Logistic Regression Analysis on Social Networking Advertisement

Divya Sai Patnana, Gone Hitesh, IpsitaSahu. N. Suresh Kumar

Department of Computer Science and Engineering, GITAM Institute of Technology, Visakhapatnam
nskgitam2009@gmail.com

Received: 14 Feb 2020 Revised and Accepted: 25 March 2020

ABSTRACT: In most of the Machine learning models tables and charts need to be plot for result analysis and predictions to be made. The result analysis is a fundamental need for any company reuired to analyse their objective of their product development. In the present work Logistic regression model is developed to illustrate social networking advertisement dataset. With this fundamental Logistic Regression Model implementation one can predict whether a user is willing to purchase their product or not. In the present paper Logistic Regression Model is designed in python to predict by evaluating accuracy.

KEYWORDS: Categorical Variables, Logistic Regression, Contrary Variables

I. INTRODUCTION

In the present work Logistic regression is implemented with the application Online Social Networking Advertisement (OSNA) analysis. Logistic Regression (LR) is one of the best supervised algorithms applied on data analysis. Logistic Regression is known for supervised classification algorithms for its best performance. The LR supervised classification algorithm produces only discrete output at the specified input feature set [1][2].

Many of the present techniques used to predict the output and analysis of the output whish are narrowly classified [3-5]. For instance narrow classification occurs in a student result in exams or a client thinking about purchase of a product. There are existing techniques which can also present discrete outputs like Ordinary Least Square Regression and Partial Least Square regression. But, Least Square regression classification technique is superior when compared with the above two existing classification techniques.

The present paper discussed about how the classification is done with Logistic regression and comparison results are discussed with Linear Regression. It is also illustrated about that how the generated output is analysed and assumptions are considered while implementing the Logistic Regression Analysis.

II. Nature of Logistic Regression

Based on the number of categories, Logistic regression can be classified as:

a) Binomial: The output is classified into two types. For instance whether a user willing to purchase or not willing to purchase i.e., either true or false. Based on the data set considered here, two features are used to predict for analysis of willingness to purchase a car in terms of Yes (value is ‘1’) or No (Value is ‘0’). In the present wok the age and salary are two features considered here and plays very important role in predicting the target value.

The Binomial regression is mathematically represented as shown in equation 1 for two numerical variables with scatter plot. Each point on the plot represents the information regarding each user. Here the activation function called sigmodal function is used to map the prediction values into probabilistic values.

$$S(z) = \frac{1}{1+e^{-z}} \quad \text{--------------} \ (1)$$

Where, s(z)= target value either 0 or 1

z is the input given to the activation function

The sigmodal function is also called as logistic function and can be represented as shown in figure 1. The functional operation is followed by decision boundary.

Decision Boundary: The function returns the probability values between 0 and 1. Select the threshold value so that the probability values are mapped into discrete target value. The values above the threshold values are considered as class 1 and below the threshold value are considered as class 0.For instance, the threshold if considered as 0.5 and if function returned 0.9, its classified as positive. If the function returned 0.2 that observation was classified as negative. If there exist multiple classes,then select the class that predicts with high probability.Hence, the combination of sigmodal function with decision boundary makes predictions for model evaluation.

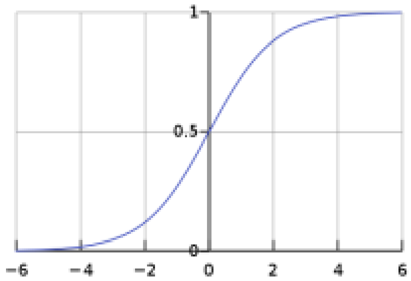


Figure 1: Sigmodal curve

b) Multinomial: Here, the target variable can have 3 or more possible types which are not ordered like “car A” vs “car B” vs “car C”. it indicates the target value will be classified as 0, 1, … n. Here the procedure of Multinomial function is same as Binomial function and it is differed in terms of classes. The procedure is shown in figure 2.

Step 1: Divide the problem into n+1 binary classification problems

Step 2: For each class repeat step 3 and 4

Step 3: Predict the probability the observations are in that single class.

Step 4: prediction = maximum (probability score of the classes)

Figure 2: Algorithm to predict the probability of observations

c) Ordinal: It is an extension model of Binomial model and used when there are multiple independent variables. The method is proposed with the models which are stored with categories. Here each category is assigned some value. For instance poor, very poor, good, and very good can be assigned with score values like 0, 1, 2, and 4respectively. Mostly, the model is used in real time applications like analysis of movie review as like, most likely, very likely, unlikely etc.

