A Pilot Study of Potential Future Entrepreneurs

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

Ron Bonnstetter, PhD
Senior Vice President
Research & Development
Target Training International

April 20, 2018

Abstract

A key area for any assessment is whether the assessment data may be used, in some form, to generate a predictive model. Following the modeling process and logic outlined in a similar previous study, see [1], this paper explores a data set of university students collected by a strategic, nonprofit partner of TTI Success Insights, the Indigo Project. A key member of Indigo Project has empirically developed an algorithm to identify potential future entrepreneurs. This paper explores the data generated by our Indigo Project to determine how successful our earlier logistic regression modeling approach to classification of serial entrepreneurs works when applied to the set of university students. The results are very solid and provide solid support for a follow on, longitudinal study to determine the potential predictive power of such a modeling approach.

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 data gathered through Indigo Project, a nonprofit TTISI partner using the TTI SI Talent Insights® assessment. Indigo Project has strong relationships with several US state universities. The data uses consists of 16,568 anonymized records of students from multiple universities across the US. There is limited demographic information available on these participants, by design.

The intent of this pilot study is to determine the strength of the relationships of the logistic regression method of classification and predictive modeling and the possibility of creating a longitudinal study of future entrepreneurs. A key member of Indigo Project has a many years of especially relevant experience in an entrepreneurial setting as well as with the use of psychometric assessments. This individual has developed an algorithm that Indigo Project believes is able to successfully identify future entrepreneurs based on their TTI SI Talent Insights® responses. The remainder of this paper is devoted to providing analytic support to the empirically derived algorithm.

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:

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

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 [2], 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 [2], 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 [2], 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.

Generation of the Data Set Under Consideration

We first set the stage for how the data came to be and then describe the process by which we design the classification possible prediction model.

A strategic non-profit partner of TTI SI, Indigo Project, works closely with several US universities with the goal of improving the educational experience of the student through the use of the TTI SI family of assessments. The assessments are use to help the students and professors better understand themselves and better understand those around them. The individuals working at Indigo Project are well versed in both the use of the assessments as tools and in the entrepreneurial world.

This talented group has taken their combined experience and developed an empirical approach to identification of potential future entrepreneurs. This empirical approach is largely based on a benchmark style identification in which the experience of the identifier has set some form of profile in Behaviors (Style Insights® portion of Talent Insights®) and Motivators (Motivation® portion of Talent Insights®). This profile is intentionally left unknown to the authors of this report.

The authors of this report received a data file with 16,568 records. The records contain two main groups of data, Behavior Characteristics which are either direct scores from the DISC portion of Talent Insights® or derived from the DISC scores, and Motivation Indicators which are either direct scores from the Motivators portion of Talent Insights® or derived from the Motivators scores.

Limited demographics information is available. There is a Gender category that shows a breakdown of 8,001 males and 8,567 females for an aproximate 48%/52% M/F split. The remaining variable of interest is a Class variable that Indigo Project created based on their aforementioned algorithm. This variable is a binary indicator of 1 for classification in the future entrepreneur group and 0 for not. This results in 5,055 individuals as belonging to the future entrepreneur group.

Logistic Regression in Classifying Potential Future Entrepreneurs

The primary focus of the early stages of generating any potential predictive model is to determine the proper modeling approach for the data and problem at hand. Our goal is to identify, with the highest possible success rate, those individuals that have been classified by a third party as belonging to a particular group. A further goal is to minimize the amount of incorrectly identified members of the non-entrepreneur group. Our indicator data is binary and we have assumed continuous explanatory variables.

We use the phrase "assumed continuous explanatory variables" to denote the fact that while the scoring algorithms behind the TTI SI assessments generate continuous data, they are reported as discrete, generally taking positive integer values. In some cases, the variable values are reported on a discrete scale ranging between 0 and 10 report to one decimal place.

Generally speaking, our goal and type of data support a decision to use logistic regression. However, this does not imply at this stage the logistic regression is the correct choice. In order to support the decision to use logistic regression, further analysis of the data is necessary. We do this in the form of plotting the log of the odds agains the category for each of the possible variables we wish to consider. For brevity, we present an example of a good log odds plot and an example of a variable we reject in this analysis.

Given that this report is to establish a baseline for moving forward on a larger project, not all relevant information is presented. For example, one can clearly see a strong negative linear relationship between the log of the odds of falling into a particular scoring bucket and the index of the bucket, see Figure 1. This is precisely the kind of relationship one wishes to see when the desire is to use logistic regression. A more complete analysis may provide a linear regression analysis complete with goodness of fit scoring such as r^2.

