IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11338

Multiple regression model for prediction of the probability of deviation from one's main aim in life

Yoshita Chakraborty[1], Prantika Baidya[2], Shubhadip Raj[3], Sulekha Das[4], Avijit Kumar Chaudhuri[5]

[1,2,3]UG-InformationTechnology, Techno Engineering College Banipur, Habra, West Bengal

[4]Assistant Professor, Techno Engineering College Banipur

[5]Assistant Professor, Techno Engineering College Banipur

Abstract: Success and failure are part of human life. Some of us may achieve our target and some may not. Those who could not achieve their desired carrier, usually opt for a suitable alternative to settle in life. It has also been noticed that many students desired to study in their dream institutions, but on failing to fulfil the necessary admission criterion had to switch to institutions that are not so desired but available to them.

This paper consists of the development of a methodology based on Multiple Regression Analysis (MRA) to predict the percentage of people who could not achieve their desired carrier/ academic institution and had to opt for suitable alternatives available to them.

Mentioned outcome(s) would simultaneously generate the percentage of the people who could achieve their target and thereby appeared to be successful in society.

Several samples of data were collected which have been used to carry out the above investigation. The first 66% of samples were used for analysis, the 34% samples were reserved for testing the accuracy of the analysis. Then 50% of samples were used for analysis and 50% samples were reserved for testing the accuracy of the analysis.

Keywords: MLR(Multiple Linear Regression), Target-career-goal, partially achieved, fully achieved, failed to achieve, Cross-fold validation, Confusion Matrix etc.

1. INTRODUCTION

Regression analysis is performed to determine the correlations between two or can be more than two variables, and to make predictions for the topic by using the relation.[4] It is widely used method of statistical technique for modeling the relationship between variables, its has numerous applications in every field as engineering, chemical science, etc[5].

Since childhood students dream about their future what they want to become to be successful in life. Focused students decide their goals at the very beginning of their student life. Most of the students have a dream institute. Not all of them can fulfil their dreams, even many hardworking students fail to get admission to their dream institution or study in their field of interest. There could be many reasons behind this, it can be a family problem, financial problem, any accident, not getting appropriate marks, etc. Not only these, **COVID-19** [13] has played a huge role in student's life. Many people have lost their job, some people face a huge loss in business which have great affect in their family including their children career. Financial crisis leads the students to not have proper facilities to study such as a good coaching, good teacher, good institute and so on. Due to Covid-situation total education system was on online mode, that too have negative impact on some number of students as they were not used to with this. In this pandemic situation many entrance exams were not took place students career has also a negative effect of it.

Many students did achieve their target but not fully, like someone's target was to study B. Tech in IIT but that person didn't get admission there, so opted for any other B. Tech institute, here that person is able to partially achieve his target.

Regression models produces results which are not biased for each variable of interest if the model is specified correctly and all potential confounding factors are included and correctly measured[6,7] . The main purpose of regression is to examine if the independent variables are successful in predicting the resulting variable.

This paper tries to predict whether a student has fully achieved his/her target career goal or not. If not, then how many students have achieved their target career goal partially.

The Multiple Linear Regression (MLR) model is used for this research. In statistics, MLR is a methodology that uses many variables to predict the outcome dependent variable. In MLR one dependent variable and several independent variables are taken, to know their relation and then accuracy is found out.

The actual objective of this project is to illustrate how multiple regression models may be fitted to our dataset.[1]

10-cross-fold validation is also done in this paper. Data are folded in 10 folds. Then MLR is applied for every fold to find out the accuracy.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11338

2.LITERATURE REVIEW

Multiple linear regression is a very popular Model in machine learning. It describes the relation between dependent variables and independent variables. Using this method, the value of dependent variables. It is used for this paper as well. Before working on this model some research works are taken as the reference.

A research work published in International Journal of Mathematics And its Application, [16] P.Ramesh Reddy1∗ and Dr.K.L.A.P.Sarma2's, "A Multiple Linear Regression Approach for the Analysis of Stress Factors of Faculty in Higher Educational Institutions".In that research work authors have focused on one's stress factors in teaching field. The objective of that paper was to address the effects of different inducing stress factors on job stress of the faculty in higher educational institutions. A total of 500 faculty members from both technical (250) and technical (250) institutions took part in the study [16]. This paper focused on the job stress, faculties have to face during their job life.

