Classification of Serial Entrepreneurs via Logistic Regression: A Case Study

Eric Gehrig, PhD
Senior Research Scientist
Research & Development
Target Training International

Octobeer 10, 2017

Abstract

Classification or identification techniques are excellent tools for initial phases of building predictive models. Many classification techniques exits and the choice of which to use is based on the underlying structure of the data in question. In this case study, the data fail to be multivariate normally distributed and hence a natural choice is the logit or logistic regression model. A logistic regression model is employed in an attempt to classify or identify a group of Serial Entrepreneurs embedded in a randomly selected control group. All data is based on the Performance TriMetrixDNA®. Volunteers completed the assessments in 2010. The study shows that, in this case, data from the Performance TriMetrixDNA® assessment provide sound classification of the target group embedded in the control group.

Introduction

A key area of validity for any assessment is predictive validity. In psychometric assessments, predictive validity is the extent to which assessment scores may be used to predict another criterion. Some examples of criterion one may wish to predict are job turnover, job performance, safety measures, and academic success, to name a few. A first step toward generating a predictive model based on a psychometric assessment and measuring the predictive validity of an assessment is to measure the ability of assessment scores to identify a targeted group of individuals.

Many classification techniques exist in the mathematical and statistical literature. One may consider linear or quadratic discriminant analysis, logit regression, probit regression, and nonparametric discriminant analysis techniques such as restricted linear discriminant analysis. Note that the previous list is not exhaustive. The choice of which technique is best suited for classification is dependent on the underlying structure of the data in question.

The data analyzed in this case study is based on a solicitation of volunteers who took the TTI Performance TriMetrixDNA® assessment in 2010. The volunteers satisfied the following criteria and are denoted *Serial Entrepreneurs*:

1. Each respondent must have previously started more than one business;
2. Each respondent must have previously started businesses in more than one field.

The group of Serial Entrepreneurs is referred to as the Target group throughout the remainder of this paper. There are a total of 76 respondents in the Target group.

A data set consisting of individuals also having taken the Performance TriMetrixDNA assessment in 2010 has been pulled from the TTI database and serves as a data set from which random samples are drawn to serve as the Control group. This data set consists of 190 individuals from whom random samples of varying sizes

are drawn and tested for inclusion in the study. Details on the random sampling and tests performed are presented below.

The remainder of this paper presents the motivation for use of logistic regression, the results of the classification exercise, and discusses the possibilities for moving forward towards an example of a full predictive model. Some of the challenges of building a full predictive model in the setting of psychometric assessments are also discussed.

A Primer on Classification Algorithms

Data classification is known in several areas of computer science, mathematics, and statistics. The underlying problem is to identify to which subgroup or category an observation belongs based on the information provided by a training data set. Some examples of classification problems are identifying spam email or a medical diagnosis based on observed patient characteristics. This paper is concerned less with the assignment of the spam email or the diagnosis and more with the training exercise that predicates the predictive model implied here.

As mentioned in the introduction, many techniques exist and may be applied to train a classification or prediction model. Linear discriminant analysis (LDA) is a very common classification technique that is used when the underlying data follow a multivariate normal distribution. To be more specific, we have two data sets to consider, the Target group and the Control group. If LDA is to apply, each of the two data sets must be multivariate normal, and, more restrictive, it is required that the Target and Control groups must share a common covariance matrix.

Another common technique is the quadratic discriminant analysis (QDA). In some cases were LDA does not sufficiently classify the groups in question, QDA may provide a more robust and accurate identification. However, the main underlying assumptions of QDA and LDA are the same. In other words, the assumption of multivariate normally distributed data with common covariance matrix is still present.

There is also a classification technique known as mixture discriminant analysis (MDA). Once again, the underlying assumption is that of normally distributed data. The main difference here is that one considered Gaussian (Normal) mixture models to model the underlying data.

Logit and Probit models are two more possible models to consider for classification problems. In general, the use of logit or probit is a choice. However, in most settings a logit model is preferable for several reasons. First, a probit model assumes the underlying cumulative distribution is that of the standard normal distribution while the logit cumulative distribution of the logistic distribution. Second, the logit model is interpretable in terms of log odds ratios. Third, probit models are more applicable to heteroskedastic problems. A final reason is that the logit model is (historically) easier to estimate than the probit model.

