

# Machine Learning Approaches to Predict Learning Outcomes in Massive Open Online Courses

Raghad AL-Shabandar, Abir Hussain, Andy Laws  
Robert Keight, Janet Lunn  
Applied Computing Research Group  
School of computing and Mathematical School  
Liverpool John Moores University  
Liverpool, UK  
R.N.ALShabandar@2013.ljmu.ac.uk  
{A.hussain,A.Laws,J.Lunn}@ljmu.ac.uk,  
R.Keight@2015.ljmu.ac.uk

Naeem Radi  
Al Khawarizmi International College  
Abu Dhabi, United Arab Emirates  
n.radi@khawarizmi.com

**Abstract**— With the rapid advancements in technology, Massive Open Online Courses (MOOCs) have become the most popular form of online educational delivery, largely due to the removal of geographical and financial barriers for participants. A large number of learners globally enrol in such courses. Despite the flexible accessibility, results indicate that the completion rate is quite low. Educational Data Mining and Learning Analytics are emerging fields of research that aim to enhance the delivery of education through the application of various statistical and machine learning approaches. An extensive literature survey indicates that no significant research is available within the area of MOOC data analysis, in particular considering the behavioural patterns of users. In this paper, therefore, two sets of features, based on learner behavioural patterns, were compared in terms of their suitability for predicting the course outcome of learners participating in MOOCs. Our Exploratory Data Analysis demonstrates that there is strong correlation between click stream actions and successful learner outcomes. Various Machine Learning algorithms have been applied to enhance the accuracy of classifier models. Simulation results from our investigation have shown that Random Forest achieved viable performance for our prediction problem, obtaining the highest performance of the models tested. Conversely, Linear Discriminant Analysis achieved the lowest relative performance, though represented only a marginal reduction in performance relative to the Random Forest.

**Keywords**—Massive Open Online Courses (MOOCs); Receiver operating characteristic (ROC), Educational Data Mining (EDM)

## I. INTRODUCTION

Massive Open Online Courses (MOOCs) have become an alternative educational platform that allows learners from dispersed geographic locations access the same quality of learning through the web[1]. Coursera, HarvardX, and Khan Academy are some examples of MOOCs. Since 2012, MOOCs modalities have received widespread usage by top Universities[1]. Investigations undertaken by such institutions indicated that the use of MOOCs have attracted many participants towards engagement in the space of courses offered, due to the removal of financial,

geographical, and educational barriers [1]. A large volume of data can be collected and captured from MOOCs platforms during Student interaction with learning activities, such as viewing of video lectures, undertaking of quizzes, posting in discussion forums, and interacting with the courseware[1],[2],[3]. Data captured from MOOCs can provide valuable information for educators by analysing the patterns present in the behaviour of learners[2], [3].

Educational Data Mining (EDM) is an emerged field of research aimed at extracting knowledge from learning processes to support decision makers[4]. Recently EDM has been used within the higher education setting to enhance teaching strategies [4]. EDM involves the use of statistics, visualization, and machine learning methods for the exploration and analysis of educational data[4],[5]. The possibility of capturing big data within MOOCs opens new horizons to educational data mining researchers who could extract deeper insights from the analysis of the data[5]. Although a prominent application of EDM is set within the online learning environment, the analysis and tracing of actionable data is challenging[5]. Learning Analytics (LA) is a new field of research that aims to improve the quality of education[4],[6]. LA is an analytics approach directed towards the analysis, measurement, and extraction of comprehensive information about the learner from various features, including cognitive, social, and psychological facets, to help the educators make decision about the learner's success and failure[4],[6]. There are various methods utilised by researchers into LA including Web analytics, Artificial Intelligence and Social Network Analysis [4]. The main feature of LA is its capacity to analyse actionable data in more objective way[6],[7]. The analysis of such big data will assist educators in drawing inferences about student performance with deeper insight[7]. Although a number of works have been reported in the literature to evaluate the learner performance in e-learning environment, it still challenging to build predicative models for MOOCs[1]. In this paper, LA tool is utilised to provide an advantage over EDM, by tracing student knowledge, precisely analysing behaviour, and measuring how such factors can affect student performances. Machine learning is an effective technique that can be applied to Learning

Analytics with the capacity to discover hidden patterns of student interaction with the MOOCs. Machine learning offers an advantage over traditional forms of statistical analysis, placing emphasis on predictive performance over provable theoretical properties and priori super-population assumptions[1]. Moreover, a key feature of machine learning is the capacity to analyse complex non-linear relationships, given that complex input variables are expected[4],[7]. Various supervised machine learning approaches have been conducted in this study to predict the learning outcome in MOOCs. The remainder of this paper is organised as follows. Section II will provide detailed information about previous works, while section III shows the methodology, which includes data descriptions, data pre-processing, data analysis, and experiment setup. The conclusion and future works are described in Section IV.

