Statistical Model of Role of Communication to Advertising the Trade Mark and Admission of a University in Thai Nguyen City

Thi Hue Tran[1], Thi Quynh Nhung Ngo[2]

[1] Faculty of International Training, Thai Nguyen University of Technology, Thai Nguyen City, Viet Nam

[2] Faculty of Postal Profession Training, Thai Nguyen College of Economics and Finance, Thai Nguyen City, Viet Nam

Abstract

The role of communication to trade mark of an university seems to be more and more effective nowadays. There are many channels of communication that impart to the reputation of a professional training agency, a university. This problem will be studied here addressed to professional traing agency in Thai Nguyen City, Viet Nam. It can not be denied that there are many factors that affect to the decision of customers, in this situation, it is their decision on choosing a university to follow their future career with which they are planning to work. This paper will show the relation among some important factors to that of universities in Thai Nguyen City. This is incomparable to universities in different cities because of the culture reason and geographical region. The result is expected to provide a scienctific scene to managements to plan for the future of educational system in the city.

Keywords: *role of communication, trade mark of a univeristy, advertising, training sector, admission.*

1. Introduction

The trade mark and the reputation of a university in Thai Nguyen City has been affected by many factors, such as an old customer which can be considered as a traditional effect; a staff who is working in the university or geographical region which can be considered as the local effect; the tendency of future careers which depends strictly on the development of the economics of the country, especially in the provinces near Thai Nguyen City; the cutting-edge technologies relating to industrial sector in the country; the passion of custumers. These factors come at once in creating a logistic model representing their relation. It is uncontroverted to predict the new outcome on the basis of the model constructed. Therefore, the model becomes much helpful to one who wants to learn the facts from this problem and may want to intervene, improve, the outcome.

To the customers, here students or probably their parents, their demands and tastes depend strongly on their tastes of the mass communication, and the mass media, which have been exploded terribly nowadays. A study of these effects is an essential evaluation and determine their role on the impart to the customers' attitude. Besides, to make the comparison between the effects of the communication factors and those caused by other factors, the study must also include other factors which chosen from the view point of statisticians who work on this study. And these factors are chosen somewhat reasonable to make their prediction being shed light on. The factors authors want to study in this work including the feedback from old customers who have studied or are studing in the current university; the people who are working in the university which is considered as a minor factor; the information the customers have gained some where through mass media; the desired future career the customers wish to work with their choice based on their knowledge of the development of economics and technology; the geographic location of the university and their hometown in the connection to the living condition of their family. Because these informations are in type of catalog, either nominal or ordinal variables, the model constructed is set up in the form of a multinomial logistic one. The quality of the constructed model is evaluated basing on certain crucial criterions which are popular used by statisticians.

The data collected to construct the model from various universities around the city. So, the geographic region factor is taken into account in this model. The reason of this consideration comes from the actual economic situation and the culture of the region. The economic condition of the city is lower than that of the surrounding cities, especially to cities in Hong River Delta, in the North of the country, which have very high competion in attracting the resource of foreign investment. However, the traditional advantage of the city in comparing to other cities in the north except capital Ha noi is presented in the ungraduate education system with various universities which have the very long history, inherited from the accomplishment of the national revolution. In fact, the number of universities and colleges existed in the whole city is ranked number 5 of all cities in the country. Besides, Thai Nguyen University (TNU) is one on the top 17 of unversities in the country which is organized into 11 reputed school and university members. Therefore, traditional customers, students who has been trained here keep a crucial role in advertising the trade mark of universities and colleges in the city. Data was collected randomly from current customers from various universities and colleges over the city, including:

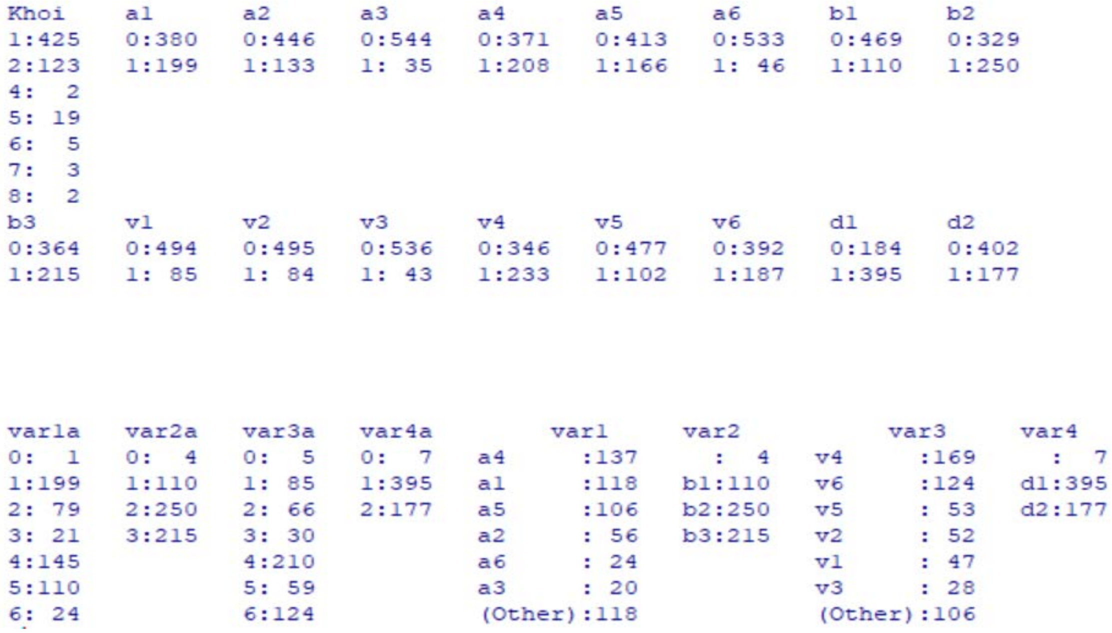
- Thai Nguyen University of Technology (a member of TNU),
- Thai Nguyen University of Education (member of TNU),
- Thai Nguyen Unversity of Agriculture and Foresty (of TNU),
- Thai Nguyen University of Information and Communication Technology (of TNU),
- Thai Nguyen University of Medicine and Pharmacy (of TNU),
- Thai Nguyen Unversity of Economics and Business Administration (of TNU),
- School of Forgein Languages (of TNU),
- International School (of TNU),
- Thai Nguyen College of Economic and Finance,
- Thai Nguyen Medical College,
- College of Commerce and Tourism,
- Thai Nguyen College of Education.

Data collected from 579 students from thoses universities and colleges. Method of collection was performed by distributing a survey to individuals who are directly the object of this study. The survey form is represented at the end of this paper.