The last reason presented above is truly a historical model. Since the probit model is based on the cumulative normal distribution, it is defined in terms of an infinite integral of the normal density function:

$$F(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt. \tag{1}$$

Modern computing power has made estimation of these types of integrals relatively straightforward. However, logit regression remains the more common choice in practice. Of further note is the fact that neither logit nor probit regression models assume the underlying data is normally distributed. Each relies on a classification (dependent variable) typically taking values in the set $\{0, 1\}$.

The choice of which model to use for the classification problem comes down to an analysis of the data and a decision based on performance. In practice, multiple models may need to be tested to determine the best model to employ for the situation at hand.

A Primer on Logistic Regression

The current case study breaks the data into two subsets, the Target group with classification equal to 1, and the Control group with classification equal to 0. In other words, our classification is a binary variable. Note that one may consider more than two classifications using the logistic regression approach.

Following [1], suppose we have a single response variable y taking values in $\{0, 1\}$ and a single, continuous explanatory variable x. The corresponding logistic regression model is of the form

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{2}$$

where the notation exp denotes the usual exponential function with base e. The function $\pi : D \mapsto [0, 1]$ where D is an appropriate domain dependent on the explanatory variable x and $[0, 1]$ is the usual unit interval in $\mathbb{R}$.

According to [1], there are two main reasons for choosing the logistic distribution in (2). First, π is an extremely flexible and easily used function, and second, π lends itself to meaningful (clinical) interpretation. To see the utility of the function π note the following transformation, called the logit transformation.

$$g(x) = \ln\left(\frac{\pi(x)}{1 - \pi(x)}\right). \tag{3}$$

Note that with a little algebra, $g(x) = \beta_0 + \beta_1 x$. This is useful in that the logit transformation of the logistic regression equation results in a linear expression with many of the desirable properties of the usual linear regression model.

One important difference between linear and logistic regression is that the error, which expresses an observations deviation from the conditional mean, is no longer assumed to be normally distributed. Again following [1], we may express the value of the outcome variable given x as $y = \pi(x) + \epsilon$.

In this formulation, ϵ may take on one of two possible values. If $y = 1$, then $\epsilon = 1 - \pi(x)$ with probability $\pi(x)$, and if $y = 0$ then $\epsilon = -\pi(x)$ with probability $1 - \pi(x)$. In summary, ϵ follows a binomial distribution with probability given by the conditional mean $\pi(x)$.

The importance of the preceding discussion is that we can now readily construct the likelihood function of the above mentioned binomial distribution. For values of $y = 1$ given x the contribution to the likelihood function is $\pi(x)$ and the contribution for values of $y = 0$ given x the contribution is $1 - \pi(x)$. Thus, for any observation x_i, the contribution to the likelihood function is given by

$$\pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1 - y_i}. \tag{4}$$

Note that (4) reduces to $\pi(x_i)$ or $1 - \pi(x_i)$ depending on the value of y_i given the choice of x_i. One assumption in logistic regression is that the observations are independent and hence the likelihood function is given by the product of the individual terms given in (4):

$$\ell(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1 - y_i}. \tag{5}$$

There is one more step involved to obtain the desired result. In all parametric regression approaches, there is an underlying optimization. This usually entails some form of differentiation. In the case at hand, (5) now requires differentiation with respect to the parameters $\boldsymbol{\beta}$ and a solution of the resulting equations. However, differentiation of products of functions is quite difficult compared to differentiation of sums of functions. This leads to a heavy computational cost. Hence, it is advantageous to construct the log likelihood function by taking the logarithm of (5) and using the appropriate properties of the logarithmic functions, namely that $\ln(f \cdot g) = \ln(f) + \ln(g)$ and $\ln(f^g) = g \ln(f)$.

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ y_i \ln(\pi(x_i)) + (1 - y_i)(\ln(1 - \pi(x_i))) \right\}. \tag{6}$$

The problem at hand is now to optimize (6) with respect to the parameters $\boldsymbol{\beta}$. It should be noted that while the parameters $\boldsymbol{\beta}$ are not explicitly

present in (6), one may substitute the definition of $\pi(x)$ from (2) into (6) to see that (6) is, in fact, a function of the parameters $\boldsymbol{\beta}$.

