**IMPLEMENTATION OF ASSOCIATION RULE MINING and regression ON e-learning platform**

**Udemy Course**

**INFO 5810 – Section 001-Group 6**

Akanksha Komirisheety - 11508345

Devendarreddy Bathini - 11602004

Harshith Battu - 11523126

Pavan sai Gundaram - 11609142

Shiva sai soma - 11600652

**INTRODUCTION:**

In modern days learning from classrooms are transformed into e learnings which helps students to upskill their knowledge on the subjects they are very interested in. Due to advancement in technology and the recent covid 19 pandemics made a lot of changes especially in learning by opening the gateways of e-platforms. So many online learning and teaching marketplaces are evolving around the world by focusing on courses related to many domains like business, Computer science, music and so on…Udemy is also one of the top online learning marketing websites which is started in 2010. It has courses related almost all the domains that someone can looking for and This website is also having an ability to have more than one course for same topic by differing the instructor, course work, length of Learning videos and cost associated with it. There will be some courses which are not associated with cost however their knowledge share might be limited. In this research paper we are trying to understand is their any association between the features like length of the course, instructor on the number of subscriptions and build a regression model to predict the subscription number by price associated with it. It is very important to know how the price of course is affecting the students enrolling into it, so that we can optimize the cost of the course effectively without no loss for both sides. Creating association is also helpful to understand how it affects the subscriptions and the precautions to take to maximize the profit.

As a case study we are taken a data of Udemy course related to 4 different domains as Business finance, graphic design, musical instruments, and web design. Gathered the data related to these domains with attributes like course id, instructor, duration, complexity, price, number of lectures, and reviews we had.

**Research Questions:**

* We need to create some association between the features which are affecting the number of subscriptions either in positive or negative way
* On the other hand, we need to perform a regression analysis to predict the number of subscriptions of a course based on other fields like price, length, complexity etc.

Our major aim is to help the business to predict the number of subscriptions so that they can customize the course to gather the most profits out of it.

**LITERATURE REVIEW:**

1. **Association Rule Mining: A Survey**

**Link**:<https://personal.ntu.edu.sg/assourav/Unpublished/UP-ARMSurvey.pdf>

Zhao & Bhowmick (2003) research on the existing techniques for association mining. There are various approaches used in association mining including Apriori Series approach, AIS algorithm, Apriori algorithm, FP-Tree (Frequent Pattern Tree) Algorithm and Rapid Association Rule Mining (RARM).

# **Prospects and limitations of e-learning application in private tertiary institutions amidst COVID-19 lockdown in Nigeria:**

# **Link:**<https://www.sciencedirect.com/science/article/pii/S2405844020323008>

# Oyediran et al. (2020) this document describes about the situation of schools during the time of COVID-19 and about the introduction of e-learning in every institution of Nigeria. The document focused on the compliant of instructors, limitations of e-learning enforcement and other social variables. The document mentioned about the systematic sampling for the collection of information from 180 staff members about above mentioned features. The document has used the multiple linear regression model for comparison between limitations and compliance. The results of this document illustrated that more compliance has occurred in universities than other education institutions and it also describes that rules have affected the compliance.

1. **Online association rule mining**

**Link:** [**https://dl.acm.org/doi/abs/10.1145/304181.304195**](https://dl.acm.org/doi/abs/10.1145/304181.304195)

Hidber (1999) in his research describes an algorithm for computation of large datasets in online applications including web applications. The algorithm is efficient in memory management when compared to DIC or Apriori. It maintains a large dataset supersets for every item.

# **Online distance learning in higher education: E-learning readiness as a predictor of academic achievement.**

# **Link:**<https://search.informit.org/doi/epdf/10.3316/informit.352722164544178>

# Toron (2020) this documentdescribes about the research problem as interaction between e-learning readiness and achievement of online course academics. The author has used the readiness scale of e-learning for data collection. The data consists of 153 new students learning English using online course. The authors had used various techniques such as Reliability analysis, Pearson correlation, linear regression analysis to predict the level of readiness on achievement in academics of online courses. The obtained result describes that one of the features self-directed learning is the highest predictors for academic results and motivation feature as well. The document also describes that as COVID-19 is spreading in vast speed and e-learning is introduced everywhere. So, e-learning readiness has to be taken care while implementation.

1. **Algorithms for association rule mining – a general survey and comparison**

**Link:** <https://dl.acm.org/doi/pdf/10.1145/360402.360421>

Hipp et al. (2000) describe the algorithms that are used in association rule mining. The research document is aimed at comparing algorithms used in association mining. The experiments were done on the algorithms and analysis of their performances done using their runtime.