2. Data Summary, Sample Space and Statistical Variables

2.1 Sample Space and Data Summary

Sample space is the collected data from the 579 students who is studying in 12 universities and colleges (called universities in common) listed above around Thai Nguyen city. This sample space was taken randomly from the population of all current students studying the city. These customers were distributed in whole city. The summary of the data is showed in Table 1.



```
Khoi     a1       a2       a3       a4       a5       a6       b1       b2
1:425    0:380    0:446    0:544    0:371    0:413    0:533    0:469    0:329
2:123    1:199    1:133    1: 35    1:208    1:166    1: 46    1:110    1:250
4:  2
5: 19
6:  5
7:  3
8:  2
b3       v1       v2       v3       v4       v5       v6       d1       d2
0:364    0:494    0:495    0:536    0:346    0:477    0:392    0:184    0:402
1:215    1: 85    1: 84    1: 43    1:233    1:102    1:187    1:395    1:177



varla    var2a    var3a    var4a          var1      var2           var3      var4
0:  1    0:  4    0:  5    0:  7    a4       :137     :  4    v4       :169     :  7
1:199    1:110    1: 85    1:395    a1       :118   b1:110    v6       :124   d1:395
2: 79    2:250    2: 66    2:177    a5       :106   b2:250    v5       : 53   d2:177
3: 21    3:215    3: 30             a2       : 56   b3:215    v2       : 52
4:145             4:210             a6       : 24             v1       : 47
5:110             5: 59             a3       : 20             v3       : 28
6: 24             6:124             (Other):118               (Other):106
```

Table 1. Data summary from the sample space.

2.2 Statistical Variables

Variables collected are explained in this subsection. Their name, type and meaning are mentioned bellow as long as with Frequency Table.

- *Khoi*: the number of years student had been trained in his/her university (denoted by UNI).

- *var1*: the first source of information they had gotten UNI. This catagorical variable has the following possible values

a1	*Since having blood relatives or friends who had been trained in UNI.*
a2	*Since having blood relatives or friends who are training in UNI.*
a3	*Since having blood relatives or friends who are working in UNI.*
a4	*Since you lived in Thai Nguyen, you had known UNI when studing in high school.*
a5	*Since the information provided in Brochure about UNI when you tried to make a choice on your future study/career at the end of your high school grades.*
a6	*Since you had been seen it by chance on social media.*

- *var2*: Distance from where your family living to UNI (a categorical ordinal variable) including the following values:

b1:	*Less than 15 kilometers*
b2:	*From 15-50 kilometers*
b3:	*Further than 50 kilometers*

- *var3:* the most important reason made you choose to study in UNI:

v1	*Since your blood relatives or friends who had been trained in UNI advised/persuaded you.*
v2	*Since your blood relatives or friends who are being training in UNI advised/affected you.*
v3	*Since your blood relatives who had not been or are not being trained in UNI advised/persuaded you.*
v4	*Since your self-decision making based on your knowledge of career opportunities and your passion.*
v5	*Since your prior calculation to the expense on which you might spend for your future study and living in UNI was less than that in other univeristies in the country which have the same program.*
v6	*Since the admission requirements of UNI are suitable for your GPA in high school.*

- *var4*: would you change your decision if you had a chance to choose the university again, the options are:
  d2: still choose to study in UNI.
  d1: would change your decision.

The variables from *a1-a6, v1-v6, d1-d2* are all normial binary catagorical variables with two possible values: 0 and 1 (chosen and not chosen, respectively). The variables from *b1-b3* are ordinal binary catagorical variables with two possible values 0 and 1 corresponding to the options chosen and not chosen. These variables represent how far it is from where students' family is living to the UNI. The mediate variables from *var1a-var4a* are in the type of nomial catagorical variables which are use to represent the obtained values from variables *a1-a6, b1-b3, v1-v6, d1-d2*. Similarly, the variables *var1-var4* are nomial catagorical variables used to summarize the values of variables *var1a-var4a*.

2.3 Data exploration

The data collected is summarized in frequencies table for a particular variables being used in the model. Let make a close look on this summation.

```
DESCRIPTIVES
Descriptives
 ------------------------------------------------------------------------------------------------------
          a1    a2    a3    a4    a5    a6    var2   v1    v2    v3    v4    v5    v6
 ------------------------------------------------------------------------------------------------------
   N         579   579   579   579   579   579    579   579   579   579   579   579   579
   Missing     0     0     0     0     0     0     0     0     0     0     0     0     0
   Mean
   Median
   Minimum
   Maximum
 ------------------------------------------------------------------------------------------------------
```

```
                                                 --------------------------------------------------
                                                  Levels   Counts   % of Total   Cumulative %
FREQUENCIES                                      --------------------------------------------------
Frequencies of a1                                 0          380     65.63040      65.63040
```

```
  1        199     34.36960      100.00000
--------------------------------------------------
Frequencies of a2
--------------------------------------------------
  Levels    Counts    % of Total    Cumulative %
--------------------------------------------------
  0        446     77.02936       77.02936
  1        133     22.97064      100.00000
--------------------------------------------------
Frequencies of a3
--------------------------------------------------
  Levels    Counts    % of Total    Cumulative %
--------------------------------------------------
  0        544     93.95509       93.95509
  1         35      6.04491      100.00000
--------------------------------------------------
Frequencies of a4
--------------------------------------------------
  Levels    Counts    % of Total    Cumulative %
--------------------------------------------------
  0        371     64.07599       64.07599
  1        208     35.92401      100.00000
--------------------------------------------------
Frequencies of a5
--------------------------------------------------
  Levels    Counts    % of Total    Cumulative %
--------------------------------------------------
  0        413     71.32988       71.32988
  1        166     28.67012      100.00000
--------------------------------------------------
Frequencies of a6
--------------------------------------------------
  Levels    Counts    % of Total    Cumulative %
--------------------------------------------------
  0        533     92.05527       92.05527
  1         46      7.94473      100.00000
--------------------------------------------------

Frequencies of var2
--------------------------------------------------
  Levels    Counts    % of Total    Cumulative %
--------------------------------------------------
                     0.00000        0.00000
  b1        110     18.99827        0.00000
  b2        250     43.17789        0.00000
  b3        215     37.13299        0.00000
--------------------------------------------------
Frequencies of v1
--------------------------------------------------
  Levels    Counts    % of Total    Cumulative %
--------------------------------------------------
  0        494     85.31952       85.31952
  1         85     14.68048      100.00000
--------------------------------------------------
Frequencies of v2
--------------------------------------------------
  Levels    Counts    % of Total    Cumulative %
--------------------------------------------------
  0        495     85.49223       85.49223
  1         84     14.50777      100.00000
--------------------------------------------------
Frequencies of v3
--------------------------------------------------
  Levels    Counts    % of Total    Cumulative %
--------------------------------------------------
  0        536     92.57340       92.57340
  1         43      7.42660      100.00000
--------------------------------------------------
Frequencies of v4
--------------------------------------------------
  Levels    Counts    % of Total    Cumulative %
--------------------------------------------------
  0        346     59.75820       59.75820
  1        233     40.24180      100.00000
--------------------------------------------------
Frequencies of v5
--------------------------------------------------
  Levels    Counts    % of Total    Cumulative %
--------------------------------------------------
  0        477     82.38342       82.38342
  1        102     17.61658      100.00000
--------------------------------------------------
Frequencies of v6
--------------------------------------------------
  Levels    Counts    % of Total    Cumulative %
--------------------------------------------------
  0        392     67.70294       67.70294
  1        187     32.29706      100.00000
--------------------------------------------------
```

Observation: From Frequencies table, we see that the variables from *v1-v6* are skewed between two possible values each might takes. Therefore, sampling process performed to get data with a balance distribution between these two values for each variable is important for the construction of an exact model taking into account of these variables to be a dependent variable. To do so, we will divide the data into two data set: training and testing data. We first construct the model on the training data, then check that model on the testing data.

