Student Marks Prediction using Linear Regression

Authors Name/s Yash Saini
Department of **Electronics and Communication Engineering**
Delhi Technological University.
New Delhi 110 042, India
Email id:-
yashsaini_2k18ec195@dtu.ac.in

Abstract—Data mining plays an important role in the business world and it helps to the educational institution to predict and make decisions related to the students' academic status. Student not giving semester exams has to settle with low semester grade, thus now a day student is not getting placed into reputed MNCs, it affects not only the students career but also on the reputation of the institute. The existing system is a system which maintains the student information in the form of numerical values and it just stores and retrieve the information what it contains. So the system has no intelligence to analyse the data. The machine learning model proposed in this paper is an advanced solution for predicting students' marks using Linear Regression technique for the prediction of semester grades based on students' previous semester grades. This experiment is conducted on 1500 students' in Delhi Technological University. Result proves that Linear Regression algorithm provides very high accuracy for predicting grades. The system aims at increasing the success graph of students who couldn't able to gives exams using Linear Regression model. It takes student previous semester grade as input the predict grade of that semester in which student scored low grade.

Keywords— *Linear Regression, Predicting student grade, Classification, Gradient Descent.*

I. INTRODUCTION

Analysing the huge amount of data to form summarized useful information is a tedious task for human kind. Data Mining is the area which analyses huge repositories of data to extract necessary or useful information. Computers can process any kind of data like numbers, texts, images and facts. This task performs the analysis based on the patterns, association, relations among all these data so as to get the information. The prediction with high accuracy in students' performance is beneficial as it helps in identifying the students with low academic achievements at the early stage of academics[6]. In universities, student retention is related to academic performance and enrolment system. The steps to assist the low academic performers with better education are:

(a) Generation of dataset of predictive variable.

(b) Data cleaning (removing unnessary data).

(c) Construction of a predictive model with the help of regression technique.

(d) Validation of the model which is developed for universities with students' performance.

Data Mining can be put in the educational field to extend our understanding of learning process by identifying the variables and evaluating them. Mining in the field of educational environment is known as the Educational Data Mining. By the means of Linear Regression It has been found that previous semester grades highly correlated with the upcoming semester grade. Data mining techniques economically offer more customized education, improved system efficiency, and reduce the education process expenses for universities. This guide us to extend student retention rate, increase academic achievements in case of student learning end result.

II. LIBRARIES

A. Numpy: -

NumPy is a python library used for working with arrays. It also has functions for working in the domain of linear algebra, Fourier transform, and matrices. It is an open source project and you can use it freely. NumPy stands for Numerical Python. It can be utilised to perform a number of mathematical operations on arrays such as trigonometric, statistical, and algebraic routines.

B. Pandas: -

Pandas is a Python package that provides fast, flexible, and expressive data structures designed to make working with structured (tabular, multidimensional, potentially heterogeneous) and time series data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python.

C. Matplotlib: -

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.

D. Seaborn: -

Seaborn is a Python data visualization library based on Matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

E. Random: -

Used in Python to randomly generate numbers.

III. Related Work

J K Jothi and K Venkatalakshmi conducted the students' performance analysis on the graduate students' data collected from the Villupuram college of Engineering and Technology. The data included five year period and applied clustering methods on the data to overcome the problem of low score of graduate students, and to raise students' academic performance [1].

Sheik and Gadage have done the analysis related to the student learning behaviour by using different data mining models, namely classification, clustering, decision tree, sequential pattern mining and text mining. They used open source tools such as KNIME (Konstanz Information Miner), RAPIDMINER, WEKA, CARROT, ORANGE, R Programming, and iDA. These tools have different compatibilities and it provided an insight into the prediction and evaluation [2].

Suyal and Mohod applied the association and classification rule to identify the students' performance. They mainly focused to find the students who need special attention to reduce failure rate [3].

Oyerinde O.D. and Chia P.A work on predicting Students' Academic performance- A Learning Analytics Approach using Multiple Linear Regression [4].

