Introduction

The last year of high school is most students' nightmare. The stress and anxiety of the preparation for the final higher education entrance examination is no joke. For the students in Nigeria, it is the Joint Admissions and Matriculation Board (JAMB) exam. All students in Nigeria that yearn for higher education, this exams is a rite of passage and gateway to their future. It is most often heard when pertaining to eduction that due to the large population of examinees, the world has found the way of testing the students via standardised exams across the globe. Determining students' academic achievement in a way, entails the educational system and standardised exams designs. More, higher education enrolment often based on the assessment of such exams and it is important to ensure that the students' performance in such exams are accurately identified.

Problem Statement

However, a students' performance should not only measurable simply by exams scores. There are various of reasons why students might be underperforming or unable to achieve their maximum potential in such standardised educational systems. This paper aims to unveil that with building a predictive model, we can take a closer look at how these factors come into play at such scenario. Millions and more of students in Nigeria across lays of social backgrounds sitting in one standardised national exam. We shall take a closer look on our dataset and what these factors are but beforehand, we present our central question: can we identify students performance early? The problem set out to be discovered more in-depth: can students performance in JAMB exams be predicted based on social disparities/ resource allocation? This moreover gives us an highlight in demographic factors' contribution to variability in JAMB scores.

1. Objectives

Throughout the paper, it is carefully selected for the suitable several models in order to reach the best prediction for this setting of dataset. Each steps and choices are explained in order to identify the choice of steps to continue. The outcome of choice was selected with particular attention paid to the trade-off between model complexity and interpretability.

2. Data

The choice of this project is “Joint Admissions and Matriculation Board (JAMB) Examination Performance Data set (2024). It is retrieved from kaggle.com. It is important to acknowledge that the dataset based the statistics of 2024 Joint Admissions to create the dataset as a stimulation. We can see in later data processing that the data are quite well-spread and bring about interest in the capture in modelling for this dataset. This dataset caught interest due to its inclusion of socio-economic status, student parental background as well as needed educational variables. This concept of estimating a student performance is essential to be acknowledged as “doing well at school” is not simply about good at studying. Tackling the concept of a good student from their background with this dataset is relatively anew of exploring.

2.1 Data Description and Exploratory Data Analysis

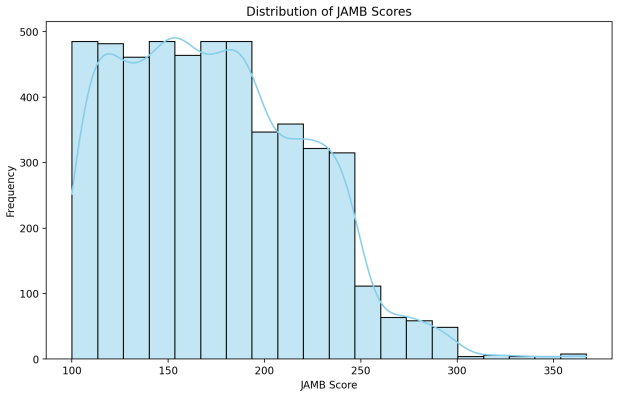
To begin with, we need to understand our dataset in details. The dataset contains 5000 students and 15 columns of informative variables where the target variable of this dataset is JAMB_Score.

Variables	Information
JAMB_Score	Student Achieved Score of JAMB ranging from 100 to 400
Study_Hours_Per_Week	Student study hours per week in hours
Attendance_Rate	Student attendance rate
Teacher_Quality	Teaching quality in rank 1 to 4 numerical scale
Distance_To_School	Distance to school in kilometres
School_Type	Private/Public school attending
School_Location	Urban/Rural of school location
Extra_Tutorials	Yes/No in receiving extra tutorials
Access_To_Learning_Materials	Yes/No to access
Parent_Involement	Low/Medium/High in parent involvement
IT_Knowledge	Low/Medium/High in IT knowledge of student
Student_ID	ID Number
Age	Student age
Socioeconomic_Status	Low/Medium/High in Socioeconomic Status
Parent_Education_Level	None/Primary/Secondary/Tertiary in Education Level of Parents
Assignments_Completed	Assignments completed in 1 to 5 numerical scale
Gender	Male/Female

Following we examine three important graphs that helps us identify key indicators of this dataset with histogram, scatter plots and box plots.