3. Model Construction

3.1 Training and Testing Data for a model with the dependent variable/respond *v6* and independent variables from *a1-a6, var2, v1-v5.*

To construct this kind of model, we used the function `set.seed(100)`in R, subdividing the data of *v6* into two subsets. In which we sample 80% of 1's values that quantity for 0's values to create the training data. Keeping the remaining data as testing data. We can summarize the work in the following code.

```
# Create Training Data
input_ones <- mdata0[which(mdata0$v6 == 1), ]  # all 1's
input_zeros <- mdata0[which(mdata0$v6 == 0), ]  # all 0's
set.seed(100)  # for repeatability of samples
input_ones_training_rows  <-  sample(1:nrow(input_ones),  0.8*nrow(input_ones))    # 
1's for training
input_zeros_training_rows <-  sample(1:nrow(input_zeros), 0.8*nrow(input_ones))   # 
0's for training. Pick as many 0's as 1's
training_ones <- input_ones[input_ones_training_rows, ]
training_zeros <- input_zeros[input_zeros_training_rows, ]
trainingData <- rbind(training_ones, training_zeros)  # row bind the 1's and 0's
# Create Test Data
test_ones <- input_ones[-input_ones_training_rows, ]
test_zeros <- input_zeros[-input_zeros_training_rows, ]
testData <- rbind(test_ones, test_zeros)  # row bind the 1's and 0's
```

Training Data has the description for the respond *v6* is given in the following table:

```
DESCRIPTIVES
Descriptives
 ----------------------------------------------------------------------------------------------------------
         a1    a2    a3    a4    a5    a6    var2   v1    v2    v3    v4    v5    v6
 ----------------------------------------------------------------------------------------------------------
   N        298   298   298   298   298   298    298   298   298   298   298   298   298
   Missing    0     0     0     0     0     0      0     0     0     0     0     0     0
   Mean
   Median
   Minimum
   Maximum
 ----------------------------------------------------------------------------------------------------------
Frequencies of v6
 ----------------------------------------------------
   Levels    Counts    % of Total    Cumulative %
 ----------------------------------------------------
   0          149     50.00000       50.00000
   1          149     50.00000      100.00000
 ----------------------------------------------------
```

3.2 The logistic model without independent variables/ predictors, OIM

For such model, only *v6* is the respond, there is no predictor. This model with only intercept helps us having a close look to regression constant. Name this model OIM.

```
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.177  -1.177   0.000   1.177   1.177
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.859e-17  1.159e-01       0        1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 413.12  on 297  degrees of freedom
Residual deviance: 413.12  on 297  degrees of freedom
```

```
AIC: 415.12
Number of Fisher Scoring iterations: 2
```

3.3 The logistic model with the richest predictors, modelv6.0

In this model, denoted modelv6.0, we consider the predictors from *a1-a6*, *var2*, and from *v1-v5*. Here, we collect all predictors we get to make the prediction on the respond, *v6*. So, by this, we expect to get some clues connecting *v6* and other predictors. If we succeed, we can be able to explain the effect of other explanatory variables on *v6*.

```
glm(formula = v6 ~ a1 + a2 + a3 + a4 + a5 + a6 + var2 + v1 +
    v2 + v3 + v4 + v5, family = binomial(link = "logit"), data = trainingData)
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.82263  -0.85544   0.02544   0.74574   2.72442
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.3190     1.2352  -0.258 0.796230
a11           1.2775     0.3489   3.662 0.000251 ***
a21           0.8855     0.3866   2.290 0.021999 *
a31           1.8244     0.6341   2.877 0.004016 **
a41           1.3358     0.3683   3.627 0.000287 ***
a51           1.7932     0.3586   5.001 5.71e-07 ***
a61           0.8405     0.5577   1.507 0.131757
var2b1       -0.5932     1.2905  -0.460 0.645730
var2b2        0.1208     1.2457   0.097 0.922732
var2b3        0.3050     1.2437   0.245 0.806250
v11          -1.2866     0.5009  -2.569 0.010212 *
v21          -1.9546     0.4404  -4.438 9.09e-06 ***
v31          -1.3459     0.6907  -1.949 0.051339 .
v41          -2.4101     0.3452  -6.982 2.92e-12 ***
v51          -1.6870     0.4201  -4.016 5.92e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 413.12  on 297  degrees of freedom
Residual deviance: 310.53  on 283  degrees of freedom
AIC: 340.53
Number of Fisher Scoring iterations: 4
```

From the R output of this model, we indeed succeeded in ruling out the variables *a6*, *var2* which seem to be redundant when we tried to include them into the model. Now, we are ready to learn from this by another try of excluding these predictors from the model.

3.4 The improvement model from the richest predictor model, denoted by modelv6

This model is constructed based on the improved belief from the modelv6.0. In fact, the R output tells us that we are right, all predictors in the new model constructed are significant.

```
glm(formula = v6 ~ a1 + a2 + a3 + a4 + a5 + v1 + v2 + v3 + v4 +
    v5, family = binomial(link = "logit"), data = trainingData)
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.75053  -0.90954  -0.02082   0.77810   2.72112
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.1707     0.2812  -0.607 0.543851
a11           1.2267     0.3409   3.598 0.000321 ***
a21           0.9505     0.3731   2.548 0.010835 *
a31           1.9621     0.6273   3.128 0.001760 **
a41           1.0635     0.3206   3.317 0.000909 ***
a51           1.8417     0.3552   5.186 2.15e-07 ***
v11          -1.2725     0.4896  -2.599 0.009353 **
```

```
v21           -1.8991     0.4340  -4.375 1.21e-05 ***
v31           -1.3252     0.6726  -1.970 0.048810 *
v41           -2.3157     0.3346  -6.920 4.52e-12 ***
v51           -1.5617     0.4038  -3.868 0.000110 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 413.12  on 297  degrees of freedom
Residual deviance: 316.11  on 287  degrees of freedom
AIC: 338.11
Number of Fisher Scoring iterations: 4
```

The comparision can simply be made to verify the advantage of modelv6 and OIM by using Chisquared test. The test is shown in the following lines. The p-value of the test is less than 2.2×10^{-16} which is very significant.

```
Analysis of Deviance Table
Model 1: v6 ~ 1
Model 2: v6 ~ a1 + a2 + a3 + a4 + a5 + v1 + v2 + v3 + v4 + v5
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       297     413.12
2       287     316.11 10   97.009 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Bisides, the percentage misclassification error of modelv6 on Testing data is not too high, nearly 10.32%, which can be trusted. The result can be discovered using the function `misClassError` in R. The confussion matrix of the model is also can be used to make this point clearer. In the out put of the function `confusionMatrix`, the columns show the actual value, and the rows show the predicted values.

```
predicted <- predict(modelv6, testData, type="response")  # predicted scores
library(InformationValue)
optCutOff <- optimalCutoff(testData$v6, predicted)[1] #0.6490757
misClassError(testData$v6, predicted, threshold = optCutOff)#0.1032
confusionMatrix(testData$v6, predicted, threshold = optCutOff)
    0  1
0 226 12
1  17 26
```

From this confusion matrix, one can find out the following measurements of the model modelv6.

$$ACURACY = \frac{True\ Positive + True\ Negative}{Total} = \frac{26+17}{26+17+226+12} \approx 0.8968.$$

$$\text{PRECISION} = \frac{True\ Positive}{True\ Positive + False\ Positive} = \frac{26}{26+12} \approx 0.6047.$$

$$\text{RECALL} = \frac{True\ Positive}{True\ Positive + False\ Negative} = \frac{26}{26+17} \approx 0.6842.$$

The concordance of modelv6 is shown below.

