ANALYSIS OF COURSE USAGE WITH HISTORY DATASET AND RATING PREDICTION USING MACHINE LEARNING ALGORITHMS

S Rohith
Department of Data Science and Business Systems SRM Institute of Science and Technology, Kattankulathur Chennai, Tamil Nadu 603203, India ,**rs2715@srmist.edu.in**

B Prasanth
Department of Data Science and Business Systems SRM Institute of Science and Technology, Kattankulathur Chennai, Tamil Nadu 603203, India **pb5385@srmist.edu.in**

Mugash priyan U
Department of Data Science and Business Systems SRM Institute of Science and Technology, Kattankulathur Chennai, Tamil Nadu 603203, India ,**mu3975@srmist.edu.in**

Abstract—**This dataset is an extensive collection comprising 10,000 online courses that have been accessed across many online learning platforms, with information given on course types, length of courses, number of students enrolled, completion rate of the courses, distribution of the platforms, price ranges and ratings. The courses can be broken down into nine broad headings, 'office tools', 'technology', and 'business' being some of the most popular. The length of these courses varies with some lasting ten hours and at the most one-hundred hours with the mean lasting for fifty-five hours. Even the enrollment figures vary greatly starting from a hundred and one to five thousand students in a single course where twenty-five hundred students on average are enrolled. The overall completion of the courses stands at 75%, however rate of completing an individual course may range from about 50% to 100%. Such courses are available on Coursera, edX, Learning LinkedIn, Udemy and other but Udemy has the highest number of users out of them all. The costs of the courses ranges from about 10 to 200 dollars with the mean being 106 while the feedback tends to margins around four on five stars which is a pretty good feedback from the users. This dataset is indicative of trends in online education and helps examine the determinants of course demand, completion and prices charged for the courses on different platforms..**

I. Introduction

The informing concepts that make up the dataset are indeed representative of the fast-changing state and trends of online education deliveries. The wide range of subjects, which are tailored for individuals with different preferences and objectives in mind can also be evidenced in the data. There are about 9 classifications which cover such topics as 'Office tools' or 'Technology', 'Business', and allite which indicates the growing need for technical and basic workplace skills requirements too. Since many people are seeking online solutions to their learning needs, especially where they are provided with self-study options, this dataset seeks to examine the versa of the courses offered and the existing market gaps in employment and learning tendencies.

The other aspect of the design of the courses is the existence of different lengths of courses with the minimum period being about 10 hours and the maximum being about 100 hours with the mean duration being about 55 hours. This enables learners to select appropriate courses depending on the time they are ready to spend and the level of understanding they would like to achieve. The enrollments of the courses varied between 101 and 5,000 students, this points to the demand for certain courses and perhaps indicates the course content or its marketing with regards to the specific skills taught. In addition, 75 % completion rate provides information on how students engaged with the course content and how much of it they retained; some reasons for high completion rates, for instance, are effective content or highly motivated learners, contrary, lower rates are indicators of something that can be rectified in the course prowadzenie or development.

The dataset is also effective in elaborating on the role played by various educational sites in enhancing the online learning process. Platforms such as Coursera, edX, LinkedIn Learning, and Udemy all have different user bases and offerings, with Udemy being the most populous in this study. Prices are also expectedly different, with ranges between 10 and 200 dollar courses, with 106 dollars being the mean rate. It hints at varying approaches towards pricing, perhaps due to the type of course, reputation of the site in question or even the tutor concerned. User feedback ratings, which tend to average 4 stars, represent a clear measure of how satisfied learners are with a given service or,

the quality they attach to it which further enriches the understanding of factors that make a course attractive or useful to its target audience.

There are also many possible ways to analyze this data set, such as studies of particular courses offered by certain platforms, or connections between price, rating, completion rate and other factors and the growth of distance education. These information are useful for researchers, educators or persons who have interest in this industry and would like to take data-informed measures to improve course development, engagement, and progression of the digital students. Understanding such parameters as what influences popularity of certain online courses is possible thanks to the data provided by this content-rich development.

II. LITERATURE REVIEW

As far as the history of science is concerned, the online learning system has developed rapidly in the last ten years, courtesy of technology, widespread internet connection, and increase in the number of people who seek for a customizable means of education. Different dimensions of e-learning have been examined with reference to the existing literature whereby, factors related to the learners' willingness to engage in learning, their willingness to complete learning as well as the impact of various intervention strategies have been evaluated. Allen and Seaman (2017) confirm that the rise of online learning in higher education has been conceivably accepted by many institutions because it has been used as a mode of delivering education and outreach services to those students who would otherwise have been considered non-traditional, offering them flexible learning. With the increase of online education, the concern for the quality of courses and the design of the services offered to the learner has grown in order to maintain high completion and high learner satisfaction (Lee & Choi, 2011).

