Volume 9, Issue 06, June 2024 ISSN (online): 2456-0006

International Journal of Science Technology Management and Research

Available online at: www.ijstmr.com

Predicting Probability of the Target Variable Using Sigmoid Function in Logistic Regression

Trapti Vishwakarma
M Tech (CSE) 4th semester
Computer Science and Engineering Department
L.N.C.T. (Bhopal) Indore Campus Indore M.P.

Nilesh Avinash Joshi
Assistant Professor
Computer Science and Engineering Department
L.N.C.T. (Bhopal) Indore Campus Indore M.P.

Abstract: ***In the proposed work we used linear regression to find the best fit line which aims at minimizing the distance between the predicted value and actual value, the line. We used students data set form a competition coaching class, which contains two variable hours study by the student and result (pass/fail). Number of hours for study is independent variable and result is dependent variables. We first calculate value of m (slop) and intercept (b). We calculate the Logit value for each independent variable using intercept and slop. Next we calculate EXP of all values logit value and we also calculate values of odds. Finally we calculate probability by using the intercept and slop value. Our objective is predict the student fail or pass by using the number of hours he/she studied Logistic Regression, which converts this straight best fit line in linear regression to an S-curve using the sigmoid function, which will always give values between 0 and 1.By the experimental analysis we predict the value of probability for unknown value of number of study hours***

Keywords: Regression, Linear, Logistic, Hours, Study, Probability

I. INTRODUCTION

Regression analysis is a way of predicting future happenings between a dependent (target) and one or more independent variables (also known as a predictor). For example, it can be used to predict the relationship between reckless driving and the total number of road accidents caused by a driver, or, to use a business example, the effect on sales and spending a certain amount of money on advertising. Regression is one of the most common models of machine learning. It differs from classification models because it estimates a numerical value, whereas classification models identify which category an observation belongs to[1,2,3].

The main uses of regression analysis are forecasting, time series modeling and finding the cause-and-effect relationship between variables. Regression has a wide range of real-life applications. It is essential for any machine learning problem that involves continuous numbers – this includes, but is not limited to, a host of examples, including:

- Financial forecasting (like house price estimates, or stock prices)
- Sales and promotions forecasting
- Testing automobiles
- Weather analysis and prediction
- Time series forecasting

As well as telling you whether a significant relationship exists between two or more variables, regression analysis can give specific details about that relationship. Specifically, it can estimate the strength of impact that multiple variables will have on a dependent variable. If you change the value of one variable (price, say), regression analysis should tell you what effect that will have on the dependent variable (sales).

Businesses can use regression analysis to test the effects of variables as measured on different scales. With it in your toolbox, you can assess the best set of variables to use when building predictive models, greatly increasing the accuracy of your forecasting [8].

Finally, regression analysis is the best way of solving regression problems in machine learning using data modeling. By plotting data points on a chart and running the best fit line through them, we can predict each data point's likelihood of error: the further away from the line they lie, the higher their error of prediction (this best fit line is also known as a regression line).

Regression analysis is a predictive modeling technique that analyzes the relation between the target or dependent variable and independent variable in a dataset. The different types of regression analysis techniques get used when the target and independent variables show a linear or non-linear relationship between each other, and the target variable contains continuous values. The regression technique gets used mainly to determine the predictor strength, forecast trend, time series, and in case of cause & effect relation. Regression analysis is the primary technique to solve the regression problems in machine learning using data modeling. It involves determining the best fit line, which is a line that passes through all the data points in such a way that distance of the line from each data point is minimized

, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

II. Classification of Linear Regression

There are many types of regression analysis techniques, and the use of each method depends upon the number of factors. These factors include the type of target variable, shape of the regression line, and the number of independent variables [5,6,7].

Below are the different regression techniques:

1. Linear Regression
2. Logistic Regression
3. Ridge Regression
4. Lasso Regression
5. Polynomial Regression
6. Bayesian Linear Regression

1. Linear Regression

Linear regression is one of the most basic types of regression in machine learning. The linear regression model consists of a predictor variable and a dependent variable related linearly to each other. The below-given equation is used to denote the linear regression model:

$$y = mx + c + e$$

Where m is the slope of the line, c is an intercept, and e represents the error in the model.

