Student Course Segmentation

Ginelle D'Souza[1] , Sneha Vijay Bhope[2]

[1]Department of Master of Computer Applications, ASM's Institute of Management and Computer Studies (IMCOST), Thane, Maharashtra, India
[2]Department of Master of Computer Applications, ASM's Institute of Management and Computer Studies (IMCOST), Thane, Maharashtra, India

***

Abstract: Over the years online learning has offered massive options for several courses, creating enormous information across learning platforms. This research paper will strive to achieve personalized course suggestions for students. The data of the students are bifurcated into - Known Information and Unknown Information. The "known information" is the data the student chooses to let us know like age, interest, languages, etc. Whereas, the unknown information is the data we can extract from the student while they participate in a course like watch time, marks obtained, number of attempts taken for examinations, etc. Course segmentation can be achieved by utilizing the well-known marketing approach – RFM (Recency, Frequency, and Monetary). Through this we will collect unknown information from the students like the recency of purchased course, frequency of course purchased, and monetary value of the courses purchased. RFM data is combined with the K-Means clustering method to produce output in the form of k-clusters of recommended courses for students. This study revolves around a population of 200 students and a course population of 150. The research aims at providing a possible solution to overburdened information by student personalization of courses.

Keywords - Artificial Intelligence, Clustering, Enormous Information, K – Clusters, Known/Unknown Information, K – Means Clustering, Online Learning, Personalized Course Suggestions, Recommended Courses, RFM Model

I. Introduction

Online learning has grown exponentially in the last year. It is deemed to be the "New Normal" considering the current scenario – The COVID-19 Pandemic. All around the world, educational institutions have increased their efforts to make learning easily available for the student. Thus, building several courses to meet the interest of the students. At times this enormous information is very difficult for students to consume. Over the years, we have witnessed a remarkable increase in personalization techniques through artificial intelligence. Through this research we will strive to achieve personalized course suggestions to students, thus eliminating the burden of course selection. The data of the students are bifurcated into - Known Information and Unknown Information. Data for the known information is collected by a survey taken through "Google Forms" to analyze interests, age, occupation, views regarding online learning, etc. Whereas, unknown information is an iterative process wherein user behavior is analyzed based on the RFM model. The analysed user data is further separated into clusters using K-Means clustering. Several data points can be generated through known and unknown information; therefore, we must reduce the dimensions of the data point such that clustering can be done based on the collected information.

II. Literature Review

A. Student Segmentation

Online learning was first witnessed in the early 2000s. Ever since, several platforms have been developed to offer education to students in the comfort of their home, and at their leisure. Universities worldwide have collaborated with learning platforms as well. Each platform offers thousands of courses ranging from a variety of domains, thus allowing a student to select from this massive variety. The identification of courses that best satisfies the learning requirements of each student is a very complex and tedious task.

This is because every student has their preference. By segmentation, we aim at grouping each student into "N" clusters based on their similarities. The segmentation can be done by considering several factors such as age, interest, qualification, aptitude, geographical location, financial background, demography, behavior, etc. Student Segmentation not only allows online platforms to offer the best courses possible to their users but also opens room for improvement. Platforms can offer courses that are not available when they analyze that there is an increase in interest and capability among the young population.

B. RFM (Recency, Frequency, and Monetary) Segmentation

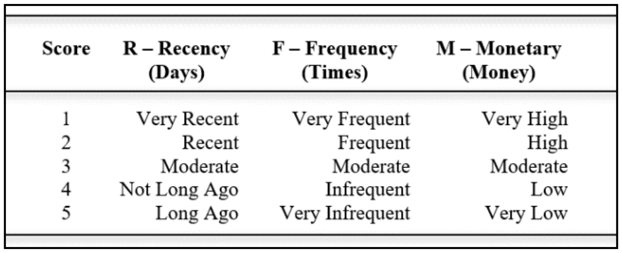
RFM is a marketing model that is used to analyze the value of user recency, frequency, and monetary. It is used to identify an organization's best customers based on certain measures. RFM analysis supports firms in making decisions such that to analyse students that are more likely to purchase another course in the future, how much revenue will be generated by each learner, and how to convert one-time course students into habitual ones. As per Jan Roelf Bult and Tom Wansbeck in an article "Optimal Selection for Direct Mail". They state that "80% of business comes from 20% of the customers". The three categories can be identified as follows:

