Performance of Logistic Regression in Classifying Graduates' Salaries

Raj Choudhary*
*School of Computer Science and Engineering, VIT Chennai, Tamil Nadu, India 600127
Email: raj.choudhary2016@vitstudent.ac.in *

***Abstract*: In statistics, the logistic model (or logit model) is a widely used statistical model that, in its basic form, uses a logistic function to model a binary dependent variable; many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model; it is a form of binomial regression. Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail, win/lose, alive/dead or healthy/sick; these are represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from logistic unit, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probit model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a constant rate, with each dependent variable having its own parameter; for a binary independent variable this generalizes the odds ratio.**

Logistic regression was developed by statistician David Cox in 1958. The binary logistic regression model has extensions to more than two levels of the dependent variable: categorical outputs with more than two values are modelled by multinomial logistic regression, and if the multiple categories are ordered, by ordinal logistic regression, for example the proportional odds ordinal logistic model. The model itself simply models probability of output in terms of input, and does not perform statistical classification (it is not a classifier), though it can be used to make a classifier, for instance by choosing a cutoff value and classifying inputs with probability greater than the cutoff as one class, below the cutoff as the other; this is a common way to make a binary classifier.

***Index Terms*: Logistic Regression, Binary Classifier, Odds ratio, Employability**

1. Introduction

In this paper, salary slab classification is attempted using the AMEO dataset which contains the employability outcomes of engineering graduates. The features include standardized test scores, academic performance metrics and other background details of the candidate. The target variable is the first salary of the candidate. Logistic Regression is used to predict the salary slabs of the graduates.

2. Methodology

Logistic Regression is used to predict the salary slab of fresh graduates' using different combinations of attributes. After this, regularized linear regression and polynomial regression were implemented to see if the performance of the model could be improved.

Logistic Regression is part of a larger class of algorithms known as Generalized Linear Model (glm). In 1972, Nelder and Wedderburn proposed this model with an effort to provide a means of using linear regression to the problems which were not directly suited for application of linear regression. Infact, they proposed a class of different models (linear regression, ANOVA, Poisson Regression etc) which included logistic regression as a special case.

The fundamental equation of generalized linear model is:

$$g(E(y)) = \alpha + \beta x1 + \gamma x2$$

Here, g() is the link function, E(y) is the expectation of target variable and $\alpha + \beta x1 + \gamma x2$ is the linear predictor (α,β,γ to be predicted). The role of link function is to 'link' the expectation of y to linear predictor.

Important Points

1. GLM does not assume a linear relationship between dependent and independent variables. However, it assumes a linear relationship between link function and independent variables in logit model.
2. The dependent variable need not to be normally distributed.
3. It does not uses OLS (Ordinary Least Square) for parameter estimation. Instead, it uses maximum likelihood estimation (MLE).
4. Errors need to be independent but not normally distributed.

3. Dataset – AMEO 2015

For every engineer, AMEO [1] dataset provides anonymised bio data information along with their respective skill scores and employment outcome information. Specifically, the following information is available for every engineer:

1. Scores on Aspiring Minds' AMCAT - a standardized test of job skills. The test includes cognitive, domain and personality assessments.
2. Personal information like gender and date of birth.
3. Pre-university information like 10th and 12th grade marks, board of education and 12th grade graduation year.
4. University information like GPA, college major, college reputation proxy, graduation year and college location.
5. The following employment outcome information is available for every engineer: First job, annual salary, First job title, First job location, Date of joining and leaving of first job.

AMEO 2015 has gained traction since its public release. Aspiring Minds annually publishes the National Employability Report, a data-driven commentary on graduates and their employability. A recent NER was based on an extension of this dataset.

4. Analysis and Result

Logistic Regression was performed on the dataset using a number of combination of attribute – (Domain, Agreeableness), (Agreeableness, Openess to experience), (Domain, Openess to experience), (10Percentage, Quant) and (Quant, ComputerProgramming). The Decision boundary for these trained model was plotted as their accuracy score was found. The best accuracy score of 36.75% was obtained from Quant and ComputerProgramming. The low level of accuracy was because of high overlapping of the datapoints.

