Predictive And Prescriptive Analytics On Courses Offered By Udemy

Rahil N Modi
Computer Science Engineering
PES University
Bengaluru, India
rahilmodi123@gmail.com

Sooryanath.I.T
Computer Science Engineering
PES University
Bengaluru, India
sooryanathit@gmail.com

Himanshu Jain
Computer Science Engineering
PES University
Bengaluru, India
nhimanshujain@gmail.com

I. Introduction

Analytics is the systematic computational analysis of data or statistics. It can be valuable in areas rich with recorded information; analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance.

The major goal of performing analysis on the data is to be able to predict the patterns in the data and to be able to make informed and reasonable decisions. Predictive analytics is used to gain future knowledge about the patterns in the data and prescriptive analysis helps an entity modify their goals in order to maximize the profits.

The data in real-world is dynamic, large and inconsistent. The need for data analytics is at its peak and is helping various organizations in the decision-making process.

Educational sector of our country is expanding and the right path to establishing a strong foundation is by learning through the right courses and by the experienced and qualified instructors. The model building process of analyzing and prescribing the right courses at the right time is the heart of this project. The various factors affecting the popularity of an online course has to be analyzed through various statistical methods. Massive open online courses (MOOC) has gained importance in recent years. Students rely on these platforms to shape their careers and hence providing the right choice of the course to a student is very much essential.

The datasets used to examine the above problem consist of details regarding various domains of courses offered in Udemy MOOC. The domains are varied from IT and software related sectors to business, finance and accounting sectors. All this data has been accumulated from the academic year 2012 to 2020. The columns mainly include information on particular courses offered for each domain, average ratings, the estimated cost and discount if any offered to them. The predictive and prescriptive analysis would help us to understand the trend followed in Udemy courses over these years and what changes can be done to attract and fascinate more students over the next course of time.

The current dataset can be used to comparatively analyse different courses opted by students in a given range of academic years. Based on exploratory data analysis and visualization a number of patterns can be recognised on which courses students opt, what are trending topics in the market, which topics require more practical exercises and how are students utilizing the knowledge gained through these topics in the industry.

Visualization mainly includes different charts and plots done using R and Python to analyze the data in the dataset and infer on summaries and statistics accumulated over time. Few basic visualisations like bar plots, histograms, pie charts help us to understand the distribution while a few other charts like whisker plots, scatter plots help us to identify outliers and correlation between attributes in the dataset. Many sophisticated charts like Coxcomb plot, strip plot, multihist, word cloud, etc also disclose few hidden patterns and characteristics.

Many other techniques like dimensionality reduction using wavelet transform, Principal Component Analysis(PCA) and Singular Value Decomposition(SVD) can be done to identify and remove irrelevant attributes and reduce the problem to lower dimensions. These techniques form the heart and soul of predictive and prescriptive analysis. Once the data is known carefully and completely a few queries would help to evaluate whether the data gained regarding the dataset is sufficient to perform analysis or further survey is required to reach a possible solution.

II. Related Works

A. Analytics Curriculum for Undergraduate and Graduate Students

1) Assumptions: This study aims to develop an analytics curriculum for undergraduate and graduate programs by identifying skill-based gaps between industry and academia and then clustering them based on methodological and semantic similarities among other criteria. [1] Specifically, we compare industry requirements and related skills for analytics jobs to three types of analytics domains, that is, descriptive, predictive, and prescriptive analytics.

2) Approach: This paper relates to data and the part of the problem we are trying to solve that is predicting a few interesting courses offered by Udemy to students at a reasonable cost. It also helps Udemy to formulate a series of courses offered as a package to students so that they can get a marginal profit and students get a whole specialization in the domain selected. This paper aims to formulate a curriculum to undergraduate and postgraduate study in a manner that best serves the interests of the students. Finally, realizing the significant difference between undergraduate and graduate students in terms of expectations and maturity, we use personality-job fit theory to recommend strategies to better promote the field to undergraduate students.

