ENTHUSIASTIC
INTERNATIONAL JOURNAL OF APPLIED STATISTICS AND DATA SCIENCE
Volume 4, Issue 1, April 2024, pp. 25-36

Characterization of Student's Performance in Massive Open Online Courses (MOOC)

Tan Ching Joe [a,1,*]

[a] Lee Kong Chian Faculty of Engineering & Science, Universiti Tunku Abdul Rahman, Jalan Sungai Long, Bandar Sungai Long, Cheras, 43000 Kajang, Selangor Darul Ehsan, Malaysia
[1] cjtan00@icloud.com *
* Corresponding author

ARTICLE INFO

Keywords
MOOC
predictive modeling
logistic regression
graph convolutional network
clustering algorithms

ABSTRACT

Massive Open Online Courses (MOOC) allow students to learn online at any time and from any location. Unfortunately, poor completion rates and a large student group make it difficult for teachers to keep track of their student's progress. Due to a lack of adequate counselling, students who perform poorly are more likely to give up. The goal of this study was to predict student's certification by analyzing data on student's learning behavior. The initial data on learning behavior was obtained from edX, a well-known MOOC platform. Based on this data, three statistical models such as logistic regression, graph convolutional network, and cluster analysis were utilized to predict student's performance. The proposed model's usefulness was demonstrated by using a testing set of data from the actual courses. Our findings showed that tracking student activity in terms of number of unique days active, watching videos, participating in forum discussions, and exploring more courseware content might help predict student's performance in MOOC and enhance completion rates.

1. Introduction

The advancement of technology, as well as the ease of communication via the internet has drastically changed the landscape of higher education instructing and learning. Massive Open Online Courses (MOOC) are the new trend of online educational platforms [1]. The advantages of MOOC are as follows: (1) content accessed by handling scalability difficulties so that course materials and recordable instructions can be accessed from any device, at any time; (2) enhance learning quality by allowing users to access and study material at their own pace; and (3) access to proactive content, free materials, and curriculum customization [2].

Over the past three years, the landscape of online education has undergone a remarkable transformation with the proliferation of MOOC. As of late 2012, prominent commercial platforms including edX, Khan Academy, Coursera, and Udacity had emerged as pioneers, offering a diverse array of academic courses sourced from esteemed institutions worldwide [3]. This expansion wasn't confined solely to North America; Europe also witnessed the rise of platforms like FutureLearn and Iversity, contributing to the global accessibility of MOOC. Moreover, the MOOC phenomenon transcended continents, with universities in China embracing the trend by either partnering with existing platforms or establishing their own [4]. This widespread adoption underscores the increasing recognition and acceptance of online education as a viable and impactful learning avenue on a global scale.

MOOC bring learners together through a collection of video lectures, readings, quizzes, peer-graded assessments, and discussion forums. MOOC self-paced learning environment provides a lot of freedom, allowing students to choose from a variety of topics. MOOC, on the other hand, have a vast number of students from all over the world. As a result, pupils have a wide range of needs, interests, motivations, and learning styles. For instance, not every student aspires to receive a course certificate or to complete the course satisfactorily. Thus, since students have different objectives in learning, MOOC serve as a good choice for everyone to meet their needs [5].

Various definitions are employed in the assessment of student performance within the realm of MOOC. Researchers commonly distinguish between completion and drop-off rates, where completion is typically defined by criteria such as final exam submission, consistent engagement, or certificate attainment. While this metric provides insight, it may not fully capture student achievement due to the diverse array of learning goals and motivations that extend beyond course completion [6].

In our study, we align with the conventional measure of course completion, focusing on students who earn a certificate. We acknowledge the inherent challenge in quantifying individual learning objectives, which may be diverse and not always explicitly stated. By emphasizing completion rates, we aim to provide a pragmatic assessment of student success within the context of MOOC, recognizing that while students may have multifaceted learning aspirations, certificate attainment serves as a tangible indicator of accomplishment.

This study aimed to delve into the analysis of MOOC data, focusing on exploring the dataset, identifying key variables, and assessing student performance. The research questions revolve around extracting valuable insights from the dataset, understanding big data analysis approaches, addressing challenges in MOOC analysis, and predicting student performance based on learning activities. Objectives include studying big data exploration in MOOC data, identifying significant information, understanding predictive modeling approaches, recognizing research challenges, limitations, and ultimately building a predictive model for forecasting student performance.

This study focused on analyzing the first academic year MOOC data from edX to predict student performance, recognizing the vast potential of MOOC data in educational research. By leveraging this data, insights can be gained into student utilization patterns and obstacles hindering optimal usage of MOOC platforms. Suggestions derived from such analyses can aid in enhancing student performance, benefiting students, educators, and institutions alike. Furthermore, insights garnered can inform improvements in learning settings, content, teaching methodologies, and platform enhancements. However, the study faces challenges such as a limited number of positive samples for analysis, response bias among participants, and restricted information due to the de-identification of the dataset by edX. These challenges necessitate careful consideration during analysis and interpretation of results, though they may limit the depth of conclusions drawn.

Several works have studied various aspects of MOOC analysis methodologies. Few studies have delved into motivations and learning behaviors, showcasing the significant impact of motivation on student engagement and performance. Motivation significantly impacts students' behavior and performance in online learning environments, with those demonstrating a mastery approach and value belief motivation showing better engagement in learning activities [3]. Research suggests that active participation in course activities and regular communication, discussion, sharing, and cooperation with peers correlate with higher likelihoods of course completion [7]. Understanding and fostering motivation are essential for promoting student success in MOOCs and other online learning platforms.