Recency - How recently has a student purchased a course

Frequency - How often a student purchases a course

Monetary Value - How much money a student spends on purchases

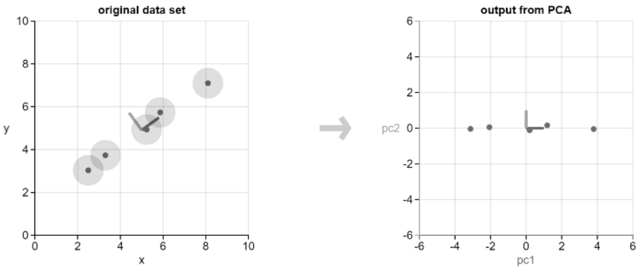
RFM model numerically ranks a student based on their behaviour in each of the above three categories. The rank scales from 1 to 5 (the higher the number, the better the result).



Score	R – Recency (Days)	F – Frequency (Times)	M – Monetary (Money)
1	Very Recent	Very Frequent	Very High
2	Recent	Frequent	High
3	Moderate	Moderate	Moderate
4	Not Long Ago	Infrequent	Low
5	Long Ago	Very Infrequent	Very Low

C. PCA (Principal Component Analysis)

Principal Component Analysis or PCA is a dimensionality reduction method that is used to reduce the dimensionality of large data, by transforming the variables into smaller ones. The transformed data will always contain the information from the large set despite the transformation. Reducing the number of variables may decrease the accuracy of the data. However, dimensionality reduction is used to trade off a small amount of accuracy for ease and simplicity of data variables. This is because smaller data is easier to explore, visualize, and analyze. Another, additional advantage of a reduced dimensional data is that it is easier for machine learning models to learn the data as much as possible, in turn giving the best results or predictions.



In the above figure, we see the original dataset consists of two dimensions namely the x and y dimensions. After processing the data through PCA, the two dimensioned dataset is reduced to a single dimension where the data is seen on only one-dimension pc1 and pc2 = 0.

D. K-Means Clustering

Clustering is a process of grouping data into categories based on the similarities within each data. In the world of the machine learning clustering is possible with the help of K-Means clustering. This algorithm identifies the similarities between the data and absorbers knowledge from the information gained. K-Means Clustering is an Unsupervised Learning algorithm, that groups similar datasets into several clusters. The letter "K" in K – Means Clustering defines the number of pre-defined clusters we need to maintain within the dataset. For example, if K = 3, there will be three clusters, and for K=10, there will be ten clusters within the dataset. To find the optimal K value, we use the "Elbow Method". This method is one of the most popular ways to find the optimal number of clusters. It uses a concept called WCSS value. WCSS stands for Within Cluster Sum of Squares.

Formula:

WCSS= ∑Pi in Cluster1 distance (Pi C1)2 +∑Pi in Cluster2distance (Pi C2)2

III. Research Methodology

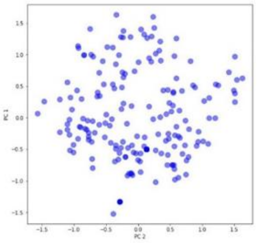
A. Survey - Google Forms (Quantitative Analysis)

A study revolving around 200 students was taken to identify and understand a student's perception of learning through online platforms. Through this survey, we observe that there is two major section of students who engage in learning platforms they are – College students and working professionals. The survey conducted consists of students and professionals majorly in the age group of 18 - 35. 61% of the participants are students out of which 98.4% among them are assertive that online learning has been beneficial. While 39% are professionals where only 66.7% agree that online has upgraded their skills. More than three fourth - 85.5% of students have agreed that online learning has been an integral part of upgrading their skills. 65.8% of students would love to carry on learning, whereas the likelihood of professionals to continue is 42.3%. A major factor in this is the "Time" involved in pursuing these courses. Considering the ability of students or a professional to be able to pick a course, we observe that there is a steep drop of comfort among students to be able to pick up a course. About 53.3% find it difficult to choose a course due to lack of "Knowledge". Professionals on the other hand find a sound place between difficulty – 40.3% and neutrality – 36.5%. When questioned about the challenges faced by online learning, a majority – 34.4% find it tedious to complete the course they begin. Another major challenge students face is the selection of a course from the massive options available across individual platforms – 29.93%.

