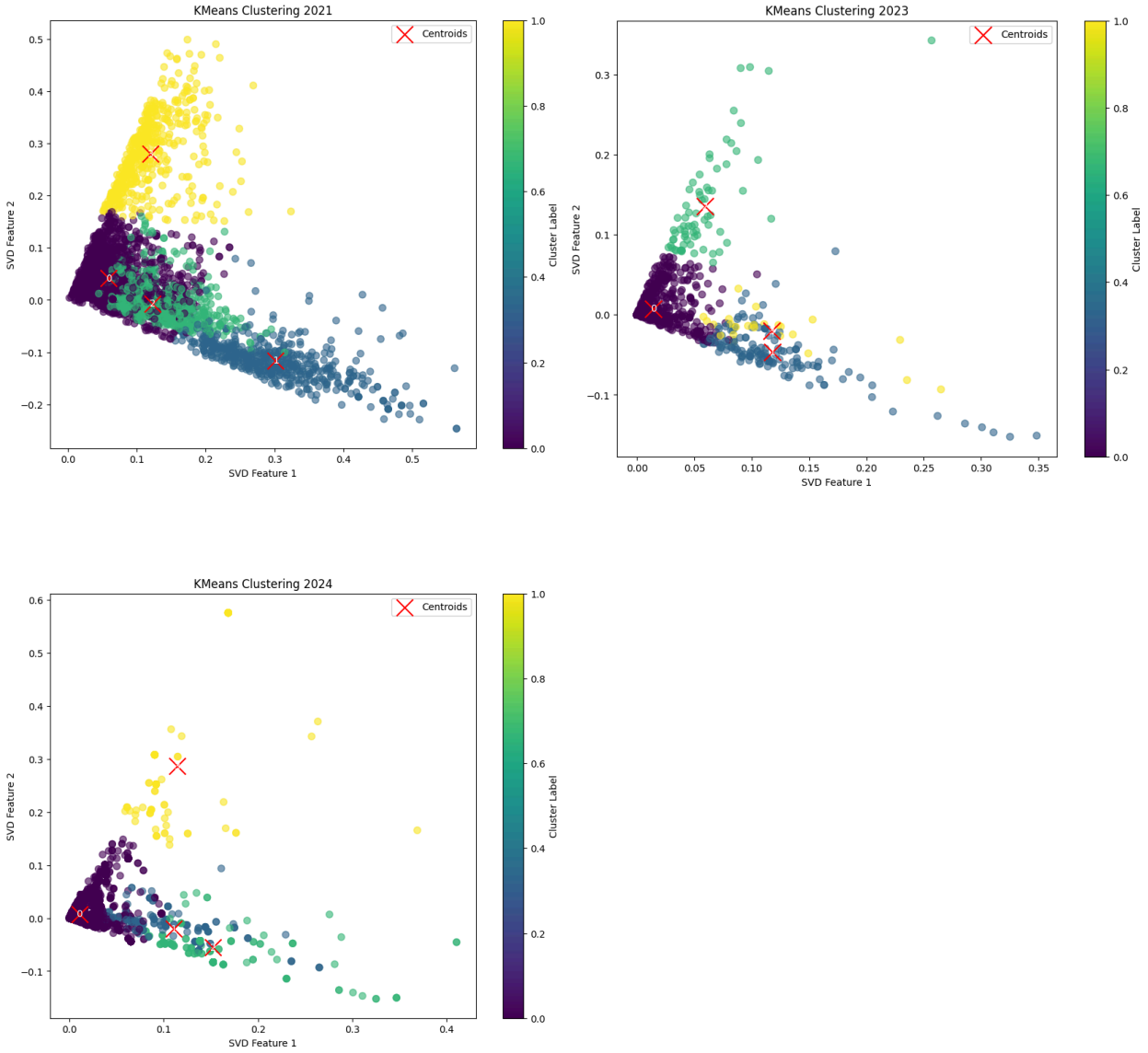


Figure 1: KMeans Clustering 2021, 2023, 2024

Table 1: Cluster Top Terms for 2021

Cluster	Top Terms
0	Sciences, Life, Social, Education, Arts
1	Data, Analysis, Machine, Learning, Science
2	Computer, Development, Science, Engineering, Software
3	Business, Leadership, Strategy, Management, Leadership_Management

Table 2: Cluster Top Terms for 2023

Cluster	Top Terms
0	Cloud Computing, Linux, SQL, Marketing, Problem Solving
1	Data Analysis, Python Programming, Machine Learning, Data Science, SQL
2	Communication, Leadership, Management, Project Management, Negotiation
3	Design, Programming, Web, Development, JavaScript