Regression analysis, a statistical technique, determines causal links between variables using mathematical models. In cases of binary dependent variables, logistic regression is employed. Studies comparing supervised machine learning algorithms for predicting student performance reveal that logistic regression demonstrates superior accuracy and precision [2]. Additionally, research comparing linear regression with other methods, such as support vector regression and neural networks, concludes that linear regression yields the lowest prediction error, indicating its effectiveness in predicting student academic performance [3].

Graph Convolutional Networks (GCN) are gaining popularity for their ability to extract meaningful feature representations from data, surpassing traditional classification methods [4]. Research indicates GCNs' effectiveness in detecting complex patterns, such as financial fraud, without explicit training data [5]. Stacking multiple layers in a GCN creates deep networks, with optimal performance observed in models with two to three layers. In education, GCNs show promise in identifying students' learning styles and forecasting their performance in online courses, outperforming baseline techniques in terms of accuracy and practicality [6]. These findings highlight GCNs' potential to enhance educational outcomes by efficiently categorizing student representations and predicting performance.

Clustering is an unsupervised learning algorithm that groups data into clusters based on similarity. The performance of a good clustering algorithm relies on its ability to uncover hidden patterns and maximize intraclass similarity while minimizing interclass dissimilarity between objects among clusters. Examples of clustering algorithms include hierarchical, k-means, and DBSCAN, with applications in categorizing student profiles based on activity levels. Studies suggest that active learners tend to complete courses, while passive learners engage intermittently [8]. Some researchers recommend using k-means and DBSCAN for predicting student performance, with DBSCAN demonstrating superior effectiveness [9].

2. Methodology

This section discusses the research methodologies used to meet the objectives of this project. The primary focus of this project was to develop a predictive model to forecast student's performance. The database used in this project was developed throughout the first academic year (fall of 2012 to summer of 2013). A total of 16 courses are included in the edX database during the first year.

2.1. Exploratory Data Analysis

Exploratory data analysis was performed by using the programming software, namely python, to seek for abnormalities, interesting trends, or patterns as well as correlations between variables with the student's performance. Student's behaviors were analyzed to relate the influence of student's learning process with their performance. Each column in the dataset was analyzed to understand the student's behaviors and information was extracted from the dataset.

2.2. Data Preprocessing

Data preprocessing is important as it resolves issues and makes datasets completer and more efficient to perform data analysis. There are a few tasks included such as variables selection, one-hot encoding process, and normalization. Based on the exploratory data analysis, among the 21 columns variables, only 9 variables (i.e., "`educational status`", "`gender`", "`duration`", "`nplay_video`", "`ndays_act`", "`nforum_posts`", "`grade`", "`viewed`" and "`explored`") showed significant information in predicting the student's performance (variable "`certified`"). Then, categorical data was converted into new binary columns to indicate the information from the original column. It is known as the one-hot encoding process. Referring to the dataset, variable "`gender`" and "`education status`" are both categorical variable, one-hot encoded each variable into binary columns representing the possible information of respective variables. Furthermore, each variable was normalized so that the values were comparable. When all the dataset values were in similar ranges, this was able to increase the accuracy in predicting student's performance.

2.3. Data Splitting and Training

This process was served to split the dataset randomly and evaluate the performance of the model later. By using the `.train_test_split()` in scikit-learn function, dataset was split into train set and test set to examine how the model was generalized to previously unknown data. Besides, Synthetic Minority Oversampling Technique (SMOTE) was performed to create extra samples for the minority class using k- nearest neighbor technique, increasing the likelihood of predicting the minority class in a training session. It is because imbalance dataset will affect the classifier performance which it skews toward the majority class instances while the minority class suffers. However, in order not to affect the accuracy as well as performance of our model, only the train set was applied with the SMOTE sampling technique. Since no information from the test data was utilized to construct synthetic observations, no information from the test data was bleed into the model training.

2.4. Data Modeling

Since the response variable ("`certified`") only showed binary outcomes, logistic regression model with all predictors in the new dataset was built. All predictors were employed to simulate the probability of the outcomes. To determine whether this model showed statistically significance in predicting student's performance, model summary was generated to interpret the model and model coefficients. Statistical test was performed with p-value approach for each of the variable's coefficient. Besides, classification report of the logistic regression model was obtained to assess the validity of a classification algorithm's predictions and Receiver Operating Characteristic (ROC) curve was also plotted to view the trade-off between the true positive rate and the false positive rate. The Area Under the Curve (AUC) was also calculated to determine how efficient a logistic regression model distinguished positive and negative outcomes at all relevant cut-offs as well as comparison between models. On the same hand, another logistic model with the variables selected using the `SelectKBest()` scikit learn function with the `f_classif` score function was built. The idea of this phase was to select the best k features based on the scoring of each variable that contributes to our response variable. Similar procedure was performed, and two models were then compared on which able to better predicting student's performance.

Then, Graph Convolutional Model (GCN) built was further verified our variables selection. In GCN, we sought to classify student's behavior in the MOOC courses. We treated the issue of classifying student's behavior as a semi-supervised node classification problem on a graph where labels were only available for a limited sample of graphs to unknown class labels. This problem can be formalized as follows. Given undirected graph, $G = (V, E)$ with V is the node representing each highly correlated variable and E is edges of the correlation coefficients. Let A be the adjacency matrix representing the graph; I be the feature matrix, which is a one hot categorical matrix representation; D as the degree matrix showing the number of edges; and W as the random initialize weight matrix for the lth neural network layer.