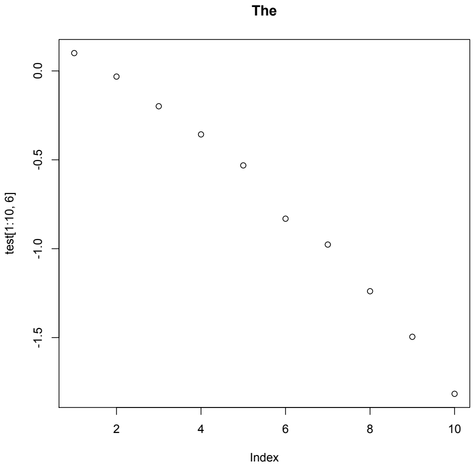

Figure 1: Theoretical Log Odds Plot

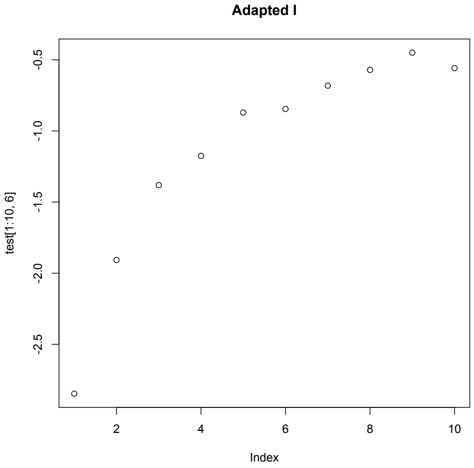

Figure 2: Adapted Influence Log Odds Plot

Compare the plot in Figure 1 with that in Figure 2. In other circumstances, the plot in Figure 2 would not necessarily be disqualifying for the Adapted Influence variable. In this case, the strength of many of the variables actually make the Influence variables somewhat unattractive. This is not necessarily a problem. Again, we refer to the fact that the study here is a pilot look at a longer vision of budding entrepreneurs. A more in depth analysis would have an a priori discussion of the most desired characteristics, then we would make decisions based on the combination of those stated desires along with the statistical analyses.

Table 1: Explanatory Variables

Tool	Variables
Behaviors	Adapted D
	Adapted S
	Adapted C
	Natural D
	Natural D
	Natural D
Motivators	Theoretical
	Utilitarian
	Aesthetic
	Social
	Individualistic
	traditional

Table 1 shows the list of explanatory variables retained after the graphical data analysis. This is again a place to point out that further analysis is required. As an example, we see that Motivators such as Aesthetic and Social are retained for the logistic regression analysis. A priori discussions may remove those variables. In this setting, no such discussions have occurred and these variables show strong linear relationships, see Figure 3.

During the data analysis previously discussed, single variable logistic regression analyses were also used to confirm that individual variables should or should not be retained. These single variable regressions agree with the visual evidence already provided. We then proceed to attempt to identify any subsets of Behaviors, Motivators, and a combination of the two that best approximate the data via logistic regression. The basic process is to take all the variables in the Behaviors set, run logistic regression and test the outcome via statistical significance.

We do this for Behaviors, Motivators, and a combined data set. In the end, we identify four variables that appear to perform best out of the possible combinations. The remaining variables are presented in Table 2

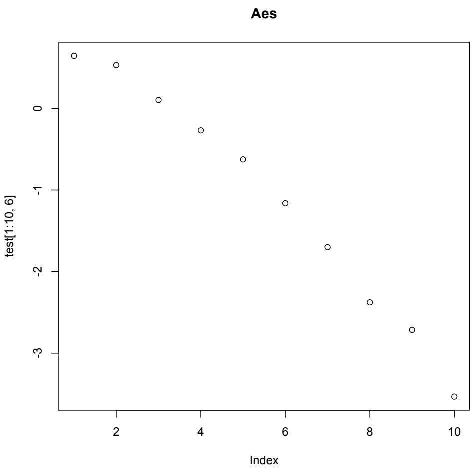

Figure 3: Aesthetic Log Odds Plot

Table 2: Explanatory Variables

Tool	Variables
Behaviors	Natural D
	Natural C
Motivators	Utilitarian
	Individualistic

The authors of this report make no claim to be experts in identification of budding entrepreneurs. However, after discussions with several colleagues at bot TTI SI and Indigo Project, it appears that Natural D and C are reasonable choices from Behaviors and Utilitarian and Individualistic are inline with an entrepreneurial mindset.