```
$Concordance
[1] 0.7964046
$Discordance
[1] 0.2035954
$Tied
[1] 2.775558e-17
$Pairs
[1] 9234
```

The following graph shows AUROC curve of modelv6 on testData.

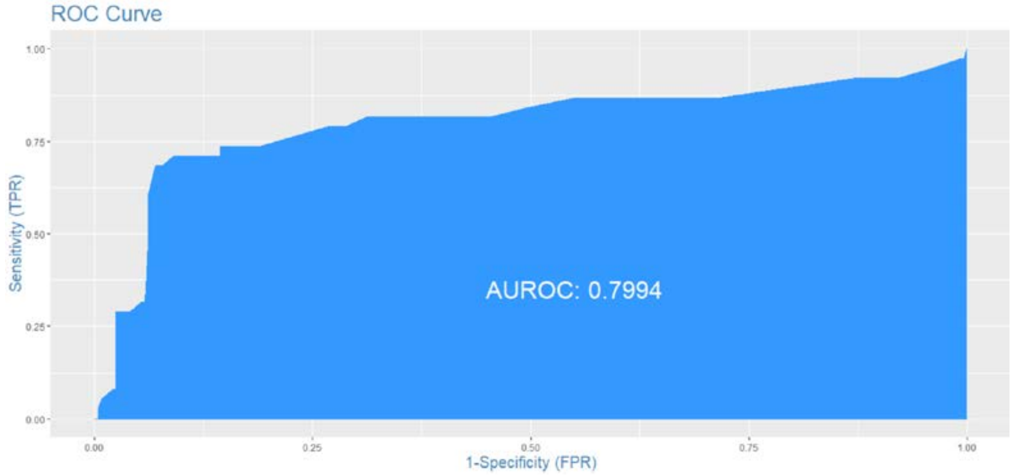


Figure 2. AUROC curve of modelv6 on testData.

Multicolinearity of modelv6 is presented with VIF

```
vif(modelv6)
      a1       a2       a3       a4       a5       v1       v2       v3 
1.360337 1.199931 1.276167 1.232960 1.401815 1.218761 1.252557 1.126430 
      v4       v5 
1.316565 1.169497 
```

Cook's distance: The Cook's distance outlines observations which exceeds the value

$$\frac{4}{size(trainingData) - \#(predictors) - 1} = \frac{4}{298 - 10 - 1} = 0.01394$$

as an outliner. Therefore, the number of outliners is about 30 observations in trainingData (making 13.42%).

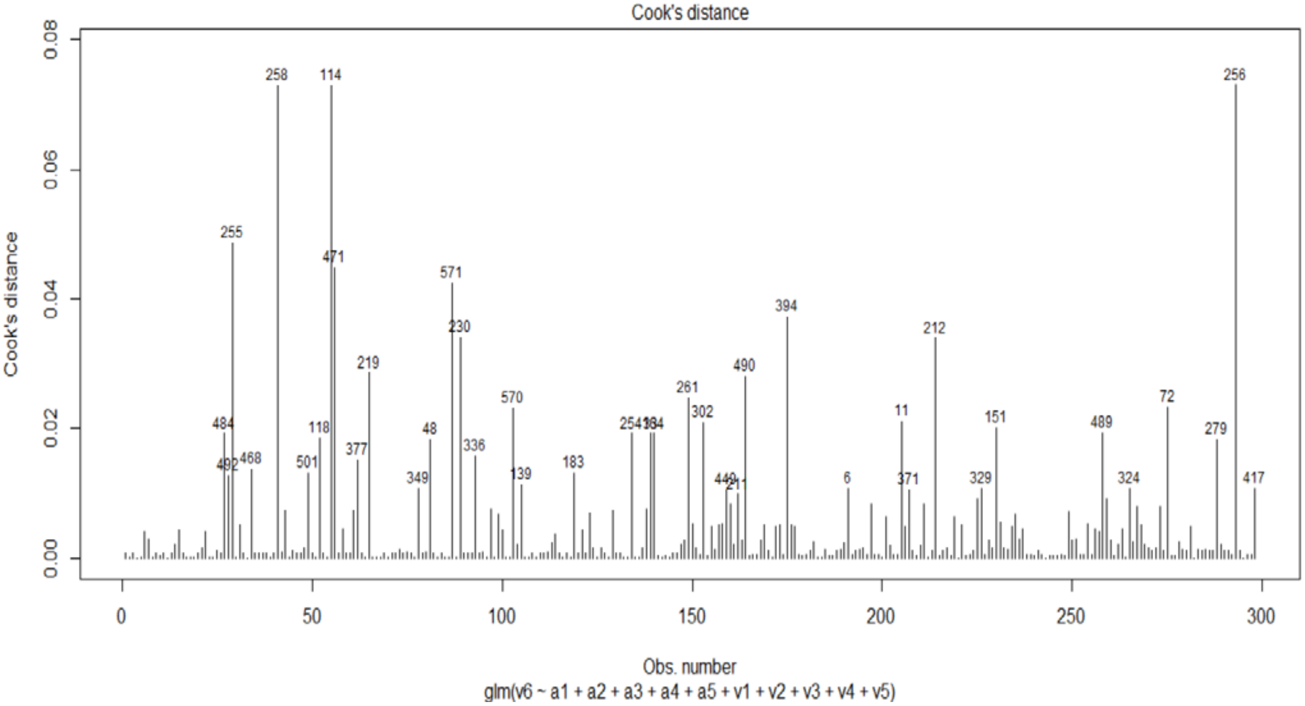


Figure 3. Cook's distance on trainingData.

The Figure 4 shows the ROC of modelv6 on trainingData. The AUROC point of 0.823 is higher than that of the model assessed on testData (which is 0.7994). This is caused by the little lack of data provided. However, the difference is not too high. So, there is no serious matter that could make the confidence to the correction of our modelv6 declined.

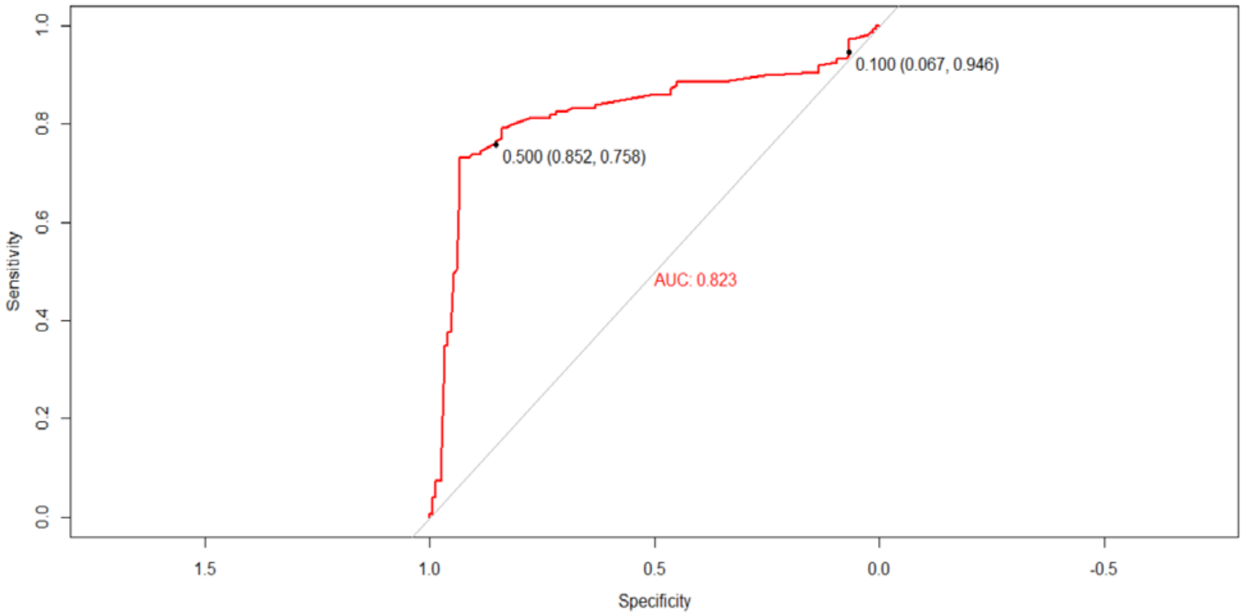


Figure 4. The ROC of modelv6 on trainingData.

3.5 The Assessments of modelv6

Binned residual plot: The Binned residual plot encloses only 7 out of 15 in total of the observations. However, those which standing out is not too far from the boundary of the confidence limit. This performance is only one of many criteria to assess the overall fit of binary regression model. We can see other criteria bellow.

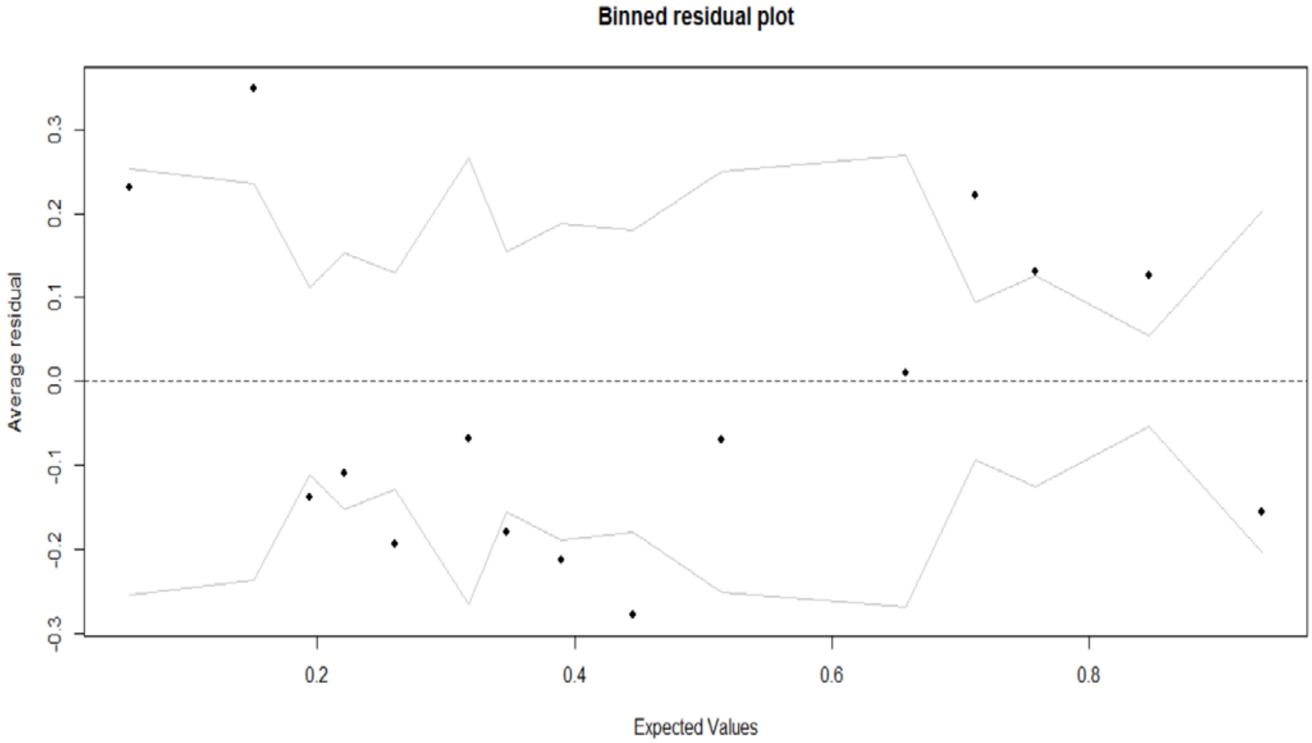


Figure 5. Binned residual plot of modelv6.

Model fitting of modelv6:

