Predicting Small-Medium-Enterprise Digital Success Using Polytomous Analysis: A Pilot Study

Gadang Ramantoko

Telkom University, Bandung, Indonesia, gadangramantoko@gmail.com

Abstract

The objective of the pilot study is to examine the predictability of digital transformation among Small and Medium Enterprises (SMEs) - using the same previous data collected – employing two step polytomous latent variable analysis technique. The dataset contains 415 observations and 13 variables. The independent latent variables are Company Presence (CompPres), User Connection (UserConn), Type of Media for Content Transmission (MediaType), User Content Personalization (UserContPerz), Store Presence (StorePres), Organizational Support (Support), Knowledge Development and Decision Making (KDDM), Marketing and Sales (MarkSale), Customers Relationship (CustRel), Internal Communications (IntComm), Ecosystem engagement (EcoEngg), and dependent latent variable is Digital Transformation Success (Success). Step one is clustering category manifest variables into classes grouped by independent and dependent latent variables. Step two, once latent variables class dataset is created, eight independent latent variables classes are selected. We then test eight one-tailed hypotheses concerning the relationship of eight independent latent variables to the successfulness of digital transformation, Success. Four out of eight variables proved to have significant impact to Success. The result of the study shows 59.3% overall prediction accuracy and 55.775% average class accuracy. Certain class has shown prediction accuracy as good as 71.6%. The applicability of this analysis technique to category dataset is confirmed and may find extension to other cases with similar category data.

Keywords: SME, Polytomous, Logistic-regression, Category data, Digital Success.

Introduction

We have previously reported the measurement of digital capability of several companies located at three business-community areas (Ramantoko, 2018). The data collected was ordinal and analyzed further using numerical method. From data point of view, a cross-sectional data collection is a portrait taken one time at the moment of investigation. However, when we dig deeper into the data, it seems that other technique of analysis may reveal another insight hidden inside the data. While the previous work measures digital maturity, the next question arises: Can we bring up a model to predict from the data, what is the probability that one SMEs member belong to a specific class of digital success?

This work is a pilot study to examine the predictability of digital transformation among Small and Medium Enterprises (SMEs) - using the same previous data collected - employing polytomous latent variable analysis technique. We test eight one-tailed hypotheses concerning the relationship of eight independent latent variables to the successfulness of digital transformation. The independent latent variables are Company Presence (CompPres), User Connection (UserConn), Type of Media for Content transmission (MediaType), User Content Personalization (UserContPerz), Store Presence (StorePres), Organizational Support (Support), Knowledge Development and Decision Making (KDDM), Marketing and Sales (MarkSale), Customers Relationship (CustRel), Internal Communications (IntComm), Ecosystem engagement (EcoEngg), and dependent latent variable is Digital Transformation Success (Success). The variables are as depicted in Appendix 1.
Methods

Polytomous Analysis of Data

Polytomous data analysis is the technique known in statistics to analyze category data. When using logistic regression model, it is also called multinomial logistic regression. Polytomous or multinomial logistic regression is an extension of binary logistic regression model. Polytomous logistic regression handles responses, $Y$, that are polytomous, i.e. the response votes-outcomes is more than 2 categories (dichotomous: occur or not occur, polytomous: multi-category response votes). While the input $X$ can be either numeric or category, or the mix of both. Dichotomous model takes the form of log of odds: $\log(\pi)/(1-\pi)$ or logit: $\logit(\pi)=\log(\pi)/\log(1-\pi)$.

When analyzing the polytomous response, we have to consider whether the scale we use is ordinal (category order) or nominal (not category order). For binary logistic model, this question does not arise. Some models fit only for ordinal response; for example, cumulative logits model, adjacent categories model, and continuation ratios models. The baseline model and logit model can be used for response ordinal as well as nominal. (psu.edu, 2018) In our case we use ordinal data (see Appendix 1). Ordinarily in response is important information as we regard it natural ordering. Using natural ordering can lead to a simpler, more parsimonious model and increase power to detect relationships with other variables. (psu.edu, 2018)

The Tool

For the tool of polytomous analysis, we use Polytomous variable Latent Class Analysis, package ‘poLCA’ in R language for categorical data written by Linzer et.al (2016). This tool serves as means to cluster category data. According to Linzer: “The basic latent class model is a finite mixture model in which the component distributions are assumed to be multi-way cross-classification tables with all variables mutually independent”. The model is called ‘latent structure analysis’ and was originally proposed by Lazarsfeld (1950). (dlinzer.github.com/poLCA, 2018) (Linzer et.al, 2016) For prediction we use logistic regression technique multinom in R package ‘nnet’.