## II. RELATED WORK

The advancement of Information and Communications Technology (ICT) has increased the growth of MOOCs applied in distance learning environments[1],[7]. Different approaches have been designed using both EDM and LA with the aim of understanding and analysing learner interaction in MOOCs efficiently[4]. LA has been used to identify dropout students[3]. For example, the University of Michigan developed Michigan Tailoring System (E2Coach)[8]. The E2Coach is an open source system aims to identify weaknesses and performance skills of physics students. E2Coach also delivers personalized learning by the customization of course material. The LA tool was implemented in E2Coach to capture and collect data about students' progress from various resources and provide indications to educators to reconstruct learning materials that match student ability and experience[8]. In[1] reference the author proposes a model to predict the latent learning behaviour in MOOCs. Various features have been considered including demographic, assessment grade, post forum, and click stream, for the purpose of obtaining more accurate prediction. The model incorporates logistic regression, support vector, and matrix factorization techniques into Dynamic factor model[1].

Other researchers focus on clustering techniques. In such works, researches cluster learners into groups, according to their patterns of behaviour[9],[10]. In reference[9], the authors employ Self Organised Map clustering to describe the learner behaviour in e learning. They have found SOM clustering is a powerful approach in terms of visualising the behavioural patterns of learners, due to the capacity to analyse high dimensional data with different type of input variables. The authors in[10] identify four different classes of learner engagement within MOOCs based on two core attributes: video lecture and assignment grades. These classes are Completing, Auditing, Disengagement, and Sampling. The Completing class represents learners who submitted assessments on time. Auditing class represents learners who did not submit assessments but watched video lecture content; Disengagement represents learners who dropout from the course; Sampling represents learners who watch video on only a single occasion[10]. In this case, the authors used clustering techniques to describe engagement activity in

MOOCs. Support vector machine (SVM) and Least Mean Square (LMS) algorithms have been used to detect the likelihood of learners' dropout rates from MOOCs over weeks where only click stream features were available[11]. A number features have been extracted from learners' historical data such as, the number of sessions, number of time viewing videos and courses[11]. Feedforward neural network have implemented in[12] to predict student attention in MOOCs, considering student sentiments. In this case, only the behavioural attributes are used to measure the performance of learners. Our work differs from the prior research works as it concentrates on the analysis of various factors affecting the learners' outcome in MOOCs. In order to discover the complex correlation between the predictor variables, we utilised two types of neural network, defined as Feedforward Neural Network (FFNN) and Self Organised Map (SOM). We used the two types of network, where SOM was used in a supervised capacity, to predict if learners would achieve certification at the end of course, or not.

## III. METHODOLOGY

### A. A.Data Description

The dataset used in this paper was obtained from Harvard University[13],[14]. Harvard University collaborates with Massachusetts Institute of Technology (MIT) to deliver high quality MOOCs. During the first year of providing the MOOCs, 15 courses have been offered by Harvard and MIT[13]. The courses cover variety of subjects, such as Computer Science, Mathematics, Humanities, Health, and Social Sciences[13]. Across all courses, 597,692 participants were registered, only 30% of registrants succeeded to achieve certification[13]. The approximate percentage of learners who viewed the main courseware content and then subsequently dropped out from the courses is reported at around 25%.[13]. The number of overall participants has markedly increased, with 1.3 million unique learners engaged in multiple courses reported at the end of 2014 [14]. Two sets of features are considered in the dataset, learner behavioral features, followed by demographic attributes[13],[14]. The primary feature of the dataset is the Click stream, which represents the number of user events relating to video lecture views, course content interaction, access to assignments, and posts in discussion forums[13],[14]. The participants' demographic information is also considered in the dataset, such as age, gender, and educational background [13],[14]. Additionally, the date of learner registration in the course and the last learner activity was also captured[13],[14]. The assignment grade is an indicator attribute to denote if a certificate from Harvard university is granted for a given student. If the weighted course mark ranges between 0.50-0.90, the registrants will gain certification, otherwise they are ineligible to obtain a certificate[13],[14]. The features denoting user exploration and viewed content are binary features discretize the percentage of exploration and course content viewing, respectively[13],[14]. If participants access more than half of the course content (chapter), the explored feature is encoded as 1, or 0 otherwise [13],[14]. The viewed content is encoded as 1 when the participants access the home page of assignments and related videos, or 0 otherwise[13],[14]. researchers have used the explored and viewed features to