The best fit line is determined by varying the values of m and c. The predictor error is the difference between the observed values and the predicted value. The values of m and c get selected in such a way that it gives the minimum predictor error.

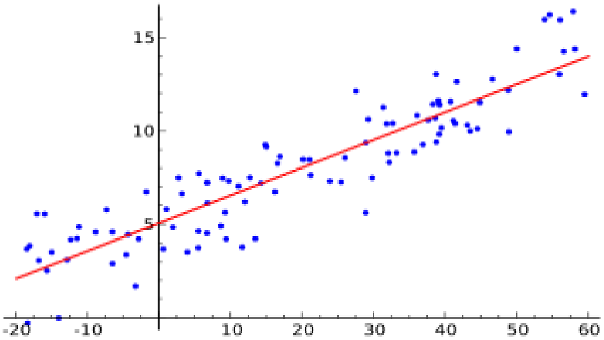


Figure 1 Linear regression

2. Logistic Regression

Logistic regression is one of the types of regression analysis technique, which gets used when the dependent variable is discrete. Example: 0 or 1, true or false, etc. Logistic regression works best with large data sets that have an almost equal occurrence of values in target variables. The dataset should not contain a high correlation between independent variables.

Logit function is used in Logistic Regression to measure the relationship between the target variable and independent variables. Below is the equation that denotes the logistic regression.

$$logit(p) = \ln\left(\frac{p}{(1-p)}\right) = b_1 + b_1X_1 + b_2X_2 \dots b_kX_k$$

Where p is the probability of occurrence of the feature.

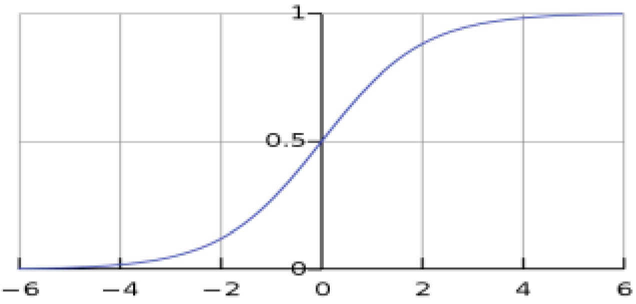


Figure 2 Logistic Regression

3. Ridge Regression

This is another one of the types of regression in machine learning which is usually used when there is a high correlation between the independent variables. It is known as a regularization technique and is used to reduce the complexity of the model. It introduces a small amount of bias known as the ‘ridge regression penalty’ which, using a bias matrix makes the model less susceptible to over fitting.

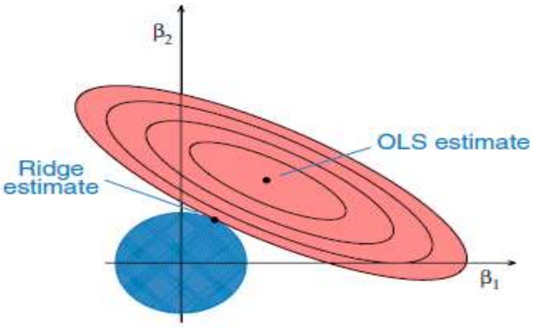


Figure 3 Ridge Regression

4. Lasso Regression

Lasso Regression is one of the types of regression in machine learning that performs regularization along with feature selection. It prohibits the absolute size of the regression coefficient. Due to this, feature selection gets used in Lasso Regression, which allows selecting a set of features from the dataset to build the model. In the case of Lasso Regression, only the required features are used, and the other ones are made zero. This helps in avoiding the over fitting in the model. In case the independent variables are highly collinear, then Lasso regression picks only one variable and makes other variables to shrink to zero.