1. **Association rule mining: models and algorithms**

**Link:** <https://link.springer.com/chapter/10.1007/3-540-46027-6_3>

Zhang & Zhang (2002) discuss on the importance of association rule mining in decision making. The document research on the differences between positive rule mining and negative association rule mining. They further research on the use of pruning algorithm for search space reduction.

# **Prediction of Student Performance Using Linear Regression:**

# **Link:** <https://ieeexplore.ieee.org/abstract/document/9154067>

# Sravani et al. (2020) this document describes about machine learning applications in the field of teaching and learning platforms improvement. The document is mainly focused on the large learning platforms such as Udemy, unacademy etc. and to analyze the student performance in the academics. The main operators for this process are features such as student previous score in academics and other related attributes. The document has derived research problem as dropout of students due to large quantity of strength in every class and complex to assist every student in the class. The authors have used the linear regression algorithm from machine learning to predict the performance of students in their academics.

# **Student Performance on an E-Learning Platform: Mixed Method Approach:**

# **Link:**<https://www.researchgate.net/publication/338898031_Student_Performance_on_an_E-Learning_Platform_Mixed_Method_Approach>

# Rackic et al. (2020) this document describes about the application and how they affect an e-learning. The authors had taken data from a e-learning platform to evaluate the performance of student based on the key points in various courses. For the evaluation they used various techniques as Multiple linear regression, K means clustering and social network analysis. The total process is conducted in the University of Novi Sad.

1. **Movie Reviews and Revenues: An Experiment in Text Regression**

**Link:** <https://aclanthology.org/N10-1038.pdf>

In this paper, we look at the issue of estimating a film's opening weekend earnings. A movie's genre, MPAA rating, and cast have all been used in previous research on this issue; very little text about the film has been included. Here, we forecast opening weekend income using the language of reviews written by film reviewers from a variety of sources. We propose a new dataset that combines movie reviews with revenue information and demonstrates how review text may serve in place of metadata for prediction, and sometimes even outperform it.

1. **Regularization Paths for Generalized Linear Models via Coordinate Descent**

**Link:** <https://hastie.su.domains/Papers/glmnet.pdf>

The algorithms calculate along a regularization path using cyclical coordinate descent. Both huge and sparsely populated situations can be effectively handled by the strategies. We find that the novel algorithms outperform competing techniques in comparative timings by a wide margin.

# **Improving Movie Gross Prediction through News Analysis**

# **Link:** <https://ieeexplore.ieee.org/abstract/document/5286056>

In this study, we employ quantitative news data to assist readers in making movie revenue predictions. Regression and k-nearest neighbor models are examined, and the results show that models utilizing only news data can perform as well as those including IMDB data. Additionally, by combining IMDB data with news data, we can improve performance. The enhancement is statistically significant, too.

1. **Predicting Movie Sales from Blogger Sentiment**

**LINK:**<https://www.aaai.org/Papers/Symposia/Spring/2006/SS-06-03/SS06-03-030.pdf?q=who-got-kicked-off-american-idol-may-5>

In this article, we investigate if using sentiment analysis techniques on weblog data improves correlation compared to using volume alone in the field of movies. Our key finding is that, when applied to a small context around references to the film in weblogs that were written before its premiere, positive emotion is actually a superior predictor of movie success.

**METHODOLOGY:**

* To predict the number of subscriptions we are using the Linear regression, as it is a numerical predicting statistical approach used to assess the sole dependent variable based on group of independent attributes. It forecast continuous values of dependent variable which is like classification with discreate labels.
* To understand association or impact of the features like course length, complexity on number of subscriptions or price we can create a correlation matrix of these attributes and check and with help of association rule mining we can establish the relation between input and output variables.

To perform the above steps, we are using the RapidMiner medium as it produces reliable outcomes without much coding involved in it.

**Diagrammatic Representation of The Pipeline of The Research:**

Validation and interpretation

Regression analysis

Data Normalization

Data Cleansing

Data Collection

Describing each block of above flow diagram is as below:

* **Data Collection:** To perform any kind of analysis at the first phase we need to fetch the data related to it so here we need to identify the source and access the data related to the four topics that we mentioned in the beginning of Udemy learning platform.
* **Data Cleansing:** Once we got the data set it is the analyst duty to perform required steps to clean the data like checking for the redundant data and delete it, identify the missing values and outliers, and perform missing value and outlier treatment to increase the model accuracy.
* **Data Normalization:** As multiple independent values are taking into consideration of predicting the dependent it is a best practice to normalize the input variables to give equal importance to all the features.
* **Regression Modelling:** Once the data is all ready, we need to apply the Multiple linear regression model on this data to assess the mathematical equation between inputs and dependent.
* **Validation:** we got the Mathematical equation which helps to understand the relation between independent and dependent variable with the help of Test data and now we need to validate it with our validation set to check the association is hold good for others.
* **Interpretation:** As the equation passes the validation now it is time for interpreting the equation in lemon terms to understand it quickly.