measure what kinds of behavioral data could affect the likelihood of certification gain. As such, the results show during the first year a certification rate of 40%, where around 60% of the certificated learners were fulfilled the criteria for explored participants[13],[14].A brief description of the dataset attributes has been explained in Table I.

TABLE I. Features Description of HarvardX

Features	Description
User-Id LOE,YOB,Gande,Grade	Demographic feature of user including User_id, sex, date of birth, GPA
Start_time_DI, last_event_DI	Date features describe start and end user interact with course.
Certified	Target binary class encoded 1/0.
Nevent nplay_video, Nchapters, nforum_post	Behavioural features including the number of click stream, play video event, interact with chapter.
Viewed, Explored	Discrete features encoded as 1/0.

### B. Data Pre-Processing

The data used in this study consists of 800,000-log file representing the completed learners' activities on MOOCs, where each row represents a single user session. Preprocessing was applied to this data, involving cleaning, example extraction, target class balancing, and scaling. The dataset has been cleaned by applying various techniques including the removal of duplicate rows, followed by the imputation of missing values with estimated numeric values. K-Nearest Neighbor algorithm was used to estimate the missing values by selecting neighboring values based on the Euclidean distance. One of the issues of capturing data in MOOCs is the large size of data. To reduce the large quantities of data, an aggregation procedure has been implemented in features. For instance, aggregating multiple learners' activities belonging to the same course during the same day to a single unit of activity results in the formation of smaller versions of the dataset. Class imbalance is another issue that occurs in the dataset. In this case, 95% of the class instances occurred with the value 'not certified', while 4% of the data occurred with the class 'certified'. To solve this problem under sampling the majority class ('not certified') was used.

### C. Exploratory Data Analysis

There are various graphical and non-graphical techniques employed for Exploratory Data Analysis (EDA), such as plot, Principle Component Analysis (PCA), histogram, and correlation matrices[15]. In the educational domain, EDA approach has been extensively used within the distance education environment [15]. For example,[16] made use of histograms to track the number of most visited page in MOOCs using different time interval, while[17]applied correlation analysis to extract information about student progress assessment and to understand the pattern of their

feedback for online courses. In this paper, EDA was applied to the dataset as a precursor to the modelling phase. The objective of data visualization is to provide an insight into the learners' behavior, in conjunction with their performances. Considerate should be noted that only behavioral data are considered when investigating the effect of patterns in learner behavior on the user certification rate. The correlation matrix was applied to measure the dependency between the behavioral data and learners' certification. Figure 1- shows the plot of correlation matrix, which indicates a positive relationship between three behavioral attributes and the target variable. The explored and click stream attributes show moderate positive correlation with the target, with a coefficient value of 0.65 and 0.64 respectively.

