2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS) | 978-1-6654-6883-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IC3SIS54991.2022.9885328

An Early Prediction of Dropouts for At-risk Scholars in MOOCs using Deep Learning

Anjali C A
Assistant System Engineer
Tata Consultancy Services
Kochi, India
anjaliachuthan1117@gmail.com

Dr. Ramani Bai V
Professor & HOD, Computer Science and Engineering
Vidya Academy of Science and Technology
Thrissur, Kerala
ramani.b.v@vidyaacademy.ac.in
ramani.research@gmail.com

***Abstract*—In 2012, the world has observed a new approach to educational systems named ”Massive Open Online Courses” which leads to a greater impact on the entire world as it transformed our traditional educational approaches. MOOCs have rapidly moved into a place of prominence within the media, in scholarly publications, and within the mind of the general public Universities. According to recent studies, more than thirty MOOC platforms are providing ethnically diverse courses to scholars across the boundaries however it was also observed that they can hardly follow the subjects until the end of the course. Such conclusions pointed out that in the future it may restrict the flourishing of such platforms and these high dropout rates may diminish the development of MOOCs. The research work focuses on developing an early prediction of a dropout system for at-risk scholars in MOOCs using the deep learning algorithm and simulating a dropout prediction model to construct a ranking system for scholars using their dropout probability for every week. Using this probability value the system could be able to predict the scholar's attitude toward the subject and chances of dropout; thereby this system would help the instructors to guide them at an early stage. The datasets of MITx and HarvardX MOOC courses were used to predict scholar dropouts by analyzing their learning patterns. The Deep Neural Network model could make predictions better than the existing technologies through hyperparameter tuning and optimization. Inducing the benefits of deep learning methods helps to construct an effective dropout prediction model and this new versatile technique helps the tutors to prioritize intervention for those dropouts and can be considered an effective solution to the early dropout students in MOOCs compared to the existing works.**

***Index Terms*—MOOCs, Dropouts, Deep learning, Early Prediction, Performance Evaluation**

I. Introduction

In recent years our online education system such as Massive Open Online Courses (MOOCs) has provided a diverse environment and attracted many participants globally. With the rise of the internet, traditional educational systems are replaced by web-based platforms and it has opened wide opportunities to improve the existing educational approaches by incorporating various learning strategies irrespective of the physical boundaries such as classrooms [4]. The main advantage of MOOCs over other platforms is that they are utilizing fewer resources yet marketing those resources to a wider range of people for instance irrespective of age anyone can learn with this MOOC platform, thus not only in the educational areas but also exhibits a rise in the marketing area. Although they provide such incredible learning platforms to the students the main drawbacks they still face are their high dropout rates. The term student dropout indicates how quality an educational institution provides services in terms of knowledge as well as other factors, this applies to online courses also. The success of a course completely depends on the percentage of students who have completed the course in a curriculum order. There can be many reasons for a student to enroll in a particular course, sometimes it can be personalized too, the main point about MOOCs is that if the scholar is learning anything from this platform that would be a significant gain to the organization[5,6]. Given the right interventions to the scholars, it's far vital to pick out the maximum likely dropout college students so that the tutors can assist them on time[7]. Certification and graduation, verification of student's identity, and being unsuitable for complex education are the other major challenges that MOOCs are facing [14]. Most of the MOOCs are hosted on servers owned and controlled through edX, and data have been saved and transferred periodically (every day and weekly, relying upon the data set) from edX.

The research work mainly focuses on analyzing the learning patterns of the students and simulating a Deep Neural Network model using parameters optimization and predicting the student dropout probability. Using these probability values, generating a ranking system for every student joining the course; thus a continuous evaluation of the student will be done, and thereby the tutor could easily identify the scholar who is more likely to drop out at an early stage. This method can be considered an efficient method as the existing works mainly focus on dropout prediction only. Also, utilizing the Deep neural network model could yield a higher prediction than the other algorithms. To compute the output DNN uses metadata as the input and forms the information through several layers. Through this procedure, the system will be able to make good predictions as every time the system is learning through individual features. Moreover, the proposed work not only acts as a dropout prediction model but at the same time it is also an early warning system resulting in a decrease in dropouts in the future; as a result, the value of MOOCs gets boosted.