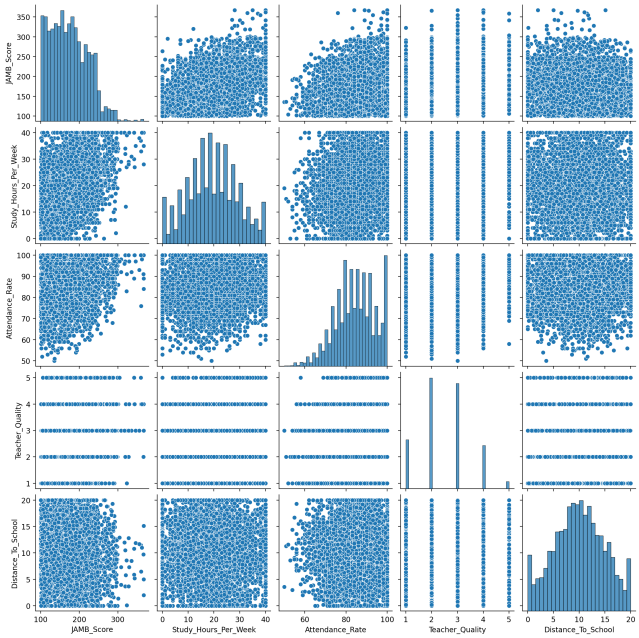
We can begin the dataset with a clear histogram our target variable of JAMB_Scores. Histogram can illustrate a overview of the target variable in distribution of our dataset as a whole. Having an overview of this is essential for us to continue on know what we are predicting. We can see that in this dataset, the students performance are highly skewed towards scores below 200 as the 50% of distribution are below 170. This is quite important to bear in mind as the training for this data is expected to have quite extreme inclination towards lower grades. It is also concerning that the model

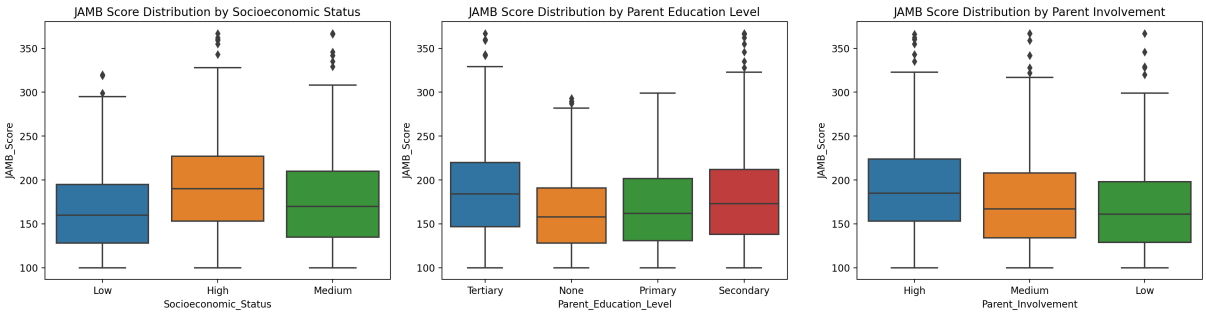
when spilt for training might encounter hardship for predicting higher values of student exam scores.



Mean	174.07
Std	47.61
Min	100
50 %	170
Max	367

Next, we see “JAMB Scores vs. Study hours per week" and “JAMB Scores vs. Attendance Rate” spotted in the scatter plot seems to be a positive relationship that follows an upward diagonal direction with very broad expansion in distribution. In “JAMB Score vs Teacher Quality" and "JAMB Score vs Distance to School", we can observe that in both scatter plots, both presents scattered and abnormal expansion across the dataset from each value. The abnormal distribution of data points can be concluded to show weak association despite the fact that it has a rough upward diagonal shape. However, we can conclude that this highlights the fact that “JAMB Scores vs. Study hours per week" and “JAMB Scores vs. Attendance Rate” have the observed linear relationship out of the four variables in our scatter plot.





However, we can see that in the scatterplot, it includes only numerical variables in the dataset. Thus, we include three important non-numerical variables as box plots for overview visualisation. The reason for the selection of individual box plots is The three variables are “Parent Education Level,” “Socioeconomic Status,” and “Parent Involvement.”

The reason of the three selected are based on examining the correlation matrix of categorical variables. Categorical variables are transformed using one-hot encoding to dummy variables with respect to JAMB Scores for the correlation calculation. We can see that a. High parent involvement b. Tertiary education c. High socioeconomic status indicate quite a relationship in contributing to JAMB Scores. However, across three factors, notable outliers shown in box plots suggest that performance can still vary quite significantly.

After the discovery and exploration in our datasets, we have found a few important patterns before the implementation of models: a. non-numerical variables contribute to the outcome of JAMB scores. b. numerical variables presents possible relationship to JAMB scores but also suffers from high variances as the datapoints are shown to be wide-spread.

3. Methods and Visualisation

Thorough the modelling process, the fitting procedure follows feature selection, model comparison, and model interpretability. Tug of war between two major focus of

models was highlighted throughout the process: production accuracy versus model interpretability. We first transformed the data into various options. Then we present five different methods in finding the best model for JAMB scores. Throughout, three methods focus on linear regression and two utilise non-linear methods for in aid of capturing the relationship of predictor to response.

3.1 Data transformation

In this project, the dataset was transformed into three:

a. *Encoded dataset* that non-numerical variables are each one-hot encoded into individual columns and columns of “Student_ID” and “Gender” are dropped for leaving out in consideration of modelling.

b. *Interactive Dataset* that variables of both categorical variables and continuous variables are introduced with an interaction effect. The removal of addictive assumption is explored in this dataset. Interaction terms added are listed in the table below:

Interaction variables	Variables X1	Variables X2
StudyHours_ParentInvolement	Study Hours per Week	Parent Involvement High
Attendance_Socioeconomic	Attendance Rate	Socioeconomic Status High
Teacher_ParentEducation	Teacher_Quality	Parent Education Tertiary
StudyHours_Attendance	Study Hours per Week	Attendance Rate

c. *Principal Components Analysis (PCA) Dataset* is explored to further conduct a principal components regression (PCR). We obtained the dataset table transformed into PCA Loadings and new PCA dataset that original data are projected onto the new principal component axes.

d. *Polynomial Dataset* where variables 'Study_Hours_Per_Week', 'Attendance_Rate', ‘Teacher_Quality' are transformed into a degree-2 polynomial.

The reason for the data transformation are extremely useful not only fitting into our training models and testing methods, but also it presents us a thorough, combined way of selecting features of importance.

3.2 Model Training

The set of course for method is followed by the core rule of trying to get the best way for understanding the level of variables in the dataset that contributes to different students/examines background. In short, the project is surrounded by deep concept around linear regression. Further expanding from linear regression, selection and regularisation is heavily and routine utilised. In between these different models, each different models also undergone resampling methods and hyper-tuning.

3.2.1 Least Square Linear Regression

We started with best subset selection with *the encoded dataset*. This is conducted via linear regression and using RSS and R^2 for comparison we found that the processed 2380 models on 4 predictors has the highest scores.

$JAMBScore = 15.65SocioStaHigh + 11.69Teach.Q + 1.76StudyHours + 1.26AttendRate + \epsilon$, with R^2 = 0.951.

What we are focusing on here is variables Socioeconomic Status High and Teaching Quality. Each change in these two variables could bring about 15.65 or 11.69 in change for students’ JAMB Score. respectively. Due to the best subset selection, we then moving on to apply interactive terms to our modelling. The variables chosen for removing addictive assumptions are based from our above *equation I.1* and is

selected from *function getBest(4)* which filters through and outputs the most optimal of all possibilities with 4 predictors.

We go through the same procedure for the *interactive dataset*. In order to satisfy the presumption of the condition for linear regression, we also plot out the residuals are normal distributed and proven true.

$$JAMBScore = 12.87ParentInvolvement + 11.39TeacherQuality + 1.74StudyHour + 1.24AttendRate + 0.17InteractAttenda.Socioeco + \epsilon, \text{ with } R^2 = 0.952$$

It is interesting to see that with the interactive terms, the R^2 only increase by 0.001.

To dive in deeper, we observe that based on the covariance matrix and scatterplots, we find that instead of a linear relationship, the variables could be carrying polynomial relationship with JAMB Scores.

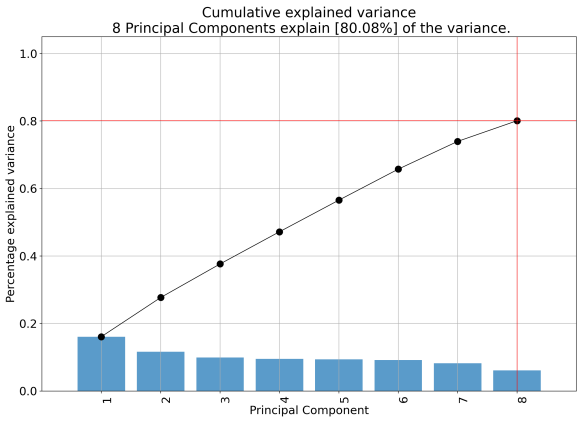
$$JAMBScore = 109SchoolPrivate + 103SchoolPublic + 0.18Interact.Attend.Scocio + 0.13Attend.Teach.Q + 0.02Interact.StudyH.Attend, \text{ with } R^2 = 0.306$$

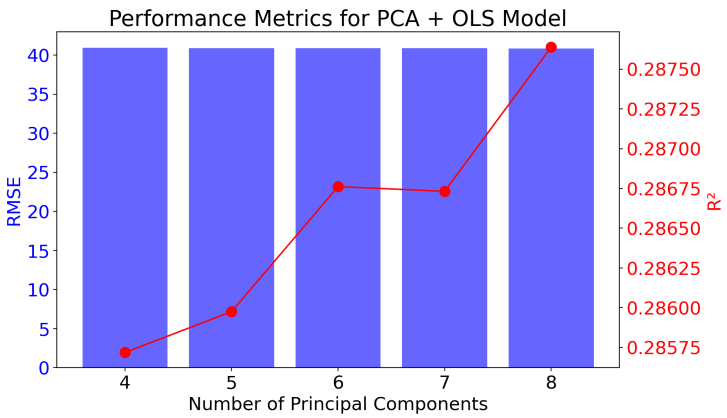
Utilising *Polynomial Dataset,* we result are quite discouraging as the RSS further increases to 7863918 with five predictors, compared to 7784222 with five predictors with interactive terms dataset and the model did not select any polynomial variables.

3.2.2 PCA and PCR

As our dataset is expanded from [5000 x 14] to [5000 x 21], we carry on to a principal component analysis. We first obtained the PCA loadings vectors in which we can have a glimpse of the positive and negative loadings for each principal component. For PC0 (the first principal component): "Study_Hours_Per_Week" contributes most

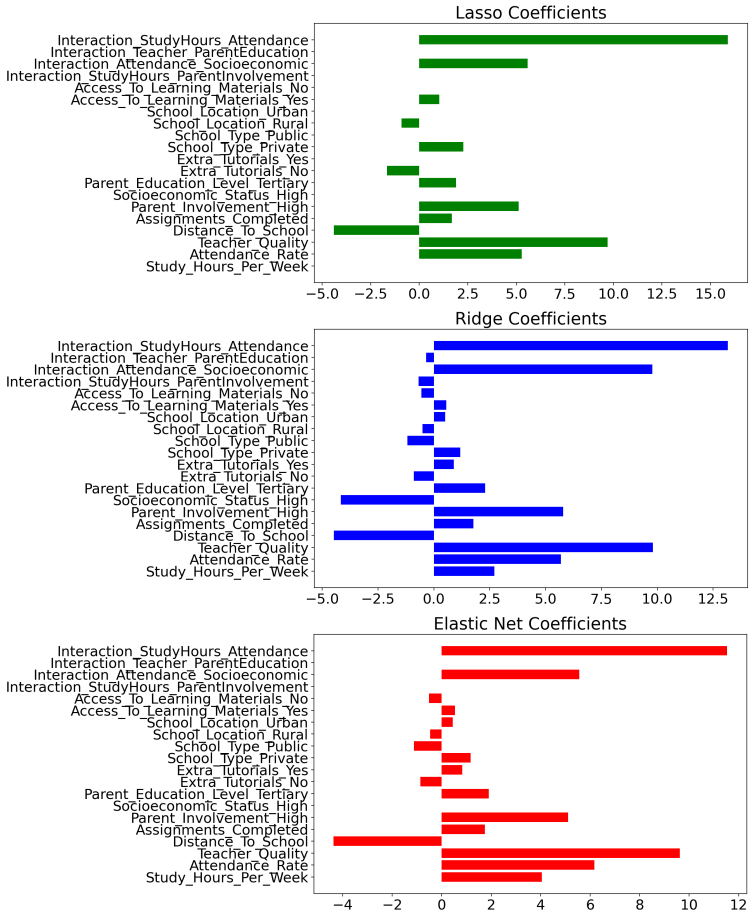
positive and significantly to PC0 whereas “School_Type_Public” has a loading of -0.084486 has most negative contribution to PC0. For PC1 (the second principal component): "Parent_Education_Level_Tertiary" has a loading of contributes strongly to PC1. The histogram presents the cumulative explained variance across the first 8 principal components which accounts for 80.08%. In following, we incorporate linear regression with PCA, using each eigenvalues and eigenvectors projected onto principal component axes. It is important here to note that PCR is not a feature selection method. The alteration and projection of the nature of PCA is constructed using each PC_n in a linear combination using *all* original features in our interactive dataset. The PCA and PCR method was conducted to find and experiment the best model that can be further implemented for student performance performance. According to the Performance Metrics for PCA + OLS Model, each components added results in marginal improvement. If we take a look at the AIC and BIC scores into account, the model with 6 component shows the lowest score which can indicate a possible fit and complexity trade-off balance. The fact that the R^2 was quite low pressure us to revalidate and thus both bootstrapping and cross-validation are used on both training and testing. The score attained was a. cross-validated RMSE score 40.34 b. bootstrapping mean RMSE score 40.47, mean R^2: 0.27. This urge the confusion as to why PCR underperforms or inability of being useful in this dataset. The discovery is explored at this paper in Discussion.





3.2.3 Lasso, Ridge, Elastic Net

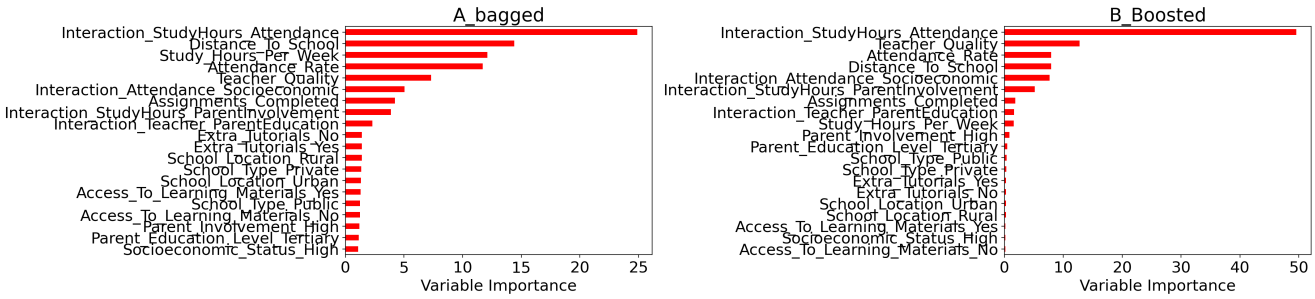
A fourth approach with the best already equipped method for linear model feature selection is undoubtedly Lasso and Ridge. We also introduced Elastic Net for comparison. The hyper-parameters are tuned using cross-validation. Most distinctive and noticeable choice of the three shrinkage models are “Interaction of Study Hours and Attendance” contributes strongly to the student JAMB scores and “Distance to School” is the least that bring about to the score. If we take a closer look at Lasso coefficients, we can actually see that it penalised 8 variables, especially the “Interaction of Study Hours and Parent Involvement” and the variable is penalised across all three models. The second highest rank of features is also attendance rate which is in accordance to the original least square model.



3.2.4 Tree-based Methods

We utilised the ensemble methods to examine the possibilities of capturing the relationship of JAMB Scores and our variables. Bagging and Gradient Boosting methods are implemented for comparison. The reason for further including tree-based methods is due to the inability of above tested models to truly capture the JAMB score to explanatory variables other than the simplest OLS. It is needed to find the best trusting models that can truly give the most accurate interpretation of our research question. The features of importance is also acquired for reference. In the process, we tried out the boosted model using *Polynomial Dataset* and does not present any significance and hence is not included here.

Taking a closer look, it is *reversely* interesting to see that only Tree-based models consider “Distance to School” to be highly informative than that of shrinkage methods. “Interaction Study Hours and Attendance” still remains the top highly informative variable.



4. Discussion

We come to the result presentation and there are a few issues we encountered throughout the process should be answered and addressed. a. PCA and PCR effectiveness b. bias and variance of models.

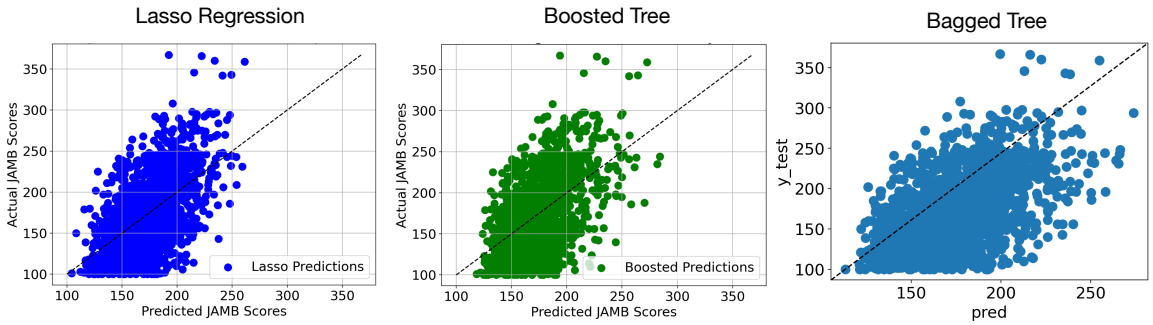
a. PCA and PCR effectiveness

The most intriguing discovery when conducting the PCA and PCR methods is that 1. it does not seem to bring about higher model performance in terms of the dimension reduction. The principal component in cumulative explained variance graph presents a steady percentage addictive relationship in stead of highly explained PC from top 1 or 2. Once fitted into PCR, our information compressor should have return a much smaller number of dimensions to pack-the-variance-up. If each of our PC explained same amount of information, the reduction is not considered useful and this is the case we are seeing now. We consider the issue to be further examined at one-hot encoding when combined with PCR. PCA based on continuous variables in attempts to find linear combinations. We are using categorical variables and the binary and sparse nature of this sparse data handling are making the PCA components more noisy than informative. As we can see each binary variables are simply offset by each other in each principal components.

Extra_Tutorials_No	-0.044158	0.000158	-0.482943	-0.098820	0.107195	0.478024	-0.107740	-0.002594
Extra_Tutorials_Yes	0.044158	-0.000158	0.482943	0.098820	-0.107195	-0.478024	0.107740	0.002594
School_Type_Private	0.084486	-0.109794	0.188092	-0.173638	0.639695	0.020148	-0.024002	0.019594
School_Type_Public	-0.084486	0.109794	-0.188092	0.173638	-0.639695	-0.020148	0.024002	-0.019594
School_Location_Rural	-0.024290	0.004828	-0.119511	0.657708	0.221896	-0.032434	0.031936	0.019478
School_Location_Urban	0.024290	-0.004828	0.119511	-0.657708	-0.221896	0.032434	-0.031936	-0.019478
Access_To_Learning_Materials_Yes	0.051315	0.005378	-0.460250	-0.147424	0.108345	-0.485172	0.124469	-0.018935
Access_To_Learning_Materials_No	-0.051315	-0.005378	0.460250	0.147424	-0.108345	0.485172	-0.124469	0.018935

b. bias and variance of models

Insofar, most of our models trained when plotting visual graph of predicted versus actual values, we observed a heavy issues of high bias.

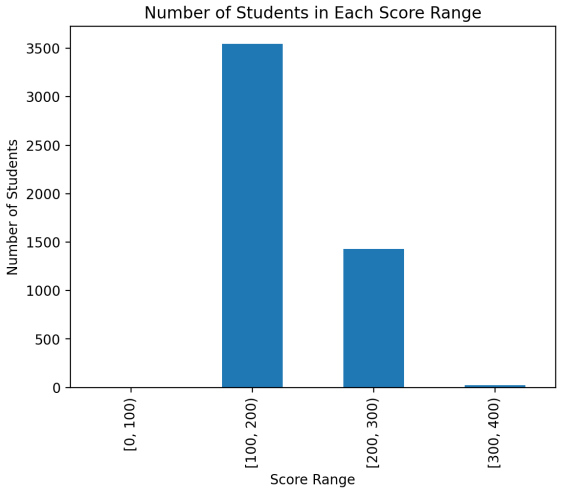


Even after hyper-tuning, feature selection and non-linear relationship, the dataset itself exists the problem of highly skewed input. If we take a look back at the original data, the min. value is 100 and max. value is 367. However the medium is only 174. Large amount of students gathered around below 200 despite of the economical or educational background. We suspected multicollinearity but based on AIC scores, we find little possibility in this dataset. Most models underestimate actual values as the students with higher scores are vey few and it could be recognised as outliers instead of normal distributed input. The models all struggles with predictions in higher scores and find limitations in capturing variance in high-scoring students.

5. Result and Conclusion

We ought to bring back the main prioritised problem statement: a. student performance b. social disparities/resource allocation.

In evaluation, we have successfully form, developed a model for predicting student performance and JAMB scores. The model can be used before the students embark on the exams. We tried out more than seven types of models, each is a step connected to the next choice of model in hope of discovering more useful information from the dataset.



Model	R^2
Ridge	0.34
Lasso	0.34
Elastic Net	0.34
PCR (6 Components)	0.27
Bagged Tree	0.27
Boosting Tree	0.32
OLS	0.95

Having the model developed, taking a step back we can discover that in average, the students in Nigeria tends to perform below the mean score which is 200 where medium is 174. Despite of different social backgrounds, many students still cannot performs well. There could be more hidden social economics situations that can be combined with our model to be discovered. If we combine the histogram of JAMB Score, our Actual versus Predicted Scores Plots of the models, the skewness of JAMB Score distribution predetermined our models to undermine high scores and prediction. The struggle of accurately obtaining high achievers could also be that of the dataset train-test spilt are unequally assigned. The graph below presents that the bins for each range of JAMB scores according to the sum of students fall in that range. It is highly likely that most training dataset are assigned to score range [100,200) or [200,300).

Mean	196
Std	50.12
Min	107
50 %	191
Max	319

Socioeconomic_Status	School_Type	School_Location	Parent_Education_Level	Extra_Tutorials	JAMB_Score
High	Private	Rural	None	No	4
				Yes	1
			Primary	No	8
				Yes	10
			Secondary	No	17
				Yes	20
			Tertiary	No	27
				Yes	32
		Urban	None	No	3
				Yes	3
			Primary	No	10
				Yes	10
			Secondary	No	22
				Yes	33
			Tertiary	No	43
				Yes	46

However, if we take a opposite point of view, what if this data is the true representation of the phenomenon? A problem statement of our paper answers: students with time, resources, guidances and the right background ought to do well, right? Apparently not so likely. Below using groupby, We have obtained the students that are given with highest quality of background which is retrieved by ['Parent_Education_Level'] == 'Tertiary') & (['Socioeconomic_Status'] == 'High') & (['School_Type'] == 'Private') & (['School_Location'] == 'Urban') &(['Extra_Tutorials'] == 'Yes')].

Feature top 4	Feature top 8
Socioeconomic_Status_High	Study Hours and Attendance Rate
Teacher_Quality	Attendance and Socioeconomic
Study_Hours_Per_Week	Teacher Quality
Attendance_Rate	Parents Involvement

The table presents this new group of 46 student's mean with 196 and min. score actually 107 which is only 7 score higher than overall min. score which is 100. Max. score is 367 which is also not in this group. This gives us big support on defending that having all resources can produce the best students in performance. Now taking into our trained model into considerations, the top four features influencing student performance are:

The selection is based on the adjustment, advisory, and result of the whole process after discussion and combination of the best OLS model with four predictors are prioritise most then comes with our lasso and ridge shrinkage methods to filter the significant contributors with interactive variables. It is a surprising dataset that shows the true meaning of "hard-work pays off!" "Distance to School" is ranked among the worst variables for contribution of most models we presents and all contributing variables are actually self-determined odds. Attendance Rate and Study Hours are among the most important variables in JAMB Score performance and it is the representation of putting in the work against all odds! As our main hypothesis was that social disparities and resource allocation can highly affect and could be the main

reason for JAMB Score performance. However, in the case for our dataset, this has been proven not the case and is highly enlightening to see such promoting result. At last, the left skewed distribution of JAMB Score then could probably be simply: student ought to study more!