Our data

Data collection is done using a questionnaire and is the same data that is being used in building Digital Maturity Index (Ramantoko et.al, 2018). Our data consists of observations as a result of sampling SMEs in area Suci, Binong and Rahayu. The data condition fits into model described in package poLCA: 1. It contains a number (17) of polytomous categorical variables (the “manifest” variables); 2. Each of manifest variable contains certain (4) possible outcomes for every individual in observation (415 observations). Each observation represents a SME’s member from the sample area where we study (business entity as unit analysis). The latent class model approximates the observed joint distribution of the manifest variables as the weighted sum of a finite number of constituent cross-classification tables. This finite number is fixed prior to estimation based on either theoretical reasons or model fit. The cross-classification between manifest variables and their appropriate latent variables follows the structure shown in Appendix 1.

Data analysis

The data analysis consists of two processes: Firstly, category clustering (cross-classification) of manifest variables; and secondly Prediction.

a. Category Clustering

First process is clustering (cross-classification) of manifest variables appropriate to each Latent Variable. In this process, we search for membership of an observation (hence a SME company) to a class. For example, Latent Variable CompPres is manifested by four polytomous variables CompPres1, CompPres2, CompPres3, and CompPres4. Each manifest variable contains 4 possible outcomes: None, Plan_to, SeldomUsed, and Operational. By applying poLCA formula, we assign 4 classes to each latent variable. The decision to take 4 classes is based on the ease of interpretation to avoid too many cross classifications among different manifest. Table 1 shows the result of clustering of latent variable CompPres. The numbers in shade show the greatest probability of manifest variables belonging to a certain class. The decision threshold is 60%. For manifest variable CompPres1, 77.61% probability of
Class1 to vote for SeldomUsed, 77.21% of Class2 to vote for Operational, 71.14% of Class3 to vote for None, and 79.8% probability of Class4 to vote for Plan_to. For manifest variable CompPres2, 78.09% probability of Class1 to vote for SeldomUsed, 100.0% of Class2 to vote for Operational, 71.39% of Class3 to vote for None (24.58%) and Plan_to (46.81%) and 87.39% probability of Class4 to vote for Plan_to (40.53%) and SeldomUsed (46.86%). Similar calculations are done for manifest variables CompPres3 and CompPres4. Summary of outcomes on Latent Variable CompPres are depicted in Table 2.

Figure 1. Estimation of the four-class latent class model CompPres using digital success data; obtained by setting graphs=TRUE in the poLCA function call. Each group of red bars represents the conditional probabilities, by latent class.

Figure 1. is a 3D image showing manifest variables [CompPres1, CompPres2, CompPres3, CompPres4]; array [1,2,3,4] are dummy numbers indicating ordinal outcomes: 1=None, 2=Plan_to, 3=SeldomUsed, and 4=Operational.

After several assessments on probabilities, we arrived at naming each class formed by manifest variables to latent variable as follows: Class1 is SeldomUsed, Class2 is Operational, Class3 is None plus, and Class4 is Plan_to. Summary of outcomes on latent variable CompPres is shown in Table 2. Note the word “plus” or “minus” accompanying the original outcomes. Those words indicate the confusion of taking decision to assign a class to original level (None, Plan_to, SeldomUsed, Operational) as the probability of a single class is less than 60%. We combine the numbers. None plus means that probability of a class belong to a vote “None” is less than 60%. We than combine with another number (probability of higher vote) so that both probabilities add up greater than 60%. That is where the additional word “plus” comes from. The same procedure applies to the additional word “minus”, where we achieve probabilities greater than 60% by adding another number or probability of lower vote.
Table 1: The result of application of poLCA clusterization on Latent Variable CompPres.

