AN ANALYSIS ON PREDICTING STUDENT DROPOUT USING MACHINE LEARNING

Thulasi Bharathi S

Department of Information Technology, St.Joseph’s College(Autonomous), Trichy

Abstract— Massive Open Online Courses (MOOCs) have transformed the way HEIs deliver educational material to the digital learners across the globe. Digital learning often deals with the high rate of dropout during a course. The new generation learners acquire the benefit of wealth of knowledge provided by the MOOC at their doorstep in contrast to the traditional way. Learners are registered in mass numbers but the HEIs are facing one big challenge—drop outs. Content Developers and MOOC providers are struggling to retain the students to complete the course. This gives a path to the researchers to do a research on the predicting students dropouts from MOOC. In this paper we are going to utilize Machine Learning techniques which can be helpful to predict the potential drop outs. In this paper classification technique based on decision tree is being utilized. The dropout risk factors are identified by using correlations and similarity measures.

Keywords— MOOCs, ML techniques, classification algorithm, dropouts.

I. INTRODUCTION

As the technology is moving towards the digital world, hence educational institutions are moving towards [1] accessing data at any time at anywhere that
(i) offering opportunity to access the proactive content at any device
(ii) providing enhanced learning in their own pace

Such technology known as MOOC, there are various famous MOOC providers like HarvardX, Coursera, Unacademy, and NPTEL are extensively used by the students and the working group of all ages to gain knowledge to withstand the competitive world [2]. In India MOOC has been growing rapidly of launching Swayam. Swayam platform is different from the other MOOCs, it provides credit based point on UGC 2016 framework and student can take the online course in Swayam. To earn the certificate, the student has to approach a local institute for the exam and earn the certificate, so it combines both the traditional education system as well as the MOOC concept of education [3]. The paper [3] also studies about the MOOC and Non-MOOC Users and found that the duration and skills in internet also plays an important role in making to students to be aware of MOOC, and also the users who are familiar with Video based learning are likely to discover new MOOCs.

Supervised Classification Models such as SVM, Logistic Regression, Random Forest and Gradient Boosting Decision Tree are used to predict dropouts of the course on the dataset provided by Edx, the above models gives an prediction to the course providers to get an insight of the individual student drop out of the course[4]. In paper LIP-Miner algorithm is designed to get the pattern of interaction of students and the instructors. This was been helpful in understanding interaction between the above two.[5]

Learning analytics in big data will be helpful to identify the behaviour and performance of the students [6]. The analysis of student action will let the researcher/MOOC provider to predict whether the student will drop out or not. This will be helpful to the increase the completion rate of the course.

II. LITERATURE REVIEW

As the MOOCs are becoming very popular and tool to gain knowledge as per convenience of time and place, a large number of students enrol in the courses and the MOOCs provider keep track of the activities of the students in log files[7] This made the researches to dive into ocean of analytics to extract better knowledge and understanding.

➢ In paper [1] the author conducted the research separating the science students (computer science, electronics engineering), from non-science students (economic, history, sports and other backgrounds). Based on the educational background of the students various factors like number of posts made by the student on the forums and also the activities was captured by the MOOC provider. Classification models were used to predict the student’s assignment grade. SVM(support Vector Machines, Linear regression, Factorization machine and Latent Dynamic factor graph was applied to the dataset to get the insight of the
behavioural patterns of the data. The author also took efforts to find the learning pattern of the student using regression model. Table 1 shows the comparison of the performance of science students vs non science students with different methods in predicting the certification earner[1].

