ISSN: 2664-9640

Predicting Early Termination from Counseling: A Binary Logistic Regression Analysis

Shepherd Chikomana*

School of Science, Zhejiang University of Science and Technology, Hangzhou, 310023, China

Abstract

This study examined the likelihood of early termination from counseling clients at a community mental health center and a sample of clients were interviewed where they responded honestly and their responses was recorded. Binary logistic regression model was built to predict the likelihood that each interviewed client will terminate their contract, based on independent variables such symptom severity and avoidance of disclose. The dependent variable in this study was 'terminate' (coded 1= terminate early and 0=did not terminate early), where the 'did not terminate' group serves as the reference/baseline category and the 'terminate early' being the target category. The two predictors in the model are continuous variables and are abbreviated as 'avdiscl ' for avoidance of disclosure and 'sympsev' for symptom severity.

Keywords

Binary logistic regression, dichotomous, odds ratio, Omnibus tests of Model Coefficients, pseudo-R-square, The Hosmer & Lemeshow.

1. Introduction

Binary logistic regression analysis is a statistical technique used to model the relationship between a binary dependent variable (also called the outcome or response variable) and one or more independent variables (also known as predictors or covariates). This type of regression is used to determine the likelihood of an event happening in various fields, including healthcare, social science, and business. According to [1], binary logistic regression can be used to estimate the likelihood of a binary outcome variable, such as the likelihood of a patient having a certain disease, based on one or several independent variables, such as age, gender, and health status.

The model is called 'logistic' because it uses the logarithm of the odds of the event occurring, instead of the probability itself. The logistic regression model assumes a linear relationship between the independent variables and the logarithm of the odds of the outcome variable [2]. Model estimates coefficients for independent variables in order to find out the magnitude and direction of the effect of these variables on the outcome variable.

The coefficients of independent variables can now be used to predict the likelihood of an outcome variable, given certain values of the independent variables. One of the advantages of binary logistic regression is having ability to handle both categorical and continuous independent variables [3]. Fundamentally, the researcher is addressing the question, "What is the probability that a given case falls into one of two categories on the dependent variable, given the predictors in the model?". As one might be inclined to ask why we don't use standard ordinary least squares regression (OLS) instead of BLR, OLS regression assumes (a) a linear relationship between the independent variables and the dependent variable, (b) the residuals are normally distributed, and (c) the residuals exhibit constant variance (Pituch & Stevens, 2016)[4]. All three assumptions are violated if the outcome variable in an OLS model is binary. And pivoting off (a), the relationship between one or more predictors and the probability of a target outcome is inherently non-linear as probabilities are bounded at 0 and 1. When modeling

ISSN: 2664-9640

a binary outcome using OLS regression, the estimation of model parameters ignores this boundedness, which has the notable effect of producing predicted probabilities that fall outside the 0-1 range. BLR estimates regression parameters by considering the fact that probabilities are bounded and 0 and 1. It also does not assume that residuals are normally distributed and exhibit constant variance.

2. Literature References

2.1. Model Estimation

Unlike OLS regression, BLR uses maximum likelihood (ML) to estimate model parameters. Maximum likelihood estimation is an iterative process aimed at arriving at population (parameter) values that most likely produced the observed (sample) data. Generally, this estimation approach assumes large samples and, aside from issues of power, smaller sample sizes can create problems with model convergence and estimation of model parameters. [Side note: With smaller samples, Exact logistic regression or Firth procedure using Penalized Maximum likelihood can be used. Unfortunately, these options are not commonly available in statistics programs.] From our sample, the model will be predicting given the independent factors of avoidance of disclosure and symptom severity and be able to display the table outlining the odds ratio pertaining to each predictor and for us to be able to draw the conclusion.