The forward propagation rule equation is shown below:

$$f(H^i, \hat{A}) = \sigma(D^{-1}\hat{A}H^{i-1}W) \quad (1)$$

where σ is the ReLu activation function, D^{-1} means the inverse of degree matrix, $\hat{A}$ representing transformed adjacency matrix $(A + I)$, W is the weight matrix and i being the number of layers.

Using the propagation rule f, these features were accumulated at each layer to generate the features of the following layer. As a result, as each layer progresses, the features grew increasingly complex. Therefore, stacking up the two GCN layers enabled us to obtain the feature representations for each node.

Lastly, cluster analysis such as k-means and DBSCAN were performed with the most suitable variables which verified in GCN. Both algorithms were used to group similar data into a single cluster and thus explored for hidden patterns and relationships through a large data collection. A random sample of 5,000 students were chosen to perform the k-means and DBSCAN algorithms.

Principal Component Analysis (PCA) was first adopted to reduce the dimensionality and increase the interpretability of the datasets, at the same time reducing the information loss. Eigenvalues and eigenvectors were used to obtain the principal components specifying a linear transformation from the original attribute space to a new space with uncorrelated attributes. Hence, we obtained a new dataset which was defined by the principal components. Then, a cumulative variance explained ratio plot was used to obtain the optimal number of principal components with a 99% cut-off threshold.

In k-means, optimal number of clusters was determined by using the "elbow" method and "silhouette" analysis and cluster plot was plotted to visualize the clusters formed using the PCA dataset. Similarly, in DBSCAN, Eps (the minimum radius) were determined by the k-distance plot and natural cluster formed were plotted using the PCA dataset. Finally, the silhouette coefficient was calculated to compare both algorithms' performance.

3. Results and Discussion

This section discusses the various information visualization to give a clear overview on the data distribution in the dataset.

3.1. Exploratory Data Analysis

In this study, edX MOOC obtained from Kaggle dataset was used. This dataset was originally from Harvard University (HarvardX) and Massachusetts Institute of Technology (MITx) for the first academic year stated from fall 2012 to summer 2013. However, in this dataset the timeline started from spring of 2012 as edX opens early registration for the online courses offered. There was a total of 16 courses offered by MITx and HarvardX on edX platform. The objective of this study was to predict the student's performance. Thus, the variable "`certified`" was chosen as the response variable. Fig. 1 shows the distribution of certification of students.

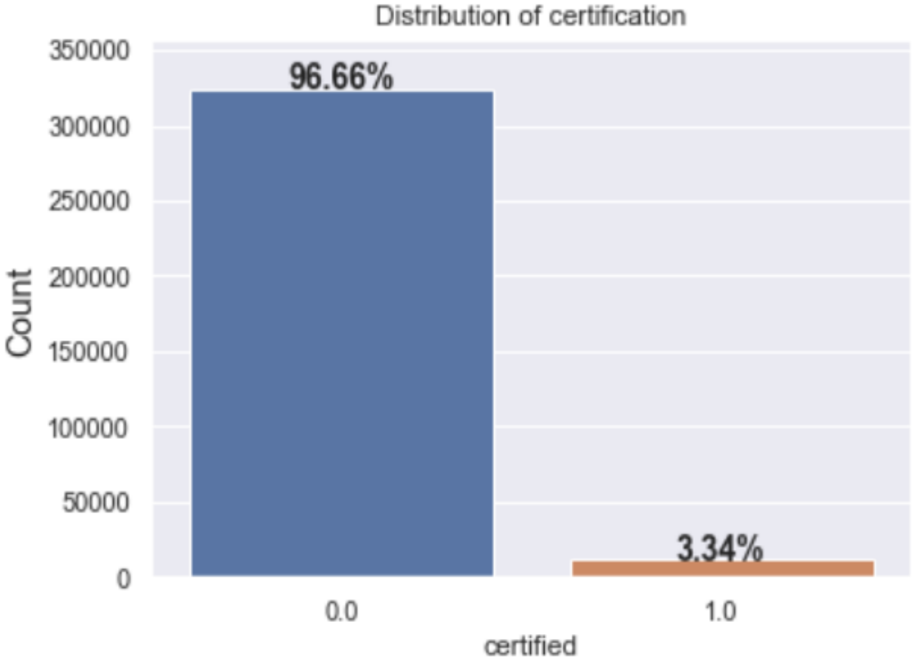


Fig. 1 Distribution of student's certification.

In Fig. 2, the heatmap shows the correlation between each variable. In the certified column, we can see that the variable "`certified`" has positive correlation with variables "`ndays_act`," "`nplay_video`," "`explored`," and "`grade`." In order, to get a clearer view on their correlation, `.corr()` function was performed to get the actual correlation coefficient of each variable with the "`certified`" variable. It helps to have better grasp the relationships between variables in data analysis and modelling.