**DATA COLLECTION AND CLEANING:**

**DATA COLLECTION:**

Our primary goal is to produce effective outcomes for an important dataset. We considered using educational datasets that are more trustworthy for our future work to make it practicable. We looked at several datasets in relation to our ideas before discovering one that is completely enough for our task. Our dataset is made up of 4 courses from the Udemy online learning site, which can be used for all the tasks we needed to complete. The columns in the dataset are ideal for classifications, predictions, and other analyses.

Additionally, the dataset contains both strong quantitative findings and qualitative data that was gathered through questionnaires. While the dataset has the capacity to reveal what is occurring, qualitative data frequently enables us to comprehend why it is occurring.

The data set is taken from the Kaggle platform which is one among the best sources of Data science or analytics practical research data.

Here’s the link for our dataset: <https://www.kaggle.com/datasets/andrewmvd/udemy-courses>

The data set consists of 12 attributes with 3678 rows with multiple courses. The attributes describe about course Id, Title, URL, Type of Course, Cost, Number of subscribers, Reviews, lectures involved, complexity of course, duration of course, publish date and Subject.

**Here is the brief explanation about features of dataset:**

**Course Id:** It is unique identifier of the specific course, can work as primary key of dataset

**Title:** It describe about Course. This dataset contains 3678 course names in it and the datatype is Nominal.

**URL:** It describe about URL of course on Udemy platform and the datatype is Nominal.

**Is Paid:** It is mainly derived variable and dependent on price attribute, and it is Nominal

**Price:** It describe about the cost associated to enrollment, it is Numeric in nature

**Number of subscriptions:** It gives the number of students enrolled into the subject. It is numeric in nature.

**Reviews:** It is a numeric variable talk about how many reviews a course got so far.

**num\_lectures:** A variable of numeric in nature and gives a detailed explanation of how many classes are involved in that course.

**Content duration:** A numeric variable which gives an information about how long the course it going to last for.

**Published Date:** It gives a brief about the day on which the course became available on Udemy platform, and the data type of this variable is Datetime.

**Subject:** A Nominal variable gives the information about the domain of the course.

**Data Cleaning:**

* Checked for duplicate records and deleted the redundant. As we see out of 3678 rows 7 are duplicate using remove duplicate operator, we deleted the redundant.
* Missing values in the numeric fields are replaced with mean and categorical variables are replaced with mode. There are around 30 records to the max have a missing value and those are replaced with the help of replace missing values operator in RapidMiner.
* Checked and conformed that the data set is ready to analyze.

Below is the simple screen snippet that designed to achieve the above results.

Diagram

Description automatically generated

**Data Visualizations:**

Educators may build e-learning on topics that interest them using the Udemy platform. Using Udemy's course construction tools, instructors may upload videos, programming source code, PowerPoint presentations, PDFs, audio files, ZIP files, and any other content that students would find valuable. Teachers now have a further avenue for involvement and communication with their students thanks to online discussion boards. The bulk of sessions include important topics including using an iPhone camera, Excel software, and AWS and Azure instruction. Through Udemy Business, a separate Udemy service, businesses may access more than 7,000 training courses on various topics. Employers may use Udemy Business to create customized learning portals for employee training. We are using a few of Power Bi's visualizations for our dataset. Power Bi is probably used to visualize the data.

**Card** :-

The total number of subscribers, one of the datapoints from our dataset, is used in this graph. Each image has a popularity score that is correlated to the year. The 12M stands for all course subscribers together.

A picture containing text

Description automatically generated

**Area chart** :-

A basic area chart is created in the region shown between the axis and the line in a line chart. Area charts may be used to highlight a pattern's overall importance and show how much it varies over time . We utilized the course title and the total number of subscriptions as our two data points in this .By taking a glance at the cart, we can see that the course "learn HTLM5 from scratch" has the biggest number of subscribers (268923). Additionally, there are 50 subscribers to the course "Forex - Trade gratley with precision - Complete Strategy."