Table 3: Cluster Top Terms for 2024

Cluster	Top Terms
0	Data Science, Project Management, Communication, Statistics, TensorFlow
1	Computer Programming, Programming, Software, Design, Development
2	Data Analysis, Python Programming, Data, Programming, Making
3	Management, Business, Leadership, Making, Decision

Figure 2: Cluster Top Terms 2021, 2023, 2024

Appendix B

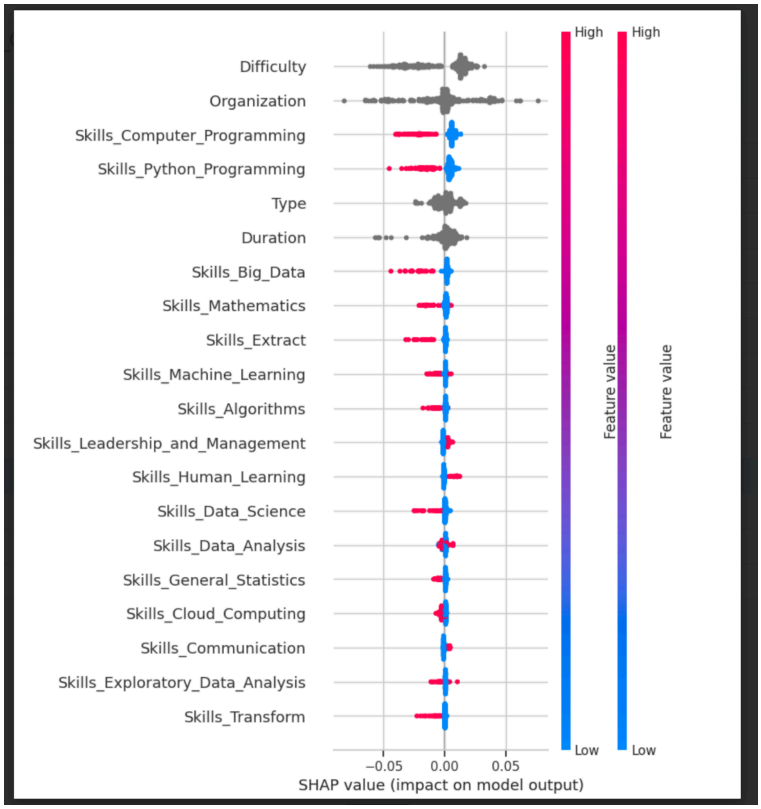


Figure 1: shap value (impact on model output)