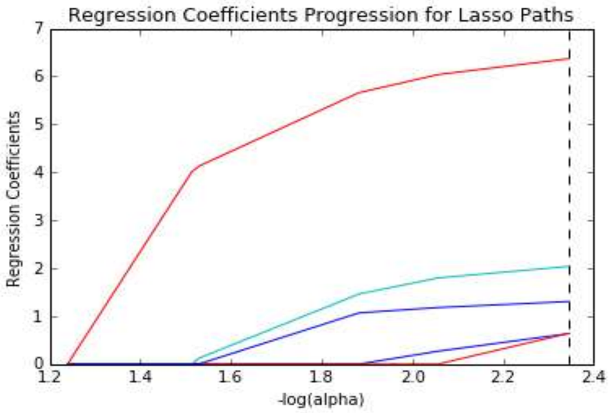
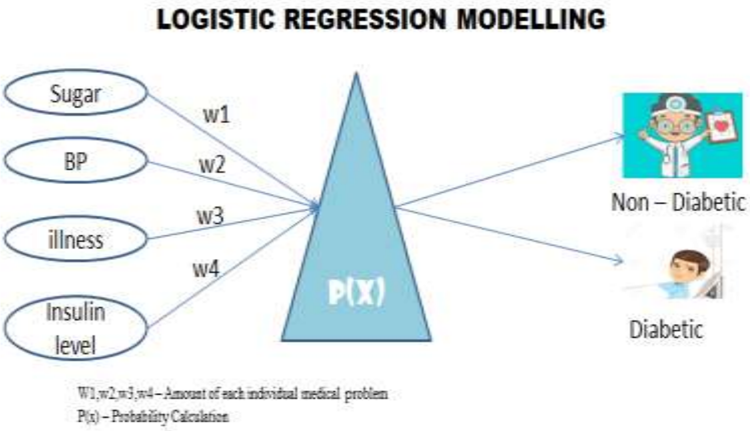


Figure 4 Lasso Regression

III. Logistic regression

Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a data set. Logistic regression has become an important tool in the discipline of machine learning. The approach allows an algorithm being used in a machine learning application to classify incoming data based on historical data. Logistic regression can also play a role in data preparation activities by allowing data sets to be put into specifically predefined buckets during the extract, transform, load (ETL) process in order to stage the information for analysis. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. For example, a logistic regression could be used to predict whether a political candidate will win or lose an election or whether a high school student will be admitted to a particular college.

An e-commerce company that mails expensive promotional offers to customers would like to know whether a particular customer is likely to respond to the offers or not. For example, they'll want to know whether that consumer will be a "responder" or a "non-responder." In marketing, this is called *propensity to respond modeling*. A credit card company develops a model to decide whether to issue a credit card to a customer or not will try to predict whether the customer is going to default or not on the credit card based on such characteristics as annual income, monthly credit card payments and number of defaults[9].



W1,w2,w3,w4 – Amount of each individual medical problem
P(x) – Probability Calculation

Figure 6 Binary Logistic Regression

IV. LITERATURE SURVEY

In 2017 Elena Bucur et al proposed "**Binary logistic regression—Instrument for assessing museum indoor air impact on exhibits**". They presented a new way to assess the environmental impact on historical artifacts using binary logistic regression. The prediction of the impact on the exhibits during certain pollution scenarios (environmental impact) was calculated by a mathematical model based on the binary logistic regression; it allows the identification of those environmental parameters from a multitude of possible parameters with a significant impact on exhibitions and ranks them according to their severity effect. Air quality (NO2, SO2, O3 and PM2.5) and microclimate parameters (temperature, humidity) monitoring data from a case study conducted within exhibition and storage spaces of the Romanian National Aviation Museum Bucharest have been used for developing and validating the binary logistic regression method and the mathematical model [4].