An extension of logistic regression that may be useful in classification problems is that of multinomial logistic regression. As a quick example, suppose that the response variable now may take on any of 3 possible values, $\{0, 1, 2\}$. In this case, one may define the conditional probabilities of each outcome category as follows:

$$P(y = 0|x) = \frac{1}{1 + \exp(g_1(x)) + \exp(g_2(x))}, \quad (7)$$

$$P(y = 1|x) = \frac{\exp(g_1(x))}{1 + \exp(g_1(x)) + \exp(g_2(x))}, \quad (8)$$

and

$$P(y = 2|x) = \frac{\exp(g_2(x))}{1 + \exp(g_1(x)) + \exp(g_2(x))}, \quad (9)$$

where

$$g_i(x) = \beta_{i0} + \beta_{i1}x_1 + \ldots \beta_{in}x_n. \quad (10)$$

In (10) the index i runs from 1 to the number of categories present (2 in this example), and n represents the number of independent variables present.

There is a similar derivation of the log likelihood function to that in (6) and a maximum likelihood estimation process is used to find the coefficients (β_{ij}).

The utility of the multinomial logistic regression technique is for a case similar to predicting the likelihood of a student with a given set of characteristics to pass a given course with a particular grade level. This process could also be useful in constructing a predictive model that would rank a group of sales employees into two categories, one category representing high performers and the other category representing low to average performers. The third category may be a random sample of the general population for differentiation purposes.

Logistic Regression in Classifying Serial Entrepreneurs

The first step in this case study is to verify that LDA, QDA, and other techniques based on normally distributed data do not apply. The first step of this case study assumes an even split between the Target and Control groups. In other words, there are 76 respondents in the Target group. To generate the even split, a random seed is chosen to ensure repeatability throughout the process and 100 random samples of size 76 are generated and tested.

The Shapiro-Wilk test for normality is used with a desired α level of .05. The interpretation is that if the p-value associated with the test statistic is less than the α level, the Null hypothesis that the data are normally distributed is rejected. Tables 1 and 2 summarizes the averages of the p-values for each of the 37 variables in question.

An analysis of the p-values in Tables 1 and 2 show that in the Target group, the Null hypothesis must be rejected for 24 of the 37 variables. Similarly, the Null hypothesis must be rejected for 20 of the 37 variables in the Control group. Therefore, in the opinion of the authors, a classification technique that relies on the assumption of normally distributed data is not the best choice for subsequent analysis.

Given the summary of classification techniques presented earlier in this case study, the choice is between logistic and probit regression. A test of the model shows relatively low levels of heteroscedasticity, hence the choice is made to proceed with logistic regression. However, an interesting and future study is to perform a similar analysis using the probit regression technique to determine whether the levels of heteroscedasticity may be accounted for in the probit approach.

The logistic regression technique is implemented in the R Statistical Package using the GLM

Table 1: Shapiro-Wilk p-values
Target Group

Variable	Shapiro-Wilk p-value
Adapted D	< 0.001
Adapted I	0.001
Adapted S	< 0.001
Adapted C	< 0.001
Natural D	< 0.001
Natural I	< 0.001
Natural S	< 0.001
Natural C	< 0.001
Theoretical	0.351
Utilitarian	0.188
Aesthetic	0.008
Social	0.303
Individualistic	0.716
Traditional	0.726
Leadership	< 0.001
Coaching/Develop.	0.001
Teamwork	0.613
Conflict Mgt	0.341
Interpersonal Skills	< 0.001
Problem Solving	0.726
Creativity/Innovation	0.011
Written Comm.	0.038
Customer Service	0.372
Flexibility	0.273
Goal Orientation	< 0.001
Planning/Organizing	0.024
Diplomacy	0.224
Personal Effectiveness	0.007
Presenting	< 0.001
Management	0.122
Negotiation	0.005
Persuasion	< 0.001
Empathy	< 0.001
Continuous Learning	0.040
Futuristic Thinking	0.020
Decision Making	0.064
Self Management	0.039