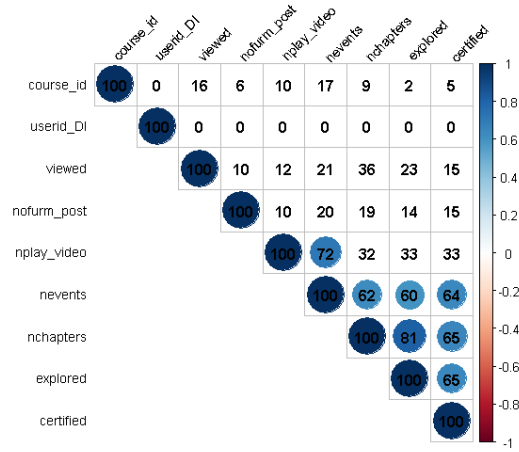


Fig.1 Correlation Matrix

### D. Experiment Setup

The method implemented in this paper follows a binary classification problem. Various linear and nonlinear supervised machine learning has been employed to predict if the learner obtain certification. Machine Learning models are Logistic Regression (LR), Linear Discriminant Analysis (LDA), Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Neural Network (MLP) and Self Organized Map (SOM). Table III illustrates a brief description of the models used in this study. In this case study, the data is segmented into a number of the subsets with records of 8000 learners in each subset. All dataset features have been considered, including both behavioral features and demographic categories, as listed in Table I. We investigate the most important features that influence the learners' performance. The Random Forest (RF) algorithm was used to rank features from the Harvard dataset[18],[19]. The algorithm computes the weight of each attribute by evaluating a loss function [18]. Hill climbing was used to select the optimal subset of features. Hill climbing is search algorithm performing a partial exploration of the power set of features to find a candidate that is close to optimal [19]. The results obtained by both RF and Hill climbing show that both indicate the same subset of features. Table II shows the features of original dataset with the resulting weight measurement. The

features with higher weights correspond to the most important features. It is clear that the click stream feature obtains the highest weight with a value of 74. Repeated k - fold cross validation was applied during the modelling to overcome the problem of overfitting by randomly partitioning the original sample of data into folds based on resampling. The cross validated training set was allocated 70 % of original dataset, the subset elements were randomly partitioned into 10 equal size subsets. For each round of cross validation, 9-fold subsets are used as the train set and single subset is used as a test sample. The cross validation procedure was repeated 3 times at each fold. A further 30% of the data, disjoint from the cross validation set, was used to evaluate the generalisation error for each classifier.

The hyperparameter optimization problem for each model is also considered in order to establish an approximately optimal configuration for each classifier[20]. The procedures applied to each model are as follows. A random search was used in RF to find the optimal number of attributes that were randomly sampled at each split, the optimum value obtained was (15, 2) for experiment 1, experiment 2 respectively. For the SVM model, the smoothing width parameter of the kernel function, denoted as sigma, was tuned through the application of grid search. The optimal value was found to be (0.05, 0.34) for both experiments. Subsequently, for the NN model, we explored both the weight decay term and the number of units present in the hidden layer, implementing a grid search to search for optimal combinations. Our results produced values of (0.01, 6) for first experiment and (0.1,3) for second experiment. Grid search was also used to tune the number of iterations and learning rate in SOM models. The result found optimal number of iteration and learning rate are (100,0.01) achieved high accuracy for both sets of experiments. In NB, the tuning was applied by set the kernel density estimation to estimate density in both experiment. The subsequent models ( LG,IDA) do not have a multi test requirement, since no hyperparameters are defined, thus model training alone was sufficient.

TABLEII. Harvard Dataset Features Weight

Feature	Weight
nevents	74.066449
nchapters	56.751807
explored	52.299048
course_id	49.759444
start_event	40.620314
nplay_video	37.324701
last_event_DI	22.642938
final_cc_cname_DI	14.322623
diseng	13.595184
viewed	11.062209
gender	10.670729
nofurm_post	5.161551
LoE_DI	5.550157
userid_DI	1.647105

TABLE III . Brief Description of ML Models

Model	Description	Architecture	Type	Algorithm
DT	Decision Tree	Recursive partition Decision rules	Nonlinear	C4.5 algorithm
RF	Random Forest	Ensemble DT	Nonlinear	Random subset Features Bootstrap
SVM	Support Vector Machine	Hyperplane kernel trick	Nonlinear	Quadratic Optimisation
NB	Naive Bayes	Bayesian Decision Rule	Linear	Maximum Likelihood Estimation
NN	Feedforward Neural Network	Units 14-3-2	Nonlinear	Backpropagation
LG	Logistic regression	Generalised Linear Model	Linear	Maximum Likelihood Estimation
LDA	Linear Discriminant Analysis	Generalized Linear Model	Linear	Maximum Likelihood Estimation
SOM	Self-Organised Map	Unit 25-3-2	Nonlinear	Competitive learning

