EAS508 - Statistical Learning and Data Mining-I

Project Report

Knowledge Enhancement & Career Advancement With MOOC

Group Members :

Manali Ramchandani

Manoj Paladi

Naresh Babu Kolli

Nitesh Padidam

Thrivikrama Rao Kavuri

May 2023

1.Abstract

Massive Open Online Courses (MOOCs) have transformed the education industry by providing accessible online learning opportunities. This project utilizes a smart framework to analyze MOOC datasets, exploring the relationship between student activity and course completion rates. Challenges encountered includes preprocessing the data by handling missing values, encoding categorical variables, normalizing numerical features, and optimizing model performance. By quantifying learner participation through video consumption patterns and employing data mining tools, significant correlations are uncovered. These findings offer valuable insights for the MOOC industry, informing course design and learner engagement strategies. Ultimately, this project enhances educational experiences and outcomes for learners worldwide.

2.Introduction

In recent years, MOOCs have gained significant attention from learners, educators, and institutions due to their potential to reach a large audience and provide flexible learning experiences. The objective of this report is to review the existing literature, identify trends, and assess the benefits and limitations of MOOCs in achieving educational goals.

This study aims to explore the dataset, analyze its various features, and ultimately build a model that can assist various research areas explored in the field of MOOCs, such as learner engagement, course completion rates, pedagogical strategies, and the impact on learners' career advancement. By synthesizing the existing knowledge, we strive to provide a comprehensive understanding of the current state of MOOCs and the challenges and opportunities they present. This report will outline the steps taken to preprocess and analyze the "MOOC" dataset, including data cleaning, outlier detection and removal, encoding categorical variables, normalizing numerical features, and optimizing model performance by evaluating range of regression models and assessing using mean absolute error and R-squared metrics. By examining previous research and developments in MOOCs, this report contributes to the broader understanding of this educational approach and its potential for shaping the future of online learning

3.Data Description

S.No	Description	Value
1	Dataset Name	Massive Open Online Courses (MOOCs)
2	Number of Observations	4,16,921
3	Categorical Variables	5
4	Numerical Variables	6
5	Informative Variables	9

3.1. Feature Description

Type	Features	Description
Categorical	Semester	Semester of course
	Viewed	Anyone who accessed the ‘Courseware’ tab
	Explored	Anyone who accessed at least half of the chapters in the courseware
	Certified	Anyone who earned a certificate
	LOE_DI	Level of education completed
Numerical	Grades	Final grade in the course, ranges from 0 to 1
	nevents	Number of interactions with the course
	ndays_act	Number of unique days student interacted with course
	nplay_video	Number of play video events of the course that student interacted
	n_chapters	Number of chapters with which the student interacted

Informative	nforum_posts	Number of posts to the Discussion Forum
	Institute	Course Provider institute
	Course_id	ID of the Course
	Userid_DI	ID of each User
	Final_cc_country	User IP address/Address
	Gender	Gender of the User
	Start_time_DI	Date of Course registration
	Last_event_DI	Date of last interaction with course
	Age	Age of the User
	Year	Year of the Course

4. Data Structuring and Cleansing

4.1. Assumptions

- At least one of the students accessed all chapters in coursework, implying that each degree requires a student to complete all courses.
- More participation, more learning, higher grade: a student who participates more in the course can learn more and receive higher grades in the course.
- Gender is not being used as a feature for analysis: Because many people do not wish to mention their gender, the gender column includes null values, and hence this cannot be called a feature analysis.
- Categorical Variables Factoring: Categorical variables are variables that are used to build a set of variables that may be classified into similar groupings. Some of the category variables in this are as follows: LoE_Di, final_cc_cname_di, course_id, institute, semester.
- The levels are used to factor course_id, institution, semester, LoE_Di, and final_cc_cname_di into numerical values: these variables are utilized for factoring numerical values by utilizing levels, which means that each degree is measured using the levels.
- LoE_Di Example with order: Less than Secondary ,Secondary ,Bachelor's, Master's, Doctorate
- These are the levels that are used for factoring this LoE_Di, and the numbers in this column are less than secondary, secondary, Bachelor's, Master's, and Doctorate.