In 2018 Choney Zangmo et al proposed **"Application of logistic regression models to cancer patients: a case study of data from Jigme Dorji Wangchuck National Referral Hospital (JDWNRH) in Bhutan".** Cancer is an uncontrolled disease caused by a damage to the cell's DNA. Cancer rates are increasing every year and is the biggest concern in many countries including Bhutan. Bhutan being a very small country loses around 100 to 200 patients every year due to cancer. They involved 3013 cancer patients who are treated at National Referral Hospital (JDWNRH) from 2010 to 2016. Those patients who cannot be treated in the country are referred out to other countries for treatment. They identify the factors that affect the survival of all cancer patients as well as male and female patients separately. The best fitted model was obtained from the analysis of deviance. The test showed that the last status of the patient and the variables of personal and clinical data are mostly significant at p-value less than0.05. In the binary logistic regression, for all cancer patients, the factors which effect the patient's last status are age, length of stay and cancer site[8].

In 2019 Miguel Saavedra et al proposed **"A new approach to study the relative age effect with the use of additive logistic regression models: A case of study of FIFAfootball tournaments".** They proposed a new approach is proposed to study the relative age effect with the use of a logistic regression additive model. The new method has been evaluated with a sample of 21,639 players involved in football tournaments organized by the Fe´de´ration International de Football Association (FIFA) between 1908 and 2012. They established that the relative age effect exists regarding player age and the year of the competition in male FIFA competitions and its effect is dynamic and complex. They proposed RAE with the use of additive logistic regression models has been proposed. The new method allows analysis of the RAE in the presence of covariates and model nonlinear relations between variables. RAE and additive logistic regression model FIFA football tournaments PLOS ONE. [10].

In 2020 Anna Borucka et al proposed **"Logistic regression in modeling and assessment of transport services"**. They proposed logistic regression and was conducted on the basis of a distribution and trade company dealing with the supply of automotive spare parts. As the most profitable group of customers is local car repair shops, it was this group that was subject to analysis. The research showed which of them (and how strongly) affect the dependent variable, which allowed for modification of strategy and implementation of new solutions increasing the number of satisfied customers. They indicated the possibility of mathematical assessment of selected elements of the company's activity influencing the quality of services provided and basing thereof to propose changes in shaping the company's strategy in the area of distribution. They presented the model allows for their broader analysis and inference showing the impact of individual predictors on the analyzed variable late delivery in the case in question [11].

In 2020 Antonio Alveset et al proposed **"The logic of logistic regression"**. They provide an intuitive introduction to logistic regression, the most appropriate statistical technique to deal with dichotomous dependent variables. They estimate the effect of corruption scandals on the chance of reelection of candidates running for the Brazilian Chamber of Deputies using data from Castro and Nunes (2014). Specifically, we show the computational implementation in R and we explain the substantive interpretation of the results. They share replication materials which quickly enables students and professionals to use the procedures presented here for their studying and research activities. They facilitate the use of logistic regression and tops read replication as a data analysis teaching tool. The absence of calculus, linear and matrix algebra, and advanced statistics limits our ability to understand more advanced data analysis techniques [12].

In 2021 Annwesha Banerjee et al proposed **"An Intelligent System for Prediction of COVID-19 Case using Machine Learning Framework-Logistic Regression".** Many scientists and medical practitioners are working hard to fight against this, in search of proper medicine and vaccine. Research is also going on in the field of machine learning and AI to predict the spread of disease and also in identification of the presence of the virus inhuman body, which will help the field of medical science. They proposed a method to identify whether a patient has risk of COVID-19 using Logistic Regression model, considering multiple symptoms. In this paper a logistic regression based model is being proposed which has considered pneumonia, diabetes, chronic obstructive pulmonary disease, asthma, hypertension, cardiovascular disease, renal disease, obesity, tobacco taking habit, and contact with other covid-19 positive one as the independent variable for covid-19 classification achieving 92% accuracy. In future we have planned for using deep learning mechanism for COVID-19 prediction [13].

In 2021 Lijalem Melie et al proposed **"Multivariate logistic regression analysis on the association between anthropometric indicators of under-five children in Nigeria: NDHS 2018".**Their aim is to assess the association between anthropometric indicators of under-five children such as stunting, underweight and wasting given that of other characteristics of children and households. The data for this study was obtained from Nigerian Demographic and health survey (NDHS) in 2018. A total of 11,314 under-five children were involved. Multivariate logistic regression model was used to determine the association between stunting, underweight and wasting given that of the estimated effect of other determinants. About half (50.7%) of the children were male, 24.1% was obtained from North West region of Nigeria, and 37.8% of them were from households having unimproved drinking water. The prevalence of under-five children with stunting, underweight and/or wasting in Nigeria was very high. The important determinants of stunting, underweight, and wasting for fewer than five children were household wealth index, women body mass index, sex of the child, anemia, mothers' age at first birth, and a diarrhea two weeks prior to the survey [14].