IV. Methodology

A. Data Preparation

Student related data were collected from result pdf released by Delhi Technological University, over 1500 students' of 2022 batch of different streams' data were collected, which include their 3rd and 4th semester grade

	roll_no	cg3	rollno	cg4
0	2K18/AE/003	4.64	2K18/AE/003	7.45
1	2K18/AE/004	4.55	2K18/AE/004	5.82
2	2K18/AE/005	7.64	2K18/AE/005	7.36
3	2K18/AE/006	9.27	2K18/AE/006	9.82
4	2K18/AE/007	6.91	2K18/AE/007	7.18
...	...	...	...	...
1541	2K18/SE/133	6.55	2K18/SE/133	6.82
1542	2K18/SE/135	7.09	2K18/SE/135	7.45
1543	2K18/SE/136	7.55	2K18/SE/136	9.09
1544	2K18/SE/137	7.91	2K18/SE/137	8.91
1545	2K18/SE/138	1.64	2K18/SE/138	2.50

1546 rows × 4 columns

Fig 1. Dataset used in the project

B. Data Cleaning

In this step only those students were selected which were required for making prediction. Students' 3rd semester marks were taken as the independent variable and students with grade too low for making predictions were removed.

C. Linear Regression Algorithm

Step 1: Scan the student data set

Step 2: Divide the dataset into training and testing 70% and 30% respectively

Step 3: Apply Gradient Descent to reduce the cost function so it can fit a line according to the training set.

Step 4: Making prediction using testing set.

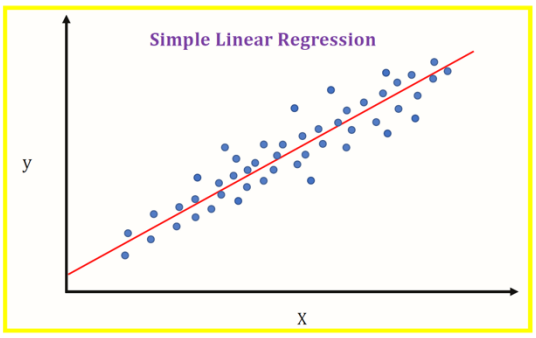
Step 5: Visualization of performance measures.

V. Linear Regression Algorithm

Linear Regression, in statistics is a linear approach to modelling the relationship between a scalar response (dependent variable) and one or more explanatory variables (independent variables) The case of one explanatory variable is called simple linear regression.

In Linear Regression a line is made closely fit to the points so that the error is minimum.

Fig 2. Linear Regression



Equation of line made fit to points on x-y axis.

A. *Equations*

- Equation of Cost function:

$$f(m,b) = \frac{1}{N}\sum_{i=1}^{n}(y_i - (mx_i + b))^2$$

Where:
yi: - training set output
xi: - training set input
N, n: - number of training instances
m: - weight (slope of the line)
b: - bias (y intercept)

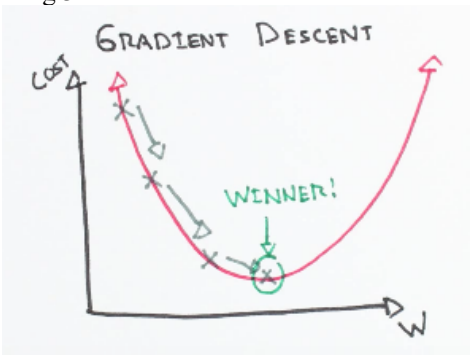
- Equation of Gradient Descent [5]:

$$f'(m,b) = \begin{bmatrix} \frac{df}{dm} \\ \frac{df}{db} \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum -2x_i(y_i - (mx_i + b)) \\ \frac{1}{N}\sum -2(y_i - (mx_i + b)) \end{bmatrix}$$

To solve for the gradient, we iterate through our data points using our new *m* and *b* values and compute the partial derivatives. This new gradient tells us the slope of our cost function at our current position (current parameter values) and the direction we should move to update our parameters. The size of our update is controlled by the learning rate.

B. *Learning Rate*

Fig 3. Visualization of Gradient Descent



The size of these steps is called the learning rate. With a high learning rate we can cover more ground each step, but we risk overshooting the lowest point since the slope of the hill is constantly changing. With a very low learning rate, we can confidently move in the direction of the negative gradient since we are recalculating it so frequently. A low learning rate is more precise, but calculating the gradient is time-consuming, so it will take us a very long time to get to the bottom.