| CompPres1: Website | None | Operational | Plan_to | SeldomUsed |
|-------------------|------|-------------|---------|------------|
| class1            | 0.0182 | 0.1041 | 0.1017 | 0.7761    |
| class2            | 0.0000 | 0.7721 | 0.1400 | 0.0879    |
| class3            | 0.7114 | 0.0000 | 0.2213 | 0.0673    |
| class4            | 0.0113 | 0.0000 | 0.7980 | 0.1907    |

|$CompPres2: SocMed$

| None | Operational | Plan_to | SeldomUsed |
|------|-------------|---------|------------|
| class | 0.0000 | 0.1986 | 0.0205 | 0.7809 |
| class | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| class | 0.2458 | 0.0536 | 0.4681 | 0.2324 |
| class | 0.0037 | 0.1224 | 0.4053 | 0.4686 |

|$CompPres3: Blog/Forum$

| None | Operational | Plan_to | SeldomUsed |
|------|-------------|---------|------------|
| class | 0.0000 | 0.1297 | 0.1849 | 0.6854 |
| class | 0.0269 | 0.6145 | 0.2947 | 0.0638 |
| class | 0.3653 | 0.0000 | 0.4308 | 0.2039 |
| class | 0.0137 | 0.0000 | 0.7284 | 0.2578 |

|$CompPres4: Mobile Apps$

| None | Operational | Plan_to | SeldomUsed |
|------|-------------|---------|------------|
| class | 0.0713 | 0.0228 | 0.3630 | 0.5428 |
| class | 0.0864 | 0.4803 | 0.3676 | 0.0657 |
| class | 0.6719 | 0.0000 | 0.2022 | 0.1259 |
| class | 0.1676 | 0.0000 | 0.8245 | 0.0079 |

Table 2: Summary of outcomes on Latent Variable CompPres.

| Class1 | Class2 | Class3 | Class4 |
|--------|--------|--------|--------|
| SeldomUsed | Operational | None plus | Plan_to |
| SeldomUsed | Operational | None | Plan_to |
| SeldomUsed | Operational | None + Plan_to | Plan_to + SeldomUsed |
| SeldomUsed | Operational | None + Plan_to | Plan_to |
| Plan_to + SeldomUsed | Operational | None | Plan_to |

Figure 2: Estimated class population shares (ECPS) of latent variable CompPres.

Figure 2 shows Estimated Class Population Shares (ECPS). The majority of class is occupied by outcomes Plan_to (42%) and SeldomUsed (33%). The number of population shares for None plus and Operational are nearly equal, 14% and 11% respectively.
We do the same clustering process for the remaining independent latent variables CustRel, EcoEngg, IntComm, MarkSale, UserContPerz, KDDM, StorePres, and dependent variable Success. All latent variables are put into a dataset which is summarized in Table 3. It can be seen from Table 3, for example that, latent variable CompPres consists of 32.9% respondents' likelihood to vote for SeldomUsed, 11.48% for Operational, 14.08% for None plus, and 41.53% for Plan_to. For the other latent variables share of population, Table 3 is self-explained.

Table 3 Summary of naming and ECPS of latent variables

| Latent Variable | Class1 | ECPS | Name | Class2 | ECPS | Name | Class3 | ECPS | Name | Class4 | ECPS |
|-----------------|--------|------|------|--------|------|------|--------|------|------|--------|------|
| CompPres        | SeldomUsed | 0.3291 | Operational | 0.1148 | None plus | 0.1408 | Plan_to | 0.4153 |
| UserConn        | Plan_to | 0.294 | SeldomUsed | 0.4966 | Operational | 0.1194 | None | 0.09 |
| MediaType       | Operational minus | 0.3772 | SeldomUsed plus | 0.3766 | Plan_to | 0.1313 | None plus | 0.1149 |
| UserContPerz    | None plus | 0.0728 | Plan_to plus | 0.344 | SeldomUsed | 0.4173 | Operational | 0.1659 |
| StorePres       | SeldomUsed | 0.3922 | Plan_to | 0.3309 | None plus | 0.062 | Operational minus | 0.2149 |
| Support         | None | 0.1941 | Operational | 0.0332 | SeldomUsed | 0.2537 | Plan_to | 0.519 |
| KDDM            | Plan_to | 0.4317 | SeldomUsed | 0.3773 | None | 0.1226 | Operational | 0.0684 |
| MarkSale        | SeldomUsed | 0.4329 | Operational | 0.2182 | Plan_to | 0.2647 | None plus | 0.0842 |
| CustRel         | SeldomUsed minus | 0.2713 | None plus | 0.1542 | Operational minus | 0.1051 | Plan_to plus | 0.4695 |
| IntComm         | Operational minus | 0.159 | SeldomUsed minus | 0.3918 | Plan_to | 0.3148 | None plus | 0.1344 |
| EcoEngg         | Operational minus | 0.1514 | None | 0.0925 | SeldomUsed | 0.383 | Plan_to | 0.3731 |
| Success         | High | 0.4299 | Triumph | 0.1478 | Moderate | 0.2632 | Low | 0.1592 |