2.2. Evaluation of Model Fit

Evaluation of model fit in binary logistic regression can be done using various measures such as goodness-of-fit tests, pseudo R-squared values, and receiver operating characteristic (ROC) curves. The Goodness-of-fit tests: tests assess the overall fit of the model by comparing the observed frequencies with the expected frequencies based on the model. With the Hosmer-Lemeshow test [5] being the most commonly used goodness-of-fit test , which divides the data into groups based on predicted probabilities and compares the observed and expected frequencies within each group. A significant result indicates poor model fit. Pseudo R-squared values: these are measures of how well the model explains the variability in the data. The most commonly used pseudo R-squared values are Nagelkerke's R-squared [6], Cox and Snell R-squared [7], and McFadden's R-squared [8]. These values range from 0 to 1, with higher values indicating better model fit. ROC curves: these plots show the trade-off between sensitivity (true positive rate) and specificity (true negative rate) for different classification thresholds.

The area under the curve (AUC) [9] is a commonly used measure of model fit, with values ranging from 0.5 (random guessing) to 1 (perfect classification). With that being said, It is important to note that none of these measures alone can fully capture the performance of a logistic regression model, and a combination of these measures should be used to assess the overall fit of the model.

3. Methodology

3.1. Data

This study was based on 45 interviewees who gave an honest assessment of whether or not they would terminate the counseling contract early. Terminate, this is to end or cancel the counseling membership of a client from a Mental Health Centre and it is our dependent variable categorized with only two outcomes which are terminate early or not terminate early. Independent factors such as avoidance of disclosure and symptom severity was recorded. In therapy, avoidance of disclosure(avdiscl) is when a client refuses or hesitates to disclose certain experiences or information to their therapist. It can be due to a number of factors, such as fear of rejection or judgment, shame or guilt or lack of faith in the therapist. In therapy, symptom severity(sympsev) refers to how a certain symptom or set is affecting a person's

ISSN: 2664-9640

quality of life and functioning. In mental health treatments, the severity of symptoms is used to track the progress of treatment and measure its effectiveness. The dataset is available for download at

https://drive.google.com/file/d/1Etmudy8b6SZRykSPxCyFzG8ANQZwv966/view

3.2. The Model

When we have a binary outcome, our desire is to predict the probability (likelihood) of membership in a target outcome Prob(Y=1; terminate early), given what we know from a set of predictors. Unfortunately, we cannot model this relationship directly using standard OLS regression, since the relationship between the predictors and probability that Y=1 is inherently non-linear (following an S-shaped curve; logistic curve). In a logistic regression, it addresses this problem by "linearizing" the relation using a logit link function. The dependent variable that we are directly predicting, therefore, are logits (a mathematical transformation of probabilities). This changes our interpretation of intercepts and slopes in our regression model since the we are speaking in terms of logits rather than probabilities. The logistic regression prediction equation can be expressed as:

$$\text{logit}(\text{p}) = \ln\left(\frac{\text{P}}{1-\text{P}}\right)$$

$$\pi(x_i) = \frac{e^{\mathbf{B_0}+\boldsymbol{B_1}avdiscl+\boldsymbol{B_2}sympsev}}{1+e^{\mathbf{B_0}+\boldsymbol{B_1}avdiscl+\boldsymbol{B_2}sympsev}}$$

Where logit(p) is our dependent variable terminate, B_0 is a constant term and I'ts calculated value from variables in the equation table is -1.139, B_1 is the coefficient of the predictor avdiscl and its calculated value from variables in the equation table is .356, B_2 is the coefficient of the predictor sympsev and its calculated value from variables in the equation table is -.317.

3.3. Assumption Checking

Before running a binary logistic regression model, there are some assumptions that are supposed to be met on the data and if all the assumptions are met, then we can analyze our data using binary logistic regression model. These includes:

The dependent/response variable is binary or dichotomous.

Dependent Variable Encoding

Original Value	Internal Value
,00	0
1,00	1

Figure 1: Dependent Variable Encoding

Logistic regression assumes that the response variable only takes on two possible outcomes. From Figure 1, we can clearly see that we have two outcomes coded 0 and 1 for 'do not terminate early' and 'terminate early' respectfully as our only two outcomes, therefore our model meets this assumption.

The Observations are Independent

This assumption states that the dataset observations should be independent of each other. The observations should not be related to each other or emerge from repeated measurements of the same individual type. In our data, all our observations of each client are independent from each other.

Little or no multicollinearity between the predictor/explanatory variables[10].