978-1-6654-6883-1/22/$31.00 ©2022 IEEE

The remainder of the paper is organized as follows.Section 2 describes the related work; Section 3 describes the materials and methods used in our research work. In Section 4, we discuss the experimental results on the MOOC platform to demonstrate the effectiveness of our prediction system. Finally, in Section 5, we make conclusions.

II. Related Works

Since 2012, Massive Open Online Course (MOOC) has been created rapidly, which picks up critical ubiquity among both scholars and teachers, and the year of 2012 is named "*The Year of The MOOC*" [1]. In spite of the awesome eagerness for and quick development of MOOC courses and stages, there has too been rising concern over a number of MOOC perspectives. One feature in specific that's troublesome to ignore is that these gigantic courses moreover have gigantic dropout rates [5,6].Some of the reasons for the low completion rates found out are course methodology, lack of social interaction, and creativity. As the number of different MOOCs continues to grow, researchers and educational technologists are proposing innovative dropout prediction solutions to heal the MOOC dropout headaches. Therefore, various machine learning (ML)techniques have been successfully applied in this particular field to obtain high dropout prediction accuracy[7]. D.F.O.Onah, J.Sinclair, and R.Boyatt proposed a method which only focused on the behavioral patterns of the students in MOOCs. They conducted a quiz for both the educational instructors and the students to analyze the trends and patterns in their social behavior[4]. Some researchers proposed different ML baseline algorithms in building the dropouts model using the KDD Cup 2015 data that contains the event log of the students[3]. A major part of studies related to student retention rate was carried out in ML areas focusing on the socio-economic characteristics. Mushtaq Hussain / Wenhao Zhu method was to find tout the dropout accuracy using their course assessment score and based on the number of clicks on the virtual environment and deployed using various machine learning techniques[10].

III. Methods and Materials

The work was conducted in two distinctive phases for week by week classification where once more the phase is partitioned into diverse modules.

- **Module I :** Week wise Scholar Classification using DNN
- **Module II :** Weekly Status Updation

A. Datasets

The dataset was obtained from the first year of MITx and HarvardX courses, it has a total of more than a thousand plus records of students and details about the courses involved, each record contains information about a scholar on the platform such as edX[13]. The dataset mainly comprises two types of data; "Administrative data" and "User provided data". Almost 18 plus attributes are contributing to the input attributes. Fig. 1 portrays those registrants whose reference is the course enlistment date and their last exercises of seven days are given underneath. This can be referred to in a couple of regions as a plot of "*hazard probabilities*" that portrays the pace of trimming down in a given period. The detailed input attributes used in the datasets are depicted in Fig. 2.

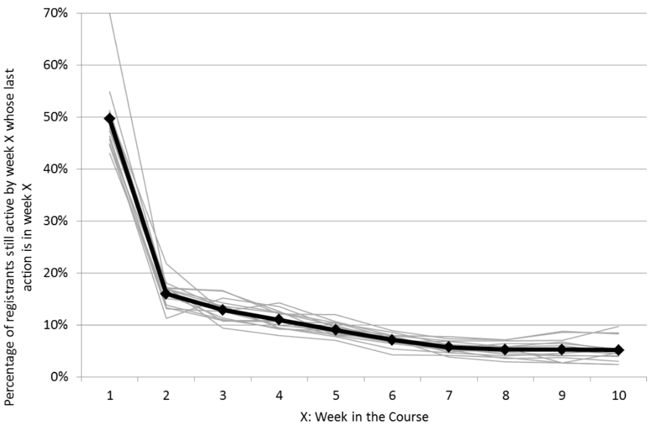


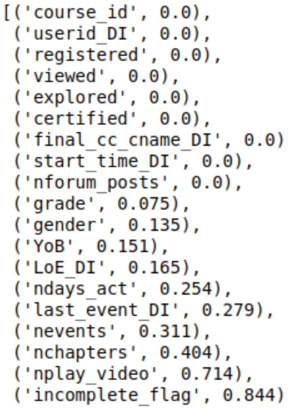
Fig. 1. Average Percentage of Last Action in a Particular Week[13]

B. Pre-processing

Replacing missing values by their mean is the maximum often used approach to fill the data set. If the type of data is an object replacing those with the most recurrently occurring value and if they are found to be outliers then deleting them. In most cases, we could see NAs and it is very much important to take care of them. The attribute 'incomplete_flag' has the highest proportion of NA however, it was found that it was due to some typographical error during the data acquisition. Thus this particular attribute can be safely removed. Here concentrating more on socio-economic factors along with scholar's information. In Fig. 2, it was observed that some of the values are not zero which indicates that those attributes have some missing data and those need to be filled inorder to proceed.