### E. Result Evaluation

The following section considers the empirical results obtained from our experiments. In this case study, machine learning has been applied to two subsets of features, whose results we denote as experiments 1 and 2 respectively. In the first experiment, we consider all dataset features, while in the second experiment we include only high weighted features, as evaluated using the Random Forest method. During the model training stage, we evaluated the fit of classifiers to the data using cross validation. Figure 2 and Figure 3 compare estimations of classifier accuracy of training set for all models over both subsets. The graph shows nearly both set of experiments have same accuracy. The Confusion Matrix was used to evaluate the performance of classifiers over the test dataset. A number of performance metrics are considered, including sensitivity, specificity, kappa, and accuracy. The metrics calculated as describe in Table IV. The accuracy performance of classifier models was enhanced by tuning their parameters explained in pervious section. The result of our experiments reported in Table VI and VII 6 after applying tuning processes. Both tables illustrate the result of classifier performances based on confusion matrix metrics. The simulation result from Experiment 1, associated with all dataset features, is shown to yield slightly higher performance than the second experiment, which used a selected features subset. The RF achieved the highest accuracy of 0.9881 in first experiment and 0.9851 in second experiment. NN, SVM, and DT give lower though compelling results, with an accuracy of 0.9856, 0.9844, and 0.983 in first experiment. Conversely, DT and SVM achieved similar accuracy with values 0.9731 in second experiment. The NN accuracy in the second

experiment is less than first experiment with a value of 0.9729. In both experiments, SOM has a lower performance than other nonlinear classifiers, achieving values of 0.9765 and 0.9569 respectively. LG and LDA classifiers achieved the lowest range of performances, with accuracies of 0.9754, 0.9656, 0.9546, and 0.9544 in first experiment, second experiment respectively. The result in both set Experiments trend to obtain high performance of accuracy with approximate values of 98% in nonlinear models (NN, SVM, DT). This is due nonlinear form of relationship between the database features and target. The selected nonlinear classifier models are powerful models. These models are capable to handle nonlinear distinguishable problems by transfer feature to high dimensional space. In order to evaluate the feasibility of classifier models for both Experiments, computational performance was considered. The figure 6 shows the speed run time measured in second for each learning algorithm .In general, time require to train all features is longer than selected features for all classifiers models. The fastest algorithms speed ware (IDA,LR) with achieving a value of (60, 25) seconds for experiment 1, experiment 2 respectively. The average run time for both (NB ,DT) ware 120 seconds in first experiment that decrees only 30 second in the second experiment. SVM algorithms does not particularly effect by size of features. As can be seen, average run time in SVM achieving value of 120 seconds when selected all features ware slightly declined on experiment 2. There is a gap in training time between the first experiment and the second experiment in both (NN , SOM )models, approximately half of time declined when train high ranking features. The RF is slowest learning algorithms compared with other algorithms, with an acquired average running time of (480, 220 ) for experiment 1 , experiment 2 respectively. A number of reasons could affect the speed of RF. The significant one is a learning algorithm. RF is ensemble learning algorithm based on bootstrap method. The number of features randomly sample at each step during the learning stage lead to grow multiple trees. To obtain a further evaluation of our classifier model, the Receiver Operator Characteristic (ROC) and area Under Curve (AUC) were considered. Figures 4 and 5 show ROC for both experiments. The curves are shown to converge to roughly the same semblance on the plot, indicating the similarity of performance across models.

TABLE IV. Performance Metrics

Metric Name	Computation
Accuracy	$(TP+TN)/P+N$
Kappa	$PR(actual)-PR(expected)/(1-PR( expected )$
Sensitivity	$TP/(TP+FN)$
Specificity	$TN/(TN+FP)$