We next randomly partition the data into five data sets of (nearly) equal size. We do this as follows. To ensure we maintain approximately the same breakdown of inclusion in the desired data set, we first split the data set by the Class variable. We then randomly create five data sets of as close to equal size as possible from each group. The amount of data is not divisible by 5, although the group of interest is. We then merge the data once again into two sets of data, the training sample (80% of the combined data) and the hold out sample (20% of the combined data). We now have 5 training sets and 5 hold out samples.

The idea here is to train the model on each of the five training sets, looking for any anomalies, such as regression coefficients for a variable changing sign or loss of significance of a variable, etc. We then apply the results of each of the training runs to the hold out sample. In other words, the model has no knowledge of the data in the hold out sample, only that of the training sample. In this way we are doing all that is possible to ensure the resulting model is doing what we claim, identifying entrepreneurs in a random set of individuals.

Again for brevity, we choose not to present the results of all five training samples and all five hold out samples. Rather we present a synopsis of the results. Table 3 presents the results of the training of the first random data set. All variables in training sample one test were significant at the 0.01 level or better. As a note, the level of significance expected for our purposes is 0.05 or better.

Table 3: Training Sample 1

Variable	Coefficient
Intercept	-22.96
Natural D	0.075
Natural C	-0.037
Utilitarian	1.538
Individualistic	2.086

When the training model is applied to the hold out sample we get the following contingency table, see Table 4.

Table 4: Contingency Table
Hold Out Sample 1

	$Y_i = 1$	$Y_i = 0$
$X_i = 1$	872	139
$X_i = 0$	131	2171

Table 4 should be interpreted as follows. The item that is in the slot of the table with $Y_i = 1$ above and $X_i = 1$ to the left is the count of the number of items that the training model predicts will be in the future entrepreneur set ($Y_i = 1$) and are actually in that set according to the data ($X_i = 1$). In this case, we have 872 individuals correctly classified as future entrepreneurs (correct according to the empirical model). This is sometimes denoted as the True Positives (TP). We have 139 False Negatives (FN), individuals we predicted to not be in the desired group, but are in that group. There are 131 False Positives (FP), individuals the model predicted are in the desired group, but are not. Finally, we have 2171 True Negatives (TN), individuals the model predicted would not be in the desired group and are not.

Table 5: Contingency Table
Percentage

	$Y_i = 1$	$Y_i = 0$
$X_i = 1$	86.25	13.75
$X_i = 0$	5.69	94.31

It should be noted that the numbers generated in Tables 4 and 5 are computed based on an assumed 50/50 split in the data. This is clearly not the case. If we randomly sampled the data at the true rate, in this case approximately 30%, our success would appear much stronger. We choose a more conservative estimate. In other words, we do not call an individual a member of the desired group unless our predicted probability for the individual to be a member exceeds 0.50, rather than the less conservative 0.30.

Further interpretation of the data in Table 5 is, in this data set the training model correctly identifies 86.25% of the individuals as members of the future entrepreneur group. Further, the training model correctly identifies 94.31% of those individuals as not part of the future entrepreneur group. Table 6 presents the TP and TN results of all five hold out samples as percentages.

Table 6: All Hold Out Samples True Results

Sample	TP	TN
Hold out 1	86.25	94.31
Hold out 2	83.78	94.35
Hold out 3	86.75	95.05
Hold out 4	86.35	95.09
Hold out 5	84.67	94.92

The results are quite solid and suggest that moving forward on a full longitudinal study to determine the adequacy of both the empirical and the analytic models is warranted.

Summary and Future Studies

This report serves multiple purposes. First, the outline of the modeling approach presented here is the start of a more in depth analysis that may be applied in many settings. The current setting is to help identify future entrepreneurs. This becomes the second of the purposes, to help align the empirical findings and provide a sound analytical background to the experience of those who developed the empirical. A third, and clearly not final, reason is to show that when a data set has sound explanatory variables combined with a well defined metric, the possibilities are many.

It is clear that this paper has not established a predictive relationship as we do not have any information on whether any of the individuals identified as possible future entrepreneurs have or will become entrepreneurs. It does suggest, however, that pursuing an academic study of this subject, over time, is likely worth the effort.

References

[1] Eric T. Gehrig. Classification of Serial Entrepreneurs via Logistic Regression: A Case Study. White Paper, October 2017.

[2] David W. Hosmer and Stanley Lemeshow. *Applied Logistic Regression*. John Wiley and Sons, Inc., 2000.