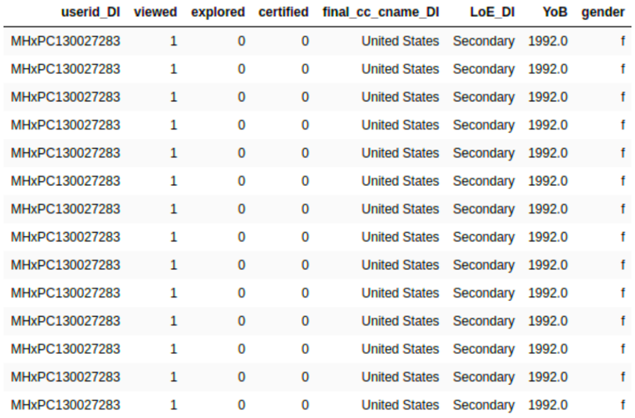
Later in analysis, it was found that some attribute features are selfly annotating, hence removing those features also [Fig. 3]. Some registrants just viewed the courses without even completing a particular section such cases also have a high chance of misleading from proper predictions. So here used an approach method to reply to our modeling capabilities rather than just take the input file as it is by incorporating information over a weekly basis.

Fig. 4 depicts the exploratory data analysis of the attribute gender over the total number of students. Thus we could able to analyze how an input factor depends on the output variable; whether it has any direct impact or not, and whether we can remove this variable or not through some graphical structures. Doing so will help in enhancing a clear picture of how an



```
[('course_id', 0.0),
 ('userid_DI', 0.0),
 ('registered', 0.0),
 ('viewed', 0.0),
 ('explored', 0.0),
 ('certified', 0.0),
 ('final_cc_cname_DI', 0.0)
 ('start_time_DI', 0.0),
 ('nforum_posts', 0.0),
 ('grade', 0.075),
 ('gender', 0.135),
 ('YoB', 0.151),
 ('LoE_DI', 0.165),
 ('ndays_act', 0.254),
 ('last_event_DI', 0.279),
 ('nevents', 0.311),
 ('nchapters', 0.404),
 ('nplay_video', 0.714),
 ('incomplete_flag', 0.844)
```

Fig. 2. Proportion of the data[13]



	userid_DI	viewed	explored	certified	final_cc_cname_DI	LoE_DI	YoB	gender
	MHxPC130027283	1	0	0	United States	Secondary	1992.0	f
	MHxPC130027283	1	0	0	United States	Secondary	1992.0	f
	MHxPC130027283	1	0	0	United States	Secondary	1992.0	f
	MHxPC130027283	1	0	0	United States	Secondary	1992.0	f
	MHxPC130027283	1	0	0	United States	Secondary	1992.0	f
	MHxPC130027283	1	0	0	United States	Secondary	1992.0	f
	MHxPC130027283	1	0	0	United States	Secondary	1992.0	f
	MHxPC130027283	1	0	0	United States	Secondary	1992.0	f
	MHxPC130027283	1	0	0	United States	Secondary	1992.0	f
	MHxPC130027283	1	0	0	United States	Secondary	1992.0	f
	MHxPC130027283	1	0	0	United States	Secondary	1992.0	f
	MHxPC130027283	1	0	0	United States	Secondary	1992.0	f
	MHxPC130027283	1	0	0	United States	Secondary	1992.0	f
	MHxPC130027283	1	0	0	United States	Secondary	1992.0	f

Fig. 3. Dataset Analysis[13]

attribute will help in better prediction. From the exploratory data analysis of input variable gender, it is very clear that both males and females are equally enrolling in the course which further implies that there can be an equal probability for both genders to get dropped out.