V. PRAPOSED APPROACH

While logistic regression seems like a fairly simple algorithm to adopt & implement, there are a lot of restrictions around its use. For instance, it can only be applied to large datasets. The dependent variable has to be binary in a binary logistic equation

- The factor level 1 of the dependent variable should represent the desired outcome
- Including non-meaningful variables may throw errors. Only include the variables that are necessary and may show a correlation
- The model should have little or no multicollinearity the independent variables should be absolutely independent of each other
- The independent variables are linearly related to the log odds

Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a data set. Logistic regression has become an important tool in the discipline of machine learning. The approach allows an algorithm being used in a machine learning application to classify incoming data based on historical data. With so many assumptions that need to be made, you may think that the equation is not versatile enough to be implemented across real-life problems but this equation has a lot of applications in the medical field and is helping people across the world with its superpower.

VI. COMPARATIVE ANALYSIS

Prediction of passing based number of hours study

Based on logistic regression model we can predict passing or chance of selection on unknown value of number of hours study. From the table we can see that for 6.0 hours of study passing chance is 0.77 percentage, for 7.5 hours of study passing chance is 0.8327 percentage, for 7.5 hours of study passing chance is 0.8484 percentage,

Table 6.3 prediction for unknown value of hours and chance of passing.

Hours	Probability of passing
6.0	0.7778
7.5	0.8327
8.0	0.8484

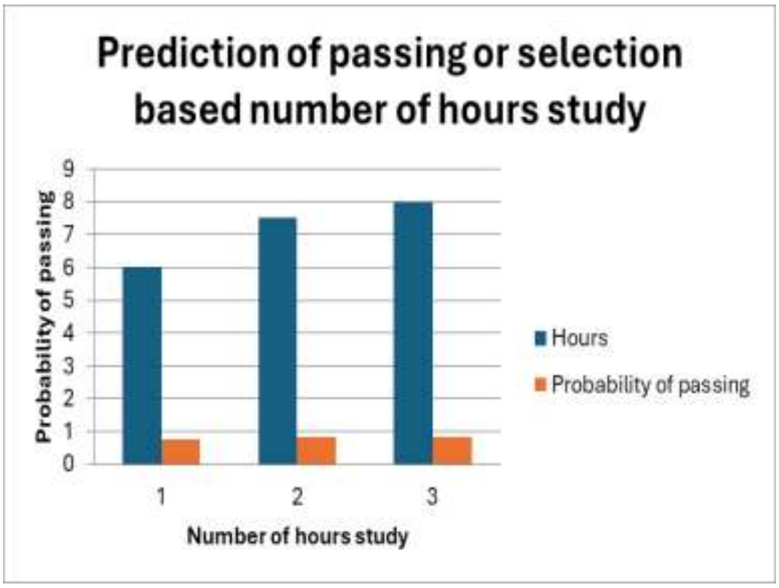


Figure 7 Predictions for unknown value of hours and chance of passing

VII. CONCLUSION

In this paper we used linear regression to find the best fit line which aims at minimizing the distance between the predicted value and actual value, the line. We used students data set form a competition coaching class, which contains two variable hours study by the student and result (pass/fail). Number of hours for study is independent variable and result is dependent variables. We first calculate value of m (slop) and intercept (b). We calculate the Logit value for each independent variable using intercept and slop. Next we calculate EXP of all values logit value and we also calculate values of odds. Finally we calculate probability by using the intercept and slop value. Our objective is predict the student fail or pass by using the number of hours he/she studied Logistic Regression, which converts this straight best fit line in linear regression to an S-curve using the sigmoid function, which will always give values between 0 and 1.By the experimental analysis we predict the value of probability for unknown value of number of study hours.