<table>
<thead>
<tr>
<th>Category</th>
<th>Method</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science</td>
<td>LRC</td>
<td>92.13</td>
<td>83.33</td>
<td>46.51</td>
<td>59.70</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>92.67</td>
<td>52.17</td>
<td>83.72</td>
<td>64.29</td>
</tr>
<tr>
<td></td>
<td>FM</td>
<td>94.48</td>
<td>61.54</td>
<td>74.42</td>
<td>67.37</td>
</tr>
<tr>
<td></td>
<td>LadFG</td>
<td>95.73</td>
<td>73.91</td>
<td>79.07</td>
<td>76.40</td>
</tr>
<tr>
<td>Non-Science</td>
<td>LRC</td>
<td>94.16</td>
<td>76.93</td>
<td>89.20</td>
<td>82.57</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>93.94</td>
<td>76.96</td>
<td>88.60</td>
<td>82.37</td>
</tr>
<tr>
<td></td>
<td>FM</td>
<td>94.87</td>
<td>80.22</td>
<td>86.3</td>
<td>83.07</td>
</tr>
<tr>
<td></td>
<td>LadFG</td>
<td>95.54</td>
<td>79.76</td>
<td>89.01</td>
<td>84.10</td>
</tr>
</tbody>
</table>

In paper [4] the dropout of students using various classification models such as Logistic regression, SVM (Support Vector Machine), random forest and GBDT was analysed. Dropout prediction is a binary classification problem where two cases are whether the student will dropout or not. The four models are used to train a part of the data and test on remaining part of the data and the result shows that GBDT has the highest accuracy on the result.

In paper [7] the author shows the capturing of raw log files of the activities of students on MOOCs. The author applied Logistic Regression Model to predict the drop out of the students on weekly basis.

In paper [8], the author discuss about the motivation of a learner in completing the course, the author uses Clustering Technique to show the relationship between the motivator vs de-motivator. The author concludes the study by taking T- tests and showed that a student with good internet skills are likely to enrol in MOOCs compared to user with internet skills.

From the survey undertaken, it is observed that various Machine Learning models have their own advantages and disadvantages, so choosing a proper model for a problem is of great importance. Also, overfitting and under fitting has to be taken into consideration while training the test data.

III. METHODOLOGY

Success percentage rate of any educational institution can be analysed by knowing the reasons for dropout student. In this study, student information on various parameters was collected through Machine Learning repository by Predicting the students dropout status whether they interested to study or not, and another parameters such as personal details, academic details, family background, social, environmental, etc. variables are essential information for the effective prediction of attributes.

A. Objectives

The present study is to classify and regressing various dropout factors in the perspective of machine learning. To attain the objectives the following methodology was followed. The main objective of the paper is to determine the factors which are related to MOOC dropout.

OB1) To explore the familiarity, awareness and issues among the college students about MOOCs in our region.
OB2) To identify the major factors which influence the students to dropout (based on demographic, internal, external factors).

OB3) To predict the drop outs who enrol in MOOCs.

A total of 3 questionnaires was prepared based on the above mentioned objectives, 

i) Q1- Familiarity, awareness and issues of MOOC among students.

ii) Q2- Pre-MOOC expectations.

iii) Q3- Post- MOOC survey

These questionnaires was prepared with the base of [9]

A. Data Collection

To identify the familiarity, awareness and issues among the students about MOOC Q1 was distributed among the students through Google Forms and the responses were recorded. The data was collected from 164 under graduate and post graduate students through a questionnaire consists of variables like issues, familiarity level. Q2 and Q3 are also distributed to the students and the data was collected in two phases. First phase is to collect the data from the students who have enrolled in MOOC during the year 2019-2020. And the second phase was collected after the completion of enrolled MOOC course to predict the students who have drop out from the course.

The data collected for the was pre-processed by removing the row containing null values; remove missing values; smoothing noisy data, selection of relevant attribute from database or removing irrelevant attributes, identifying or remove outlier values, from data set, and resolving inconsistences of data. It has been done through Jupyter iPython notebook. The data has been explored by different graphical and non-graphical techniques utilized for Exploratory Data Analysis (EDA). All the variables were labelled.

B. Data preparation:

The data is partitioned sequentially into training, validation, and test sets. For training this model, we use 60% of the data while 25% is used for validation. The remaining 15% of the unseen data by the model is used for testing. Correspondingly, out of 164 samples from the dataset, 96 samples are used for training, 41 for validation, and the remaining for testing for the Q1. Supervised learning approach is used to train the network. For Q2 and Q3 Correlation based Feature selection (CBFS) is the process done to select a subset of input data analysis and future prediction by eliminating irrelevant information in predictor attributes. Q2 and Q3 were given to the students as Pre-MOOC and post-MOOC studies. In this phase, the CFS approach is used to identify feature subset which is highly correlated with class and minimum correlated with attributes by using best first search method between two dataset’s.