For the betterment of classification accuracy used TomeLinks sampling method[16] so that most of the samples from the majority classes are removed. The parameters changes are sampling_strategy='auto'. The standard score of sample x is calculated as[3]:

$$z = (x - u)/s \tag{1}$$

where u is the training sample's mean value and s is the training sample's standard deviation.

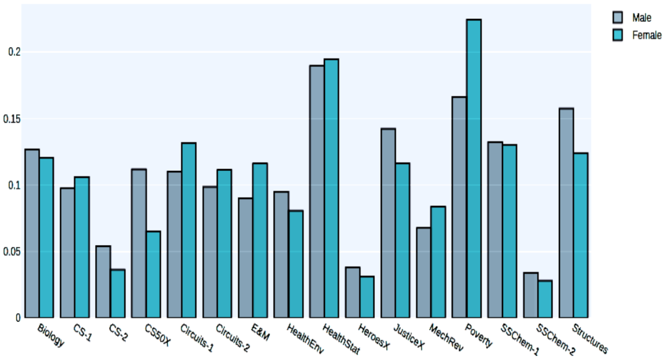


Fig. 4. EDA : Total proportion of students explored by gender

C. Module I: Week wise Scholar Classification using DNN

The research work proposes to utilize the Improvised DNN model to develop the dropout forecast model and, in advance, produce the predicted individual scholar's dropout likelihood[Fig. 5]. Creating such frameworks can distinguish those researchers who are more likely to drop out before the course completion and subsequently can give satisfactory steps to bring them back.

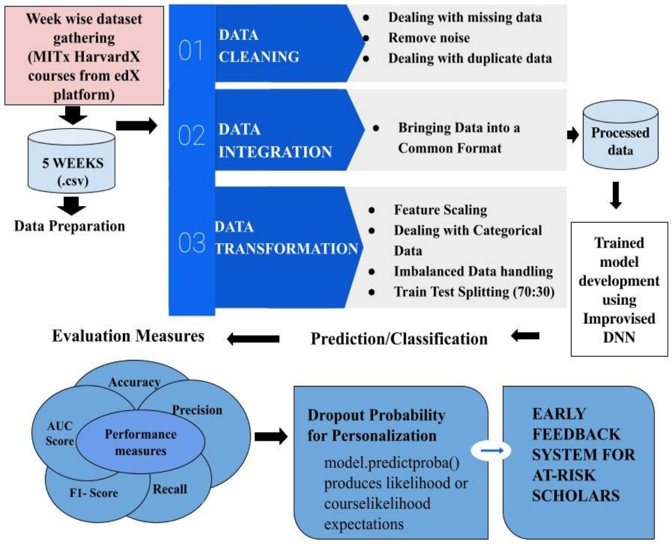

Week wise dataset gathering (MITx HarvardX courses from edX platform)
5 WEEKS (.csv)
Data Preparation
01 DATA CLEANING
Dealing with missing data
Remove noise
Dealing with duplicate data
02 DATA INTEGRATION
Bringing Data into a Common Format
03 DATA TRANSFORMATION
Feature Scaling
Dealing with Categorical Data
Imbalanced Data handling
Train Test Splitting (70:30)
Processed data
Trained model development using Improvised DNN
Prediction/Classification
Evaluation Measures
Accuracy
Precision
AUC Score
Performance measures
F1- Score
Recall
Dropout Probability for Personalization
model.predictproba() produces likelihood or courselikelihood expectations
EARLY FEEDBACK SYSTEM FOR AT-RISK SCHOLARS

Fig. 5. System Architecture

To way better assess the forecast execution of the week-wise calculations, the data were isolated into 70% for training and 30% for testing. Analyzing at-risk student's weekly-based activities as a first step toward building the dropout prediction model performance toward intervention personalization for at-risk students engaging in MOOCs [11]. An Improvised DNN was built using hyper parameter optimization and by adding layers one by one into the predetermined neural network was able to create a sequential model which yields better predictions[Fig. 6]. The Dense class is used to define fully connected layers and here used the benefits of the dense class. Before the activation function of the previous layer, a batch normalization layer is introduced.