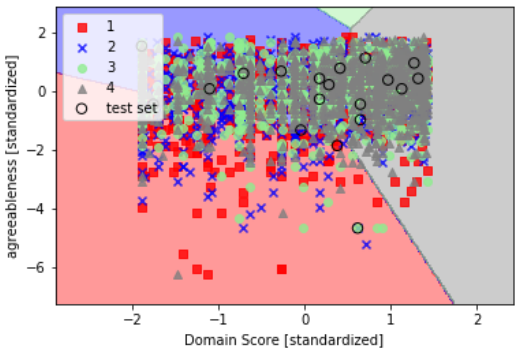

Figure 1 Accuracy 32.50%

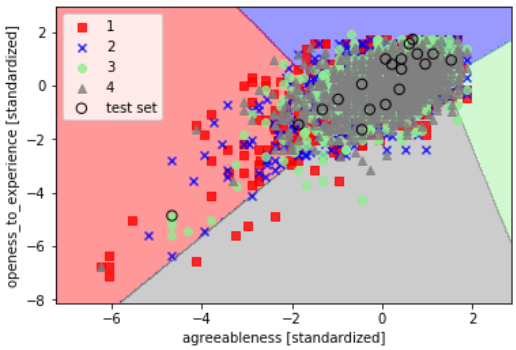

Figure 2 Accuracy 26.08%

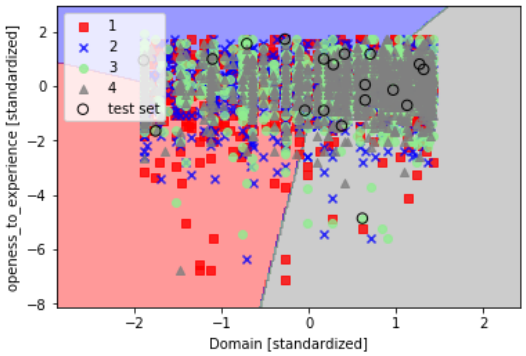

Figure 3 Accuracy 32.33%

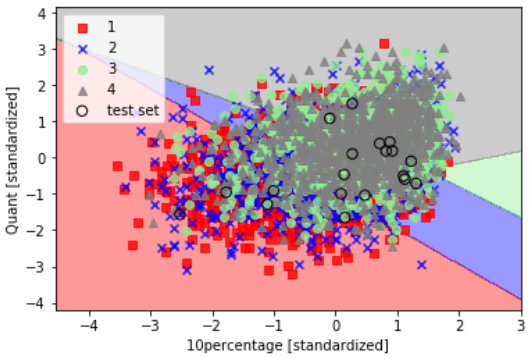

Figure 4 Accuracy 36.00%

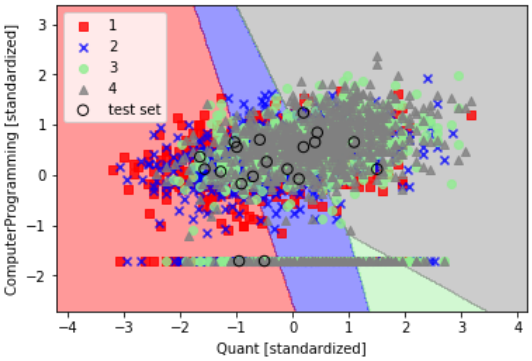

Figure 5 Accuracy 36.75%

5. Conclusion

Hence, we confirm that simple logistic regression will not be able to classify the salary of the graduates' accurately because of the presence of

high dense data points. To be able to make better classification, complex learning models or dimensionality reduction is to be used.

6. Reference

[1]. V. Aggarwal, S. Srikant, and H. Nisar, "Ameo 2015: A dataset com- prising amcat test scores, biodata details and employment outcomes of job seekers," in *AMEO 2015: A Dataset Comprising AMCAT Test Scores, Biodata Details and Employment Outcomes of Job Seekers*. ACM, 2016.