```
> PseudoR2(modelv6, which = c("CoxSnell","Nagelkerke","McFadden"))
  CoxSnell Nagelkerke   McFadden
```

```
 0.2778580  0.3704774  0.2348228
>chisq.test(trainingData$v6,predict(modelv6))
        Pearson's Chi-squared test
data:  trainingData$v6 and predict(modelv6)
X-squared = 228.79, df = 85, p-value = 3.98e-15
```

The AIC score of modelv6: (AIC standing for Akaike Information Criteria)

```
AIC(modelv6)=338.11, AIC(modelv6.1)=338.39, AIC(modelv6.0)=340.53.
```

Wald Test for assessing the importance of a particular variable of modelv6: The output below shows the result of Test.

```
> wald.test(b=coef(modelv6),Sigma=vcov(modelv6), Terms=2)
Wald test:
----------
Chi-squared test:
X2 = 12.9, df = 1, P(> X2) = 0.00032
> wald.test(b=coef(modelv6),Sigma=vcov(modelv6), Terms=3)
Wald test:
----------
Chi-squared test:
X2 = 6.5, df = 1, P(> X2) = 0.011
> wald.test(b=coef(modelv6),Sigma=vcov(modelv6), Terms=4)
Wald test:
----------
Chi-squared test:
X2 = 9.8, df = 1, P(> X2) = 0.0018
> wald.test(b=coef(modelv6),Sigma=vcov(modelv6), Terms=5)
Wald test:
----------
Chi-squared test:
X2 = 11.0, df = 1, P(> X2) = 0.00091
> wald.test(b=coef(modelv6),Sigma=vcov(modelv6), Terms=6)
Wald test:
----------
Chi-squared test:
X2 = 26.9, df = 1, P(> X2) = 2.2e-07
> wald.test(b=coef(modelv6),Sigma=vcov(modelv6), Terms=7)
Wald test:
----------
Chi-squared test:
X2 = 6.8, df = 1, P(> X2) = 0.0094
> wald.test(b=coef(modelv6),Sigma=vcov(modelv6), Terms=8)
Wald test:
----------
Chi-squared test:
X2 = 19.1, df = 1, P(> X2) = 1.2e-05
> wald.test(b=coef(modelv6),Sigma=vcov(modelv6), Terms=9)
Wald test:
----------
Chi-squared test:
X2 = 3.9, df = 1, P(> X2) = 0.049
> wald.test(b=coef(modelv6),Sigma=vcov(modelv6), Terms=10)
Wald test:
----------
Chi-squared test:
X2 = 47.9, df = 1, P(> X2) = 4.5e-12
> wald.test(b=coef(modelv6),Sigma=vcov(modelv6), Terms=11)
Wald test:
----------
Chi-squared test:
X2 = 15.0, df = 1, P(> X2) = 0.00011
```

Since these ouputs, we conclude that the predictors which cause the important effects to the response are terms numbered 2, 5, 6, 8, 10, 11 corresponding to the predictors *a1*, *a4*, *a5*, *v2*, *v4*, *v5*, especially *v4* which shows the most important contribution to the response *v6*.

Odd Ratio of Estimated Coefficients of modelv6:

```
> exp(coef(modelv6))
(Intercept)         a11         a21         a31         a41         a51 
 0.84310569  3.40990812  2.58710716  7.11423273  2.89659997  6.30711475 
        v11         v21         v31         v41         v51 
 0.28012635  0.14970734  0.26574459  0.09869834  0.20977746 
```

Confident Intervals to the Estimates at Level 0.05 of modelv6:

```
> confint.default(modelv6)
                 2.5 %       97.5 %
(Intercept) -0.7217229  0.380396993
a11          0.5584573  1.894913358
a21          0.2193601  1.681720529
a31          0.7326941  3.191500658
a41          0.4351863  1.691888937
a51          1.1455818  2.537774844
v11         -2.2321865 -0.312842524
v21         -2.7497628 -1.048383101
v31         -2.6435225 -0.006916751
v41         -2.9715716 -1.659802717
v51         -2.3530733 -0.770342733
```

Odd Ratio of Estimates and Confident Intervals at Level 0.05 of modelv6:

```
> exp(cbind(OR = coef(modelv6), confint(modelv6)))
Waiting for profiling to be done...
                    OR      2.5 %     97.5 %
(Intercept) 0.84310569 0.48375525  1.4636685
a11         3.40990812 1.77781685  6.7941717
a21         2.58710716 1.26218491  5.4801834
a31         7.11423273 2.12774926 25.3743163
a41         2.89659997 1.56733079  5.5279965
a51         6.30711475 3.21896796 13.0123059
v11         0.28012635 0.10420912  0.7222113
v21         0.14970734 0.06171577  0.3417062
v31         0.26574459 0.06621000  0.9786275
v41         0.09869834 0.04994275  0.1860934
v51         0.20977746 0.09216716  0.4521132
```

Interpretation of Odd Ratio of Estimates: *For instance, the predictor a1 with odd ratio 3.40991 interpretes the odd ratio of customer chosing a1 (value of a1 is 1) who choose v6 (value v6 is 1) is equal to 3.40991 times to customer not choosing a1 but choosing v6. This fact tells us that there is a relation between a1 and v6. And this points out that UNI must be thankful to students (old customers) who has finished their study here for their recommendation to their friends or relatives to trust in the training process in UNI for their future careers. This also shows the good reputation of UNI in the region.*

3.5 Model with response *v5*, denoted modelv5

```
glm(formula = v5 ~ a2 + a4 + a5 + v6 + v4, family = binomial(link = "logit"), 
    data = trainingData)
Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.9125  -1.0438  -0.0195   0.9126   1.8508  
Coefficients:
            Estimate Std. Error z value Pr(>|z|)   
(Intercept) -0.08652    0.28438  -0.304  0.76093   
a21          0.74721    0.38491   1.941  0.05223 . 
a41          0.80329    0.37971   2.116  0.03438 * 
a51          0.93710    0.38159   2.456  0.01406 * 
v61         -0.97420    0.38781  -2.512  0.01200 * 
v41         -1.20041    0.37673  -3.186  0.00144 **
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 224.58  on 161  degrees of freedom
Residual deviance: 200.96  on 156  degrees of freedom
AIC: 212.96
Number of Fisher Scoring iterations: 4
```

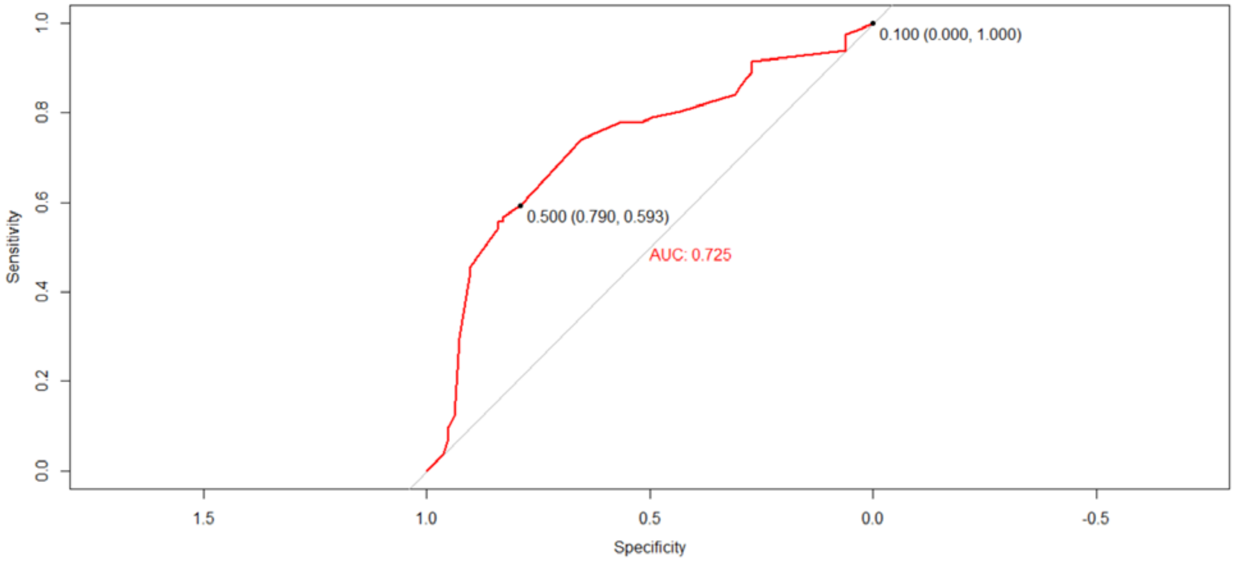


Figure 6. ROC of modelv5 on trainingData.

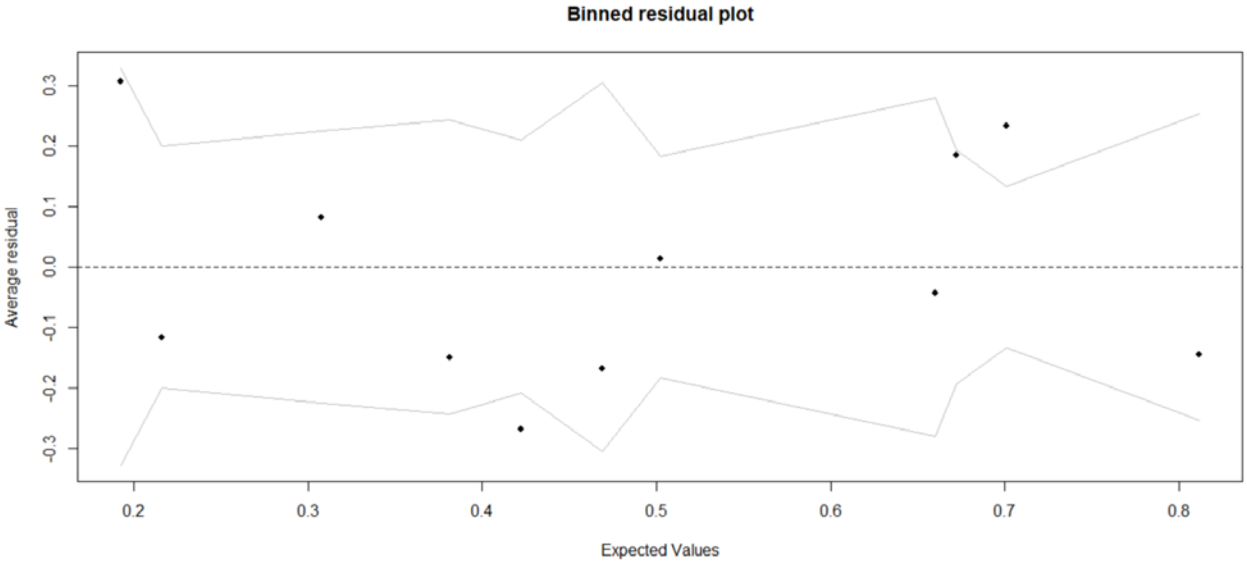


Figure 7. Binned Residual plot of modelv5.

MissClassification Predicted Error of modelv5