Dropout is used on the network's hidden layers to reduce overfitting. Dropout rates of 0.2, 0.3, and up to 0.6 were utilized to improve the model's performance. With a dropout rate of 0.2 can get better outcomes. Different activation functions

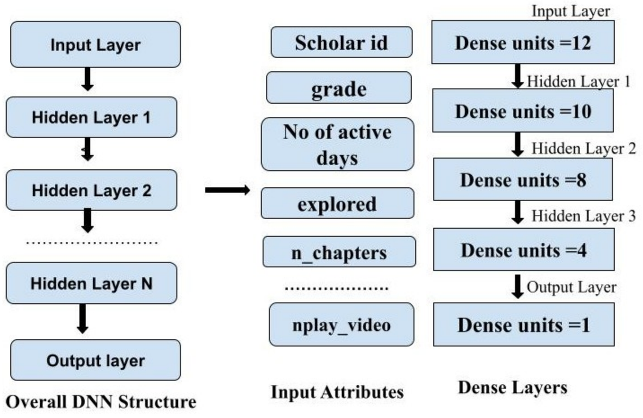


Fig. 6. Improvised DNN model

(sigmoid,tanh, and ReLU) were used in the hidden layers, and it was shown that ReLU can process results better.

Hidden Layer- RELU[17]:

$$Relu(x) = max(0, x) \tag{2}$$

Output Layer- Sigmoid[17]:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T}x} \tag{3}$$

Deep learning optimizes model tuning and selection for its claim, saving a significant amount of human effort and time. For each week, an improvised deep neural network was described; in addition, forward and backward propagation methods were utilized to train the weights, and batch gradient descent was used to reduce the computational complexity in this research. This analysis goes even further, proposing that predicted dropout probability be used to customize and focus interventions.

TABLE I
SPECIFICATIONS OF IMPROVISED DNN

Specifications	Parameters used
Model	Sequential
Dropout	0.2
Activation function	ReLu
Loss Function	Binarycross-entropy
Normalization	Batch Normalization
Batch size	30
Epochs	100
Optimizer	Adam

D. Module II : Weekly Status Updation

After using dropout expectation models to identify which MOOC learners which are more likely to drop out each week, advanced analysis can be done to see how the intervention might be tailored to every individual. This section of the research project intends to calculate each student's dropout likelihood for each week. Making use of Deep learning algorithms could predict a student's likelihood of dropping out on a particular course.The **model.predict_proba()** function in Keras generates likelihood or course likelihood expectations for the input tests. Taking this strategy allows the following:

i.For driven decision-making, the tutors can provide efficient feedback to the students, and thereby they can make their studies even more fruitful.
ii.A ranking technique has been proposed to prioritize intervention, by giving more concentration to those learners who have a very high chance of dropout from courses.

E. Evaluation Criteria

The Confusion matrix together with the Area under Curve (AUC) is assessed which gives an understanding of the proposed technique and its potential for detailed classification.

The classification models value and efficiency were measured utilizing the conventional measurements of precision, accuracy, recall, and F1 score. Accuracy is the calculation of the model's predictions and all overall expectations.

IV. EXPERIMENTAL RESULTS

We have implemented the proposed Early Feedback Prediction System using Python 3.8 programming language with IntelR Core i5-8300H CPU processor @ 2.30GHz and Windows 10 RAM 8 GB.

Performed week wise classification of scholars in MOOCs.For each week of a course considered the task of predicting, for any student currently active in that particular week with the help of a binary outcome variable that indicates whether the scholar will complete that particular week or not. Conducted this evaluation for five-week for all the students.

A. Analysis of Hyper Parameterized Results

The customized Deep Neural Network model was run at different dropout rates, say 0.2,0.3,0.4,0.5,0.6, and found out a dropout rate of 0.2 can able to produce better results and yield better classification. The results of this particular experiment are depicted in Table I and Fig 7. In Fig 7, the x-axis represents every week of scholars,and the y-axis represents the dropout values.