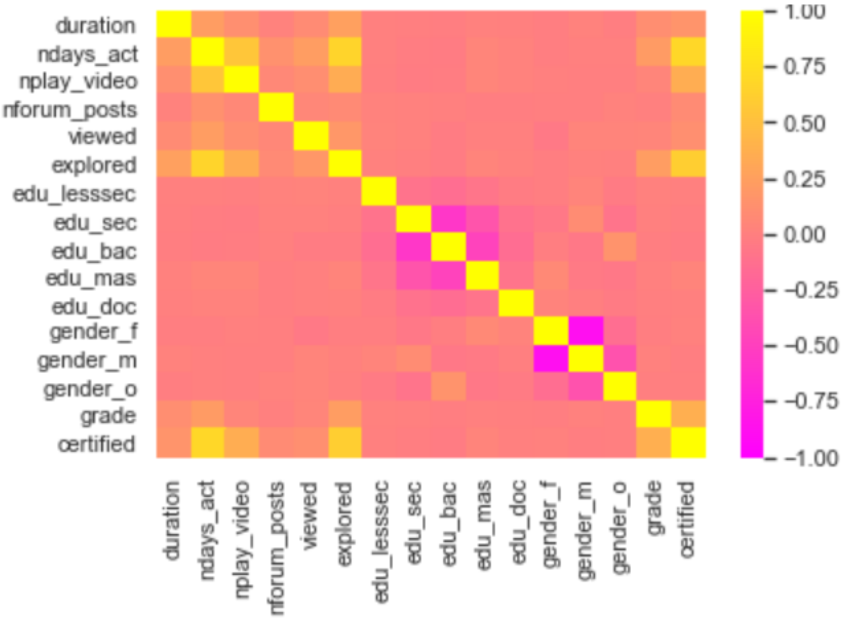


Fig. 2 Heatmap of different variables.

In Fig. 3, the clustermap of randomly chosen 5,000 students to obtain insights based on the similarity of the students and the variables. From the column dendrogram which highlighted with the yellow rectangle, these students were classified as male and bachelor in their education status. In the duration spent between this group of students, the lighter color bars show that they spent more time in the course, and also light bar on the number of unique days active showing that this group of students did active more frequently and did explored at least half of the content. However, these students did not pass the assessment but did get certified.

Furthermore, for the light blue rectangle in the clustermap, this group of students was bachelor in their education status but females. There are light color bars in the duration row showing that these students spent more time in the course and accessed more than half of the content. Besides, they did active frequently although they did not show good performance in their assessment, but they were certified.

In addition, the green highlighted rectangle shows that this groups of students were male with master educational status. They did not spend long duration in the course and active less frequently but did explored at least half of the content.

For the pink rectangle which highlighted in the clustermap, this group of students was also male but classified as secondary in their education status. Although the row "duration" also shows lighter bar meaning that students also spent longer time in the course, they did not active frequently and did not access more than half of the course content. Thus, no evidence on earning a certificate for this group of students. Thus, they still got certified. With this analysis, the variable "explored," "duration," and "ndays_act" can be concluded that they show significant information on predicting student's performance.

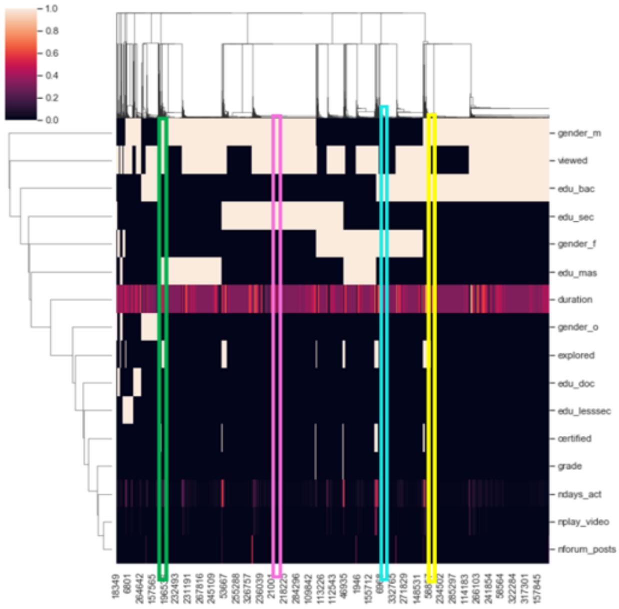
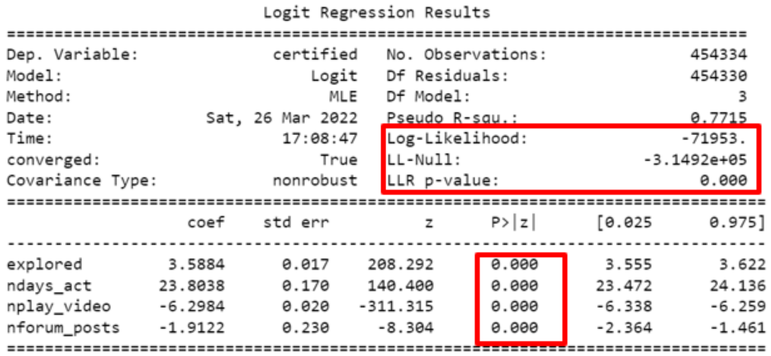


Fig. 3 Clustermap of 5,000 randomly chosen students.

3.2. Logistic Regression

The goal of logistic regression was to determine the model that best described the relation between the dependent and independent variables. It is a well-suited model when we need to predict binary outcomes. Since our dependent variable "`certified`" resulted in binary outcomes (i.e. 0: certified, 1: uncertified), logistic regression model was the first proposed model to predict the student's performance.