b. Prediction

The second process involves prediction. For prediction we exclude latent variables CompPres, UserConn, and MediaType for risk of multi collinearity. We learn the possibility of multicollinearity by studying the model which change erratically with the inclusion of the three mentioned variables. They produce redundancy to StorePres and CustRel. The basis dataset for prediction is ‘global_classPred’, i.e. the result of binding all latent variables’ classes data into one dataset. In this prediction, we test the relationship between each and all independent latent variables with dependent variable Success as follow:

H1: UserContPerz has a significant impact on Success;
H2: StorePres has significant impact on Success;
H3: Support has significant impact on Success;
H4: KDDM has significant impact on Success.
H5: MarkSale has significant impact on Success.
H6: IntComm has significant impact on Success.
H7: EcoEngg has significant impact on Success.
H8: CustRel has significant impact on Success.
H9: UserContPerz, StorePres, Support, KDDM, MarkSale, CustRel, IntComm, EcoEngg, Success, simultaneously, have significant impact on Success.

For analyzing category response data, we use multinomial generalized linear technique ‘multinom’ in nnet package in R (cran.r-project.org, 2017). Multinom technique is basically logistic regression for category response whose outcome is greater than 2. Multinomial Logistic Regression models how multinomial response variable Success (Y) depends on a set of explanatory variables shown in hypotheses H1 up to H11 (X=[X_1, X_2, ..., X_9]). This is a GLM where the random component assumes that the distribution of Y is Multinomial (n,π), where n is number of observation and π is a vector with probabilities of “success” for each category. The systematic components are discrete explanatory variables and are linear in the parameters, e.g., β_0 + βx_1 + ... + βx_9. The result is as follows:
Table 4 Result of running package nnet::multinom.

| Number of predictors: 8 |
|------------------------|
| Residual Deviance: 790.0017 |
| AIC8: 940.0017 |
| Log likelihood: -395.001 (75 df) |
| Pseudo R-Square8: 0.34813594 |

==== ANOVA ====
Analysis of Deviance Table (Type II tests)
Response: Success
---
Signif. codes: 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

| LR          | Chisq | Df | Pr(>Chisq) |
|-------------|-------|----|------------|
| CustRel     | 21.648| 9  | 0.0100643 *|
| EcoEngg     | 12.573| 9  | 0.1828921  |
| IntComm     | 5.512 | 9  | 0.7875486  |
| MarkSale    | 30.330| 9  | 0.0003854 ***|
| UserContPerz| 33.324| 9  | 0.0001173 ***|
| KDDM        | 6.833 | 9  | 0.6545149  |
| Support     | 5.435 | 9  | 0.7948685  |
| StorePres   | 17.733| 9  | 0.0383954 * |

From Table 4, we see that variables CustRel, MarkSale, UserContPerz and StorePres have P-value (Pr(>Chisq)) lower than 0.05. We can conclude that within 0.95 confidence level, hypotheses H1, H4, H5, and H7 are accepted, while H2, H3, H6, H7, having P-value greater than 0.05, are rejected. Redo the prediction using only 4 accepted predictors provides us with the result depicted in Table 5. Anova test verifies hypotheses H9, and then H9 is accepted.