Linear Regression Vs Logistic Regression

Linear regression gives continuous number values,unbounded and measured on an interval or ratio scale.It assumes only linear relationships(which are straight-lined) between dependent and independent variables.Linear Regression is uncertain to outlierswhich are data that are at extremes which can be based on a single variable or more It works effectively only for independent data.Whereas Logistic regression gives the output using sigmoid function to return a probability estimate to map to classes.It’s used when the data is linearly classifiable and the outcome is contrary [6][7].Linear Regression understands the consumer behaviour deeply, and understands influencing factors to generate accurate results. They are also used in estimating the forecasts. It is also used in analysis of finance and advertising on sales. On the other side Logistic regression predicts the target based on the observations and the symptoms [8][9].

III. Design

The below figure 3 is the block diagram which represents our work model in a hierarchical manner for analysis of a logistic regression model. It explains the various steps required for analysis of a logistic regression model that has been demonstrated in our report. Initially it splits the data into training and test data set. The training data is trained by fit() method. And the test data is used to build the model with a logistic function which is then used to find the confusion matrix for calculating the accuracy. In this way the architecture of the model is shown in simple tree structure.

The present work is tested on dataset contains 400 users data with the attributes like age and salary. Out of which 300 datasets are used for training and remaining are used as test dataset. The Logistic Regression is used to derive the relationship between categorical output variables and where there are more categorical predictor variables. If the mean of the variables is computed with respect to the categories then the categorical output appears as linear in the middle and as curve at the end. Such a shape is called sigmoidal shape of s-shape which is shown in figure 1. The x-axis represents the input given to the sigmoid function and the y-axis which have the values ranging between 0 and 1.

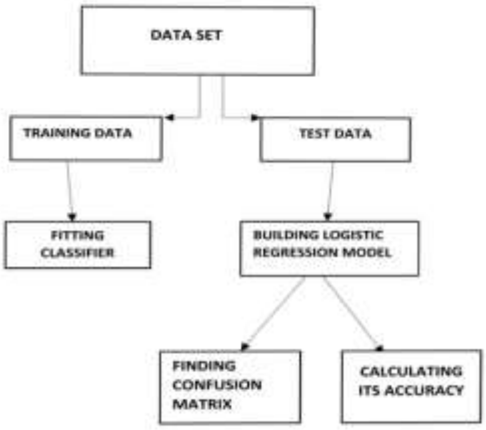


Figure 3: Block diagram of present model

It is difficult to implement with linear regression as the errors are not normally distributed and they do not follow linear trend. The problem can be solved by Logistic regression. Logistic regressionsolves these problems by transforming logistic function to logit function [9][10].The logit is the natural logarithm (ln) of odds of *Y*, and odds are ratios of probabilities (x) of *Y* happening (i.e., a user buys a brand new SUV car) to probabilities (1 – x) of *Y* not happening.

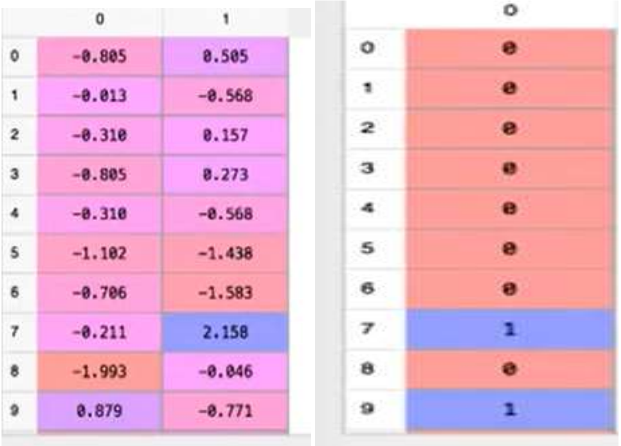
A two-predictor logistic model was fitted to the data to test the research hypothesis regarding the relationship between the likelihood that a user buys a car based on his or her age and salary.

IV. Results

In figure 4 showing 10 users result analysis. The the first two columns of figure 4 represents the x_test which consists of age and gender used to predict if the user buys a car or not. And the last column represents the y_pred which consists if a user buys a car or not I.e. dichotomous variable.

In figure 5 depeicts the observations made after executing Lostic Regression model on the database. Each point in the figure 5 represents the users with age and estimated salary. The x-axis represents the age and y-axis represents the estimated salary.Most of the users who are older and with a higher estimated salary bought the car and only few of the users who are younger and with a higher estimated salary bought the car.The goal is to classify the right users into right categories.