4.2. Outlier Identification & Treatment

- Inserted a column Incomplete flag which takes value 1 for the records that have null values for nevents but have non-null values for ndays_act, nforum_posts, or nchapters. These values are considered as outliers
- As per the data declaration, the users with explored field value as 1 cannot have viewed as 0. Hence, set of data with viewed=1 and explored=0 are outliers.
- Set of data with explored as 1 and nchapters less than half of the maximum and explored as 0 and nchapters greater than half of maximum.
- Set of data with last_event_di less than start_event _di are outliers.
- Set of data with certified 1 and grade not meeting the threshold.
- Age < 9 are considered as outliers based on level of degree values like Less than Secondary, Secondary , Bachelor’s, Master’s, Doctorate.

- ndays_act greater than day difference between last_event_di and start_event _di are considered as outliers and difference is added as new column n_days.

5. Data Validation

Post outlier treatment, Format validations are done for all the informative variables and range validations for factored categorical and numerical variables to ensure data quality. Correlation among variables is a beginning step to understand the data and that calculation for this data is not to be done on numerical analysis but rather on rank analysis as there are both categorical and numerical data. Some of the techniques that are admissible are Spearman's rank correlation and Kendall's tau correlation.

5.1. Spearman's rank correlation

It is a non-parametric measure of the strength and direction of monotonic association between two variables. It is based on the ranks of the data rather than the actual data values and is calculated as

$\rho = 1 - (6\Sigma d^2 / (n(n^2-1)))$ where

ρ - Spearman's rank correlation coefficient

Σd^2 - sum of squared differences between the ranks of corresponding data points

n - number of data points

5.2. Kendall's tau correlation

It is a non-parametric measure of the strength and direction of association between two ranked variables. It assesses the similarity of rankings between two variables without assuming any specific distributional properties of the data.

$\tau = (P - Q) / \text{sqrt}((P + Q + T) * (P + Q + U))$ where

τ represents Kendall's tau correlation coefficient

P represents the number of concordant pairs

Q represents the number of discordant pairs

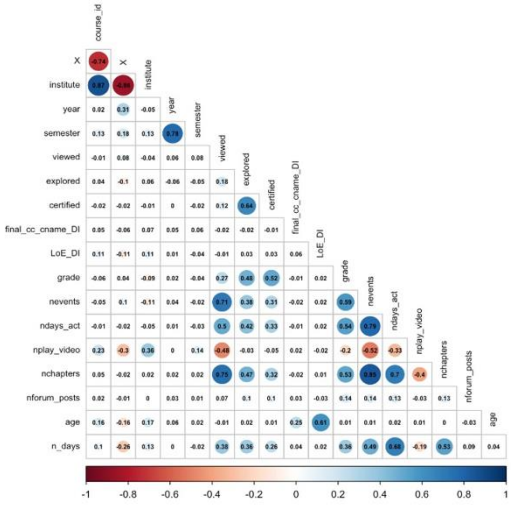
T represents the number of ties in variable X

U represents the number of ties in variable Y

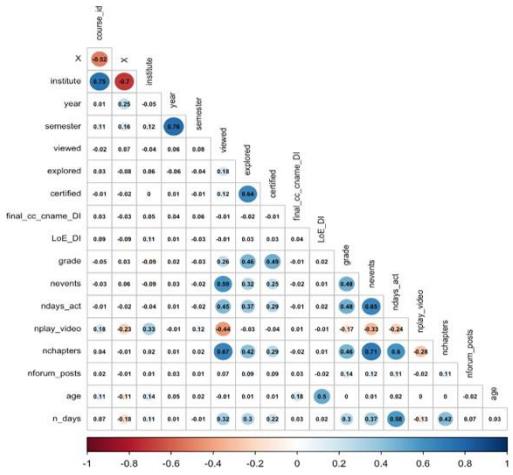
5.3. Computational Challenges

As data is huge, this calculation is highly computing. Therefore, random samples of 10 percent of data are considered for about 10 times and then average values are considered as correlation.

As shown below, both correlation values are almost near. As spearman considered total data, those correlation values are considered down the road for feature selection and analysis.