C. *Math Behind Gradient Descent*

Gradient descent starts from a random point on the cost function and iteratively moves the points in the negative direction of the gradient of the function until the point reaches a local minima.

As a mathematical formula, it would look like:

$$\text{Repeat until convergence } \{$$

$$\theta_j \leftarrow \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta)$$

$$\}$$

Here, θj is the cost function and, in each iteration, θj is moved in the opposite direction to the gradient of θj. This will continue iteratively until the cost function is minimized.

VI. Performance Measures

A. *Mean Absolute percentage Error: -*

MAPE, is a measure of predicting accuracy in statistics, Ut usually expresses the accuracy as a ratio defined by the

Formula: -

$$\mathrm{M} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|,$$

Where At is the actual value and Ft is the predicted value.

B. *Mean Square Error: -*

Measures the average of the squares of the errors — i.e. , the average squared difference between the estimated values and the actual value.

Formula: -

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

Where Yi is the actual value and Y' is predicted value.

C. *Root Mean Square Percentage Error: -*

RMSE is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results.

The formula is:

$$RMSE = \sqrt{(f-o)^2}$$

Where

- f = forecasts (expected values or unknown results),
- o = observed values (known results).

D. R2 Score: -

"(total variance explained by model) / total variance." So if it is 100%, the two variables are perfectly correlated, i.e., with no variance at all. A low value would show a low level of correlation, meaning a regression model that is not valid, but not in all cases.

E. Mean Bias Error (MBE) : -

It is mean of difference of actual values and calculated values.

Formula: -

MBE = (y_test – y_predicted).mean()

VII. Result

TABLE I.

Performance Measures	***Percentage/ Value***
MAPE	5.9%
MSE	0.33
RMSPE	9.08%
R2 Score	0.79
MBE	-0.0007

Fig 4: - Result table

A. It can be seen from the result that based on MAPE the model has approx 94% accuracy.

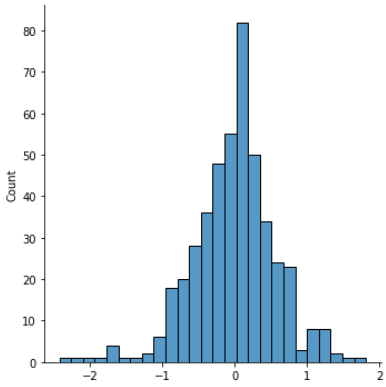
B. Based on RMSE it has approx 91% accuracy

C. Based on MBE it has approx 99.99% accuracy.

D. The R2 score is 0.79 out of the best score of 1 which implies good accuracy of the model.

E. Histogram plot of difference of actual values and predicted values is more concentrated at zero which shows that predicted values are very close to actual values

Fig 5.



Histogram plot.

VIII. Acknowledgment

We would like to acknowledge and appreciate the efforts and contributions of Mrs Juhi Jain of Computer Science Department of Delhi Technological University of New Delhi, India towards this research.

References

[1] J.K.Jothi and K.Venkatalakshmi, "Intellectual performance analysis of students by using data mining techniques", International Journal of Innovative Research in Science, Engineering and Technology, vol 3, Special iss 3, March 2014.

[2] Nikitaben Shelke and Shriniwas Gadage, "A survey of data mining approaches in performance analysis and evaluation", International Journal of Advanced Research in Computer Science and Software Engineering , vol 5, iss 4, 2015K.

[3] Sayali Rajesh Suyal and Mohini Mukund Mohod, "Quality improvisation of student performance using data mining techniques", International Journal of Scientific and Research Publications, vol 4,iss 4, April 2014.

[4] Oyerinde O.D. and Chia P.A work on predicting Students' Academic performance- A Learning Analytics Approach using Multiple Linear Regression .

[5] Ting Hu, Qiang Wu, Ding-Xuan Zhou, "Convergence of Gradient Descent for Minimum Error Entropy Principle in Linear Regression", IEEE Transaction on Signal Precessing (Volume: 64, Issur: 24, Dec. 15, 15 2016).

[6] Baradwaj, B. K., & Pal, S. (2012). Mining educational data to analyze students' performance. arXiv preprint arXiv:1201.3417.