[17] A new multiple regression model for predictions of urban water use S. Alireza Eslamian, S. Samuel Li ∗, Fariborz Haghighat Department of Building, Civil and Environmental Engineering, Concordia University Montreal, QC H3G 1M8, Canada. This paper is about shortages of freshwater in many region across the globe. They have collected the data and used MLR to predict the daily use of water.

[18] Stepwise Multiple Regression Method to Forecast Fish Landing Intan Martina Md Ghania,*, Sabri Ahmadb a,bDepartment of Mathematics, Faculty of Science and Technology, Universiti Malaysia Terengganu, Malaysia . They have also used Multipile Linear Regression to select the suitable controlled variables in forecast fish landing. Data have been collected from Fisheries Annually Statistics of Department of Fisheries Malaysia. The response variable is marine fish landing that is Y while fishermen that is X1 , fishing boat that is X2 and fishing gears licensed that is X3 were controlled variables.

All of these papers have used Multiple linear regression .That helps us to implement MLR in our project.

3.METHODOLOGY

3.1 Dataset

For this paper, data were collected from a group of students from different institutions and different streams by using the online questionnaire method. The dataset for this study is extracted to predict whether a student has fully achieved his/her target career goal or failed to achieve based on the following attribute information (in the following Table 1)

Table 1. Description of the Dataset

Sl. No.	Feature	Description	Range of values
1.	Family Income (per year)		<1,00,000; 100,000 - 200,000; 200,000 - 300,000; 300,000 - 400,000; 400,000 - 500,000; 500,000 - 600,000; 600,000 - 700,000; 700,000 - 800,000; 800,000 - 900,000; 900,000 - 10,00,000; >10,00,000
2.	Reason for not being able to fully achieve their target career goal		Financial problem, due to pandemic, due to accident, couldn't score well in the exam, personal problem, family problem
3.	Time Spent with parents		1-4 hrs., 4-8 hrs., 8-12 hrs., 12-16 hrs., 16-20 hrs., 20-24 hrs.
4.	Achievement of target career goal		Fully achieved, Partially Achieved, Fully Achieved

String Value	Converted numerical value
1.Fully achieved	1.0
2.Partialy achieved	0.6
3.Failed to Achieve	0.3

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11338

4. Financial problem	0.5
5.Due to pandemic	0.8
6.Due to some accident	0.6
7.Couldn't score well in the exam	0.9
8. Personal problem	0.4
9. Family problem	0.7
10. 1-4 hrs	2.5
11. 4-8 hrs	6
12. 8-12 hrs	10
13. 12-16 hrs	14
14. 16-20 hrs	18
15. 20-24 hrs	22
16. <100,000	90000
17. 100,000-200,000	150000
18.200,000-300,000	250000
19. 300,000-400,000	350000
20. 400,000-500,000	450000
21.500,000-600,000	550000
22. 600,000-700,000	650000
23.700,000-800,000	750000
24.800,000-900,000	850000
25.900,000-10,00,000	950000
26.>10,00,000	150000000

	Family Income	Time spent with parents
Mean	379516.13	8.383
Standard Deviation	1338892.088	5.834

3.2 Research Method

The processing of the data can use statistical techniques, mathematics, machine learning and artificial intelligence to dig up information that can be used[8,9].

For this research Multiple Linear Regression (MLR) model is used. Regression models are used to describe the relationships between variables, it allows to estimate if independent variable(s) changes then dependent variable also changes.

In MLR, there is one dependent variable and two or more independent variables.

In this research work, three independent variables (x_0, x_1, x_2) are considered to predict one dependent variable(y_i).

Where,

y_i = "how much have you achieved your target career goal?"

x_0 = "why your target career is not fully achieved?"

x_1 = "family income"

x_2 = "time spent with parents"

The equation for Multiple Linear Regression (MLR):

$y_i = = a + b_0x_0 + b_1x_1 + b_2x_2$

Where,

y = dependent variable

a = intercept

b_0 = coefficient of first independent variable x_0.

b_1 = coefficient of second independent variable x_1.

b_2 = coefficient of third independent variable x_2.

x_0 = First independent variable ("why your target career goal is not achieved?")

x_1 = Second independent variable ("family income (per year)"

x_2 = Third independent variable ("spent time with both of the parents together (in student life)".