Spearman Correlation

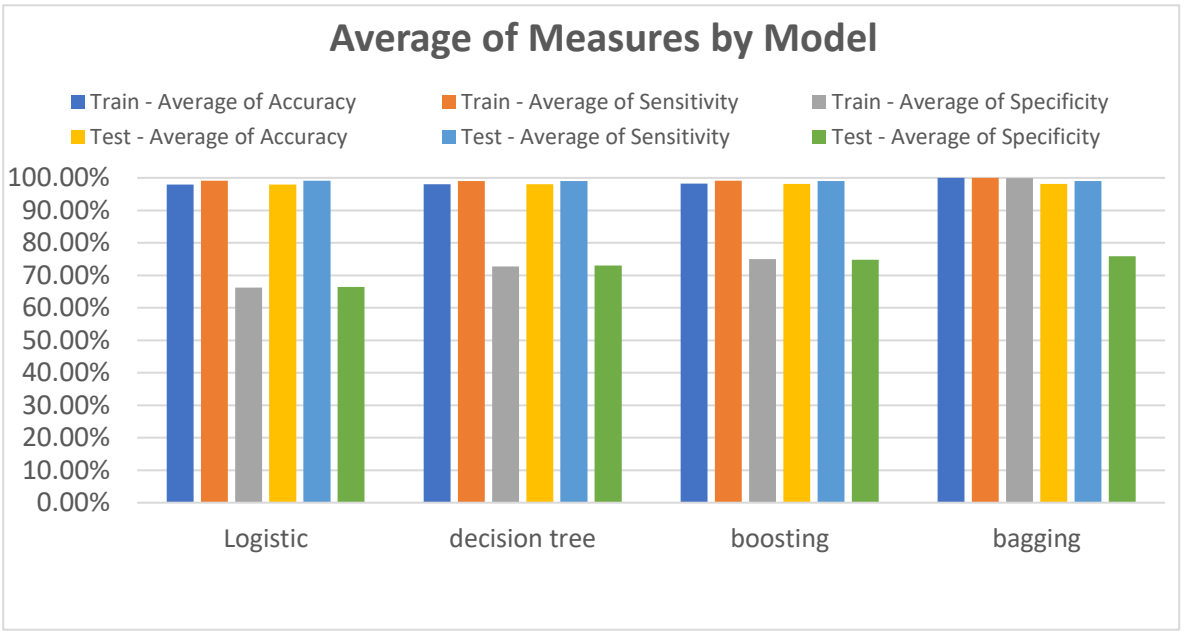


Kendall Tau Correlation

6. Analysis and Predictions

6.1. S: Analysis on how activity of student effects the completion of course?

Here activity is not a single variable but instead a function group of variables, these variables will be identified through correlations. As the question asks if he will pass or fail, classification models will be utilized. We will try to analyse what criteria will determine in qualification of course. Following features are identified from the heat map: **explored, ndays_act, nchapters, n_days, viewed.** Decision tree, Logistic Regression, Bagging tree, Random Forest Boosting models were used to analyze the data.



Observations:
Almost all the models have similar accuracy in both test and train. However, specificity values are low compared to others in both train and test.

Results:
From the above, it was concluded that the Bagging Tree was the best model for predicting and Logistic Regression was an appropriate model for interpretability.

Implications and Recommendations:
Based on our analysis, a recommendation of interactive model is made where it tells the user his status (can pass) and suggests what activities he should focus on to succeed

6.2. M: Analysis on how number of videos watched describes the participation of the learner

Participation of learner in a course is housed by metric nevents in the data which is measure of interactions with course. From the correlation matrices above, it can be noticed that negative correlation is observed between nplay_video and nevents. Increase in nplay_video might have an impact on features nchapters, ndays_act etc. that in turn amplifies participation. Hence, Impact of nplay_video feature can be better explained considering all these features.
From correlation heat map, these features are considered for analysis : ndays_act, nforum_posts, nchapters, nplay_video. Multiple Regression models are build using the above features. Trained and performed cross validation and randomness with seeds-(508,240,620

Results :

Model	RMSE Train	RMSE Test	R Squared Train	R Squared Test	Features
LR	0.958	0.955	0.084	0.083	nplay_video
MLR	0.534	0.536	0.715	0.711	ndays_act, nforum_posts, nchapters, nplay_video
Ridge	0.533	0.539	0.708	0.716	ndays_act, nplay_video, nchapters, nforum_posts
Lasso	0.533	0.539	0.714	0.723	ndays_act, nplay_video, Nchapters