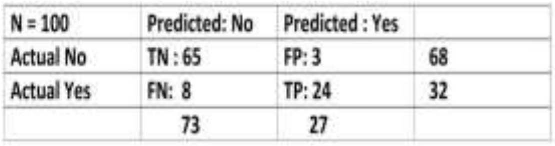
The model is used to predict if each user has a value 0 or 1.The red colour point represents value 0 and the green colour point represents 1 where 0 means buying a car and 1 means not buying a car.As the logistic regression is a linear classifier,it is a straight line with two dimensions.



	0	1
0	-0.805	0.505
1	-0.013	-0.568
2	-0.310	0.157
3	-0.805	0.273
4	-0.310	-0.568
5	-1.102	-1.438
6	-0.706	-1.583
7	-0.211	2.158
8	-1.993	-0.046
9	0.879	-0.771

	0
0	0
1	0
2	0
3	0
4	0
5	0
6	0
7	1
8	0
9	1

Figure 4: Test values and Target value



N = 100	Predicted: No	Predicted : Yes	
Actual No	TN : 65	FP: 3	68
Actual Yes	FN: 8	TP: 24	32
	73	27	

Figure 5 (a) Confusion matrix

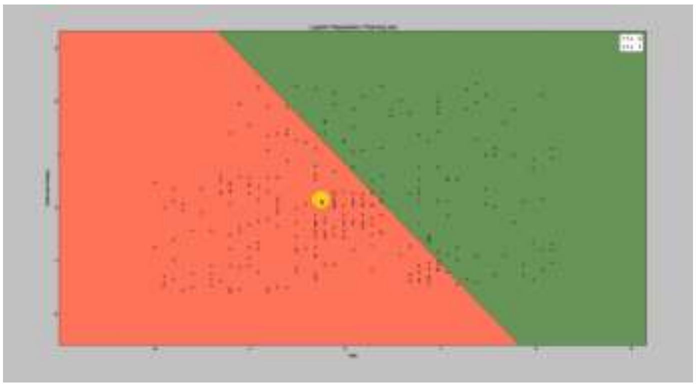


Figure 5 (b): Confusion matrix in the form of Graphics

The total number of correct predictions is derived from the sum of true positive and true negative. The total number of incorrect predictions is the sum of false positive and false negative. Then, the accuracy can be calculated as TC / TD. Where, TC isTotal Number of Correct Predictions and Td is Total Data in Test Data Set.Here, for our test data set is evaluated as follows,

Accuracy = 89 / 100 = 0.89

The opposite of accuracy is Misclassification Rate. The Misclassification Rate is derives as the ratio of TIC and TD. Where TIC is Total Number of Incorrect Predictions and TD is represented as Total Data in Test Data Set. Here, in the present work Misclassification rate is evaluated as,

Misclassification rate = 11 / 100 = 0.11

V. Conclusion

In this paper, it is explained that logistic regression can be a powerful analytical technique for use with a dichotomous variable outcome as it is more informative than linear regression. This led to researchers and readers what to expect from a report that uses the logistic regression techniques. What tables, charts, or figures should be included? What assumptions should be verified? We also demonstrated the application of logistic regression on a data set of users on social network for which a company analyses the sales of its cars. It is observed that Logistic regression model successfully analyzed the user will to purchase a car or not

VI. REFERENCES

[1] Peng, C.-Y. J, Lee, K Land G.M. Ingersoll, GM (2002) “An introduction to logistic regression analysis and reporting,” The Journal of Educational Research, 96(1),3-14

[2] AbhilashaTyagi et al., “Sentiment Analysis using Logistic Regression and Effective Word Score Heuristic”, International Journal of Engineering & Technology, Vol 7, 2.24, 2018, 20-23.

[3] Lo, Y.W, Potdar, V, “A review of opinion mining and sentiment classification framework in social network”, 3rd IEEE International Conference on Digital EcoSystems and Technologies, DEST 2009.

[4] Peter D. Turney, “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews”, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 417-424.

[5] Sun B., Ng V et al., “Analysis Sentimental Influence of Posts on Social Network”, 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design (CSCWD)

[6] Duda, R.O. Hart, P.E.(1973), Pattern classification and scene analysis, Wiley, New York

[7] Pang B, Lee L, S. Vaithyanathan Thumbs up? Sentiment Classification using Machine Learning Techniques.

[8] Syamala Rao et al., “A Comprehensive Survey of Financial Data Modelling Processes & Data Cleaning Methods Using Composite Coefficient”, Volume 12 Issue 01-Special Issue, 2020.

[9] N. Suresh Kumar et al., “Integrating and Organization of Multidimensional Virtual Citizen Database with Extinction and Limited Access”, IJCA, Vol 85 No 2, Jan 2014.

[10] Renuka et al, “Statistical Accuracy of Authentication with Biometrics”, International Journal of Engineering and Advanced Technology, vol 8, issue 4, April 2019, pp1040-1043.