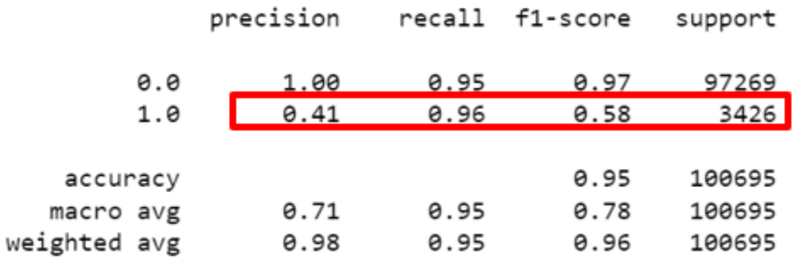
Logistic regression full model with all predictive variables were built. After fitting the full model, model summary table was generated to interpret the model coefficients. Fig. 4 shows the regression results of final model. According to the p-value of each coefficient, variables "`explored`," "`ndays_act`," "`nplay_video`," and "`nforum_posts`" coefficients show significant effects on the student's performance. The likelihood test (LR test) is the goodness of fit test which compare the null model and final model. The log-likelihood is maximum value of the log-likelihood function and the LL-Null is the maximal log-likelihood function when only an intercept is present. The LR test revealed that model improved when fitted with the variables "`explored`," "`ndays_act`," "`nplay_video`," "`nforum_posts`" as log-likelihood is higher than the LL-null. The LLR p-value shows value of 0 indicated that this model was able to create meaningful representation and we have 100% confident that these results are valid.



```
                           Logit Regression Results
==============================================================================
Dep. Variable:              certified   No. Observations:               454334
Model:                          Logit   Df Residuals:                   454330
Method:                           MLE   Df Model:                            3
Date:                Sat, 26 Mar 2022   Pseudo R-squ.:                  0.7715
Time:                        17:08:47   Log-Likelihood:                -71953.
converged:                       True   LL-Null:                   -3.1492e+05
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
explored         3.5884      0.017    208.292      0.000       3.555       3.622
ndays_act       23.8038      0.170    140.400      0.000      23.472      24.136
nplay_video     -6.2984      0.020   -311.315      0.000      -6.338      -6.259
nforum_posts    -1.9122      0.230     -8.304      0.000      -2.364      -1.461
================================================================================
```

Fig. 4 Regression results of final model.

Fig. 5 shows the classification report on final model. Accuracy and confusion matrix were calculated for this model to evaluate the model performance and accuracy. The accuracy of the final model showed 95% and confusion matrix of $\begin{bmatrix} 92600 & 4669 \\ 150 & 3276 \end{bmatrix}$. Nevertheless, this was against our objectives of building this model because we were more interested in predicting the certified students and having high accuracy in predicting certified students. Thus, F1-score was used as another metric to determine whether the model met our objective. The certified students group showed a F1-score of 0.58 which was then used to further compare with another logistic model.

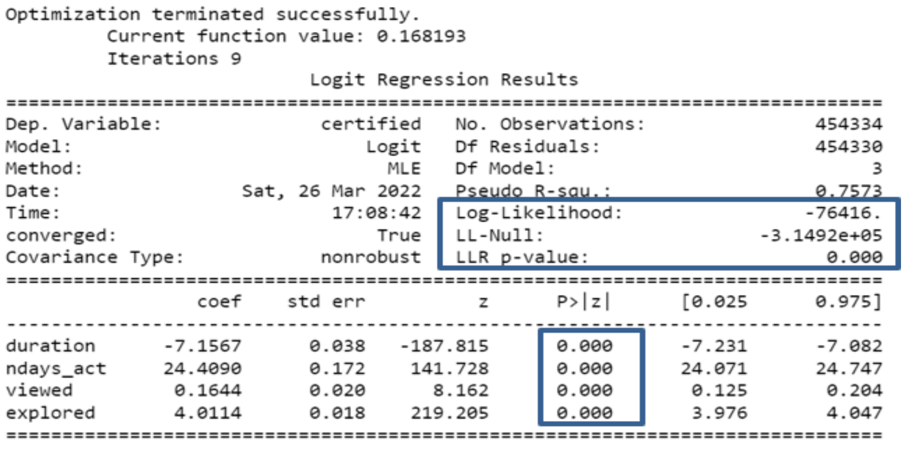


```
              precision    recall  f1-score   support

         0.0       1.00      0.95      0.97     97269
         1.0       0.41      0.96      0.58      3426

    accuracy                           0.95    100695
   macro avg       0.71      0.95      0.78    100695
weighted avg       0.98      0.95      0.96    100695
```

Fig. 5 Classification report of final model.

On the other hand, we built another logistic model by using the scikit-learn function `SelectKBest()`. Since our final model showed that four variables were able to perform prediction, we then cchose the first four highest score variable with the `f_classif` function. This function used the ANOVA F-test assumptions in choosing the best four variables. With this, we obtained the top four highest score variables, which were "`duration`," "`ndays_act`," "`viewed`," and "`explored`."

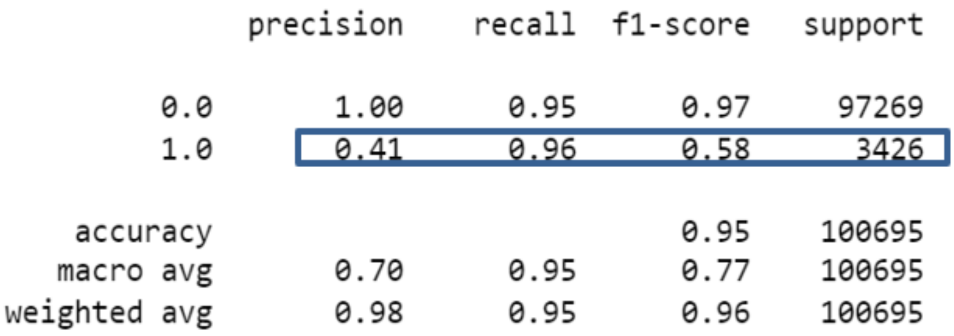
Fig. 6 shows the KBest model regression results. The LR test revealed that model improved when fitted with the variables "`explored`," "`ndays_act`," "`viewed`," and "`duration`" as log-likelihood was higher than the LL-null. The LLR p-value showed value of 0, indicating that this model was able to create meaningful representation. According to the p-value of each coefficient, all coefficients also showed significant effects on the student performance.



```
Optimization terminated successfully.
         Current function value: 0.168193
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:              certified   No. Observations:               454334
Model:                          Logit   Df Residuals:                   454330
Method:                           MLE   Df Model:                            3
Date:            Sat, 26 Mar 2022   Pseudo R-squ.:                  0.7573
Time:                        17:08:42   Log-Likelihood:                -76416.
converged:                       True   LL-Null:                   -3.1492e+05
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
duration      -7.1567      0.038   -187.815      0.000      -7.231      -7.082
ndays_act     24.4090      0.172    141.728      0.000      24.071      24.747
viewed         0.1644      0.020      8.162      0.000       0.125       0.204
explored       4.0114      0.018    219.205      0.000       3.976       4.047
==============================================================================
```