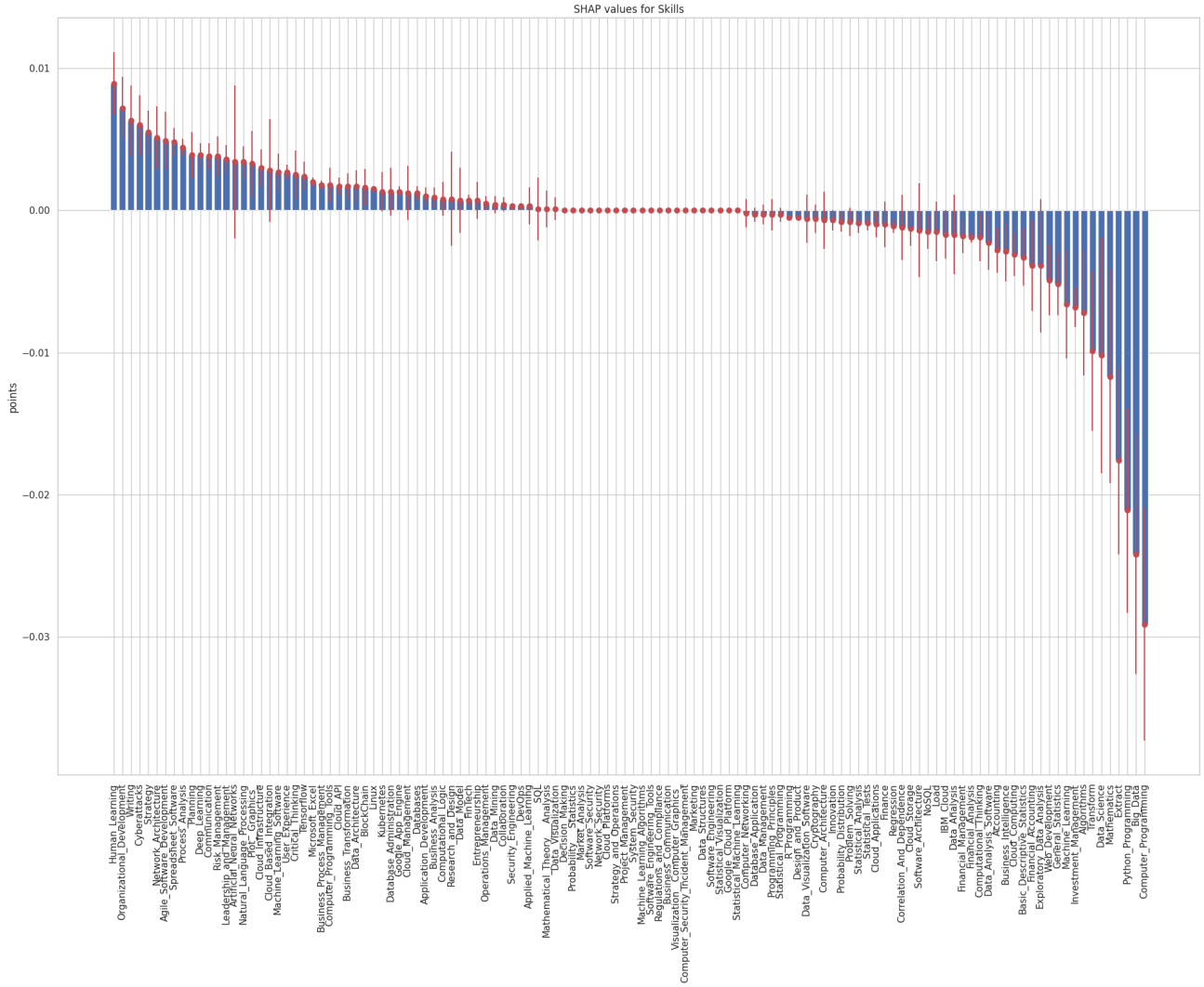
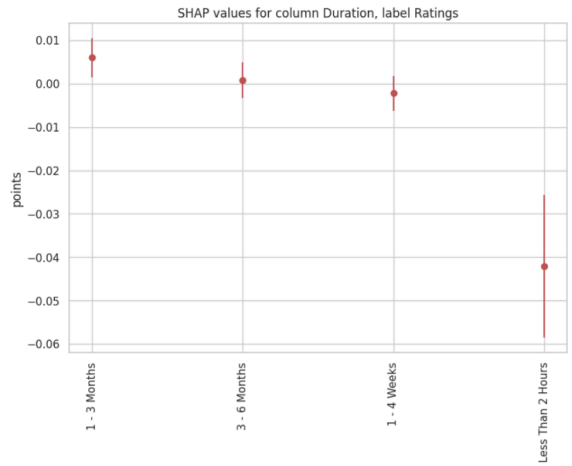
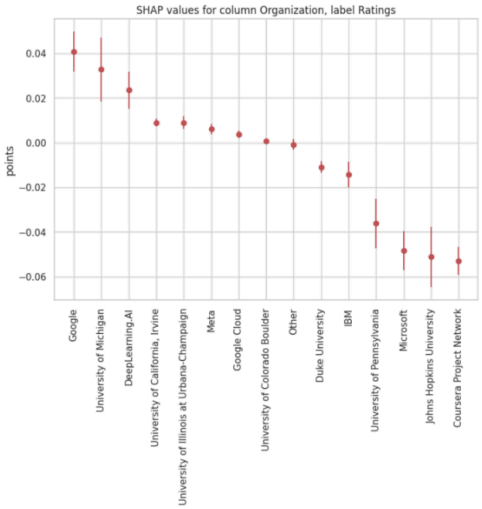


Figure 2: SHAP values for each skills

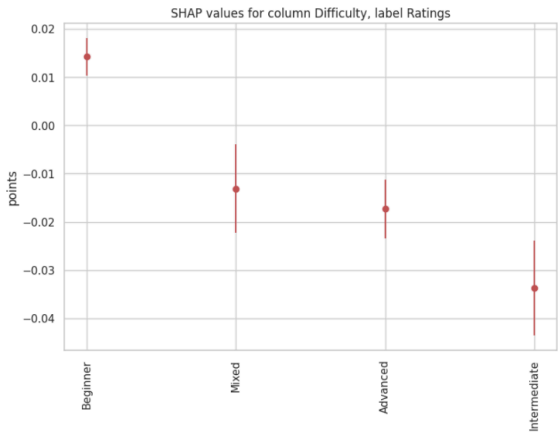


col	gain	gain_std	count
1 – 3 Months	0.0060	0.0045	96
3 – 6 Months	0.0008	0.0041	142
1 – 4 Weeks	−0.0022	0.0040	65
Less Than 2 Hours	−0.0421	0.0164	9