Table 5 Result of rerunning package Table 5 using only CustRel, MarkSale, UserContPerz, and StorePres.

| Number of predictors: 4 |
|------------------------|
| Residual Deviance: 824.3051 |
| AIC4: 902.3051 |
| Log likelihood: -412.153 (39 df) |
| Pseudo R-Square4: 0.29827109 |

==== ANOVA ====
Analysis of Deviance Table (Type II tests)
Response: Success
---
Signif. codes: 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

| LR          | Chisq | Df | Pr(>Chisq) |
|-------------|-------|----|------------|
| CustRel     | 23.481| 9  | 0.005201 **|
| MarkSale    | 42.053| 9  | 0.0000032135 ***|
| UserContPerz| 44.988| 9  | 0.000009276 ***|
| StorePres   | 19.908| 9  | 0.018489 * |
We observe from Table 4 and Table 5 that rerunning the same script `nnet::multinom` incorporating only the significant latent variables improves the model (shown by decreasing AIC from AIC8: 940.0017 to AIC4: 902.3051) but worsen the variability (shown by decreasing R² from 0.35 to 0.3).

Results and Discussion

Up to this moment, we have shown how to make prediction from category dataset. The dataset contains both category input as well as response. To do prediction, first, we must calculate the probability of each respondent to indirectly vote for a certain latent variable outcome based on its votes in manifest variables. The manifest independent latent variable has four ordinal outcomes or votes: None, Plan_to, SeldomUsed, and Operational. Each of these outcomes represents a digital condition or state of the progress the respondent is currently in for the case asked in the questionnaire. The answers of 415 respondents are organized into single category dataset. We call it category dataset because it contains all category input (independent manifest variables) and category response (dependent manifest variables). All independent manifest variables form 12 independent latent variables, while all dependent manifest variables form 1 dependent latent variable. (See appendix 1) Using polytomous technique, we calculate the likelihood a single respondent will be associated with a certain outcome of its appropriate latent variable. Note that the outcomes of a certain latent variable are built by a combination of votes by the respondents manifested during the fill in of the questionnaire. For example, in latent variable 1, respondent 1 vote will have a combination of none for manifest variable 1, SeldomUsed for manifest variable 2, Plant_to for manifest variable 3, and none for manifest variable 4. Respondent 2 will vote for different combination. And it continues up to combination of the 415th respondent. Latent variables formed by manifest variables are all category. The latent variable outcomes resulted from manifest variables can take 11 forms, that are None, None plus, Plan_to minus, Plan_to, Plan_to plus, SeldomUsed minus, SeldomUsed, SeldomUsed plus, Operational minus, Operational, Operational plus. Note also the additional word minus and plus accompanying the original four manifest outcomes. The summary of the dataset is summarized in Table 3. For prediction we omit 3 variables for risk of multicollinearity. When running package `poLCA`, we learn that the result varies from one run to other run. To guarantee consistency, we run 500 iterations for each latent variable, and take the majority of occurrence as the final data. The example case for latent variable `CompPres` is shown in Table 6. In this example, we let `poLCA` calculate the likelihood for 5 classes. We can see that different iteration reveals different data set. However, nearly 50% (231) occurrence is respecitve to maximum likelihood -1642. For this illustration we take the likelihood -1642 as dataset.

Table 6  The example case for latent variable `CompPres` set for 5 classes, run for 500 iterations and grouped by maximum log-likelihood.

| maximum | `CompPres` Class 1 | `CompPres` Class 2 | `CompPres` Class 3 | `CompPres` Class 4 | `CompPres` Class 5 | occurrence |
|---------|------------------|------------------|------------------|------------------|------------------|------------|
| -1653   | 0.147468872798014 | 0.283410604225341 | 0.409786228541508 | 0.026794422357781 | 0.132536852199359 | 1          |
| -1650   | 0.1481832211363584 | 0.323534256972792 | 0.389331261945287 | 0.0441702120001382 | 0.0947810479451998 | 6          |
| -1649   | 0.13757242298835 | 0.32078099941368 | 0.404599828484324 | 0.10420928927824  | 0.03209240529203  | 30         |
| -1648   | 0.123239585359932 | 0.376982381122659 | 0.332091968586869 | 0.0679363150594389 | 0.098575214953697 | 10         |
| -1647   | 0.139143414057928 | 0.257569455709333 | 0.385894021646871 | 0.11989677254806  | 0.097506095917822 | 14         |
| -1646   | 0.120972583291179 | 0.385225432291388 | 0.250782397396015 | 0.12804722808224  | 0.114971812665372 | 12         |
| -1644   | 0.0993616640929366 | 0.127624574851284 | 0.39633553630622  | 0.247148612519398 | 0.129529777763759 | 19         |
| -1643   | 0.0924567289374879 | 0.213707583655146 | 0.26949120271544  | 0.310599877553122 | 0.113704689060402 | 95         |
| -1642   | 0.0775096232321089 | 0.15086778148233  | 0.290427272724826 | 0.36703705209128  | 0.11412819802912 | 231        |
| -1641   | 0.0975285951621973 | 0.0406697180559412 | 0.41849519506453 | 0.328793394258419 | 0.11451336161699 | 80         |