Fig. 6 KBest model regression result.

Fig. 7 showed the classification report on KBest model. The accuracy of the KBest model showed 95% and confusion matrix of $\begin{bmatrix} 92576 & 4693 \\ 149 & 3277 \end{bmatrix}$. The 95% of accuracy was interpreted using this model, students who were uncertified could be predicted with 95% accuracy. In this KBest model, the certified students group also showed a F1-score of 0.58, meaning that both logistic models performed similarly. We further confirmed the results with the ROC curve.



```
              precision    recall  f1-score   support

         0.0       1.00      0.95      0.97     97269
         1.0       0.41      0.96      0.58      3426

    accuracy                           0.95    100695
   macro avg       0.70      0.95      0.77    100695
weighted avg       0.98      0.95      0.96    100695
```

Fig. 7 Classification report on KBest model.

Fig. 8 shows the ROC curve for the comparison of KBest model and final model. Based on this figure, we further confirmed that both models performed similarly as the ROC curve overlapped with each other. The AUC for both models were 0.95. With this, we can conclude both models performed similarly although they were different with two predictors. These two models shared the variables “`ndays_act`” and “`explored`”. The similar performance of both models was due to the coefficient of variable “`ndays_act`” as it acted as the largest coefficient in both models.

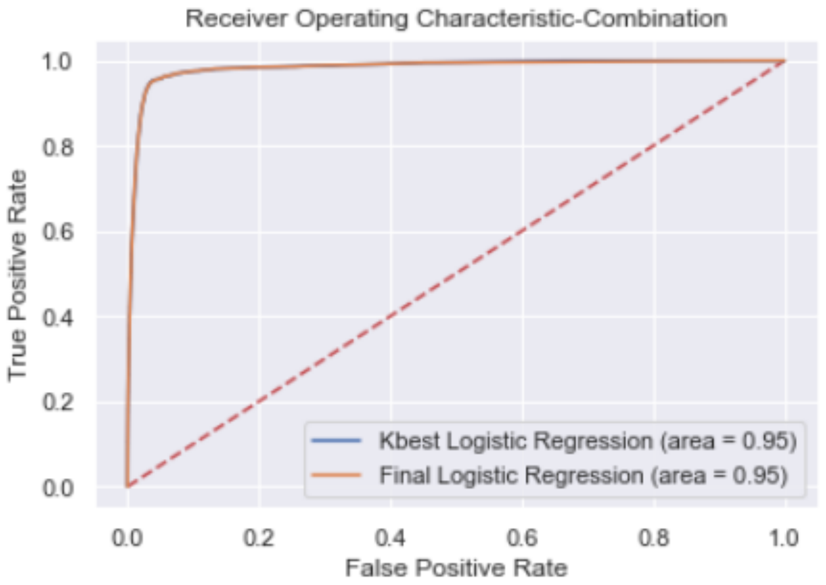


Fig. 8 Comparison of ROC curve.

3.3. Graph Convolutional Network

Moving on to the semi-supervised leaning model which is the GCN. As we compared both final logistic regression model and KBest logistic regression model, both models performed similarly. Thus, GCN model was then used to verified both model’s result. Figure 9 shows the feature representation of the random highly correlated variables.

The findings from Fig. 9 provide valuable insights into the underlying structure of the data and the predictive power of different variables in determining student performance. The clear distinction between the two groups, as identified by the plot, highlights the potential significance of certain variables in predicting student outcomes.

The right group, consisting of nodes 0, 1, 5, and 7, exhibits distinct characteristics represented by the variables “`ndays_act`,” “`nplay_video`,” “`explored`,” and “`nforum_posts`.” These variables likely played a crucial role in determining student performance within this group. The presence of these variables as key predictors suggested that student engagement with the course material, as indicated by activities such as active days, video interactions, course exploration, and forum participation, strongly influences their performance outcomes.

Conversely, the left group, comprising nodes 2, 3, 4, 6, 8, 9, and 10, might exhibit different patterns and behaviors that contributed to student performance. Overall, the identification of key predictors such as "`ndays_act`," "`nplay_video`," "`explored`," and "`nforum_posts`" underscored the importance of student engagement and participation in MOOCs as significant determinants of performance outcomes.