3) Results: Defining a curriculum for undergraduate and graduate students taking descriptive, predictive, and prescriptive analysis into consideration. Personality job-fit theory adds on to this to make wonderful predictions and prescriptive analysis.[1]

B. Machine Learning Approaches to Predict Learning Outcomes in Massive Open Online Courses

1) Assumptions: The implementation and methodology take into account only behavioral data when investigating the effect of patterns in learner behavior on the user certification rate, and latent variables like emotional state are ignored. Two set experiments have been performed. In the first set of experiments, all features from the dataset were included. For the second, a subset of features is segmented by their weights/importance in accordance with the target variable and selected.

2) Approach: Related to educational data mining (EDM) on courses taken/offered, no_of_tests, duration of lectures, correlation/relationship between clickstreams, and courses completed. Knowledge extraction to enhance teaching strategies. Also emphasizes on introduction about the domain of learning Analytics that discusses crowd behaviors, experiences that contribute towards making significant decisions on MOOCs. In general, nonlinear classifiers have better accuracy in both experiments than the linear classifier. This indicates the nonlinear form of correlation between the predictor features and target in the learner dataset. Two set experiments have been performed. In the first set of experiments, all features from the dataset were included. For the second, a subset of features is segmented by their weights/importance in accordance with the target variable and selected.

3) Results: Choosing or adapting boosting methods and feature segmentation by ranking / assigning weights tends to deliver a better performance in case of predictive analysis, class balancing plays a major role in performing a predictive analysis, functional form between the target and predictor features need to be in line with the choice of models chosen for prediction (linear or non-linear). The simulation results in experiments indicate that Random Forest (RF) and Support Vector Machine (SVM) achieved ideal performance, with the accuracy values of 0.9881 and 0.9851 respectively.[2] Latent variables like the emotional state of a learner also need to be considered to perform a more accurate predictive analysis.

4) Limitations: The result shows average run time of machine learning models is much longer in performed experiments.[2] The learner emotional states of students are not considered here, which can be inferred from their interaction with online courses over time.

C. MoocRec: Learning Styles-Oriented MOOC Recommender and Search Engine

1) Assumptions: It uses (Felder and Silverman Learning Style Model) FSLSM to recommend courses. This model indexes only computer science courses. [3]

2) Approach: Massive Open Online Courses provide a large number of courses in different domains. A specific domain has multiple courses. Selecting the most suitable course based on factors like learning styles, individual needs, course quality makes a difference in effective learning. MoocRec model has been developed for personalized learning. It performs content analysis of MOOC data to find the best course which satisfies the learning style of the user. Recommending courses based on specific topic parameters is implemented in the MoocRec model for course recommendations.

3) Results: The major takeaway for this model is the use of topic modeling and text processing to filter most appropriate courses based on the user's search query.

4) Limitations: The current version of MoocRec is only limited to the Felder and Silverman learning style model. similarly, MoocRec currently indexes only computer science courses from edX and Coursera platforms.[3]

III. Proposed Problem Statement

MOOC aims to provide access to numerous amount of courses at a reasonable cost, here we try to address the methods by which the course instructors or the course providers can increase the number of course recipients or subscribers in an intuitive manner by introducing short and crisp lecture series and offering courses at a reasonable discount when compared to the competitive course providers, probably by increasing the assessment contents etc. What is the optimal number of lecture hours, assessments per course and duration of the courses to increase the ratings and the number of recipients for a course, such questions can be answered with appropriate analytics done on the set of features from the training dataset.

Capture the features for the courses with high subscriber count and ratings and analyse the reason (causation) behind the factors which are responsible for such numbers (a heuristic approach). Given the key performance parameters or indicators (KPI) for a course, how many subscribers may choose it (A regression analysis)? Suggest the user with top 10 highly rated/attempted courses when given with a category of course. A correlation analysis based on how every single KPI affects the overall ratings for a given course. how to predict the optimal cost for a course given its features like total lecture hours and assessments, this information will help the user set up the near to optimal cost for a course.