ISSN: 2664-9640

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	avdiscl	,960	1,042
	sympsev	,960	1,042

a. Dependent Variable: terminate

Figure 2: Collinearity Statistics

In binary logistic regression, multicollinearity refers to the presence of high correlations between predictor/explanatory variables. "Little or no multicollinearity" means that the variables included in the regression model are not highly correlated with each other. Figure 2 above, shows the Collinearity Statistics and since all our tolerance values are greater than 0.1 meaning our assumption is not violated and also, we can check the VIF values and see that all our predictor values are less than 10 confirming that there is no multicollinearity in our dataset.

Multicollinearity can cause problems in binary logistic regression, such as unstable or unreliable estimates of regression coefficients, inflated standard errors, and difficulties in interpreting the coefficients. Therefore, it is important to check for multicollinearity before fitting a binary logistic regression model.

The sample size is sufficiently large (at least 10-20 observations per independent variable)[11].

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	45	100,0
	Missing Cases	0	,0
	Total	45	100,0
Unselected Cases		0	,0
Total		45	100,0

a. If weight is in effect, see classification table for the total number of cases.

Figure 3: Case Processing Summary

Logistic regression assumes that the sample size of the dataset if large enough to draw valid conclusions from the fitted logistic regression model. As a rule of thumb, you should have a minimum of 10 cases with the least frequent outcome for each explanatory variable. Figure 3 above shows the total sample size of 45 clients which is a reasonably good sample size number.

There are no outliers

Logistic regression assumes that there are no extreme outliers or influential observations in the dataset. By Using the Mahalanobis distance to check for outliers we see that our data doesn't have any outliers since the minimum value is greater than 0.001, therefore the assumption of no outliers has been met.

Now that we have tested our data and see that it meets all the assumptions for modelling the data using binary logistic regression model, we can start our analysis.

ISSN: 2664-9640

4. Analysis and Result Discussion

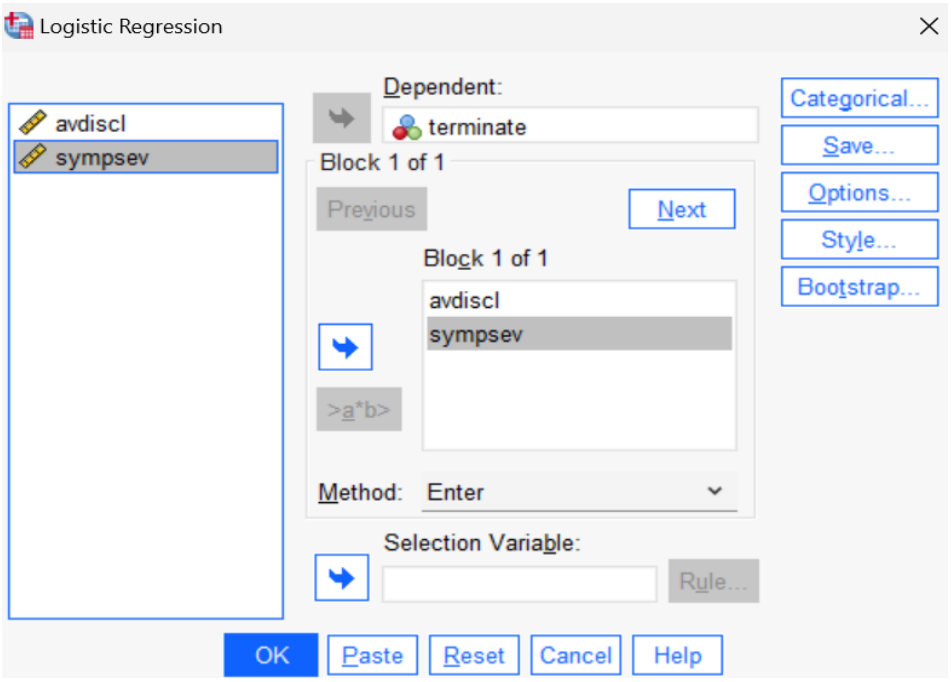

Figure 4: Logistic regression window

The diagram above (see Figure 4) shows the steps on how to execute the binary logistic regression model using SPSS.