```
predicted <- plogis(predict(modelv5, testData))  # predicted scores
library(InformationValue)
optCutOff <- optimalCutoff(testData$v5, predicted)[1] #0.83941
misClassError(testData$v5, predicted, threshold = optCutOff)#0.0552
```

Concordance of modelv5

```
> Concordance(testData$v5, predicted)
$Concordance
[1] 0.6476671
```

```
$Discordance
[1] 0.3523329
$Tied
[1] 5.551115e-17
$Pairs
[1] 8316
```

Sensitivity and Specificity of modelv5 are bad, this seems to be expected from Confusion Matrix with poor data.

```
> sensitivity(testData$v5, predicted, threshold = optCutOff)
[1] 0
> specificity(testData$v5, predicted, threshold = optCutOff)
[1] 0.9949495
> confusionMatrix(testData$v5, predicted, threshold = optCutOff)
    0  1
0 394 21
1   2  0
```

Every assessment seems to be perfect to modelv5 except the sensitivity and specificity which has a main reason from poor data collected. We could expect to amend this by gathering more data.

3.6 Model with response v4, denoted modelv4

```
Call:
glm(formula = v4 ~ a1 + a2 + a3 + a4 + a5 + a6 + v1 + v2 + v5 +
    v6 + v3, family = binomial(link = "logit"), data = trainingData)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6964  -0.8196  -0.0142   0.6753   3.9420
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.01411    0.28566  -0.049 0.960598
a11          0.92812    0.31810   2.918 0.003526 **
a21          1.12479    0.35730   3.148 0.001644 **
a31          1.19598    0.62122   1.925 0.054203 .
a41          1.42387    0.31539   4.515 6.34e-06 ***
a51          1.37642    0.33847   4.067 4.77e-05 ***
a61          1.30259    0.48380   2.692 0.007094 **
v11         -2.35255    0.45818  -5.135 2.83e-07 ***
v21         -2.57233    0.43346  -5.934 2.95e-09 ***
v51         -1.43126    0.37055  -3.863 0.000112 ***
v61         -2.16586    0.31225  -6.936 4.02e-12 ***
v31         -1.73444    0.56731  -3.057 0.002233 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
 Null deviance: 515.70  on 371  degrees of freedom
Residual deviance: 378.34  on 360  degrees of freedom
AIC: 402.34
Number of Fisher Scoring iterations: 5
```

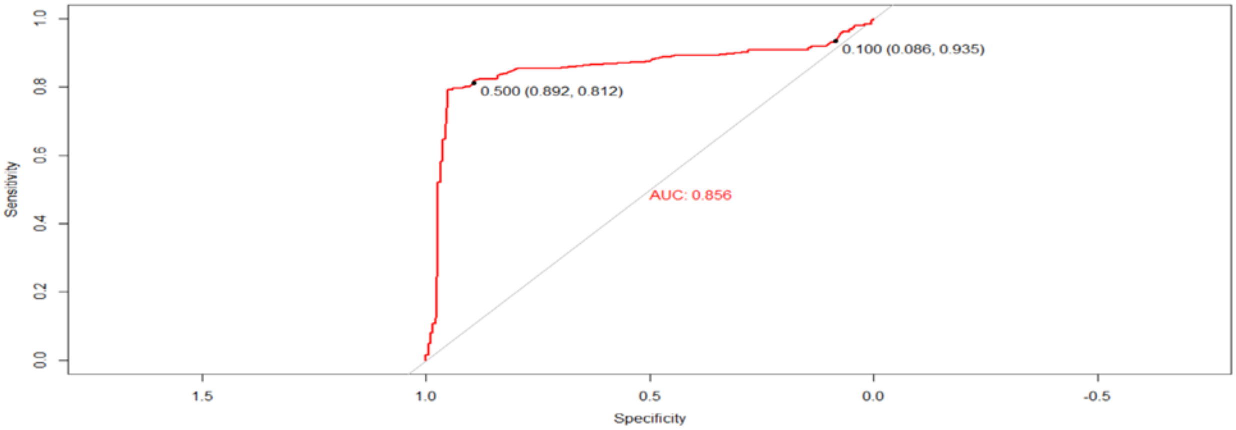


Figure 8. ROC of modelv4 basing on trainingData.

MissClassification Predicted Error

```
>predicted <- plogis(predict(modelv4, testData))  # predicted scores
>optCutOff <- optimalCutoff(testData$v4, predicted)[1] #0.59955
>misClassError(testData$v4, predicted, threshold = optCutOff)#0.1111
```

Binned Residual Plot

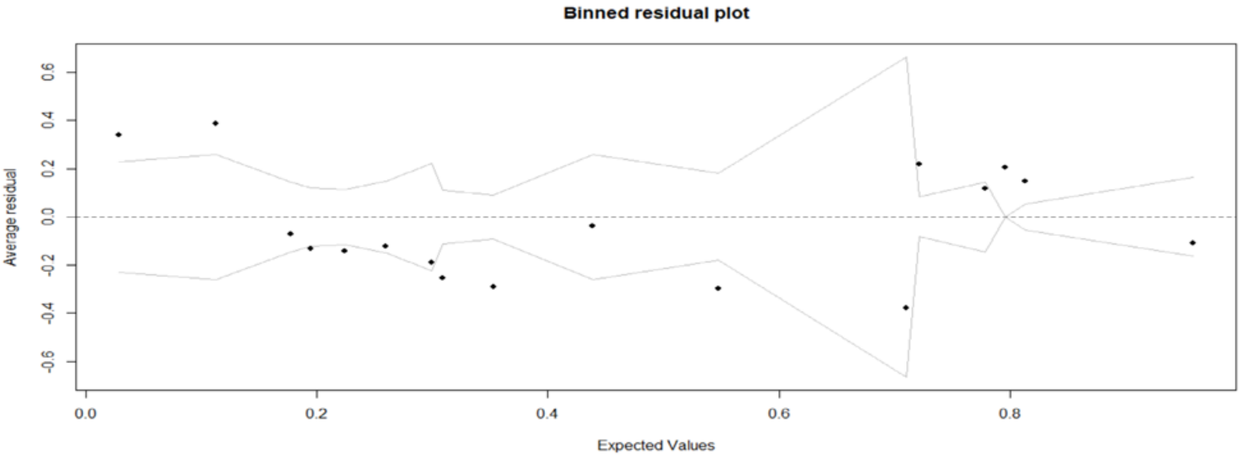


Figure 9. Binned residual plot of modelv4.

Multicolinearity VIF