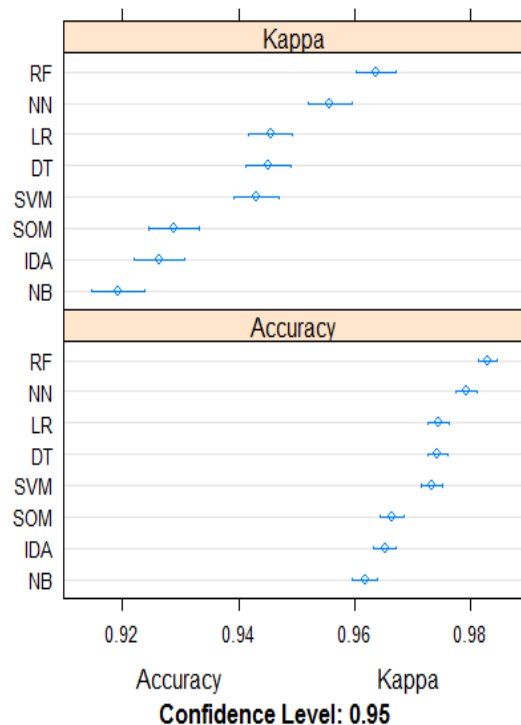


Fig.2. Estimation Accuracy Classifier Experiment 1

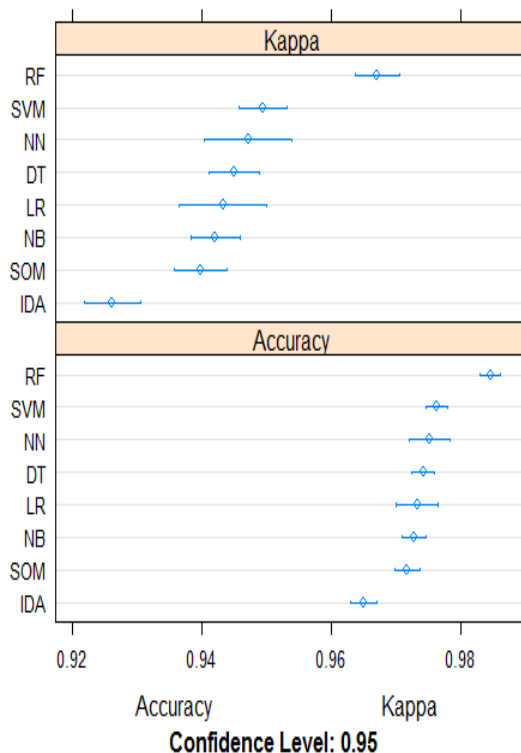


Fig.3. Estimation Accuracy Classifier Experiment 2

TABLE VI. Classification Performances for Experiment 1 (All Features)

Model	Acc.	Kappa	Sens.	Spec.	AUC
DT	0.9831	0.9661	0.9775	0.9894	0.9789
RF	0.9881	0.9762	0.9846	0.9920	0.9973
SVM	0.9844	0.9686	0.9811	0.9880	0.9939
SOM	0.9765	0.9448	0.9693	0.9761	0.9726
NB	0.9794	0.9397	0.9775	0.9615	0.9939
NN	0.9856	0.9712	0.9811	0.9907	0.9856
LG	0.9754	0.9586	0.9728	0.9867	0.9946
IDA	0.9656	0.9312	0.9657	0.9655	0.9942

TABLE VII. Classification Performances for Experiment 2 (select features)

Model	Acc.	Kappa	Sens.	Spec.	AUC
DT	0.9731	0.9461	0.9693	0.9774	0.9978
RF	0.9851	0.9515	0.9882	0.9615	0.9978
SVM	0.9731	0.9463	0.9728	0.9734	0.9916
SOM	0.9569	0.9136	0.9512	0.9632	0.9569
NB	0.9621	0.9199	0.9500	0.9611	0.98726
NN	0.9729	0.9523	0.9728	0.9801	0.99427
LG	0.9546	0.9111	0.9524	0.9592	0.9881
IDA	0.9544	0.9086	0.9464	0.9632	0.98685