TABLE II
COMPARISON OF DNN MODEL OVER DIFFERENT DROPOUT RATES

Dropout	Accuracy
0.2	0.9853
0.3	0.9826
0.4	0.9733
0.5	0.9656
0.6	0.9664

Since the yearly data set contains a total of 45,000 plus

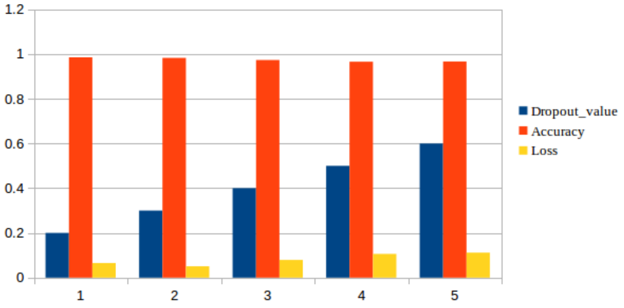


Fig. 7. Improvised DNN Model Results With Different Dropout Rates

scholars information, for getting a year static view performed DNN classification along with other ML models. The yearly data analysis model was trained using various machine learning algorithms like Logistic Regression(LR), Naive Bayesian(NB),Random Forest(RF) and Deep Neural Network(DNN) and it was found that the model predicts better for DNN which is being depicted in Fig. 8 and it provides a clear picture of the results with respect to all of the evaluation measures used in the study.

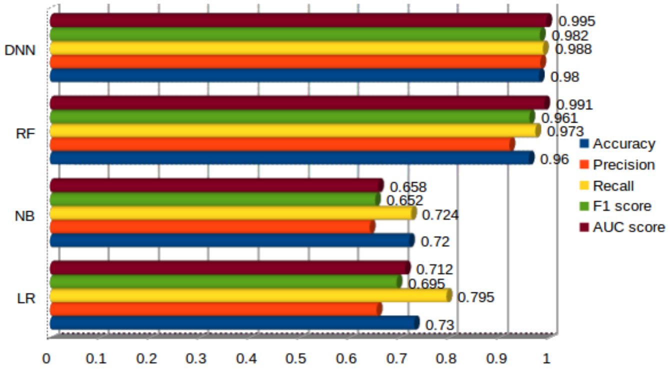


Fig. 8. Consolidated Results - Over a year

B. Week Wise Classification Performance

Dropout prediction models on a weekly basis were built and performed week-wise classification of scholars in MOOCs over five weeks. For each week considered the task of predicting, any particular student currently active in that particular week. This can be achieved with the help of a binary outcome variable that indicates whether the scholar will complete that particular week or not. Table III gives the results of scholars over five weeks.

During week wise classification highest accuracy of 99% was obtained for week 2. Week 2 yields better performance than other weeks concerning the evaluation measures mentioned in Table III. Fig. 9 depicts the ROC curves along with the AUC score of each week. The higher value of the AUC Score was 99.2% for Week 5, and the lower value obtained was 91.7% for Week 3.

TABLE III
COMPARISON OF DNN MODEL OVER 5 WEEKS

	Week 1	**Week 2**	**Week 3**	**Week 4**	**Week 5**
Accuracy score	0.985	0.99	0.97	0.98	0.98
Recall score	0.91	0.96	0.90	0.94	0.93
Precision score	0.86	0.90	0.85	0.87	0.82
F1 score	0.88	0.91	0.86	0.89	0.87
AUC Score	0.932	0.986	0.917	0.958	0.992

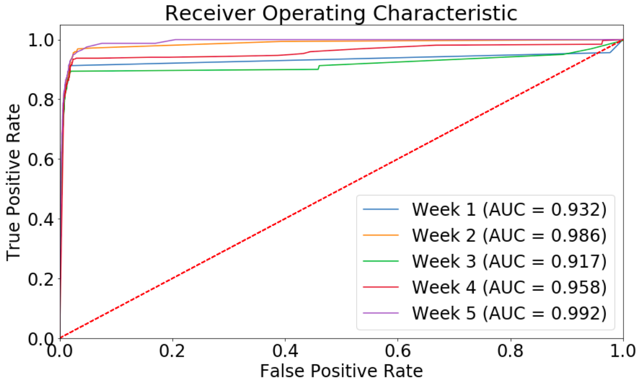
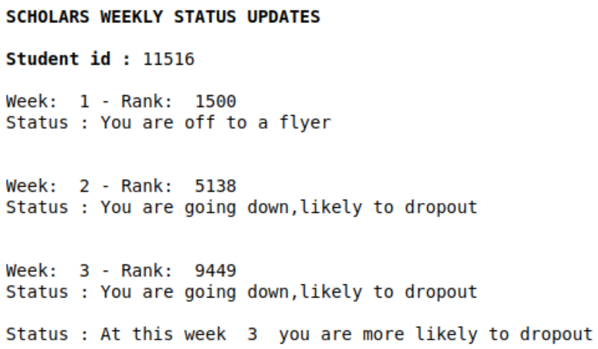