C. Results and discussion:

By the classification technique it has been identifies that only 6.7% of the students are familiar about MOOC. Out of this 6.7% many of the students are aware about various MOOC providers which shows a good sign and good interest towards learning. Though 60% of the students use their mobile
phones for more 50 hrs /week, the students were not aware about massive Open Online Course which shows lack of awareness among student. And also they have issue with language skills fig 1

Fig 1: Issues in taking Online Course

The students who were aware and enrolled in MOOC were facing some issues and challenges like

- Course could be taken in native language
- Fee concession can be given.
- Not interested to study alone.

In analysing Q2 and Q3 science stream students who were enrolled in MOOC courses filled the questionnaire. To evaluate the individual class accuracy, we use confusion matrix with individually predicted class accuracy. A sample can be classified either as a true positive (TP), false positive (FP), true negative (TN), or false negative (FN), as shown in Table 3

<table>
<thead>
<tr>
<th>Prediction</th>
<th>GROUND TRUTH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>POSITIVE SAMPLES</td>
</tr>
<tr>
<td></td>
<td>TP</td>
</tr>
<tr>
<td></td>
<td>FN</td>
</tr>
</tbody>
</table>

Here, TP represents the correctly classified samples for dropout prediction. Thus, the dropout prediction accuracy can also be represented as:

\[
P(d)_{acc} = \frac{TP}{TP + FP} \times 100
\]

\(P(d)_{acc}\) is actually a precision in this case, which represents how many selected samples from the test dataset are dropouts. The test dataset contains 100 total number of positive ground truth samples, while the number of negative ground truth samples is 15. These ground truth samples correspond to 15% of the test dataset. It is also worth mentioning that, positive samples are the dropout students while the negative samples are those students who completed the registered MOOC.

<table>
<thead>
<tr>
<th>TABLE 4: DROP OUT VS LACK OF TIME</th>
</tr>
</thead>
<tbody>
<tr>
<td>RESPONDANTS</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>MALE</td>
</tr>
</tbody>
</table>

TABLE 3: CONFUSION MATRIX
Out of 100 only 15.4% of students have completed the MOOC course. CFS shows that there was a positive correlation between course completion and academic credits awarded as shown in fig 2.

![Course Completion Diagram](image)

**FIG 2: COURSE COMPLETION**

**TABLE 5: DROP OUT VS LANGUAGE PROBLEMS**

<table>
<thead>
<tr>
<th>RESPONDANTS</th>
<th>DROPOUT DUE TO LACK OF LANGUAGE PROBLEMS</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>YES</td>
<td>NO</td>
<td>TOTAL</td>
</tr>
<tr>
<td>MALE</td>
<td>30</td>
<td>24</td>
<td>54</td>
</tr>
<tr>
<td>FEMALE</td>
<td>32</td>
<td>14</td>
<td>46</td>
</tr>
<tr>
<td>TOTAL</td>
<td>62</td>
<td>38</td>
<td>100</td>
</tr>
</tbody>
</table>

From the above table 4, it also shows that 60% of the students face problems on completing the course because of the lack of language skills and lack of time management.

**IV. CONCLUSION AND FUTURE WORK**

The main purpose of the study was to investigate the major factors causing the dropout of students from MOOC in our region. Based on the review of the related literature, basic questions were formulated to indicate the nature of assumed relationships among various parameters considered in this study. To verify the stated assumptions, the Study had employed Machine Learning techniques like Classification – supervised learning to find about the familiarity, awareness and issues among the students. And also the pre-MOOC and POST- MOOC study also predicts that lack of time management and lack of language skills are the major factors MOOC drop outs in our region. This analysis can be useful for the institution to training the students with skills that were lacking. So that they can upgrade themselves to the technology.
In order to fully identify, understand and explain the reasons for MOOC student dropouts, there is a need to investigate more precise and accurate predictive solutions. A deeper understanding of the dropouts, therefore, could help course designers and instructors improve MOOCs and undertake timely interventions.

References: