IRSTI 55.01.11

Diana Bairamova[1]
[1]Suleyman Demirel University, Kaskelen, Kazakhstan

PREDICTING COURSE GRADES OF STUDENTS' ACADEMIC PERFORMANCE USING THE LIGHTGBM REGRESSOR.

Abstract. In the modern world, using all available opportunities and technologies, special attention should be paid to the development of the education system of students, since education serves as the basis for the development of the future generation. Nowadays, thanks to the use of available Artificial Intelligence methods, it is possible to predict various events, anomalies or other important things. With the help of machine learning, it is possible to predict at an early stage of a student's education whether he will finish the course successfully or not. In this study, it is proposed to predict the final score which student will receive at the end of the course using a number of predictors as an assessment for the first quiz and 3 types of tasks using the LightGBM regressor, which is a high-performance algorithm with gradient boosting. The results of using the LGBM regressor using GridSearchCV allowed to determine the best settings of hyperparameters from three selected tree-like boosting methods: 'dart', 'gbdt', 'goss'. The GOSS method was determined to be the best of the three methods listed with an estimate of R2 score in 0.81, which is 0.24 more than the R2 score of the Linear Regression forecast of – (0.57).

Keywords: Machine learning, grades prediction, outliers' identification, LGBM Regressor, Linear Regression.

***

Аңдатпа. Қазіргі әлемде барлық қолжетімді мүмкіндіктер мен технологияларды қолдана отырып, оқушылардың білім беру жүйесін дамытуға ерекше назар аудару керек, өйткені білім болашақ ұрпақтың дамуына негіз болады. Қазіргі уақытта қолжетімді жасанды интеллект әдістерін қолдану арқылы әртүрлі оқиғаларды, ауытқуларды немесе басқа маңызды нәрселерді болжау мүмкіндігі бар. Машиналық оқытудың көмегімен оқушының курсты сәтті аяқтайтынын немесе аяқтамайтынын ерте кезеңде болжауға болады. Бұл зерттеуде оқушының курстың соңында алатын қорытынды бағасы болжанады, және градиентті бустинг арқылы жоғары өнімді алгоритм болып табылатын lightgbm регрессорының көмегімен бірінші викторина мен тапсырмалардың 3 түрі бойынша бағалау ретінде бірқатар болжаушылар арқылы алады. Gridsearchcv көмегімен регрессорды қолдану нәтижелері таңдалған үш ағашты бустинг әдісінің ең жақсы гиперпараметр параметрлерін анықтауға мүмкіндік беретін келесі

әдістемелер болып табылады: 'dart','gbdt','goss'. GOSS әдісі бойынша модельдің R2 көрсеткіш 0.81 құрайды, сызықтық регрессия R2 (0.57-ды) құрайды, ол 0.24 - ға дейін жоғары болып табылатын тізімделген үш әдістің ең жақсы нұсқасы болып анықталды.

Түйін сөздер: Машиналық оқыту, бағалауды болжау, шығарындыларды анықтау, LGBM регрессоры, Сызықтық регрессия.

***

Аннотация. В современном мире используя все доступные возможности и технологии особое внимание стоит уделять развитию системы образования учащихся, так как образование служит основой развития будущего поколения. В наше время, благодаря использованию доступных методов Искусственного Интеллекта есть возможность предсказания различных событий, аномалий или других важных вещей. При помощи машинного обучения, можно предсказать на ранней стадии обучения студентом окончит ли он курс успешно или нет. В данном исследовании предсказывается итоговая оценка, которую ученик получит в окончании курса используя ряд предикторов как оценка по первой викторине и 3 видам заданий при помощи LightGBM регрессора, который является высокопроизводительным алгоритмом с градиентным бустингом. Результаты использования регрессора с помощью GridSearchCV позволили определить наиболее лучшие настройки гиперпараметров из трех выбранных древовидных бустинговых методов - 'dart','gbdt','goss'. Метод GOSS был определен лучшим из трех перечисленных методов с оценкой R2=0.81, что на 0.24 больше оценки R2 Линейной Регрессии – (0.57).

Ключевые слова: Машинное обучение, прогноз оценок, определение выбросов, LGBM Регрессия, Линейная Регрессия.

I. Introduction

In the age of technology, we have at hand many opportunities to improve everything around us, perfect existing systems or create new ones. All this can allow us the world of Artificial Intelligence, which is simply unlimited and has a lot of useful opportunities for the development of learning systems and for other aspects of life. The education system in the age of accessible technologies should help both students and teachers to respond to any gaps or shortcomings in the learning process of students. Machine learning algorithms can allow detect which students need the help of a teacher. For example, whether the student will pass the course successfully or what final grade the student will receive at an early stage of training in order to help the student and the teacher determine which students need to be helped the most. In this study, a number of experiments are carried out on the available real data collected by scientists from

the North American University in March 2020 [6] to determine the final grade that the student will receive during the graduation of the subject. In their study [6], authors clustered students into 2 groups of students who need help in learning using the K-means clustering algorithm. To do this, the authors collected 12 predictors such as the time of passing various tasks, grades and the number of posts and content read on the online learning platform. In contrast to this study, the authors analyzed one of 2 groups of students who need help in learning process, while in this study, the assessment of the course is predicted using an effective method with the method of gradient boosting.

With the help of available Machine Learning algorithms, we can predict the final assessment of a student's course at an earlier stage of student training, which can help increase the student's learning performance not at the end of training when the student will not be able to change his situation, but in the middle of his training. Earlier, scientists in their study [7] revealed that the notification system whether a student completes the course successfully or not was able to increase the performance of the course by 23%.

Nowadays, it is very important to focus on increasing the motivation of students in their studies, since many courses at universities, colleges or other training events have switched from an offline format to an online one. This can negatively affect the level of motivation for learning, so predicting the course assessment can be a good incentive to maintain the effectiveness of education. In this study, before training the data by the machine learning algorithms, the collected data were analyzed to obtain a higher accuracy of predicting the course assessment.

The analysis of the collected data (more information in the subheading "Data description and preprocessing") showed that the data have outliers that may prevent us from achieving a good prediction result.

Thanks to the use of ready-made function - np.where() in the NumPy library, all founded outliers in the data were replaced with an average values. In addition to analyzing and processing data, an important step in predicting student grades is to identify the significance of the available predictors in the data. To identify significant predictors, the study used a table of ordinary least squares (OLS), where there is a p-score near each predictor, which indicates the significance of the predictor. If this value is more than 0.05 [10], the predictor is not a significant (dependent) predictor, and vice versa, if it is less than this value, the predictor is a significant predictor. As a result of the resulting OLS summary table, it was revealed that all predictors are dependent on the prediction score, since their p-value was close to 0, which explains the significance of the predictors [10].

In this study, two methods of predicting estimates were used: Linear Regression and LightGBM Regressor. Experiments (in the Results section) show that the regression of LGBM using the boosting – "GOSS" method largely exceeds the level of accurate prediction of Linear Regression. LGBM is a very effective method [11] that can be used to predict numerical values, including

predicting student grades. This is a regressor that is based on decision trees and has various boosting methods for increasing the gradient.

II. Literature review

In their study [1], authors used three different methods of ensemble learning as the LGBM gradient enhancement method, the random forest (RF) and the extreme gradient enhancement method (XGB), which learns to predict the target value based on estimates of simpler models. Using the combined method of the three collected classifiers allowed scientists to three classes of academic performance of students predict with 97% accuracy of the model, where class A includes students with a high level of academic performance, class B includes students with a lower level and the last class C [1] includes students with a low level of academic performance. This method, assembled from three classifiers, was used before the end of the training course at the University of Saudi Arabia (Umm al-Qura), which helped to increase the level of students to 98%, where the proportion of students from class C decreased from 12% to 1%, which is a positive impact of the use of an effective method of three classifiers [1].

Earlier, the researcher Mahesh Gadhavi in his research [2] uses two grades on internal exams to predict the final grade that a student will receive at the end of the course. In his work, the scientist used the Linear Regression method to predict the final score, however, in order to determine the final score, two intermediate scores are used as a variable that are converted from the original form (score A, B, C, and D) to a non-numeric format (percentage format) for the accuracy of the final score. To determine the final grade for the course, only one average value is taken from the two grades used as a predictor variable. The researcher notes that in order to obtain a more accurate prediction estimate, the input data were normalized to 100% [2] in order to get the correct final estimate. This solution of obtaining a final grade [2] based on the results of two intermediate grades was used at the Chandaben Mohan Bhai Patel Institute of Computer application, where to obtain a final grade for the course, a student must pass two internal exams for the course, which is used as one average value as a predictor variable.

Later in their work [3], scientists used the results of Linear regression prediction errors as the input layer of a Neural model consisting of an input layer (36), an output layer (1) and a hidden layer (1). A hybrid model based on Linear Regression and a Neural Network algorithm was trained on 70% of students and on the remaining 30% of student's hybrid model was trained. The hybrid model was trained from 35 selected significant predictor variables [3], which showed a strong correlation between the predictor data and the predicted estimate. As a result of the study, the hybrid model showed a coefficient of determination R2=1 [3], with precise variable predictors significantly affecting the final grade, which include not only the grades received during the course, but also the number of hours spent studying the subject, experience in various computer programs, spending time on homework, and so on.

To determine the success of the course [4] and increase the level of motivation of engineering students, scientists from the University of Chicago have identified the 3 most popular courses where students receive the lowest grades and subsequently students drop out or change their specialty or college. The first year is Mathematics, the second year is Programming and the third year is Physics [4], the data of which were used by researchers to determine the academic performance of students. As a result, scientists have proposed an effective Bayesian model that differs in high performance from a simple Bayesian model. In addition to academic performance as a preliminary exam score, the researchers used demographic data of students, which are statistically significant predictors for the model [4].

In the study [5], scientists found out that one of the most effective methods for predicting course grades is the SVM vector machine of all three machine learning algorithms used as the Naive Bayes method, KNN, SVM. The authors note the superiority in using the support vector machine in that with it we can get a more optimal classifier for different types of classes, since it creates 10 different classifiers and selects the most optimal of all available ones. The authors note that for KNN algorithms and the Naive Bayes method [5], the final evaluation of the model is predicted to be the same as it was received earlier, although in real life the opposite happens and students may end up with a much higher or lower score than they received earlier. The SVM method coped well with this phenomenon [5] compared to other algorithms.

III. Methodology

III.1 Data description and preprocessing.

In the study, the following predictors collected by scientists from North American
University [6] were used as data for predicting the assessment:

- Quiz assessment
- Assignment 1 assessment
- Midterm assessment
- Assignment 2 assessment
- Assignment 3 assessment

The collected data contains 486 rows with 8 columns of data about each student with grades of various tasks (quizzes), including student id, final score and the group to which the student belongs. Grades are presented in numerical int64 format, 3 groups as “Good”, “Satisfactory”, “Unsatisfactory” denoting academic performance on the course is the type of object. After analyzing the collected data, it was revealed that there is a negative data asymmetry for all three estimates of tasks, as we can see in the figures (1-3), the peak of the data distribution falls on the data distribution to the left, with a negative skew value.

Comparing the grades for the three types of completed tasks: Assignment 1 assessment, Assignment 2 assessment, Assignment 3 assessment we can say that

if a student receives a high grade for one of the assignments, then to a greater extent he receives a good grade for the other two assignments (Figure 4).

However, as we can see from figure 4, there are cases when a student receives a high grade on one assignment, but has a low grade on another assignment. As in the example in figure 4, we see how the student received a high score on assignment 2 - (>80%) and a low score on assignment 1 – 0%.

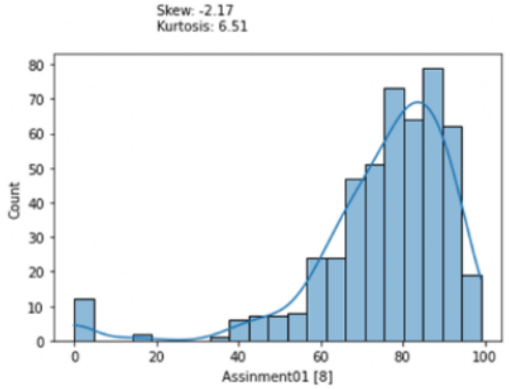

Figure 1. Distribution of data by predictor 'Assignment 1 assessment'

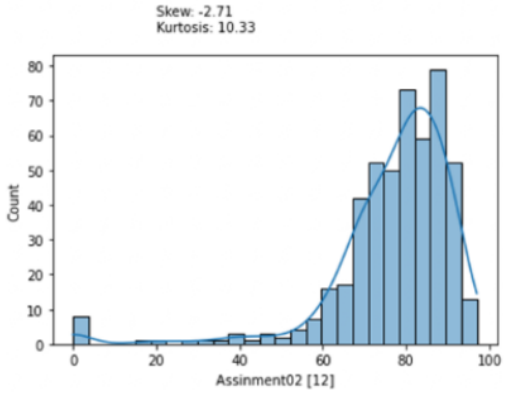

Figure 2. Distribution of data by predictor 'Assignment 2 assessment'

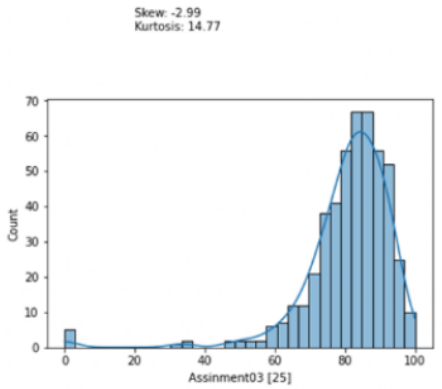

Figure 3. Distribution of data by predictor 'Assignment 3 assessment'

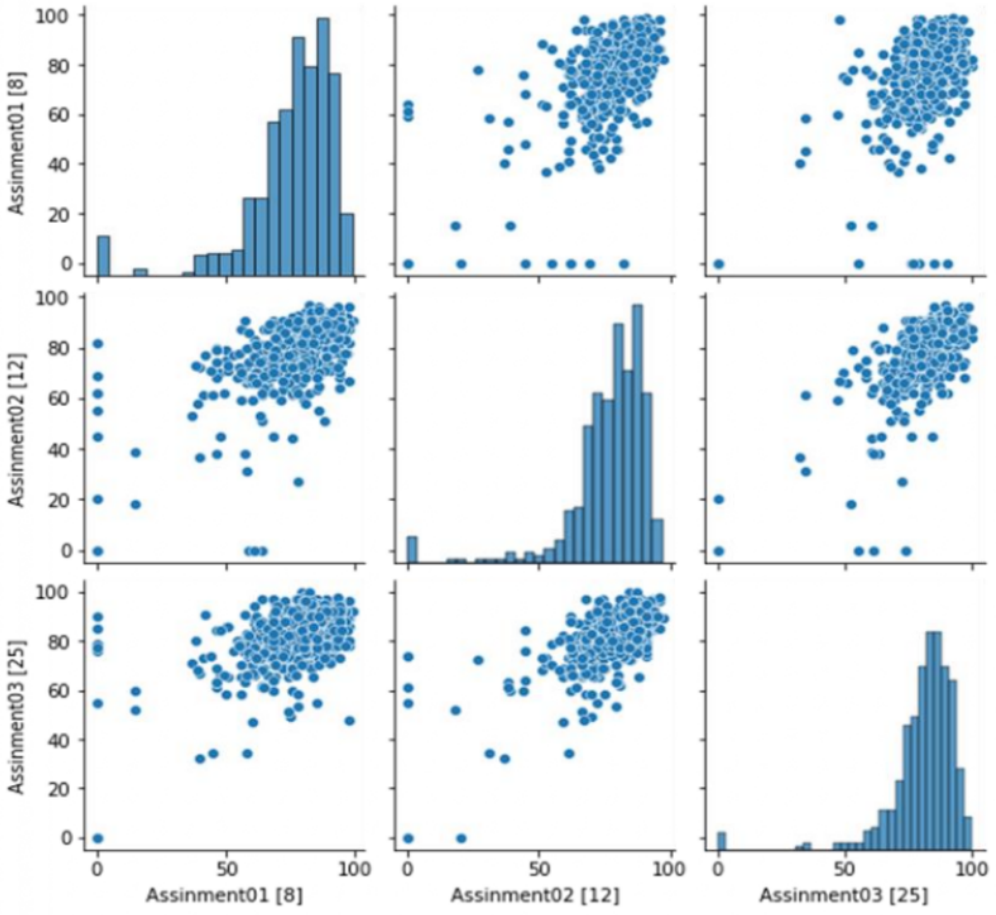

Figure 4. Pair plot distribution of 3 predictors from data

After analyzing the distribution of data from the three tasks in the data, an analysis of outliers in the data was carried out. Outliers are points or values in the data that differ significantly from the main distribution of the available data. For each of the selected predictors, the values of outliers in the data were determined using the well-known statistical method IQR - interquartile range, which calculates the range of emissions selected using the method – quantile (). It finds the 25th and 75th percentile of the available data [8], and then finds the difference between them, which represents outliers in the data that negatively affect the estimation of the coefficient of determination of the forecast [9]. Before finding out which predictors have hidden outliers that negatively affect the training of the model, you can easily determine using the ready-made describe() function, where we see the minimum, average and maximum values for each of the predictors. Thus, it was revealed that there are outliers in each of the available predictors for the forecast (Figures 5-8). To illustrate the outliers in the data, two types of graphs were used as a diagram with boxes and whiskers, a histogram. In the first type of diagram (figures 5-7), we see points in the graphs

that are our outliers since they are not included in the main data range. In the second form of the diagram (figure 8), we see single outliers on the left side of the diagram, which mean outliers in this predictor.

III.2 Proposed method

Before starting data training, work was done with outliers in the data, since the impact of outliers on training is large and can negatively affect the assessment of the coefficient of determination of the model, also work with outliers was done before scaling the data using the - StandardScaler function. Detecting in the dataset outliers was determined using the interquartile range Q1 which calculates values below which 25% of data points are located and Q3 - calculates all values below which 75% of data points are located [8]. Then, to find outliers, the difference between the values found is calculated from Q3 and Q1. After setting the formula for calculating outliers in the data, the following logic was used to work with outliers [9], namely, to begin with, all the found outlier points in the data were replaced with missing values using the specified formula (as if these values were not in the data set) and then these empty values were replaced with average values for each of the predictors: Quiz, Assessment 1-3, Midterm results:

1. Determination of values above/below the calculated difference between the 25 and 75 quantile of data.
2. Replacing the average value of each of the predictors.

Comparing the distribution of data after working with emissions before the introduction of working with outliers, the maximum and minimum values differed significantly from the average for every predictor in the data, however, after the introduction of the function of finding and replacing outliers, the maximum and minimum values did not differ significantly from the average for each of the predictors. Further, after identifying outliers and replacing them with average values, the StandardScaler data scaling method was used, which allows you to normalize the data and bring them into one range from the minimum to the maximum ranges from data. Having applied the necessary measures with data processing and preparation before starting data testing, the set was divided into training and test data sets with 33% of the test set and random_state = 1. The value of the prediction in the data is the final grade that the student will receive at the end of the course, the other factors were used as predictors to determine the final grade.

Applying Linear Regression with the obtained low coefficient of determination (R2 score), a boosting algorithm – LGBM [11] with an increase in the gradient was applied, with three known types of parameters:

- GOSS
- GBDT
- DART

In order to determine the best of the three parameters, the method of searching for optimal hyperameters as GridSearchCV [13] was used. The method of

searching for optimal parameter values showed that the GOSS type [12] is the best type, which is the best indicator of all three selected types of gradient increase. To visually see the best results obtained after applying the search for optimal parameters, the method was used .cv_results_ which returns the results obtained from the conducted search for optimal parameters using GridSearchCV [13]. By applying the GOSS boosting method [12] to the model, the best results were obtained in predicting student grades using all predictors.

IV. Results

In total, 34 outliers were identified in the data, which were replaced by the average values of the predictors. From the graph (5-8), there are outliers in the data, which have been replaced with the average value in each of the predictor.

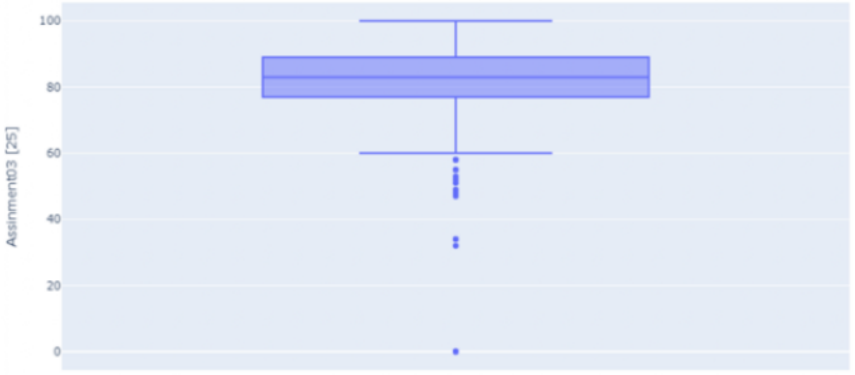

Figure 5. Outliers identification for 'Assignment 3 assessment' indicator

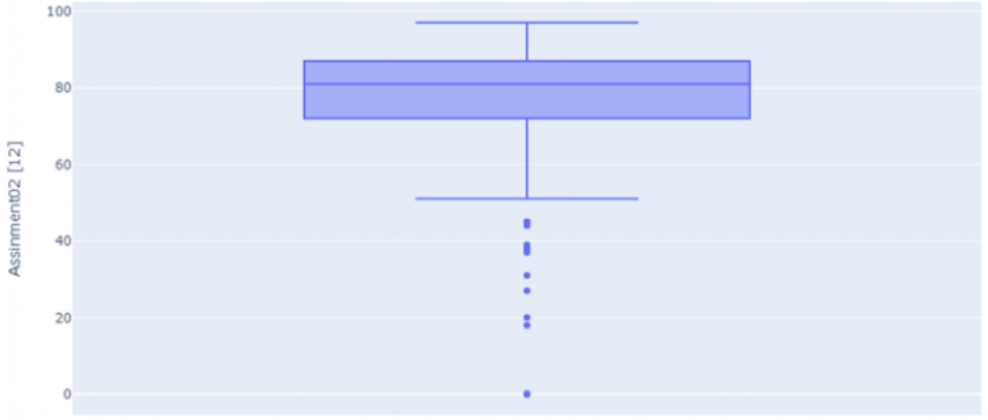

Figure 6. Outliers identification for 'Assignment 2 assessment' indicator

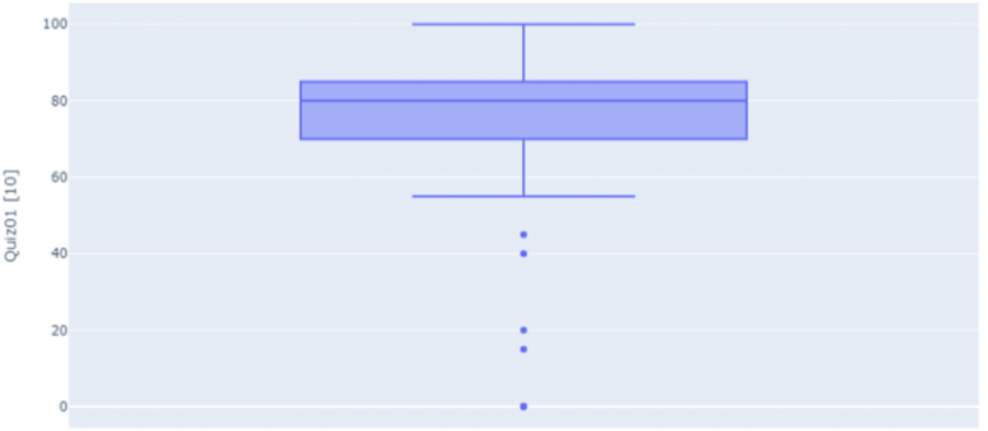

Figure 7. Outliers identification for 'Quiz assessment' indicator

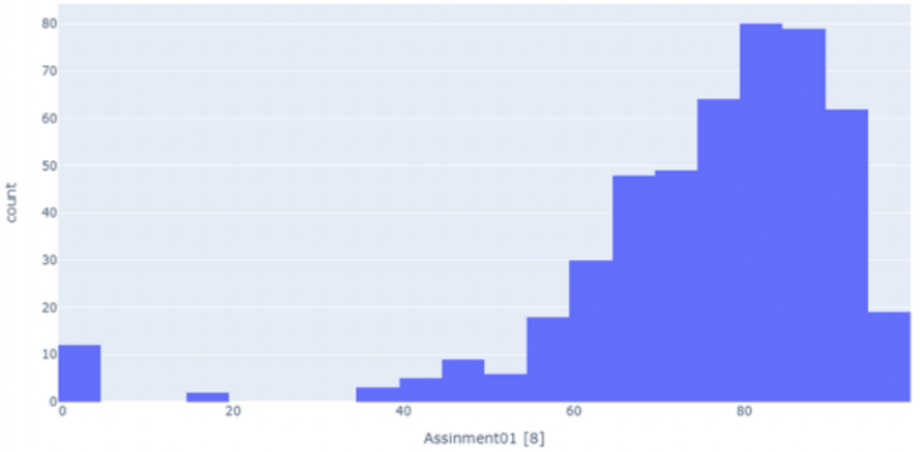

Figure 8. Outliers identification for 'Assignment 1 assessment' indicator

In this research, two models were used to predict the final grade that a student receives at the end of training. The results of the first model, Linear Regression, showed that it was necessary to improve the model or use another model, since the R2 score of the model was = 0.57, which is a very low indicator for the forecast.

From the resulting summary table after applying Linear Regression (figure 9), the p-value for each of the available predictors explains the significance or influence of the predictor on the variable of the result (final assessment) for the subject. For each of the predictors, all the variables are statistically significant, since their p-value values are less than 0.05 [10], which shows how much the variables affect the final grade on the subject. In this case, all predictors are significant and were used in the subsequent training of the model.

OLS Regression Results

Dep. Variable:	Course Grade	R-squared:	0.575
Model:	OLS	Adj. R-squared:	0.571
Method:	Least Squares	F-statistic:	129.9
Date:	Sun, 19 Mar 2023	Prob (F-statistic):	7.58e-87
Time:	16:46:38	Log-Likelihood:	-1658.2
No. Observations:	486	AIC:	3328.
Df Residuals:	480	BIC:	3353.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-23.5350	4.633	-5.080	0.000	-32.638	-14.432
Quiz01 [10]	0.1735	0.039	4.484	0.000	0.097	0.250
Assinment01 [8]	0.0942	0.034	2.784	0.006	0.028	0.161
Midterm Exam [20]	0.4981	0.037	13.442	0.000	0.425	0.571
Assinment02 [12]	0.2151	0.046	4.674	0.000	0.125	0.306
Assinment03 [25]	0.3254	0.050	6.502	0.000	0.227	0.424

Omnibus:	420.526	Durbin-Watson:	2.105
Prob(Omnibus):	0.000	Jarque-Bera (JB):	10398.390
Skew:	-3.705	Prob(JB):	0.00
Kurtosis:	24.415	Cond. No.	2.47e+03

Figure 9. OLS Regression Result after Linear Regression training

After applying the effective gradient enhancement method – LGBM [11], three different parameters of the boosting model were tested as: GOSS, DART and GBDT [12]. Of the three selected boosting methods, the best indicator fell on the first type - GOSS, which was determined using the method of searching for optimal parameters – GridSearchCV [13] with the highest indicator of the mean test score, which is the average of the test results trained and tested for each test combination from cv_results_.

Using the GOSS method made it possible to increase the R2 score of the model estimation forecast by 0.24 (for more information, see Table 1).

Table 1. Regression Algorithms Results

Name	R2	MAE	MSE	MAPE	Estimation of variance
Linear Regression	0.52	4.93	78.47	3.69	0.52
LGBM Regressor (GOSS)	0.81	4.06	33.57	3.07	0.81

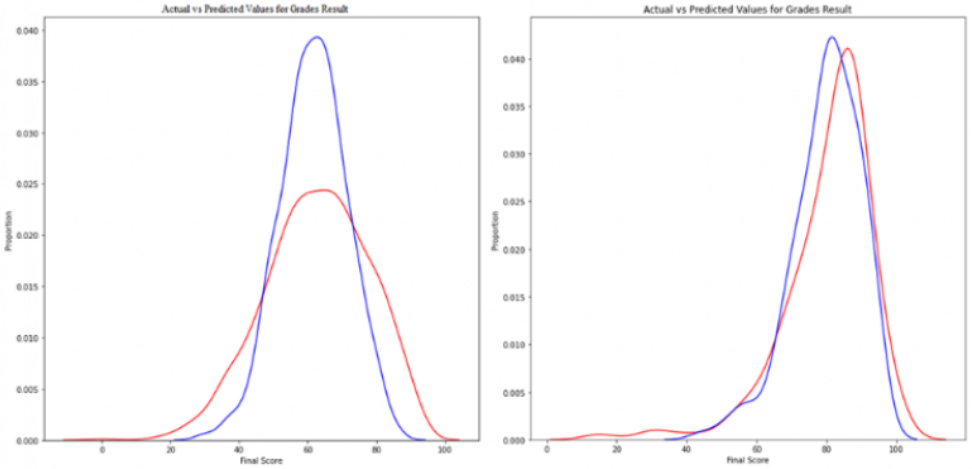

Figure 10. Actual and Predicted Values after usage Linear Regression (left side) and after LGBM Regressor (GOSS) (right side)

As you can see in Figure 10 on the left side, after using Linear Regression with a R2 score of up to 0.57, we get more error between the actual data and those that are received at the output after the forecast. On the other right side, the figure shows the results of comparing current and predicted data after using the LGBM regressor with R2= 0.81. After using the best of the three types of regressor - GOSS, the standard error decreased from 78 to 33, which is an effective method for determining student grades.

V. Conclusion

This study is devoted to one of the most important things at the moment - the development of the future generation, namely the prediction of student academic performance for some cources. The study used predictors such as: Quiz

assessment, 1-3 Assignment assessment and Midterm results and can predict before completion of the course which final grade student receive at the end.

Before using models to predict student grades, all predictors were analyzed, which showed that all predictors are statistically significant and the data also contains 486 outlier points. Using the method of finding outliers – IQR [8], 25 quantiles and 75 quantiles were determined, which served as the beginning in the search for outliers. After finding all the outliers, the search function was used .where() with which outliers were found, followed by replacing them with the average value in each of the predictors. Finding outliers allowed us to correctly apply the method - StandardScaler of normalization and standardization of data, which allows to scale data into one range.

After processing and analyzing the data, the Linear Regression algorithm would be applied in the study, which is a classic option for the task of predicting a numerical value rather than a class. The Linear Regression algorithm showed a very low R2 score=0.57 with a large root-mean-square error of 78. Further, to increase the prediction level, one of the most effective boosting methods for increasing the gradient was used - a LGBM regressor with the necessary hyperparameters determined by searching for optimal parameters – GridSearchCV [13]. This algorithm made it possible to increase the estimate of the model's R2 score and reduce the root-mean-square error from 78 to 33.

As you can see in figure 11, the best results between the actual and predicted data was obtained after using a gradient enhancement regressor (LGBM) with the GOSS boosting type.

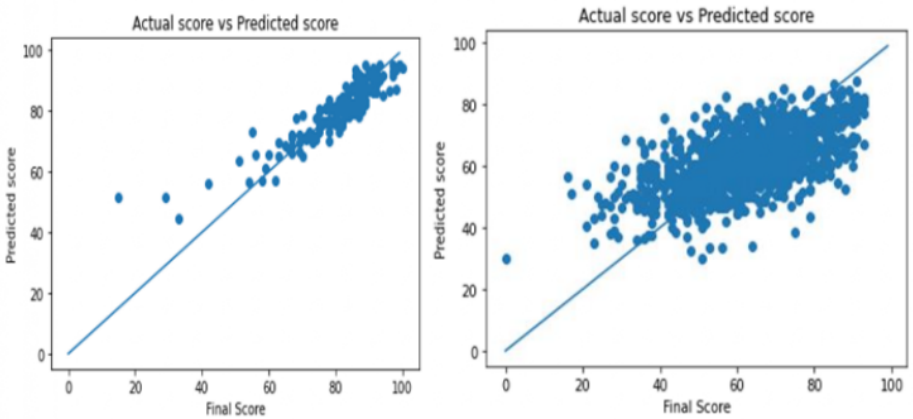

Figure 11. Actual and Predicted Values after usage Linear Regression (left side) and after LGBM Regressor (GOSS) (right side)

References

1 Lassaad K. Smirani, Hanaa A. Yamani, Leila Jamel Menzli, Jihane A. Boulahia, 'Using Ensemble Learning Algorithms to Predict Student Failure and Enabling Customized Educational Paths', Scientific Programming Towards a Smart World 14 April 2021, Volume 2022

2 Mahesh Gadhavi, 'Student final grade prediction based on linear regression', Indian Journal of Computer Science and Engineering, Vol. 8 No. 3 Jun-Jul 2017

3 Zoe Kanetaki, Constantinos Stergiou, Georgios Bekas, Christos Troussas, Cleo Sgouropoulou, University of West Attica, Athens, Greece 'A Hybrid Machine Learning Model for Grade Predictionin Online Engineering Education', iJEP – Vol. 12, No. 3, 2022

4 Ashkan Sharabiani, Fazle Karim, Anooshiravan Sharabiani, Mariya Atanasov, Houshang Darabi,'An Enhanced Bayesian Network Model for Prediction of Students' Academic Performance in Engineering Programs', 3-5 April 2014, Military Museum and Cultural Center, Harbiye, Istanbul, Turkey 2014 IEEE Global Engineering Education Conference (EDUCON) Page 837

5 Timothy Anderson, Stanford University Randy Anderson, California State University, Fresno, 'Applications of machine learning to student grade prediction in quantitative business courses', Global Journal of Business Pedagogy, Volume 1, Number 3, 2017

6 A. Moubayed, M. Injadat, A. Shami, and H. Lutfiyya, "Student Engagement Level in an e-Learning Environment: Clustering Using K-means", American Journal of Distance Education, 34:2, pp. 137-156, Mar. 2020

7 Cameron I. Cooper "Using Machine Learning to Identify At-risk Students in an Introductory Programming Course at a Two-year Public College". Advances in Artificial Intelligence and Machine Learning; Research 2 (2) 407-421, (July 2022).

8 Xiang Wan, Wenqian Wang, Jiming Liu & Tiejun Tong, 'Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range', BMC Medical Research Methodology volume 14, 135 (19 December 2014)

9 David F. Andrews, Daryl Pregibon, 'Finding the Outliers that Matter', Journal of the Royal Statistical Society: Series B (Methodological), (1978), 40, No.1, pp. 85-93

10 Marta D'Alonzo , Laura Martincich , Nicoletta Biglia , Alberto Pisacane , Furio Maggiorotto , Giovanni De Rosa , Filippo Montemurro , Franziska Kubatzki , Piero Sismondi , Riccardo Ponzone, 'Clinical and radiological predictors of nipple-areola complex involvement in breast cancer patients', European Journal of Cancer Volume 48, Issue 15, October 2012, Pages 2311-2318

11 Ritha Nyirandayisabye, Huixia Li, Qiming Dong, Theogene Hakuzweyezu, François Nkinahamira, 'Automatic pavement damage predictions using various machine learning algorithms: Evaluation and comparison', Results in Engineering, Volume 16, December 2022, 100657

12 Joao Miguel Mendes Ribeiro Agulha, 'Ensembling Neural Networks for Regression', Faculty of engineering of the University of Porto, July 24, 2021

13 Ghulab Nabi Ahmad, Hira Fatima, Shafiullah, Abdelaziz Salah Saidi, Imdadullah, Institute of Applied Sciences, Mangalayatan University, Aligarh, Uttar Pradesh 202, 'Efficient Medical Diagnosis of Human Heart diseases Using Machine Learning Techniques with and without GridSearchCV', Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques, Volume 10, 2022 80173