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11338

Cross-Validation

Cross validation is a technique used to assess and evaluate the performance of machine learning algorithms. This can be done by partitioning a dataset into two subsets one for training dataset and the other for testing dataset. In every iteration of cross validation randomly partition of the given original data set is done. In the study, 10 cross fold validation is used, where the dataset has been partitioned into 10 equal partitions and each is used for validation and the remaining training dataset is used for finding out the accuracy.

Confusion Matrix

In machine learning, a confusion matrix is a methodology that uses a table for describing the performance of a model. The evaluation can be done in terms of correctness by computing statistical measures which are the True Positives (TP), True Negatives (TN), False Positive (FP) and False Negatives (FN).[2] It contains information about actual and predicted classification done by a model. Performance of such models is commonly evaluated using data in the matrix.[11]

TABLE:

	Predicted: No	Predicted: Yes
Actual: No	TN	FP
Actual: Yes	FN	TP

The target variable has Yes and No value. As per the Table the columns represent the predicted value and the rows represent the actual value:

True Positive (TP): In true positive, predicted value matches the actual value.
The actual value was yes and the predicted value was also yes.
True Negative (TN): In true negative, actual value matches the predicted value.
The actual value was No, and the predicted value was also No.
False Positive (FP): In false positive, the actual value doesn't match the predicted value.
The actual value was No but predicted value was Yes.
False Negative (FN): In false negative, the actual value does not match the predicted value.
The actual value was Yes but the predicted value was No.

4.ACCURACY

1.To find the accuracy first the Multiple regression is used.
2.Data are cross-folded to find out the accuracy.
3.To find the accuracy of this prediction, the predicted y values are compared with the actual y values and checked how many of them are approximately equal. Predicted y values which are approximately equal to the actual values are counted in a variable.
Accuracy (%) = (matched y values/total number of y values) *100
For the binary classification, evaluation metrics include accuracy, sensitivity and specificity. The metrics are defined as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Specificity} = \frac{TN}{TN+FP}$$

Where TP=True positive: Student achieved target-career-goal correctly classified, TN= True negative: Students failed to achieve correctly classified, FP= False positive: Students achieved target-career-goal incorrectly classified, FN= False negative: Students failed to achieve incorrectly classified