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	45	100,0
	Missing Cases	0	,0
	Total	45	100,0
Unselected Cases		0	,0
Total		45	100,0

a. If weight is in effect, see classification table for the total number of cases.

Figure 5: Case Processing Summary

Dependent Variable Encoding

Original Value	Internal Value
,00	0
1,00	1

Figure 6: Dependent Variable Encoding

The first section of the output shows Case Processing Summary Highlighting the cases included in the analysis. In this study we have a total of 45 respondents as shown in Figure 5.

The Dependent variable encoding shows the coding for the criterion variable. In this case those who will choose to not terminate early(0.00) are classified as 0 whereas those that choose to terminate early(1.00) are classified as 1.(see Figure 6)

ISSN: 2664-9640

Block 0: Beginning Block

Classification Table[a,b]

Observed			Predicted terminate ,00	Predicted terminate 1,00	Percentage Correct
Step 0	terminate	,00	25	0	100,0
		1,00	20	0	,0
	Overall Percentage				55,6

a. Constant is included in the model.

b. The cut value is ,500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-,223	,300	,553	1	,457	,800

Figure 7: Block 0, Beginning Block

The next section of the output, headed Block 0 , as shown on Figure 7, is the result of the analysis without any of our independent variables used in the model. Therefore, this will serve as a baseline later for comparing the model with our predictor variable included. he output is presented in blocks. Block 0 contains the results from a null (i.e., intercept-only) model.

The next portion of the output is headed as Block 1. This part of the output is our main focus when interpreting results, as it is based on our regression model including our predictor(s).

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	26,174	2	<,001
	Block	26,174	2	<,001
	Model	26,174	2	<,001

Figure 8: Omnibus Tests of Model Coefficients

The Omnibus Tests of Model Coefficients (see Figure 8) contains results from the likelihood ratio chi-square tests. These test whether a model including the full set of predictors is a significant improvement in fit over the null (intercept-only) model. In effect, it can be considered an omnibus test of the null hypothesis that the regression slopes for all predictors in the model are equal to zero (Pituch & Stevens, 2016)[4]. The results shown here indicate that the model fits the data significantly better than a null model, $\chi^2(2)=26.174$, $p<0.001$.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	35,653[a]	,441	,590

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than ,001.

Figure 9: Model Summary

ISSN: 2664-9640

The Model Summary above (see Figure 9) contains the -2 log likelihood and two "pseudo-R-square" measures. The pseudo-R-squares are interpreted as analogous to R-square (being mindful that they are not computed in the same fashion as in OLS regression). These are more descriptive indices and are useful for evaluating overall model fit. The -2 Log Likelihood (i.e, -2LL) is referred to as the model *deviance*. And values closer to 0 indicate closer model fit (less discrepancy between the model and the data) whereas higher values indicate worsening fit (greater discrepancy between model and the data). [Discrepancy is reflected in the difference between conditional probabilities for group membership based on the model and actual group membership]

As noted previously, the likelihood ratio (LR) chi-square value in Figure 8 is equal to the difference in deviances (i.e., -2LL) between the model containing a complete set of predictors and reduced model containing only the intercept. We can compute the deviance for the intercept only model using the LR chi-square and the deviance of the full model:

$$Deviance(\ null\ model) = 35.653 + 26.174 = 61.827$$

The Cox & Snell and Nakelkerke R-squares are 'pseudo-R-square' values, since they are not computed the same way as R-square in the context of OLS regression. In OLS regression, R-square is interpreted as the proportion of variation accounted for in the DV as a function of the predictors. The pseudo-R-square values here are generally designed to represent proportionate change/improvement in model fit relative to the intercept-only model.

Below is the computation of Cox & Snell pseudo-R square (see Figure 10). Unlike R-square in OLS regression, the upper bound is less than 1.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	35,653[a]	,441	,590

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than ,001.

Figure 10: Model Summary on Cox & Snell R Square Calculation

$$CS = 1 - exp\left(\frac{deviance_{full}\ -\ deviance_{null}}{n}\right) = 1 - exp\left(\frac{35.653\ -\ 61.827}{45}\right) = 1 - e^{-.5816} = .441$$

*The 'exp' in the first two expressions above are equivalents to the 'e' in third expression. This was done to make the equation more readable.