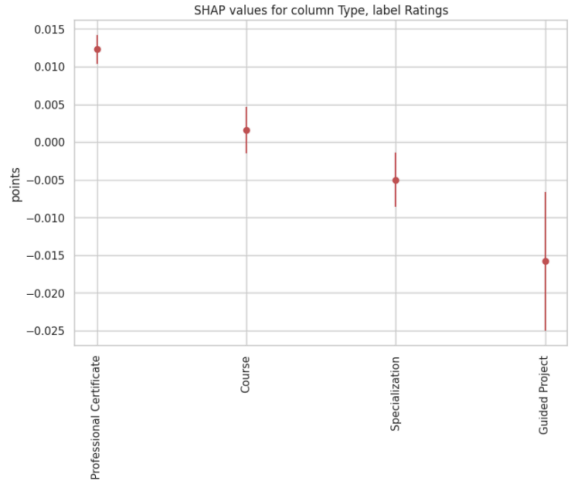
col	gain	gain_std	count
Google	0.0408	0.0090	33
University of Michigan	0.0328	0.0143	8
DeepLearning.AI	0.0235	0.0084	12
University of California, Irvine	0.0090	0.0018	5
University of Illinois at Urbana-Champaign	0.0090	0.0029	7
Meta	0.0061	0.0023	5
Google Cloud	0.0037	0.0017	34
University of Colorado Boulder	0.0007	0.0010	7
Other	−0.0008	0.0023	120
Duke University	−0.0109	0.0025	9
IBM	−0.0142	0.0057	35
University of Pennsylvania	−0.0362	0.0110	11
Microsoft	−0.0484	0.0088	6
Johns Hopkins University	−0.0512	0.0135	11
Coursera Project Network	−0.0531	0.0063	9

Difficulty



col	gain	gain_std	count
Beginner	0.0142	0.0039	202
Mixed	−0.0131	0.0092	20
Advanced	−0.0173	0.0061	13
Intermediate	−0.0337	0.0098	77

Type



col	gain	gain_std	count
Professional Certificate	0.0123	0.0019	42
Course	0.0016	0.0031	148
Specialization	−0.0050	0.0036	113
Guided Project	−0.0158	0.0092	9