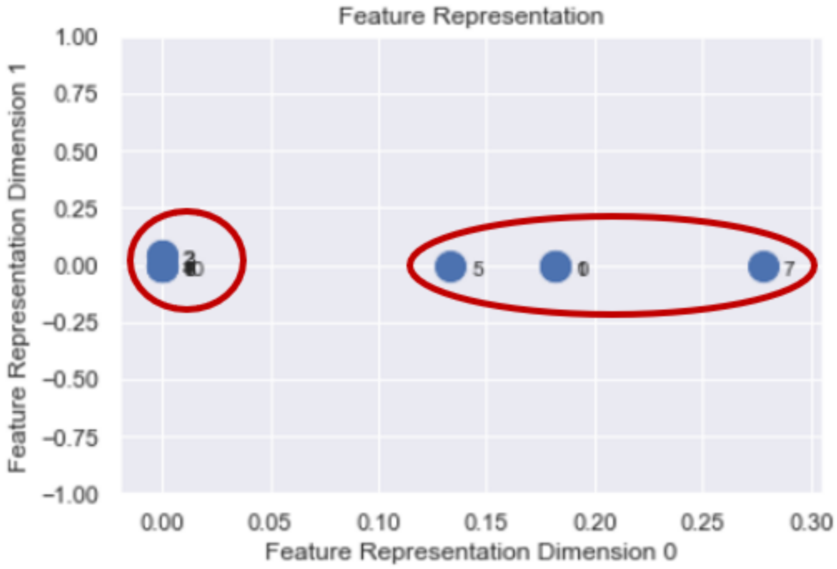


Fig. 9 Feature representation plot.

3.4. Clustering Analysis

In this study, we focused on two unsupervised models which were k-means algorithm and DBSCAN algorithm. A random sample of 5,000 students were used to construct both algorithms.

PCA was performed to reduce our data dimension. PCA was able to help us in compressing the high dimensional data into a 2-dimensional form, while at the same time maintaining high information gain. Based on the cumulative explained variance plot shown in Fig. 10, with 99% set as a cut-off threshold two principal components were selected to visualize the dataset.

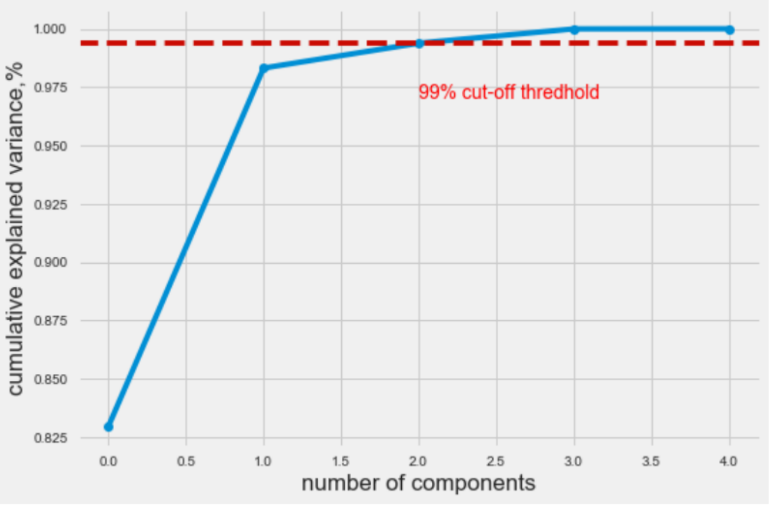


Fig. 10 Cumulative explained variance plot on PCA model.

With this, we performed the k-means algorithm. From the cluster plot in Fig. 11, two clusters were formed, with the first cluster being orange and second cluster being blue. The two clusters formed showed that the k-means algorithm met our objectives of clustering certified and noncertified students. According to the statistical analysis, first cluster represented the certified students, and the second cluster represented the noncertified students. For the first cluster students, they appeared to be more active in the platform and did explored at least half of the course content.

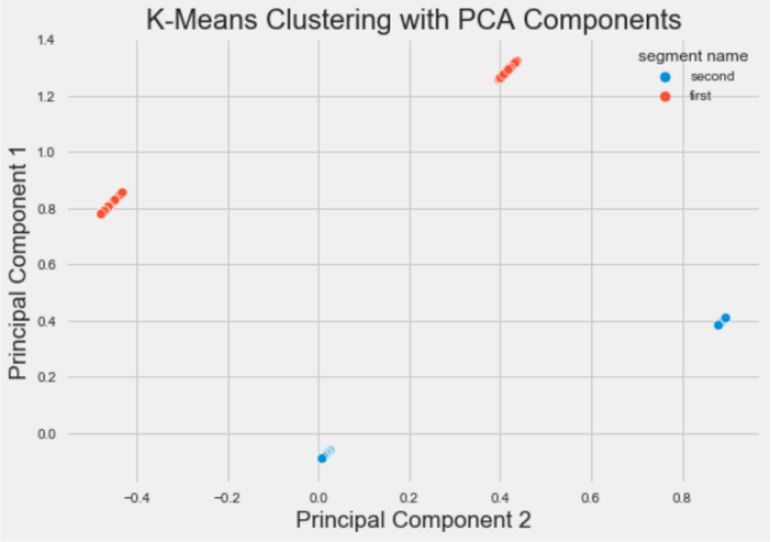


Fig. 11 Cluster plot on k-means-PCA model.

Another unsupervised machine learning model is DBSCAN algorithm. In this algorithm, natural clustering was performed without the needs of specifying the number of clusters. Single-level density method was used to determine ideal value of epsilon value. Three natural clusters were formed with label 0, 1, and -1.

From the cluster plot in Fig. 12, almost half of the data points were considered as noise. Thus, they only showed noise in the cluster plot. Based on the statistical analysis on DBSCAN-PCA model, noise cluster appeared to be the certified student's group as it showed that students in this cluster were active on more days and did explore more of the course contents. It can be further explained due to the imbalance dataset in our study. Since DBSCAN algorithm forms natural clusters, thus it will place the minority samples as the noise cluster, showing that they do not belongs to any of the clusters.