Second, once the category latent variable dataset has been built, we do the prediction, again using polytomous technique. This polytomous technique consistently regard the input as well as response variable as being category. If we run `nnet::multinom` individually, we get significant relationship for all pairs of for latent input-response with response being latent variable `Success`. If we run simultaneously all independent variables to dependent variable `Success`, we find out that significant relationship is only applicable for 4 exogenous variables: CustRel, MarkSale, UserContPerz and StorePres. These 4 exogenous variables reveal better model given lower AIC compare to regression using all (8) exogenous variables. We inspect the 4 exogenous model and we have as well better
accuracy. The accuracy of prediction (full train dataset) is given in Table 7. We see from Table 7 that the model provides best prediction for Low. Second best is High while it is not very good result for Triumph and Moderate. Overall error is 40.7%, and Averaged class error is 44.225%.

|      | High  | Low  | Moderate | Triumph | Error |
|------|-------|------|----------|---------|-------|
| Actual|       |      |          |         |       |
| High  | 144   | 16   | 4        | 16      | 20.0  |
| Low   | 7     | 53   | 5        | 0       | 18.5  |
| Moderate | 54   | 23   | 26       | 6       | 76.1  |
| Triumph | 31   | 4    | 3        | 23      | 62.3  |

Table 7 Evaluation of prediction result.

In addition to logistic regression, we check variable importance using random forest technique in R-package Random Forrest. Both technique shows acceptable agreement. Random forest outputs MarkSale, UserConPerz, StorePres, IntComm, CustRel as 5 most important variables in predicting Success.

Next, we examine the distribution of each latent variables group. Figure 3 shows the distribution of exogenous latent variable CustRel grouped by Success. From the figure we see that 87 votes for Plan_to plus, 57 votes for SeldomUsed minus, 9 votes for none plus, and 27 votes for Operational minus contribute to SuccessHigh. Same reasoning can be done for SuccessLow, SuccessModerate and SuccessTriumph. And the same reasoning also is applicable throughout all exogenous latent variables. The extended visualization is depicted in Appendix 4. The results suggest that each Success-outcome is formed by unique or nearly orthogonal combination of each latent-variable-outcome. The interpretation of the result is that if a member of SMEs vote profile is known, then we can expect the company to belong the a certain Success outcome whose accuracy is given by Table 7.

![Figure 3 Contribution of votes to Success-outcomes.](image)

Conclusions

This study has examined the use of polytomous analysis technique to predict digital successfulness of SMEs. The analysis is carried out in two steps. Step one is clustering of category latent variables by their appropriate manifest variables. The result of step one analysis is put into category dataset with four outcomes for each latent variable. We used to employ numeric regression for this type of problem by regarding ordinal data as numeric. However, we see the data as category instead of numeric. The prediction demonstrates moderate overall error, but it predicts sufficiently accurate for Success-high and Success-low.

References

Ramantoko, G., Fatimah, L.V, Pratiwi, S.C and Kinasih, K. (2018) Measuring Digital Capability Maturity: Case of Small-Medium Kampong-Digital Companies in Bandung, Pertanika J. Soc. Sci. & Hum. 26 (T): 215 - 230
Polytomous (Multinomial) Logistic Regression. (2018, October). Retrieved 2018, from https://onlinecourses.science.psu.edu/stat504/node/172/ Drew A. Linzer, J. L. (n.d.). Polytomous Variable Latent Class Analysis (R package). Retrieved 2018, from http://dlinzer.github.io/poLCA/

Linzer, D.A. & Lewis, J.B., poLCA: An R Package for Polytomous Variable Latent Class Analysis, Journal of Statistical Software MMMMMM YYYY, Volume VV, Issue II., http://www.jstatsoft.org/