	Coefficients	Units
(Intercept)	-0.003	
ndays_act	0.828	days
nplay_video	-0.05	centiseconds
nchapters	0.005	increments of 1
nforum_posts	0.003	increments of 1

Observations:

Based on accuracies of models, it can be noticed that there is not much increase in accuracy with increase in complexity of model. Considering the problem, interpretation of model is more important than predicting using the model. So , increasing the complexity of model is not likely. Therefore, Multiple Linear Regression is considered as best fit for consideration and also RMSE values for train and test data are closer explaining that overfitting is avoided
From MLR, the coefficients of model can be seen.

Implications and Recommendations:

Based on impact of the metrics on participation, course design can be improvised such that both the parties can be benefitted. Adding the number of videos in the same chapter frequently allows student to interact with website. Quality and length of video can be crisp in this way which is an encouragement to student.

6.3. A: Analysis of Learner's Chapter Completion in the Course

The objective of this analysis is to predict the number of chapters completed by learners in a course based on specific features. By understanding the factors influencing chapter completion, we aim to improve course design and enhance learner engagement and progress.

To achieve our objective, we conducted an analysis using regression models to predict the number of chapters completed. The features considered for this prediction were 'viewed', 'nevents', 'nplay_video', and 'ndays_act'. We evaluated the performance of four regression models: Linear Regression, Multi Linear Regression, Ridge Regression, and LASSO Regression.

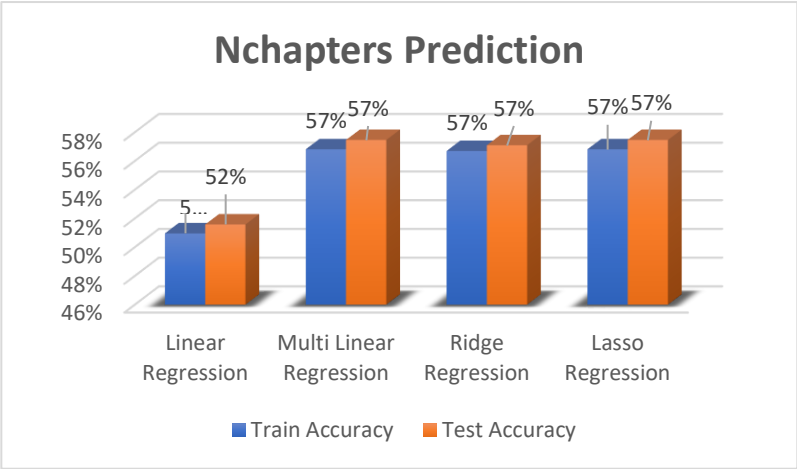
Observations:

Model Selection: Among the four regression models tested, the Multi Linear Regression model emerged as the best model for predicting and interpreting the number of chapters completed by learners.

Accuracy and RMSE Values: The Multi Linear Regression model demonstrated good accuracy for both the training and test data sets. The root mean squared error (RMSE) values for the model were relatively low and consistent, indicating a reliable and robust performance.

Consistency Across Seeds: The analysis revealed that the model's performance remained largely consistent even when the data's randomness, represented by different seeds or randomizations, was changed. This highlights the model's robustness and consistent predictive capability.

Results:



Seeds	Train RMSE	Test RMSE	Train Accuracy	Test Accuracy
508	2.85	2.81	57%	57%
200	2.84	2.86	57%	57%
600	2.84	2.86	57%	57%
80	2.85	2.82	57%	57%
800	2.83	2.88	57%	56%

Implications and Recommendations:

The website can enhance the course design, personalize the learning experience, and provide appropriate interventions to improve learner engagement, progression, and chapter completion.

In conclusion, analysis of learner chapter completion in the course highlights the effectiveness of the Multi Linear Regression model in predicting and interpreting the number of chapters completed based on the selected features. These insights will contribute to a more successful and fulfilling learning journey for the users.

6.4. R: Analysis of User Posting Frequency in the Course

The objective of this analysis is to predict how frequently users have posted in the course based on their activity and engagement. By understanding the factors that influence user posting frequency, we aim to improve user engagement and participation in the course.