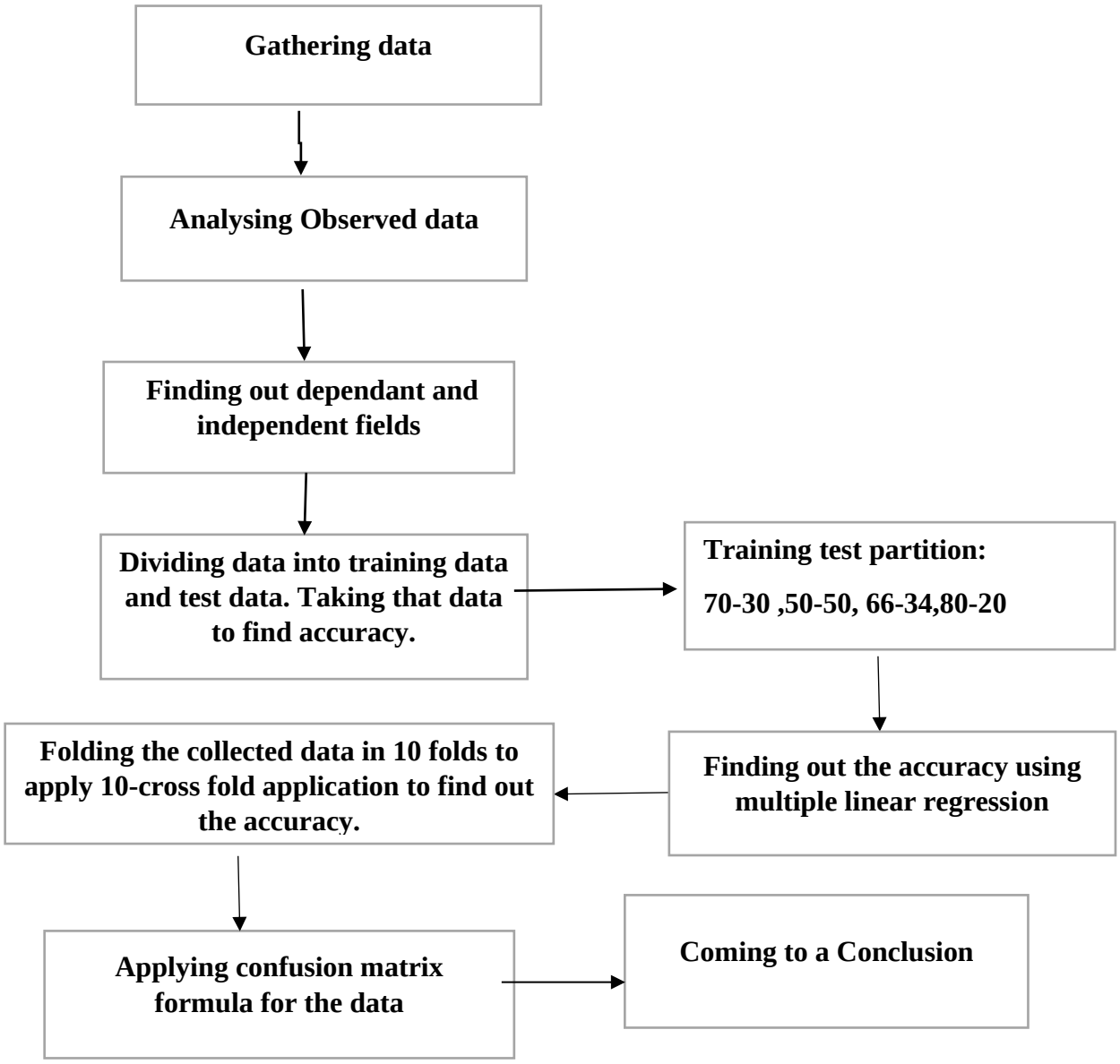
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11338

5.FLOWCHART OF THIS RESEARCH:



6.RESULT

For this research python programming language is used. In this paper, MLR model is used, for finding accuracy, the dataset is divided into 70-30, 50-50,66-34, 80-30 train-test data, and 10-fold cross-validation is also performed. The result shows that k-fold Cross Validation performed better than the random validation of 70-30, 50-50, 66-34, 80-20 train-test size[12].

The accuracy of each train-test split is mentioned in the following Table.

Training – Testing Partition	Accuracy
70-30	76.31%
50-50	74%
66-34	74.41%
80-20	68%

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11338

10-fold cross-validation	Accuracy
1st fold	92.30%
2nd fold	76.92%
3rd fold	84.92%
4th fold	92.61%
5th fold	91.66%
6th fold	91.33%
7th fold	90%
8th fold	86%
9th fold	91.66%
10th fold	90%

Confusion Matrix:

Train-test partition	Accuracy	Sensitivity	Specificity
70-30	84%	96%	0.0
80-20	74%	87%	0.0
66-34	86%	95%	0.0
50-50	62%	67%	0.0

10-fold cross-validation	Accuracy	Sensitivity	Specificity
1st fold	100%	100%	0.0
2nd fold	81%	90%	0.0
3rd fold	83%	91%	0.0
4th fold	91%	95%	0.0
5th fold	83%	91%	0.0
6th fold	85%	85%	0.0
7th fold	75%	95%	0.0
8th fold	75%	85%	0.0
9th fold	81%	81%	0.0
10th fold	60%	83%	0.0

In the confusion matrix the specificity is 0% for each and every case. As in this paper among 300 students the number of students who have failed to achieve their target was very minimum. But in this prediction the number of students who have failed to achieve their target career is zero. That's why the value of TN is also zero. As the value of TN is zero the specificity is zero.