Figure 3: SHAPE values for duration, organization, difficulty, and type

Appendix C

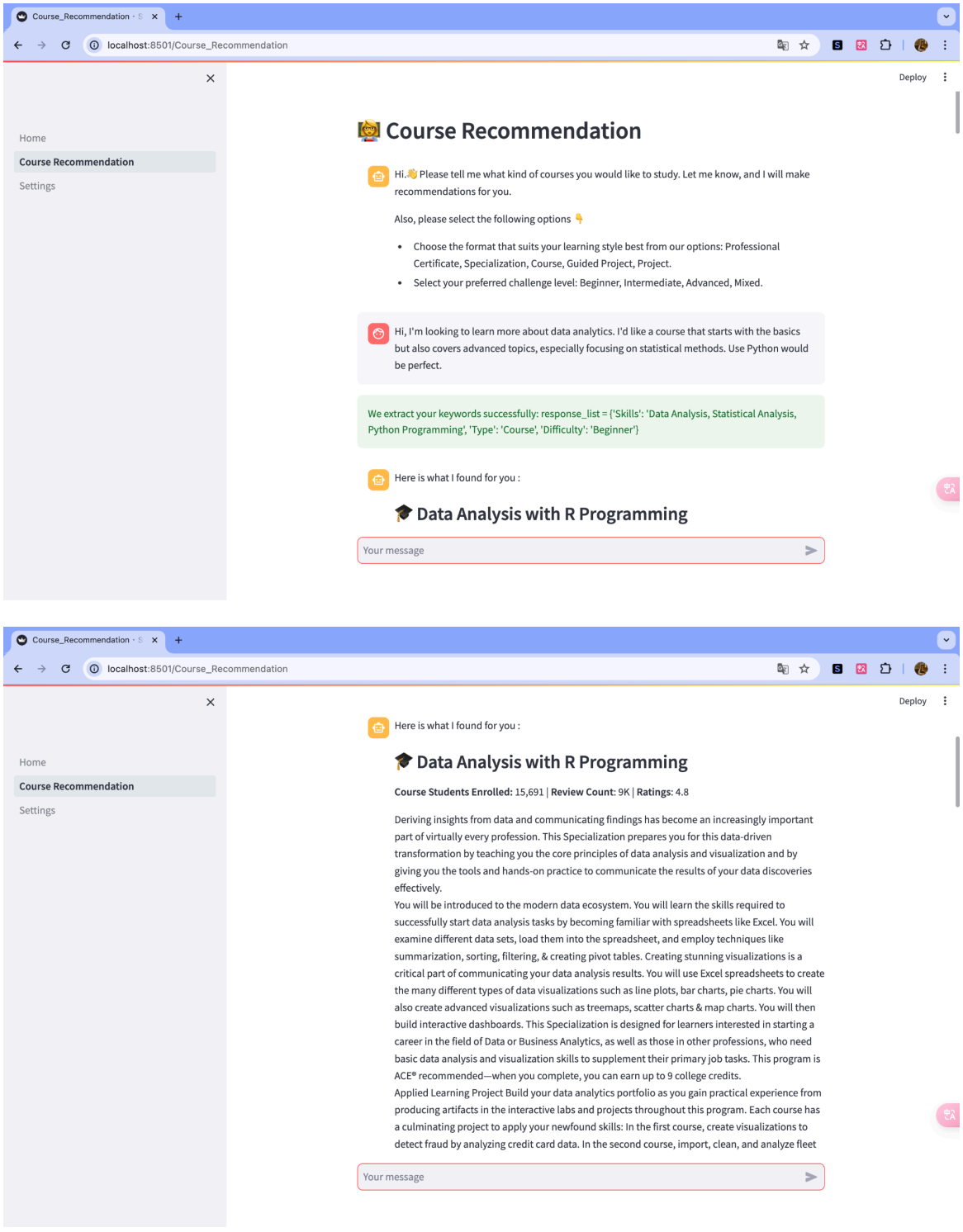


Figure 1: Screenshot of Course Recommendation System

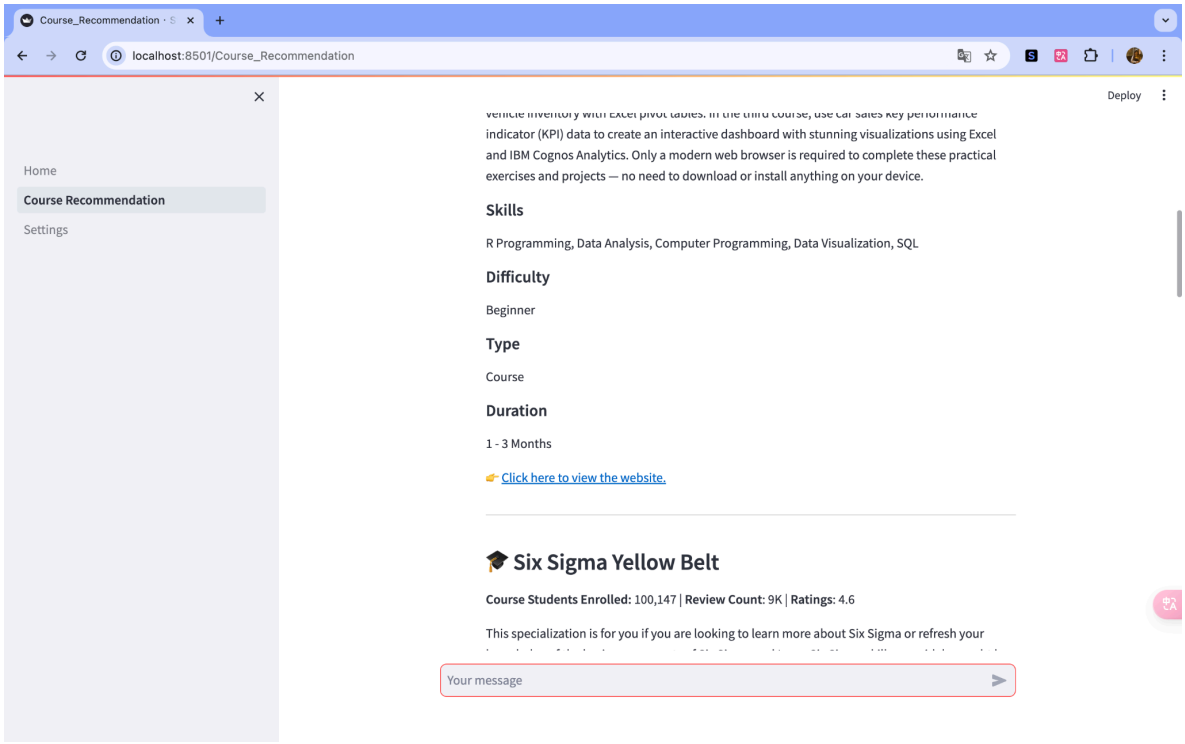


Figure 2: Screenshot of Course Recommendation System (Cont.)