One area of research that is particularly notable drew on relevant literature about the issue of course completion rates which has remained a thorn in the flesh of course management systems designed for e-learning. Research by Kizilcec et al. (2013) also provides evidence for the suggestion that numerous aspects, including but not limited to the duration of the course, the way the content is presented and the motivation of the learner, can contribute to high rates of dropout among learners who enroll in Massive Open Online Courses or MOOCs. This implies that such variables as course length and course activities may affect students' dedication towards the completion of a course, which is supported by course length and completion rates in the current data. Moreover, engagement techniques which include interactive content, feedback, and social interaction with others have been associated with higher completion and satisfaction rates, with learners (Hew, 2016). These techniques emphasize the role of course development in making sure that online courses are able to capture and maintain the attention of the learner.

The pricing structure is yet another vital factor regarding the number of people taking up a course as well as an issue of course availability. It has been established that learner price sensitivity is not constant, and depends on the characteristics of the learner, the value he/she attributes to the course, or the level of the course's marketers (Su & Hurd, 2021). For example, Coursera and edX type of platforms work with some of the leading universities, which can result in higher charges from learners. On the other hand, Udemy courses are likely cheaper in the sense that the courses are on a wider range hence different people of different classes can take up the courses. Yuan and Powell (2013), point out that due to the fact that competitive pricing of some of the platforms enhances their global reach, this has led to their increased usage since they have made the courses offered to the public at lower prices. This explains the observed differences in pricing in the dataset and further indicates that the price of a course can affect the number of people getting enrolled as well as the number of those who complete the course.

Another pertinent aspect addresses how platform reputation and users' ratings contribute to the success of the course. User reviews and ratings are particularly important for prospective students as they reflect the quality of the course, the competency of the instructor, and the overall experience (Bishop & Verleger, 2013). Studies by Kashif Khalil & Ebner (2014) also show that more enrollments are seen in higher-rated courses and more learnt enrollees complete the course. This has been attributed to social proof where learners make decisions based on what others have done. Online learning platforms, on the other hand, have established a reputation, for instance, LinkedIn Learning and Coursera, in providing course even of better quality, often with reputable organizations enhancing learners' trust, and yielding more subscriptions. This phenomenon can be observed in the dataset, where it can be noted that the rated average of courses available on highly reputed platforms is higher proving the effect of platform reputation on the courses offered within them.

To summarize the literature, coursework completion, cost of courses, and the reputation of the provider are determinants of learner involvement and the effectiveness of a course in online education. With this examination of these elements in the dataset, the study aims to extend the existing body of research to answer the question how online courses can be made more learner-friendly. Moreover, these findings can help explain the competitive dynamics of the industry of online education and inform the platforms on how to better their courses and prices to win more learners.

III. PROPOSED METHODOLOGY

In order to effectively evaluate the database and provide appropriate solutions regarding E-learning courses, the indicated procedure includes a number of stages. These stages are meant to examine the possible dependencies between different characteristics of a course, such as its length, number of participants, completion rates, cost, study platform and user review. Trends that may increase or undermine the interest of users in any of the offered course in a positive or negative way have also to be considered. The methodology will include the investigation of those interrelations and the development of relevant recommendations concerning the courses and their delivery systems by means of data cleansing, exploratory data analysis, and statistical modeling.

Data Cleaning and Preprocessing

To begin with, the cleaning and preprocessing of the data in the dataset will be done in order to ensure that the results produced in the later analyses are accurate and reliable. This stage comprises the maintenance of missing information, data format standardization, and the detection and correction of data errors. Cases of duplicate entries will be recorded and eliminated where they are found. Data quality will be assessed and any necessary transformations made to the data in readiness for the analysis stage which follows. Qualitative variables of interest (e.g. type of course, type of platform) will be transformed into numerical values, and quantitative variables will be normalized if required.

Exploratory Data Analysis (EDA)

EDA will be performed to explore the basic structure and distribution of the dataset under consideration. Summary statistics will be computed for the metrics of interest, e.g., enrollment, duration, completion rate, price, ratings. In this context, different graphs such as histograms, box plots and scatterplots will be used for the purpose of detecting patterns and relationships in the data. For example, the data on ratings and enrollment will be examined for various platforms and sites in an effort to infer about the popularity of these platforms and the preferences of learners. Finally, EDA will also look for any relationships between course-related variables. For example: the price of courses and the number of enrollments, duration of courses and their completion rates, etc.