Nagelkerke provided an adjustment to Cox & Snell providing an index ranging from 0 to 1. Below is the computation of the Nagelkerke pseudo-R-square.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	35,653[a]	,441	,590

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than ,001.

Figure 11: Model Summary on Nagelkerke pseudo-R-square

ISSN: 2664-9640

$$Nagelkerke = \frac{CS}{1 - exp\left(-\frac{deviance_{null}}{n}\right)} = \frac{.441}{1 - exp\left(-\frac{61.827}{45}\right)} = .590$$

*These versions of the Cox & Snell and Nagelkerke pseudo-R-squares (see Figure 11) are provided by Field (2018)[12].The -2*Log likelihood (also referred to as "model deviance" [4] is most useful for comparing competing models, particularly because it is distributed as chi-square .

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	2,024	7	,958

Figure 12: Hosmer and Lemeshow Test

The Hosmer & Lemeshow test is another test that can be used to evaluate global fit. A non-significant test result (see Figure 12; p=*.958*) is an indicator of good model fit.

Classification Table[a]

Observed			Predicted terminate ,00	Predicted terminate 1,00	Percentage Correct
Step 1	terminate	,00	21	4	84,0
		1,00	4	16	80,0
	Overall Percentage				82,2

a. The cut value is ,500

Figure 13: Classification Table

The classification table provides the frequencies and percentages reflecting the degree to which the model correctly and incorrectly predicts category membership on the dependent variable. On Figure 13, we see that $100\% \times \left(\frac{21}{21+4}\right) = 84\%$ of cases that were observed not terminate early were correctly predicted (by the model) to not terminate early. Of the 20 cases observed to terminate early, $100\% \times \left(\frac{16}{4+16}\right) = 80\%$ were correctly predicted by the model to terminate early. [As you can see, sensitivity refers to accuracy of the model in predicting target group membership, whereas specificity refers to the accuracy of a model to predict non-target group membership.] The overall classification accuracy based on the model was $100\% \times \left(\frac{21+16}{45}\right) = 82\%$.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
Step 1[a]	avdiscl	,356	,118	9,148	1	,002	1,427	1,133	1,798
	sympsev	-,317	,123	6,582	1	,010	,729	,572	,928
	Constant	-1,139	1,500	,577	1	,448	,320		

a. Variable(s) entered on step 1: avdiscl, sympsev.

Figure 14: Variables in the Equation

The 'Estimate' column (see Figure 14)contains the regression coefficients. For each predictor, the regression slope is the predicted change in the log odds of falling into the target group (as compared to the reference group on the dependent variable) per one unit increase on the predictor (controlling for the remaining predictors). [Note: A common misconception is that the regression coefficient indicates the predicted change in *probability* of target group

ISSN: 2664-9640

membership per unit increase on the predictor – i.e., p(Y=1|X's). This is WRONG! The coefficient is the predicted change in log odds per unit increase on the predictor].

Nevertheless, you can *generally* interpret a positive regression coefficient as indicating the probability (loosely speaking) of falling into the target group increases as a result of increases on the predictor variable; and that a negative coefficient indicates that the probability (again, loosely speaking) of target membership decreases with increases on the predictor. If the regression coefficient = 0, this can be taken to indicate changes in the probability of being in the target group as scores on the predictor increase.

The Odds Ratio (OR) column contains values that are interpreted as the multiplicative change in odds for every one unit increase on a predictor. In general, an odds ratio (OR) > 1 indicates that as scores on the predictor increase, there is an increasing probability of the case falling into the target group on the dependent variable. An odds ratio (OR) < 1 can be interpreted as decreasing probability of being in the target group as scores on the predictor increase. If the OR=1, then this indicates no change in the probability of being in the target group as scores on the predictor change.

The 95% confidence interval for the Odds ratio can also be used to test the observed OR to determine if it is significantly different from the null OR of 1.0. If 1.0 falls between the lower and upper bound for a given interval, then the computed odds ratio is not significantly different from 1.0 (indicating no change as a function of the predictor).