Chart

Description automatically generated

**Clustered bar chart :-**

A bar chart is the industry standard for comparing a given figure across different categories. We are using the data points course title and price to extract the visual. The course "The entire webdeveloper masterclass: beginner to advance" costs the most, $215, as can be seen in the table. therefore it is evident that the course "Introduction to HTML course" costs $20 the least.

Chart, bar chart

Description automatically generated

**Table :-**

A table is a logically ordered grid of connected data that is divided into rows and columns. It might also have a row for headers and totals. Tables are a useful tool for numeric analyses comprising many entries for a specific area. I have tried to maintain the pricing, the total number of subscription and subject in this table. We note that the total number of subscribers to 4 topic courses is 11759120, and the sum of the prices for the 4 subject courses is 242930.

Chart

Description automatically generated

**Pie chart :-**

Pie charts show the relationships between various components. In order to create these visualizations, we added together the subjects and total number of subscribers. According to this graph, web development has the largest percentage of subscribers at 67.87%. The topic of musical instruments has the fewest subscribers (9.2%).

Chart, pie chart

Description automatically generated

**Multi-row card :-**

One data point is presented per row on multi-row cards. To make this graphic, we blended the subject and course levels. In this card it shows us the levels of subject .

A picture containing table

Description automatically generated

**Funnel chart :-**

A process with phases and a sequential flow of objects through one phase to the next may be visualized using funnels. We utilized the content length and level as our two data points for visualization in this. Looking at the figure below, we can see that the expert level has the shortest course time (0.17k), while the novice level has the longest course duration (3.93k).

Chart

Description automatically generated

**Doughnut chart :-**

Pie charts and doughnut charts are comparable. They display how several components relate to one another. The center is empty and can include a description as the only change. Two data points—subjects and overall lecturers—were shown. The overall number of web development instructors is at 42.96%, according to the graph below, whereas the total number of lecturers for musical instruments is at 17.66%.

Chart, diagram

Description automatically generated

**Dashboard :-**

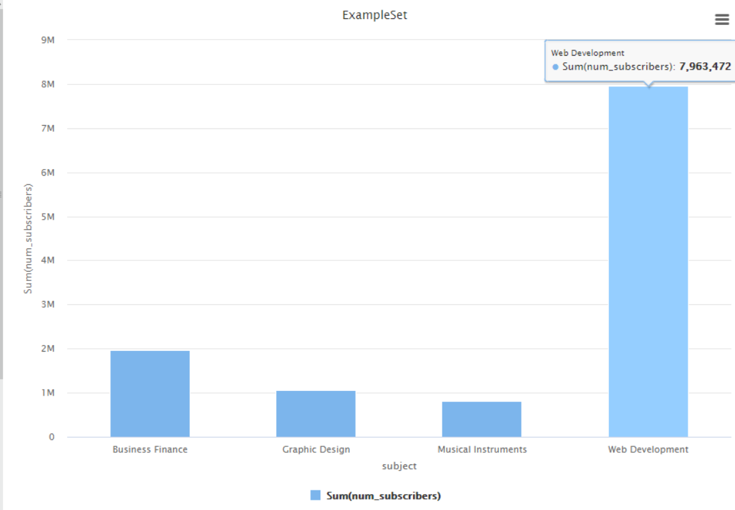
A single page, also known as a "canvas," on which visuals are used to tell a story is referred to as a "dashboard." Because the complete story would not fit on one page, the important elements of the story are presented on a visually appealing dashboard. Readers can find further information in the reports that are attached.

Graphical user interface

Description automatically generated with low confidence

**EXPERIMENT AND DATA ANALYSIS:**

From the visualization and data analysis we can see there are around 11.7M subscribers are there for complete 4 courses and that can be divided by each subject in the below way.



Highest is for web development and followed by Business, Graphic design and musical instruments.

Table

Description automatically generated

From the above we conclude that we have 3679 different courses in which 3369 are paid once which are having 8M subscriptions with average subscription is 2429 per course with average cost associated is $72. Similarly, 310 free courses are available and as expected it is having more average subscription that is 11534.

Average course duration is 2 hours in free versions, and it is double for paid versions that is 4 hours because in paid version we have detailed explanation of each topic whereas in free version it is mostly designed to grab the attention of users and promote them to use paid versions to get in detailed knowledge.

If we further divide the same KPI’s at subject level it is clearly seen that Web development is in first position in terms of subscriptions, cost, and duration of course this says that there is a huge demand of this course compared to others. Similarly graphic design and Business finance are almost similar in nature however the associated cost of Business Finance is bit higher as it mostly required by professional who works or thinking to work on finance department. Average lectures per course are also in the same flow as described above. The above statements are drawn from the below table.