7.CONCLUSION

Our proposed MLR model is able to predict students partially achieving their target career goal. Accuracy is determined by counting the predicted dependent values matching with the actual dependent values. The results were obtained by considering 3 features from the dataset, which are family income, the reason for not achieving the target goal fully, time spent with parents. All the features were considered which can affect the dependent variable, the dependent variable in our research is how much target career goal is achieved, whether it's fully achieved, partially achieved or failed to achieve.

8.FUTURE SCOPE

For this research work only one model is used. Other Machine Learning Method can be used for increasing the accuracy.

9.REFERENCES:

[1] J. C. DeFries 1 and D. W. Fulker 1's "Multiple Regression Analysis of Twin Data".
[2] Performance Analysis of Text Classification Algorithms using Confusion Matrix Maria Navin J R, Pankaja R.

DOI: 10.17148/IJARCCE.2022.11338

[3] "An Enhanced Random Forest Model for Detecting Effects on Organs after Recovering from Dengue" Avijit Kumar Chaudhuri1 , Arkadip Ray2 , Prof. Dilip K. Banerjee3 , Dr. Anirban Das4.

[4] Uyanik[i], Nese Gulerii's "A study on Multiple Linear Regression Analysis Gulden Kaya"

[5] International Journal of Instrumentation and Control Systems (IJICS) Vol.8, No.2, April 2018 DOI : 10.5121/ijics.2018.8201 1 MULTIPLE LINEAR REGRESSION ANALYSIS FOR PREDICTION OF BOILER LOSSES AND BOILER EFFICIENCY Chayalakshmi C.L1 , D.S. Jangamshetti2 and Savita Sonoli3.

[6] Liu.,Zhiyong's "Quality Reporting of Multivariable Regression Models in Observational Studies Review of a Representative Sample of Articles" Published in Biomedical Journals

[7] Gerhard T. Bias: considerations for research practice. Am J Health Syst Pharm 2008; 65:2159–2168.

[8] Bintang Dewi Fajar Kurniatullah *1 , and Yuventius Tyas Catur Pramudi2's "Estimation of Students' Graduation Using Multiple Linear Regression Method" Journal of Applied Intelligent System, Vol. 2 No. , April 2017.

[9] D.H Kamagi dan S. Hansun, "Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan mahasiswa," Jurnal ULTIMATICS, Vol. VI, No. 1. 2014.

[10] Yunus Koloğlu, Hasan Birinci, Sevde Ilgaz Kanalmaz, Burhan Özyılmaz's "A Multiple Linear Regression Approach For Estimating the Market Value of Football Players in Forward Position".

[11] A. K. Santra, C. Josephine Christy's "Genetic Algorithm and Confusion Matrix for Document Clustering", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, January 2012

[12] Sitefanus Hulu1 , Poltak Sihombing2 , Sutarman2's "Analysis of Performance Cross Validation Method and K-Nearest Neighbor in Classification Data", International Journal of Research and Review Vol.7; Issue: 4; April 2020

[13] P. Arumugam, V. Kadhirveni, R. Lakshmi Priya, Manimannan G's "Prediction, Cross Validation and Classification in the Presence COVID-19 of Indian States and Union Territories using Machine Learning Algorithms", International Journal of Recent Technology and Engineering (IJRTE)

[14] M. U. Sarwar, M. K. Hanif, R. Talib, A. Mobeen, and M. Aslam, "A survey of big data analytics in healthcare," Int. J. Adv. Comput. Sci. Appl., vol. 8, pp. 355-359, 2017

[15] E. W. Steyerberg, Clinical prediction models. Cham: Springer International Publishing, 2019, pp. 297-308.

[16] P.Ramesh Reddy1∗ and Dr.K.L.A.P.Sarma2's, "A Multiple Linear Regression Approach for the Analysis of Stress Factors of Faculty in Higher Educational Institutions"

[17] A new multiple regression model for predictions of urban water use S. Alireza Eslamian, S. Samuel Li ∗, Fariborz Haghighat Department of Building, Civil and Environmental Engineering, Concordia University Montreal, QC H3G 1M8, Canada.

[18] Stepwise Multiple Regression Method to Forecast Fish Landing Intan Martina Md Ghania,*, Sabri Ahmadb a,bDepartment of Mathematics, Faculty of Science and Technology, Universiti Malaysia Terengganu, Malaysia