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	avdiscl	,356	,118	9,148	1	,002	1,427	1,133	1,798
	sympsev	-,317	,123	6,582	1	,010	,729	,572	,928
	Constant	-1,139	1,500	,577	1	,448	,320		

a. Variable(s) entered on step 1: avdiscl, sympsev.

Figure 15: Variables in the Equation Interpretation

On diagram above (see Figure 15), Avoidance of disclosure(*avdiscl*) is a positive and significant (b=.356, s.e.=.118, p=.002) predictor of the probability of early termination, with the Odds Ratio indicating that for every one unit increase on this predictor the odds of early termination change by a factor of 1.427 (meaning the odds are increasing).

Symptom severity(*sympsev*) is a negative and significant (b=-.317, s.e.=.123, p=.010) predictor of the probability of early termination. The Odds Ratio indicates that for every one unit increment on the predictor, the odds of terminating increase by a factor of .729 (meaning that the odds are decreasing)

5. Conclusion

Binary Logistic Regression was used to test the likelihood of early termination from counseling from a sample of clients from a mental health centre. A preliminary analysis suggested that the assumption of multicollinearity was met (tolerance=.96).

The model was statistically significant, x^2 *(2, N=45) =26.174, p<.001* indicating that the model fits the data significantly better than a null model and can model the likelihood of early termination correctly. The model explained between *44.1%* (Cox and Snell R square) and *59%* Nagelkerke R square) of the variance in the dependent variable and correctly classified *82.2%* of the cases. As shown in the table 1 below, symptom severity and avoidance of disclosure both contributes to the model.

ISSN: 2664-9640

	B	*SE*	*Wald*	*df*	*p*	*OR*	*CI for OR*	
							LL	UL
avdiscl	0.36	0.12	9.15	1	0.002	1.43	1.13	1.80
sympsev	-0.32	0.12	6.58	1	0.010	0.73	0.57	0.93
Constant	**-1.14**	**1.50**	**0.58**	**1**	**0.448**	**0.32**		

Figure 16: Logistic Regression predicting the likelihood of early termination

As the summary shown above (see Figure 16), Avoidance of disclosure (*avdiscl*) is positive and significant and the odds for a client to terminate early because of avoidance of disclosure are *1.43* times higher than those for not terminate early with a *95%* CI of *1.13* to *1.8*. Symptom severity (*sympsev*) is negative and significant and the odds for a client to terminate early because of symptom severity are 0.73 times lower than those for not terminate early with a *95%* CI of *0.57* to *0.93.*

References

[1] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. John Wiley & Sons.

[2] Allison, P. D. (2012). Logistic regression using SAS: theory and application. SAS Institute.

[3] Menard, S. (2002). Applied logistic regression analysis (Vol. 106). Sage.

[4] Pituch, K.A., & Stevens, J.A. (2016). Applied multivariate statistics for the social sciences (6th ed). New York: Routledge.

[5] Hosmer, D.W., Hosmer, T. and Lemeshow, S. (1980) A Goodness-of-Fit Tests for the Multiple Logistic Regression Model. Communications in Statistics, 10, 1043-1069.

[6] https://doi.org/10.1080/03610928008827941

[7] Nagelkerke, N.J.D. (1991) A Note on a General Definition of the Coefficient of Determination. Biometrika, 78, 691-692.

[8] https://doi.org/10.1093/biomet/78.3.691

[9] Cox, D. R., & Snell, E. J. (1989). Analysis of binary data (2nd ed.). Chapman and Hall.

[10] McFadden, D. (1974) Conditional Logit Analysis of Qualitative Choice Behavior. In: Zarembka, P., Ed., Economic Theory and Mathematical Economics, Academic Press, New York, NY, 105-142.

[11] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

[12] Hair Jr, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). Multivariate data analysis (7th ed.). Pearson Education Limited.

[13] Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. Journal of clinical epidemiology, 49(12), 1373–1379. https://doi.org/10.1016/s0895-4356(96)00236-3

[14] Field, A. (2018). Discovering statistics using IBM SPSS statistics (5th ed). Los Angeles: Sage.