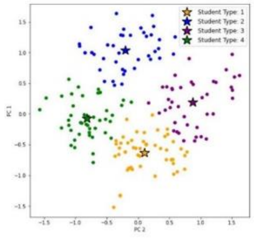
Through this study, we can identify several key points which will allow us to build a predictive model best suited for both our target sections – Students and Professionals. For initial clustering the participants have been asked their interest, thus forming the base of the model. Additional features can be used as dimensions to increase the accuracy of the model such as the challenges, age, time, money, etc.

B. Student Segmentation (Clustering)

Through the survey taken we get 8 features namely age, interests, occupation (Student/Professional), challenges faced, etc. By processing the data further, it is possible to additionally create 24 other features that will be helpful to cluster individual students. Since the total number of features we are dealing with is 32, the need for principal component analysis arises such that we are enabled to dimensionally boil down the data. Through principal component analysis, we boil down the 32 features into 6 features. In other words, principal component analysis is used for feature extraction of each record in the dataset. It transforms the data in a specific way, such that the least important features are dropped while retaining the most important features. By analyzing the data produced by principal component analysis, it is observed that features 2 and 1 prove to provide the best possible results. Thus, these features are used to cluster the students. For this study, we have used the "Udemy Dataset – Kaggle" to recommend a course. Within the dataset, there are over 1,500 courses mainly from 4 major subjects being – Web development, Business and Finance, Musical Instruments, and Designing. Therefore, we will cluster the students into 4 categories. Clustering or segmentation will be done by K – Means Clustering, where the number of student clusters will be 4 i.e., K=4. Initially, 4 centroids (center of a cluster) will be generated randomly. Based on the distance of a data point to a centroid the datapoint will be assigned to either of the 4 clusters. Once all the data points are assigned to clusters, the next iteration begins by selecting a new centroid within each of the 4 clusters. The centroids of a cluster can be computed by taking the average of the data points within each cluster. Once the new centroids are computed the data points are reassigned to new clusters. This iteration will carry on until there is no change to the centroids and the data points within each cluster.



Data Before K – Means Clustering



Data After K – Means Clustering

C. Enhanced Clusters Through RFM – (Subsequent to study III. B)

Once clustered the students can be recommended courses of their based on their interests and choices. While they are actively in the process of completing their course, learning platforms can analyse their study pattern. This involves how frequently they not only purchase but also learn for the courses they have purchased. The recency of visited courses and the money spend for each course. By analysing this data, online learning platforms can build much more accurate clustering models based on student buying capabilities, time investment patterns, subject matter enthusiasm, etc. RFM enhances the clustering model by providing an increased inaccuracy. Thus, resulting in more personalized course recommendations

IV. Conclusion

Every student has their interest and capabilities. Therefore, it makes no sense to offer courses that do not match the requirement of individual students. Educators are expected to understand not only a student's interest or capabilities but also their behavior and approach to studying. Thus, after recognizing these aspects educators would be in a better position to offer courses that best suit an individual. This study proposes a technique to enhance personalized courses for each student. For an initial stage of personalization clustering concerning an individual interest may prove to be optimal. As the student completes their courses additional attributes indicating their behavior can be used to give the utmost precise course recommendations. The study proposes an alternative method to highly trained recommendation systems, which can not only be used by well-established learning platforms, but also by startups.

V. References

[1] Yash Kushwaha and Deepak Prajapati: "Customer Segmentation using K-Means Algorithm": IJCRT.

[2] Yanuar Rafi Rahadian and Bambang Syairudin: "Segmentation Analysis of Students in X Course with RFM Model and Clustering": Jurnal Sosial Humaniora (JSH) 2020, special edition.

[3] Victor Powell and Lewis Lehe: "Principal Component Analysis"

[4] Andrea Sindico: "Customer Segments with PCA"