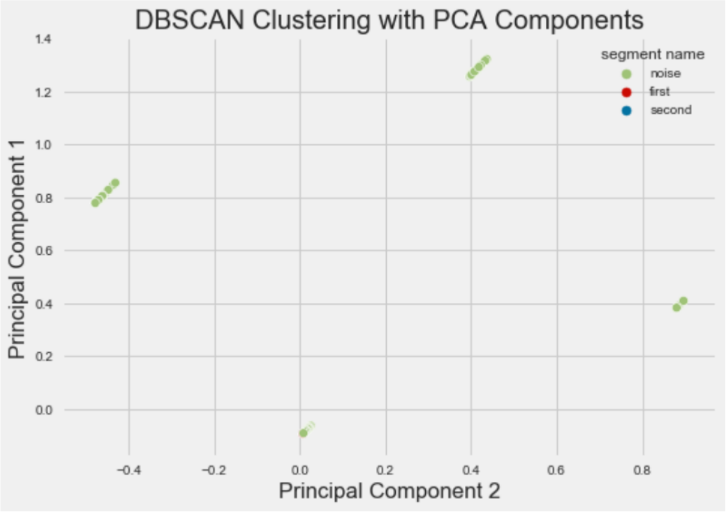


Fig. 12 Cluster plot for DBSCAN-PCA model.

4. Conclusion

Various ways of mining student records for useful data attributes to predict academic achievement were discussed in this research. To establish the essential variables that influenced student's performance, we analyzed the certification rate, as well as social and physiological aspects. We conclude from this research that the key factors that directly influenced students receive certification are closely related to their involvement in MOOC courses. Exploring more than half of the courseware contents, watching more of the lecture videos, and actively participating in more unique days in the MOOC courses have been the most important factors affecting the student's certification. Social aspects, such as communicating with other students in forum discussions, on the other hand, been a minor but significant influence impacting certification.

This type of study can assist both teachers and students improve their MOOC performance. Teachers can gain a better understanding of their student's learning patterns by extracting hidden information about them. Furthermore, the established predictive model aids in the early prediction of students' performance, allowing teachers to distinguish between weak and strong pupils. As a result, teachers may take appropriate action at the appropriate moment, with effective planning and decision-making, to increase educational quality. This type of study builds a foundation for an early warning system, which can help to enhance the teaching-learning process overall. Similar datasets can be used to calculate student dropout rates in the future. With a larger data collection and more attributes, better insights may be gained.

References

[1] P.G. de Barba, G.E. Kennedy, and M.D. Ainley, "The Role of Students' Motivation and Participation in Predicting Performance in a MOOC," *Journal of Computer Assisted Learning*, Vol. 32, No. 3, pp. 218–231, Jun. 2016, doi: 10.1111/jcal.12130.

[2] G. Hughes and C. Dobbins, "The Utilization of Data Analysis Techniques in Predicting Student Performance in Massive Open Online Courses (MOOCs)," *Research and Practice in Technology Enhanced Learning*, Vol. 10, pp. 1–18, Jul. 2015, Art. no. 10, doi: 10.1186/s41039-015-0007-z.

[3] R. Al-Shabandar, A.J. Hussain, P. Liatsis, and R. Keight, "Analyzing Learners Behavior in MOOCs: An Examination of Performance and Motivation Using a Data-Driven Approach," *IEEE Access*, Vol. 6, pp. 73669–73685, 2018, doi: 10.1109/ACCESS.2018.2876755.

[4] B. Xu and D. Yang, "Motivation Classification and Grade Prediction for MOOCs Learners," *Computational Intelligence and Neuroscience*, Vol. 2016, pp. 1–7, 2016, Art. no. 2174613, doi: 10.1155/2016/2174613.

[5] H. Nen-Fu, H. I-Hsien, L. Chia-An, C. Hsiang-Chun, T. Jian-Wei, and F. Tung-Te, "The Clustering Analysis System Based on Students' Motivation and Learning Behavior," in *2018 Learning With MOOCS (LWMOOCS)*, 2018, pp. 117–119, doi: 10.1109/LWMOOCS.2018.8534611.

[6] R. Conijn, A. Van den Beemt, and P. Cuijpers, "Predicting Student Performance in a Blended MOOC," *Journal of Computer Assisted Learning*, Vol. 34, No. 5, pp. 615–628, Oct. 2018, doi: 10.1111/jcal.12270.

[7] J.L. Santos, J. Klerkx, E. Duval, D. Gago, and L. Rodríguez, "Success, Activity and Drop-Outs in MOOCs an Exploratory Study on the UNED COMA Courses," in *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge*, 2014, pp. 98–102, doi: 10.1145/2567574.2567627.

[8] P. Duffy, "Engaging the YouTube Google-Eyed Generation: Strategies for Using Web 2.0 in Teaching and Learning," *The Electronic Journal of e-Learning*, Vol. 6, No. 2, pp. 119–130. [Online]. Available: https://academic-publishing.org/index.php/ejel/article/view/1535/1498

[9] L. Breslow, D.E. Pritchard, and J. Deboer, "Studying Learning in the Worldwide Classroom: Research into edX's First MOOC," *Research & Practice in Assessment*, Vol. 8, pp. 1–25, Jun. 2013. [Online]. Available: https://www.rpajournal.com/dev/wp-content/uploads/2013/05/SF2.pdf