REFERENCE

1. Francis Bach Self-concordant analysis for logistic regression Electronic Journal of Statistics Vol. 4 (2010) 384–414ISSN: 1935-7524 DOI: 10.1214/09-EJS521
2. Jill C. Stoltzfus, PhD " Logistic Regression: A Brief Primer From the Research Institute, St. Luke's Hospital and Health Network", Bethlehem, PA. Received January 22, 2011; revisions received April 6 and May 9, 2011; accepted May 9, 2011.
3. Abdalla M. EL-HABIL An Application on Multinomial Logistic Regression Model Head of the Department of Applied Statistics Faculty of Economics and Administrative Sciences Al-Azhar University, Gaza – Palestine Pak.j.stat.oper.res. Vol. VIII No.2 2012 pp271-291.
4. Elena Myftaraj (Tomori) , Eglantina Zyka Ruzhdie Bici Identifying Household Level Determinants Of Poverty In Albania Using Logistic Regression Model Myftaraj et al. OIDA International Journal of Sustainable Development 07:03 (2014) Ontario International Development Agency. ISSN 1923-6654 (print) ISSN 1923-6662 (online).
5. Branimir MilosavljeviT, and Predrag DašiT Binary Logistic Regression Modeling of Idle CO Emissions in Order to Estimate Predictors Influences in Old Vehicle Park Hindawi Publishing Corporation Mathematical Problems in Engineering Volume 2015, Article ID 463158.
6. Joseph Adwere Boamah "Predicting social trust with binary logistic regression" Copyright statement: Authors retain the copyright to the manuscripts published in AABRI Research in Higher Education Journal Volume 27 - January, 2015 journals.
7. Miftar Ramosacaj1 Application of Logistic Regression in the Study of Students' Performance Level (Case Study of Vlora University) ISSN 2239-978X ISSN 2240-0524 Journal of Educational and Social Research MCSER Publishing, Rome-Italy Vol. 5 No.3 September 2015
8. Choney Zangmo and Montip Tiensuwan Application of logistic regression models to cancer patients: a case study of data from National Referral Hospital (JDWNRH) in Bhutan ICAPM 2018 IOP Publishing IOP Conf. Series: Journal of Physics: Conf. Series 1039 (2018) 012031 doi :10.1088/1742-6596/1039/1/012031.
9. Lian Niu "A Review of the Adoption of Logistic Regression in Educational Research: Common Issues, Implications, and Suggestions" Educational Review. Advance online publication. doi: 10.1080/00131911.2018.1483892
10. Miguel Saavedra-Garcı´aID , Marcos Matabuena, A new approach to study the relative age effect with the use of additive logistic regression models: A case of study of FIFA football tournaments (1908-2012) https://doi.org/10.1371/journal.pone.0219757 July 16, 2019 RAE and additive logistic regression model: FIFA football tournaments.
11. Anna Borucka Logistic regression in modeling and assessment of transport services Open Eng. 2020; 10:26–34 © 2020 A. Borucka, published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0License.
12. Antonio Alves Torres Fernandes "The logic of logistic regression" Received in October 19, 2019. Approved in May 7, de 2020. Accepted in May 16, 2020. Rev. Sociol. Polit., v. 28, n. 74, e006, 2020.
13. Annwesha Banerjee Majumder, An Intelligent System for Prediction of COVID-19 Case using Machine Learning Framework-Logistic Regression IOCER 2020 Journal of Physics: Conference Series 1797 (2021) 012011 IOP Publishing doi:10.1088/1742-6596/1797/1/012011.
14. Lijalem Melie Tesfaw and Haile Mekonnen Fenta Multivariate logistic regression analysis on the association between anthropometric indicators of under-five children in Nigeria: NDHS Tesfaw and Fenta BMC Pediatrics (2021) 21:193https://doi.org/10.1186/s12887-021-02657-5 Tesfaw and Fenta BMC Pediatrics (2021) 21:193.