Hypothesis Testing

When it becomes critical to measure the strength of the observed effects, hypothesis testing will be performed. For instance.

Platform vs. Enrollment: How significant are enrollment numbers on different platforms, and are certain platforms more suitable for onboarding learners than others, will be addressed by conducting hypothesis tests.

Price vs. Completion Rate: Research will test the validity of the course prices in accordance with the completion rates and will help to understand the effects of such prices on the users' active involvement in the course.

Category vs. Rating: This test will check if there are any differences in course ratings across different category levels, which will assist in determining how satisfied learners may be in different subjects.

Analysis of Correlation and Regression

As part of the analysis expected in this research, correlation analysis shall be performed with the aim of establishing the relatedness of some key variables such as duration, enrollment, completion rate, and ratings in terms of the strength and direction of the relationships. Following this, regression models will be used to express such relationships in quantitative terms. Finally, a Multiple regression Analysis would be conducted to assess which factors among the following affect completion rates and ratings the most, including course length, enrollment, cost, and platform, among others. Linear regression may apply if the associations are found to be linear, otherwise nonlinear type of regression or logistic regression may be employed depending on the nature of the data.

Applications of Cluster Analysis for Course Segmentation

A cluster analysis will also be used in this research to separate courses into distinct categories according to various traits such as target category, length of the course, amount charged for the course, and the course's provision medium. This categrization is achieved through the use of the unsupervised approach to machine learning, which entails the detection of groupings of entities that are naturally present in the high-enrolled, high-rated courses as opposed to the low enrolled and low-rated courses. Such an understanding would suggest that there is potential to offer recommendations for new strategies for new courses based on where they fit, e.g. higher course fees can be charged for the popular, well-rated courses while the courses in the new categories might need promotional efforts.

Insight Monetization

As a result of the statistical analysis and clustering performed, recommendations will be made to improve course design, pricing and platform activity engage. It will be explained how these factors such as course pricing and its duration relate to student retention levels, and how enrollment and rating is affected by the reputation of the platform and its category. Moreover, practical approaches will be suggested to boost completion rates by changing some aspects of a course or taking new approaches to involve the audience.

Validation and Reporting

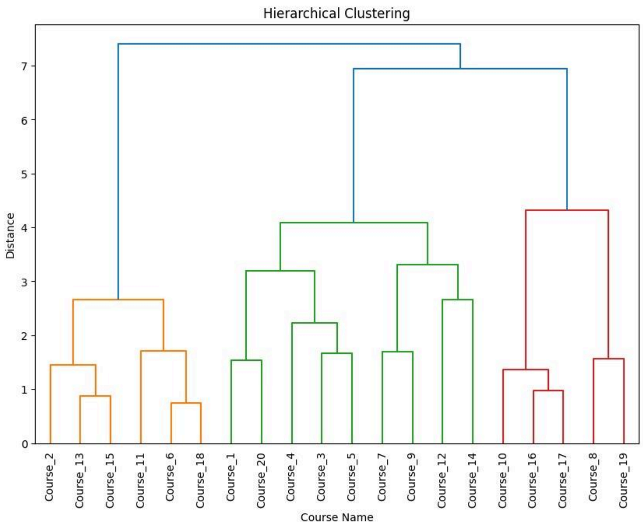
The results arrived at after the analysis will be confirmed in order to ascertain their reliability. Where applicable, the methods of cross-validations will be employed to affirm the strength of the regression models. The results will also be presented in a report that will include graphical elements, any statistically significant findings, and an overview of the concerns regarding online education, as the data in the report will consist of the online education dataset collated.