Fig. 9. ROC-AUC Score of Different Weeks

C. Early Feedback System for At-Risk Learners

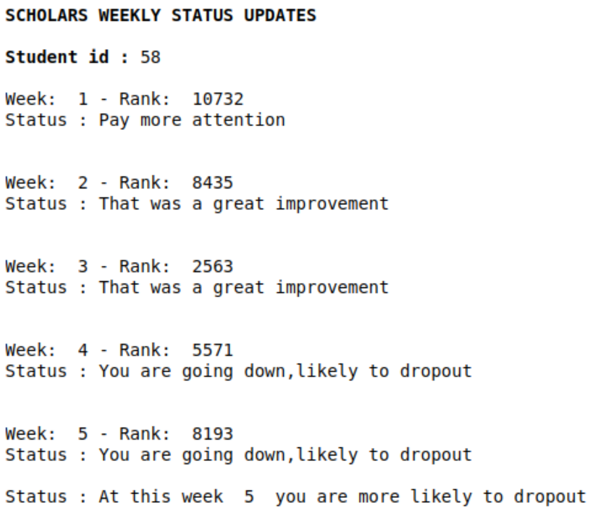
Created a ranking system, and a randomly generated rank will be assigned to the scholar initially, and then if the rank goes high he/she might be losing touch with the subject or vice versa. Whenever the file is being called, it randomly outputs the status of different students (provided the assumption student id in every week is unique). Therefore can generate the status level of every scholar over weeks and thus can find out those at at-risk. Since its a ranking system the student with a lower rank has better performance than the one with a higher rank.



```
SCHOLARS WEEKLY STATUS UPDATES

Student id : 11516

Week:  1 - Rank:  1500
Status : You are off to a flyer


Week:  2 - Rank:  5138
Status : You are going down,likely to dropout


Week:  3 - Rank:  9449
Status : You are going down,likely to dropout

Status : At this week  3  you are more likely to dropout
```

Fig. 10. Status Prediction :Scholar Id: 11516

If the scholar got the status You are going down, likely to drop out more than two times(in 2 weeks) continuously then he/she has a high chance of retention rate. At this point,



```
SCHOLARS WEEKLY STATUS UPDATES

Student id : 58

Week:  1 - Rank:  10732
Status : Pay more attention


Week:  2 - Rank:  8435
Status : That was a great improvement


Week:  3 - Rank:  2563
Status : That was a great improvement


Week:  4 - Rank:  5571
Status : You are going down,likely to dropout


Week:  5 - Rank:  8193
Status : You are going down,likely to dropout

Status : At this week  5  you are more likely to dropout
```

Fig. 11. Status Prediction :Scholar Id: 58

educators can offer assistance to personalize and prioritize intervention for at-risk students. Fig. 10 and Fig. 11 are showing the week-wise status of the scholars participating in MOOCs. Using the dropout probability which was derived from the DNN, a rank will be generated for each scholar and if the rank goes up every week, then the model predicts at which week the student is likely to drop out which is depicted in both figures[Fig. 10 & Fig. 11]. However, if the rank value goes down by each week, then the student can complete the course. Thus the model can able to find the at-risk scholars in MOOCs.

V. Conclusion & Future Scope

MOOC is an incredible tool for the development of skill and knowledge obtainment. MOOC education has brought a new revolution in how education has been conveyed to the people and can be considered an environmentally friendly method of teaching. The study's main objective was to take advantage of the power of the deep learning algorithms in building the dropout prediction models. The research was conducted in two phases; in phase 1 a DNN model was built using hyper parameter tuning and then a dropout probability was inherited from the model for five weeks for every student. In phase 2, using this dropout probability model, the system could be able to find the scholars who are most likely to drop out using a ranking system. If the rank is going up sequentially for every week means the student is more likely to drop out; however vice versa the scholar has the least chance of dropout. Thus with this approach, the system could help the instructors to guide the scholars so that a high dropout rate could be diminished to some extent using the personalized intervention.