Table

Description automatically generated

When looking at the total number of reviews associated with the topic, the tree map is the next type of visualization employed. It should be no surprise that web development is the most discussed topic. On the other hand, the overall amount of reviews is a strong indication that more people are interested in enrolling for the topic. This indicates that more people will be able to enroll in the courses and pay for them if the lessons and subjects are tailored to their specific needs. This indicates that, as time goes on, there will be a large increase in the profit margins.

Chart, treemap chart

Description automatically generated

We can also see that the course is mainly divided into 4 types like all level, beginner, intermediate, expert levels, and considerable amount of people are enrolled in All Level course. which indicates that the selected topic will be covered in depth by these individuals as they progress through the various levels. This indicates that there will be a significant increase in revenues due to the increased number of persons enrolling.

Table

Description automatically generated

If we further divide the above with subject level it is shown as below.

Chart, bar chart, waterfall chart

Description automatically generated

**Regression Analysis: (Work in progress)**

Our Main focus is to fit a regression model to predict the number of subscriptions based on the remaining independent variables. This can be achieved with the help of RapidMiner and below is the basic design screenshot that can be used to get the desired output.

Diagram, schematic

Description automatically generated

The detailed explanation and result analysis will be given in the final submission.

**GitHub Link:**

below is the link where we can find the dataset related to the project and this also contains all the documents used or prepared so far to make progress in the project.

<https://github.com/harshithraj08/IMPLEMENTATION-OF-ASSOCIATION-RULE-MINING-AND-REGRESSION-ON-E-LEARNING-PLATFORM.git>

**Tools Required:** Excel/RapidMiner/Power BI.

**CONTRIBUTION:**

|  |  |
| --- | --- |
| **GROUP MEMBER** | **CONTRIBUTION** |
| Akanksha Komirisheety | Investigation of 4 articles – Literature Review and some portion of EDA |
| Devendarreddy Bathini | Introduction, Methodology, Data cleansing, portion of EDA |
| Harshith Battu | Data visualizations, Dashboard creation, Regression setting |
| Pavan Sai Gundaram | Investigation of 4 articles – Literature Review, portion of EDA |
| Shiva Sai Soma | Investigation of 4 articles – Literature Review, portion of EDA |

All the team members are cooperative, Interactive with the continuous efforts of all we were able to make this progress within short period.

**REFERENCES:**

Zhao, Q., & Bhowmick, S. S. (2003). Association rule mining: A survey. Nanyang Technological University, Singapore, 135.

# Oyediran, W. O., Omoare, A. M., Owoyemi, M. A., Adejobi, A. O., & Fasasi, R. B. (2020). Prospects and limitations of e-learning application in private tertiary institutions amidst COVID-19 lockdown in Nigeria. *Heliyon, 6*(*11*), e05457.

Hidber, C. (1999). Online association rule mining. *ACM Sigmod Record*, *28*(*2*), 145-156.

# Torun, E. D. (2020). Online distance learning in higher education: E-learning readiness as a predictor of academic achievement. *Open Praxis*, *12*(*2*), 191-208

Hipp, J., Güntzer, U., & Nakhaeizadeh, G. (2000). Algorithms for association rule mining—a general survey and comparison. *ACM sigkdd explorations newsletter*, *2*(*1*), 58-64.

Zhang, C., & Zhang, S. (Eds.). (2002). *Association rule mining: models and algorithms*. Berlin, Heidelberg: Springer Berlin Heidelberg.

# Sravani, B., & Bala, M. M. (2020, June). Prediction of student performance using linear regression. In 2020 *International Conference for Emerging Technology (INCET)* (pp. 1-5). IEEE.

# Rakic, S., Tasic, N., Marjanovic, U., Softic, S., Lüftenegger, E., & Turcin, I. (2020). Student Performance on an E-Learning Platform: Mixed Method Approach. *International Journal of Emerging Technologies in Learning, 15*(*2*).

Joshi, M., Das, D., Gimpel, K., & Smith, N. A. (2010, June). Movie reviews and revenues: An experiment in text regression. In *Human language technologies: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics* (pp. 293-296).

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(*1*), 1.

Zhang, W., & Skiena, S. (2009, September). Improving movie gross prediction through news analysis. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology* (Vol. 1, pp. 301-304). IEEE.

Mishne, G., & Glance, N. S. (2006, March). Predicting movie sales from blogger sentiment. In *AAAI spring symposium: computational approaches to analyzing weblogs* (pp. 155-158).