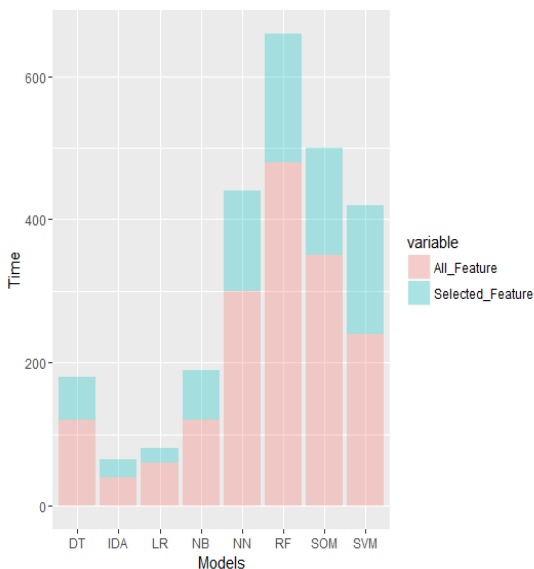


Fig. 6. Comparing Experiments Computational Time

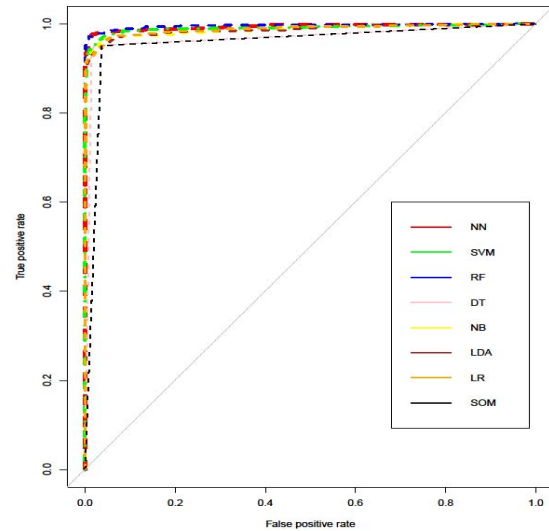


Fig. 4. Roc Curve for Experiment 1

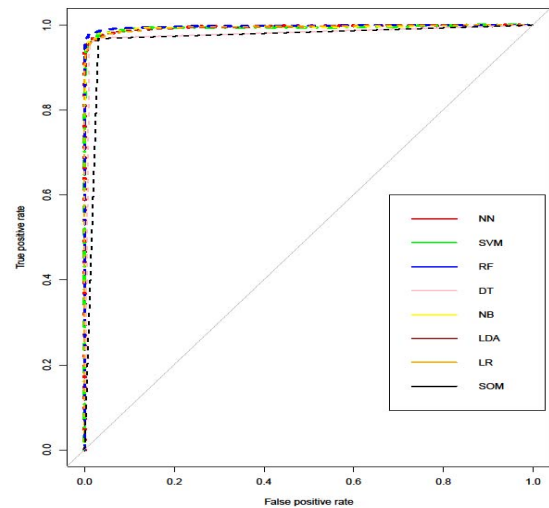


Fig. 5. Roc Curve for Experiment 2

## F. Discussion

The experiments in this study aimed to predict the performance of student participations in MOOCs. A series of *data* pre-processing methods were undertaken, including data scaling, imputing of null values, and class balancing. The correlation matrix was used to measure the interaction between attributes. The results revealed a moderate linear relationship between the target outcome and both the click stream and explorer features, exhibiting coefficients of 0.57 and 0.64 respectively. In this paper, two types of experiments have been conducted. In the first set of experiments, all the features are used, and passed to the ML, while in the second set of experiments only high ranked features are used. We compared the result for both set of experiments, observing a number of similarities between them in terms of performance metrics. In the first set of



experiments achieved a close to ideal specificity for nonlinear classifiers, ranging from 0.99 to 0.97. Conversely, nonlinear classifiers in the second set of experiments showed a marginally lower specificity with values between 0.98 and 0.96. Linear classifiers obtained less specificity for both set of experiments, with values bounded between 0.96 and 0.95. However, LR in first experiment obtained better specificity, with a value of 0.98. RF, NN, and SVM classifiers also obtain the highest sensitivity in first set of experiments with a value of 0.98, whereas NN and SVM obtain slightly lower sensitivity in second set of experiments. 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. The ROC curve was also used to derive both an AUC and to choose a suitable decision threshold value for the true negative and false negative rates of each classifier. Overall, in both sets of experiments an AUC of 0.90 was obtained for all classifiers. RF in both experiments presented the highest AUC at around 0.99, whereas SOM achieved the lowest AUC with values of 0.97 in both set of experiments. Average run time of machine learning models in first experiments compared to models of second experiment, the results demonstrate a significant difference of speed run time. The fastest models were linear models in both set experiments. The nonlinear model require a half amount of time to train models in the second set of experiments.

#### IV. CONCLUSION AND FUTURE WORK

This study was undertaken to examine the effectiveness of machine learning approaches for the behavioral analysis and prediction of student outcomes within MOOCs. Behavioral features were used in conjunction with demographic features to predict whether learners gained certification in MOOCs. In this work, two set experiments have been applied. In the first set of experiment, all features from the dataset were included. For the second, a subset of features were considered which selected using a RF approach. Various binary machine-learning approaches have been applied over both experiments to predict the learning outcomes relating to a Harvard dataset.

The simulation results in both experiments indicate that RF and SVM achieved ideal performance, with the accuracy values of 0.9881 and 0.9851 respectively. Other classifier models gave lower performance, for instance NB showed a value of accuracy 0.9794, and 0.9621 for both set of experiments. The results show that machine learning is a viable approach to our problem, providing an exceptional capability to distinguish between success and failure outcomes. Two set of experiments have been compared in term of computational performance. The result shows average run time of machine learning models is much longer in first experiment than second experiment. Future work will investigate passive engagement within MOOCs in terms of the effect on learning outcome. The learner emotional states of students are considered to be a latent variable, which can be inferred from their interaction with online courses over time. We will construct a robust

predictive model, taking into account the latent learner engagement as unlabeled data within MOOCs. Semi-supervised machine learning approaches will considered including Low density speartion and Generative models.