Future considers can explore methods to advance the deep learning forecast execution in MOOCs and can conduct similar experiments on more online courses to examine whether deep learning can be valuable in other instructive areas.

References

[1] Wu, Dongen,Yuxiang Zheng,Pengyi Hao,Tianxing Han, and Cong Bai *"Classmates Enhanced Diversity-Self-Attention Network for Dropout Prediction in MOOCs."*,International Conference on Neural Information Processing, pp. 609-620. Springer, Cham, 2021

[2] Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E. and Nshimyumukiza *"Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. Computers and Education"*,Artificial Intelligence, p.100066

[3] Noraffandy Yahaya,Waleed Mugahed Al-Rahmi *"Massive Open Online Courses (MOOCs): Data on higher education"*,International Journal of Knowledge Engineering Technology 2020

[4] Rahila Umer, Teo Susnjak, Anuradha Mathrani, and Suriadi Suriadi, *"Prediction of Students Dropout in MOOC Environment"*,International Journal of Knowledge Engineering 2019

[5] D.F.O.Onah, J.Sinclair and R.Boyatt, *"Dropout Rates Of MOOCs : Behavioural Patterns"*, The University of Warwick (UK)2016

[6] Parr, C. (2013), *"MOOC Completion Rates "Below 7%""*, Available at: http://www.timeshighereducation.co.uk/news/mooc-completion-rates-below-7/2003710.article. [Accessed: 13/01/18]

[7] Jordan, K. (2013), *"MOOC Completion Rates"*, Available at: http://www.katyjordan.com/MOOCproject.html [Accessed: 18/02/18]

[8] Fisnik Dalipi, Ali Shariq Imran , Zenun Kastrati, *"MOOC Dropout Prediction Using MachineLearning Techniques: Review and Research Challenges"*,2018 IEEE Global EngineeringEducation Conference (EDUCON)

[9] Yunfan Chen,Ming Zhang, *"MOOC Student Dropout: Pattern and Prevention"*,ACM TUR-C '17, May 12-14, 2017, Shanghai, China

[10] Mushtaq Hussain , Wenhao Zhu , Wu Zhang , and Syed Muhammad Raza Abidi, *"Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores"*, 2 October 2018 School of Computer Engineering and Science

[11] Jason Browniee, *"Logistic Regression Tutorial for Machine Learning"*, Available at:https://machinelearningmastery.com/logistic-regression-tutorial-for-machine-learning/, [Accessed :12/09/2019]

[12] Wanli Xing , Dongping Du, *"Dropout Prediction in MOOCs: Using Deep Learning for Personalized Intervention"*, Journal of Educational Computing Research 2018

[13] HarvardX 2014, *"HarvardX Person-Course Academic Year 2013 De-Identified dataset, version 3.0"*,Available at:https://dataverse.harvard.edu/dataverse/mxhx,[Accessed: 14/05/14]

[14] S. S. A. Hamid, N. Admodisastro, N. Manshor, A. Kamaruddin, and A. A. A. Ghani, *"Dyslexia adaptive learning model: student engagement prediction using machine learning approach,"*in Advances in Intelligent Systems and Computing, R. Ghazali, M. Deris, N. Nawi, and J. Abawajy, Eds., pp. 372–384, Springer, Berlin, Germany, 2018.

[15] H. Coates, Student Engagement in Campus-Based and Online Education, University ConnectionsRoutledge, London, UK, 2006

[16] Nabila Amir; Fouzia Jabeen; Sidra Niaz *"A Brief Review of Conditions, Circumstances and Applicability of Sampling Techniques in Computer Science Domain"*,2020 IEEE 23rd IEEE International Multi-Topic Conference

[17] IanGoodfellow, YoshuaBengio and AaronCourville, *"Neural Networks and Introduction to Deep Learning"*,IEEE 2018