To achieve our objective, we conducted an analysis by considering various features and employing regression models. After analyzing correlations, we identified several potential features that could help define a regression model. These features included grade, nevents, ndays_act, and nchapters. We tested Multiple Linear Regression (MLR) and Ridge Regression models using these features to predict user posting frequency.

Observations:

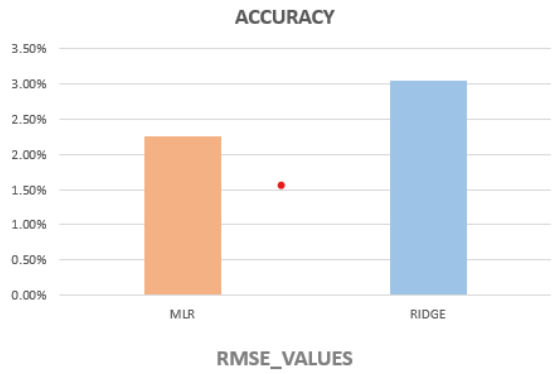
Based on our analysis, the following below observations were made:

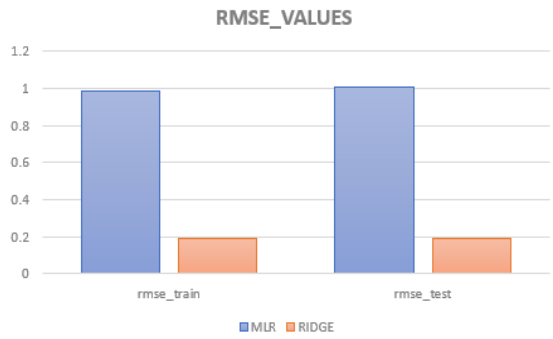
RMSE Values: The root mean square error (RMSE) values for both the MLR and Ridge regression models were relatively low. This indicates that the models were able to make reasonably accurate predictions of user posting frequency based on the selected features.

Low Accuracy: Unfortunately, none of the regression models given accuracy greater than 4%. This suggests that the selected features alone are not sufficient for accurately predicting user posting frequency in the course.

Low Posting Frequency: It was observed that only 1% of users have posted in the given dataset. This low posting frequency could be a potential reason for the low accuracy of the regression models.

Results:





Implications and Recommendations:

Analysis of user posting frequency revealed the limitations of the selected features in accurately predicting user behavior. However, by incorporating additional relevant features and implementing strategies to motivate user contribution, the website can enhance user engagement and increase the likelihood of receiving answers for posted doubts. This, in turn, will contribute to a more interactive and enriching learning experience for all users.

6.5. Analysis of User Engagement and Time Spent on MOOC Platform

The objective of this analysis is to understand the relationship between the time spent by users on a MOOC platform and their level of engagement, as measured by the number of days they are active on the platform. By examining this relationship, we aim to provide insights that can help improve user engagement and optimize course design. To achieve our objective, we performed regression models on the data using three different approaches:

Observations:

Based on our analysis, the multiple linear regression model performed the best among the three models. It exhibited high R-squared values, indicating good predictive capability for the target variable (ndays_act). Although there were slight variations in the root mean squared error (RMSE) values between the training and test sets, the multiple linear regression model was overall the most effective in predicting the time spent by users on the platform.

Results:

Implications and Recommendations:

The multiple linear regression model performed the best in forecasting the number of days users are active on the platform, according to our examination of user engagement and time spent on the MOOC platform. The research highlights the value of regular content releases and a good learning curve to increase user engagement. Implementing these suggestions and utilizing tailored recommendations can help the website increase user engagement and give users a more enjoyable and satisfying learning experience.

7.Results and Conclusions:

- The project's goal is to keep the business on course by achieving the SMART objectives.
- Data available from the industry has been good but some recommendations are made for the future so that some other perspectives of improvement can be pursued.
- We have conducted the most thorough analysis possible for each target. All these models and analyses describe the underlying patterns of user behavior and preferences, which are critical for improving the current business model.
- Data cleaning was the most challenging part as they contain mixed data types and class imbalances. All these issues we addressed in the analysis.
- Presence of mixed data types limited the model approaches in the analysis.

8.References:

- Dataset
- "Introductory Statistics with R" by Peter Dalgaard