Table 2: Shapiro-Wilk p-values
Control Group

Variable	Shapiro-Wilk p-value
Adapted D	< 0.001
Adapted I	0.001
Adapted S	0.003
Adapted C	0.001
Natural D	0.001
Natural I	< 0.001
Natural S	0.005
Natural C	< 0.017
Theoretical	0.416
Utilitarian	0.147
Aesthetic	0.405
Social	0.396
Individualistic	0.489
Traditional	0.349
Leadership	< 0.012
Coaching/Develop.	0.045
Teamwork	0.004
Conflict Mgt	0.377
Interpersonal Skills	< 0.001
Problem Solving	0.175
Creativity/Innovation	0.259
Written Comm.	0.050
Customer Service	0.007
Flexibility	0.358
Goal Orientation	0.042
Planning/Organizing	0.399
Diplomacy	0.037
Personal Effectiveness	0.391
Presenting	0.001
Management	0.384
Negotiation	0.005
Persuasion	0.201
Empathy	< 0.019
Continuous Learning	0.111
Futuristic Thinking	< 0.001
Decision Making	0.310
Self Management	0.050

functionality. The same random samples generated during the test for normality are then run through the logistic regression. An analysis of the output is performed by first visually inspecting the predicted values generated by the GLM functionality. In some cases, probabilities identically equal to 0 or 1 were generated. By the definition of logistic regression, a probability of 0 or 1 is only possible in a limit and hence there is a flaw in these underlying data sets.

Given the relatively small Control group to work with, there are limitations in the random selection process. A future version of this study seeks to generate a much larger Control group in order to avoid this pitfall. For purposes of this analysis, any predicted data sets containing predicted probabilities of 0 or 1 are removed from consideration.

The results of the logistic regression classification exercise are presented in Tables 3, 4, and 5. It is interesting to note the slightly different progression in correct identification percentages. Adapted Behaviors appear to be better at classifying the Control group when compared to Natural Behaviors and Motivators.

Table 3: Logistic Regression Analysis Target

Factors Used	% Correctly Classified
Null Set (n^{-1})	50.00%
Adapted Behaviors	68.9%
Motivators	69.1%
Natural Behaviors	74.1%
All but DNA	76.3%
DNA Soft Skills	84.2%
DNA & Adapted	84.9%
All but Motivators	85.5%
All Factors	88.2%

Table 4: Logistic Regression Analysis Control

Factors Used	% Correctly Classified
Null Set (n^{-1})	50.000%
Motivators	65.8%
Natural Behaviors	66.7%
Adapted Behaviors	73.4%
All but DNA	76.3%
DNA Soft Skills	85.5%
DNA & Adapted	85.5%
All but Motivators	86.8%
All Factors	86.8%

Table 5: Logistic Regression Analysis Overall

Factors Used	% Correctly Classified
Null Set (n^{-1})	50.000%
Motivators	67.4%
Natural Behaviors	70.4%
Adapted Behaviors	71.0%
All but DNA	76.3%
DNA Soft Skills	84.2%
DNA & Adapted	85.5%
All but Motivators	85.9%
All Factors	87.5%

Also interesting to note is that in all three cases, Target group, Control group, and the overall correct identification percentage, continuing to add additional information from Behaviors to Motivators to DNA Soft-skills shows a pattern of improvement. In all three cases, the final classification using all factors ranks first or tied for first.

A final note is that the above analysis applies only to this data set and would not necessarily be predictive if applied to another data set. There is no variable selection approach employed to the preceding analysis and an inspection of the statistical significance of the resulting regression model shows only a small handful of the 37 variables as significant. In this setting, it is only desired to determine whether differentiation capabilities exist in the data using the logistic regression technique. The above charts confirm that data resulting from the TTI Performance TriMetrixDNA does show said differentiation capability in this study.

Summary and Future Studies

This cases study shows that the TTI Performance TriMetrixDNA assessment data differentiate a Target group of serial entrepreneurs from a sample of the general population. The ability to classify based on these variables in this data set is strong, but has its shortcomings. The data set is relatively small with only 76 respondents. The Control group overall only contained 190 records, limiting the number and quality of ran-

dom samples to generate control groups for comparison. The study focuses on a 50/50 split in the data sets studied when many different splits would likely prove more informative.

Follow on studies plan to rectify the aforementioned issues by utilizing larger samples of data on Behaviors and Motivators. An exploration of DNA Soft-skills is likely to accompany such analysis, although the control data set of 190 observations is the population taking the Performance TriMetrixDNA for the time frame in question. It remains to be seen whether inclusion of data from individual DNA respondents is useful in expanding the data set while providing relevant information.

Further follow on studies plan to extend the logistic regression technique to include a predictive model building approach and subsequent predictability studies as relevant data comes available.

References

[1] David W. Hosmer and Stanley Lemeshow. *Applied Logistic Regression.* Wiley, New York, 2000.