#### REFERENCES

- [1] J. Qiu *et al.*, "Modeling and Predicting Learning Behavior in MOOCs," *Proc. Ninth ACM Int. Conf. Web Search Data Min.*, pp. 93–102, 2016.
- [2] A. Ramesh, D. Goldwasser, B. Huang, H. Daum, and L. Getoor, "Learning Latent Engagement Patterns of Students in Online Courses," in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence Learning*, 2014, pp. 1272–1278.
- [3] M. Kloft *et al.*, "Predicting MOOC Dropout over Weeks Using Machine Learning Methods," *Knowl. Manag. E-Learning*, vol. 4, no. 3, pp. 60–65, 2014.
- [4] R. S. J. D. Baker and G. Siemens, "Educational Data Mining and Learning Analytics," in *Cambridge Handbook of the Learning Sciences*, 2014.
- [5] D. M. D. M. West, "Big Data for Education: Data Mining, Data Analytics, and Web Dashboards," *Gov. Stud. Brookings, US Reuters*, no. 1, 2012.
- [6] J. Fiaidhi, "The next step for learning analytics," *IT Professional*, vol. 16, no. 5, pp. 4–8, 2014.
- [7] D. Gašević, C. Rose, G. Siemens, A. Wolff, and Z. Zdrahal, "Learning Analytics and Machine Learning," *Proc. Fourth Int. Conf. Learn. Anal. Knowl. - LAK '14*, pp. 287–288, 2014.
- [8] K. D. Mattingly, M. C. Rice, and Z. L. Berge, "Learning analytics as a tool for closing the assessment loop in higher education," *Knowl. Manag. E-Learning*, vol. 4, no. 3, pp. 236–247, 2012.
- [9] U. F. Alias, N. B. Ahmad, and S. Hasan, "Student behavior analysis using self-organizing map clustering technique," *ARPJ J. Eng. Appl. Sci.*, vol. 10, no. 23, pp. 17987–17995, 2015.
- [10] R. F. Kizilcec, C. Piech, and E. Schneider, "Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses," *Lak '13*, p. 10, 2013.
- [11] J. He, J. Bailey, and B. I. P. Rubinstein, "Identifying At-Risk Students in Massive Open Online Courses," *Proc. 29th AAAI Conf. Artif. Intell.*, pp. 1749–1755, 2015.
- [12] D. S. Chaplot, E. Rhim, and J. Kim, "Predicting student attrition in MOOCs using sentiment analysis and neural networks," *Work. 17th Int. Conf. Artif. Intell. Educ. AIED-WS 2015*, vol. 1432, pp. 7–12, 2015.
- [13] A. D. Ho *et al.*, "HarvardX and MITx: The First Year of Open Online Courses, Fall 2012-Summer 2013," *SSRN Electron. J.*, no. 1, pp. 1–33, 2014.
- [14] A. D. Ho *et al.*, "HarvardX and MITx: Two Years of Open Online Courses Fall 2012-Summer 2014," *SSRN Electron. J.*, no. 10, pp. 1–37, 2015.
- [15] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 40, no. 6, pp. 601–618, 2010.
- [16] M. Wen, D. Yang, C. P. Ros, C. P. Rosé, and C. P. Rose, "Linguistic Reflections of Student Engagement in Massive Open Online Courses," *Proc. 8th Int. Conf. Weblogs Soc. Media, ICWSM 2014*, pp. 525–534, 2014.
- [17] C. S. Ong, J. Y. Lai, and Y. S. Wang, "Factors affecting engineers' acceptance of asynchronous e-learning systems in high-tech companies," *Inf. Manag.*, vol. 41, no. 6, pp. 795–804, 2004.
- [18] D. W. Aha and R. L. Bankert, "Feature selection for case-based

classification of cloud types: An empirical comparison,” *Proc. AAAI’94 Work. CaseBased Reason.*, pp. 106–112, 1994.

- [19] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, “Variable selection using random forests,” *Pattern Recognit. Lett.*, vol. 31, no. 14, pp. 2225–2236, 2010.
- [20] D. R. Tobergte and S. Curtis, *Machine learning with R*, vol. 53, no. 9, 2013.