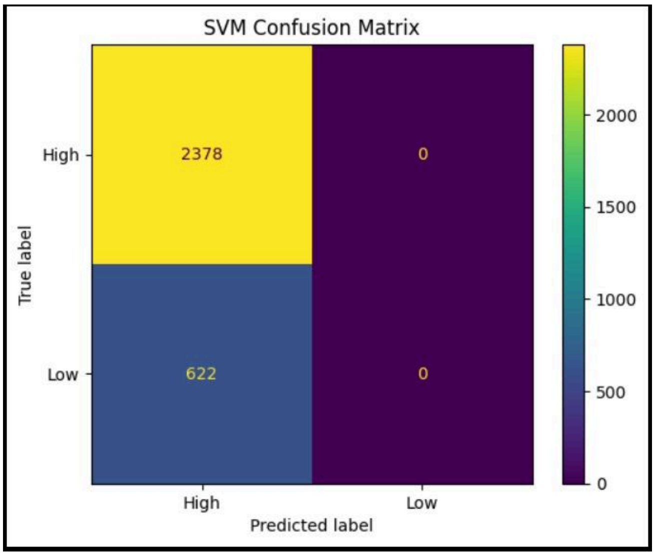
IV. RESULT AND DISCUSSION

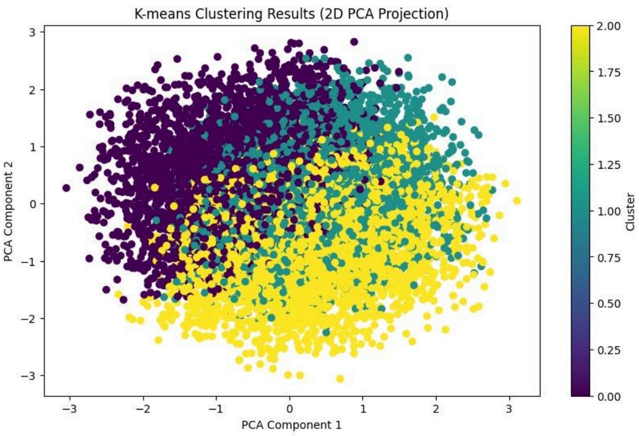
A range of patterns can be derived from the global data on the distribution of online courses in various subjects over numerous resources. Approximately, the average number of hours is around 55 and the rate of course completion is at an impressive figure of 75.1%, which indicates a decent level of participation from the learners. Course fees shown a wide disparity ranging from 10 dollars to two hundred dollars, the average course fee is around one hundred and six dollars. It seems that satisfaction is more or less the same in all places, with a mean google rating of nearly four over five. Student enrollment into the platforms like Udemy, LinkedIn Learning, edX runs into millions with Udemy being slightly ahead with total enrollments of approximately 6.5 million students. There are differences as in some countries, however, the completion rates roughly average 75% for all the platforms, with the exception of edX where the completion rate is slightly higher than that. There is also little variety in the prices charged, with most range falling between one hundred and five and one hundred and eight dollars, and which suggests some level of competition among the platforms, while the average course rating remains low at about four, which means students generally have a good experience.

On the other hand, on this course as well, the closer attention is paid to the categories of offered courses, the more interesting trends become. By contrast, the courses oriented on programming have the highest average completion rates (75.6%). Also courses, which are Data Science and Marketing, provide the highest average scores among other categories which means that such subjects not only have high enrollment but also satisfies the students registered extremely well. The pricing also differs from one category to another; quite high average prices such as $110 are found in areas like Data Science and Programming, which may well be due to both factors: the nature of the content and the skills offered are highly sought after in the market.

The information available on course analytics databases offers substantial resources for the development and training of machine working models to assess users engagement patterns and course effectiveness. Structured data on, course length, number of enrollments, completion rates, information about the platform, pricing and ratings is available for all of those courses developed and presented giving room for comprehensive data analysis. The data set helps to analyze algorithms to learned content optimization and strategic pricing as well as factors that enhance student engagement and satisfaction. The analysis of such variables also in turn assists in machine learning methods like regression analysis forecasting the course ratings or other successful course offerings on the given platform. Also, because the data includes both categorical and continuous data, and therefore the possibilities of employing simple and complex predictive models using supervised and unsupervised learning techniques is very high. Implementing such gut feeling measures through the use of data would help in offering more personalized elearning experiences and even bringing about tactical overture towards the elearning business. Simply put, the use of this dataset in conjunction with machine learning can help revolutionize education in a win-win situation for education service providers and consumers.







References

1. **Machine Learning for course analytics** mines machine learning algorithms, including regression models and neural networks, for accurate crop price predictions. https://ieeexplore.ieee.org/document/8955065
2. **Review of course Price Forecasting Methods Using Machine Learning**
An overview of machine learning methods applied to crop price forecasting, focusing on recent trends, challenges, and model comparisons. https://www.sciencedirect.com/science/article/pii/S2352914819300888
3. **Application of Machine Learning Models to Analytics Market Analysis**
This paper details the use of regression, classification, and clustering techniques foranalyzing agricultural market trends and price predictions. https://www.frontiersin.org/articles/10.3389/fpls.2019.01443/full
4. **Machine Learning in course analysis: A Review**
This review covers a wide range of machine learning applications in agriculture, including crop price prediction and yield forecasting. https://link.springer.com/article/10.1007/s00170-018-3075-7