Perform various visualizations to find the strength of the relationship between various features of a course and find out what are the most commonly offered courses based on their Titles, what are the catchy phrases that must be present in the title of a course to attract the learners and so on. Provide insights on what a learner looks for and what must a course provider look to provide.

IV. Approach

The approach mainly leverages between analytics on the MOOCs dataset and the Machine learning aspect to answer the questions posed in the problem statement section. The existing works on the MOOCs datasets focus on predicting the probability if a student would complete a course or not based on the history of dropouts from a course, whereas here we focus more on what features of a course particularly attract the target audience, right from the title of the course till the cost incurred in taking up and finishing a course and also to provide the course instructors/providers with a window through which they can gain profit by obtaining more number of subscribers and above par user ratings.

We start by analyzing the dataset features and the strength of the relationship between them, few statistical measures which reflect the shape of the datum, visualize the emergence and rise of UDEMY courses over the years and inspect the courses with dominant attribute values such as best rated courses and most opted courses and finally find how well the features complement each other and many more of those are covered in the exploratory data analysis and visualizations, we provide our intuitive insights on why such behaviours are observed over the time.

We categorize a course into a domain of development studies based on the technicalities of the title it possesses, for example, a course with the title “Networking With Python Tools” would be categorized into the domains of Networks and Python programming language, so that the scope of analytics can be drilled down based on the domain of the courses in a more refined way, achieved using basic Natural Language Processing (NLP). The title of a course also plays a role in the way it attracts the audience, we provide the course providers with a set of most frequently used words in the title and so on.

The major portion of the course analysis includes suggesting top courses based on the category of programming languages, frameworks, topics covered in the course, etc. Sorting among these thousands of courses becomes a challenge. Hence, we use different techniques of NLP to extract relevant keywords from the title of the course and filter based on Parts Of Speech (POS) to get the most appropriate keywords that describe the topic of the course in greater detail. Based on these keywords we filter the best courses that the user might be interested in and might take those courses.

The relation and impact between the explanatory and target feature are modelled using the Simple Linear Regression (SLR) and Multiple Linear Regression (MLR) techniques, the possible case would be to model the relationship between the total subscriber count and the other numerical explanatory variables when the model is presented with a set of features along with the domain of the course then the modelling would be more effective as we consider the instances that belong to the same domain as that of the presented test instance.

Some analysis based on time series data also can be done to find the relative courses offered from the academic year 2012 to 2020. Based on the trend few new courses can be combined to benefit students in terms of knowledge and increase gross profit too. The average rating for each course predicts the user preference based on which a prescriptive analysis can be done to prescribe a few top 10 courses in each domain which are most popularly chosen by people. Based on advice from industrialists and a few leading companies few specialisations could be provided as a pack to give them a cumulative specialisation certificate on completion of a few courses.

References

[1] Paul, Jomon A., and Leo MacDonald. ”Analytics Curriculum for Undergraduate and Graduate Students.” Decision Sciences Journal of Innovative Education 18.1 (2020): 22-58.

[2] R. Al-Shabandar, A. Hussain, A. Laws, R. Keight, J. Lunn and N. Radi, ”Machine learning approaches to predict learning outcomes in Massive open online courses,” 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, 2017, pp. 713-720, doi: 10.1109/IJCNN.2017.7965922.

[3] S. Aryal, A. S. Porawagama, M. G. S. Hasith, S. C. Thoradeniya, N. Kodagoda and K. Suriyawansa, ”MoocRec: Learning Styles-Oriented MOOC Recommender and Search Engine,” 2019 IEEE Global Engineering Education Conference (EDUCON), Dubai, United Arab Emirates, 2019, pp. 1167-1172, doi: 10.1109/EDUCON.2019.8725079.