```
> vif(modelv4)
      a1       a2       a3       a4       a5       a6       v1       v2 
1.366346 1.222299 1.170244 1.386124 1.366863 1.128307 1.197648 1.162895 
      v5       v6       v3 
1.155669 1.258128 1.110323 
```

ROC of modelv4 on testData with AUC = 83.74%.

Concordance

```
> Concordance(testData$v4, predicted)
$Concordance
[1] 0.8316489
$Discordance
[1] 0.1683511
$Tied
[1] 2.775558e-17
$Pairs
[1] 7520
```

Sensitivity and specificity of modelv4

```
> sensitivity(testData$v4, predicted, threshold = optCutOff)
```

```
[1] 0.7659574
> specificity(testData$v4, predicted, threshold = optCutOff)
[1] 0.925
```

Confusion Matrix of modelv4

```
> confusionMatrix(testData$v4, predicted, threshold = optCutOff)
    0  1
0 148 11
1  12 36
```

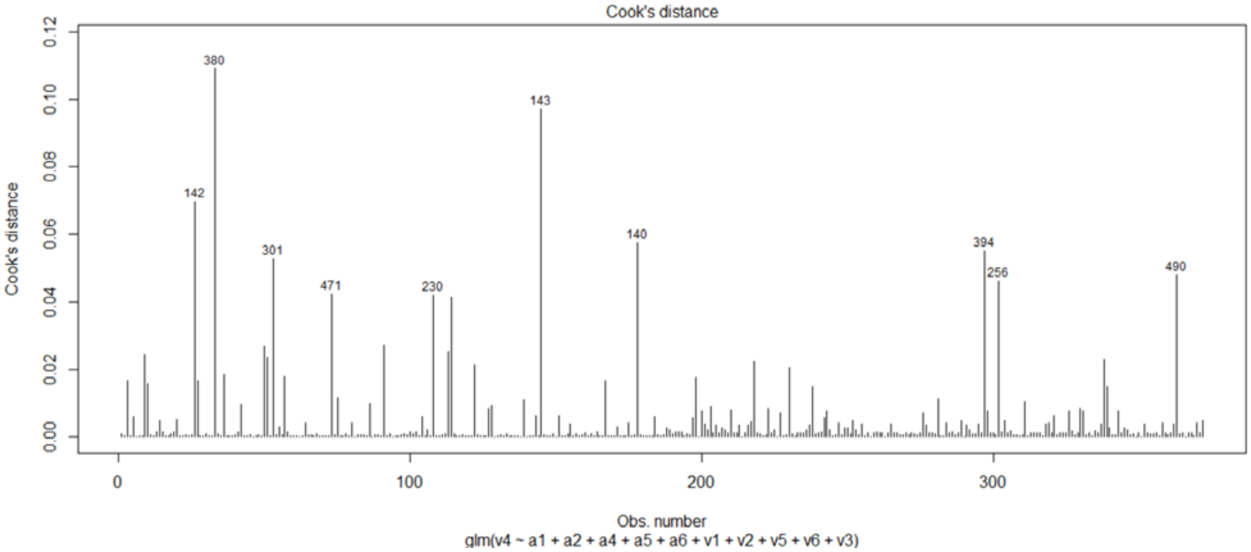


Figure 10. Cook’s Distance of modelv4.

Modified modelv4 with no predictor a3, denoted by modelv40

The reason to consider this model is the fact that the significance of the predictor a3 is weak, so in the output of Estimate p-value of modelv4.

```
Call:
glm(formula = v4 ~ a1 + a2 + a4 + a5 + a6 + v1 + v2 + v5 + v6 +
    v3, family = binomial(link = "logit"), data = trainingData)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5006  -0.8426   0.0021   0.6669   3.8428
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.1250     0.2739   0.456 0.648131
a11           0.8137     0.3062   2.657 0.007878 **
a21           1.0791     0.3493   3.089 0.002008 **
a41           1.3036     0.3029   4.304 1.68e-05 ***
a51           1.2652     0.3287   3.849 0.000119 ***
a61           1.2513     0.4776   2.620 0.008792 **
v11          -2.2815     0.4491  -5.080 3.77e-07 ***
v21          -2.4790     0.4267  -5.810 6.25e-09 ***
v51          -1.4525     0.3687  -3.939 8.17e-05 ***
v61          -2.1291     0.3115  -6.835 8.23e-12 ***
v31          -1.5101     0.5564  -2.714 0.006645 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
  Null deviance: 515.70  on 371  degrees of freedom
Residual deviance: 382.07  on 361  degrees of freedom
AIC: 404.07
Number of Fisher Scoring iterations: 5
```

Goodness-and-fit of modelv4 and modelv40: a model is good if the McFadden score is from 0.2 to 0.4. Therefore, these two models are good (p.35-[6]).

```
> pR2(modelv4)#The value under “McFadden” if the pseudo-R2.
fitting null model for pseudo-r2
         llh      llhNull           G2     McFadden         r2ML         r2CU 
-189.1684807 -257.8507512  137.3645410    0.2663644    0.3087540    0.4116720 
> pR2(modelv40)
fitting null model for pseudo-r2
         llh      llhNull           G2     McFadden         r2ML         r2CU 
-191.0372620 -257.8507512  133.6269784    0.2591169    0.3017739    0.4023652 
```

Besides, we can trust on **CoxSnell, Nagelkerke indices** for Goodness-and-fit of these models.

```
> PseudoR2(modelv4, which = c("CoxSnell","Nagelkerke","McFadden"))
  CoxSnell Nagelkerke   McFadden 
 0.3087540  0.4116720  0.2663644 
> PseudoR2(modelv40, which = c("CoxSnell","Nagelkerke","McFadden"))
  CoxSnell Nagelkerke   McFadden 
 0.3017739  0.4023652  0.2591169 
```

4. Comparison between the Different Methods

Here we make the comparison between the methods KNN, logistic regression, Random Forest, and Gradient Boosted Machine. Modelv4 is considered.

```
set.seed(818)
(train(v4~a1+a2+a3+a4+a5+a6+v1+v2+v3+v6+v5,
      data = trainingData,
      preProcess = c("center", "scale"),
      method = "knn"))
(train(v4~a1+a2+a3+a4+a5+a6+v1+v2+v3+v6+v5,
      data = trainingData,
      preProcess = c("center", "scale"),
      method = "glm"))
(train(v4~a1+a2+a3+a4+a5+a6+v1+v2+v3+v6+v5,
      data = trainingData,
      preProcess = c("center", "scale"),
      method = "ranger"))
(gbm <- train(v4~a1+a2+a3+a4+a5+a6+v1+v2+v3+v6+v5,
      data = trainingData,
      method = "gbm",
      preProcess = c("center", "scale"),
      verbose = F))
```

The output reveals the following scores of each methods aforementioned.

```
> (train(v4~a1+a2+a3+a4+a5+a6+v1+v2+v3+v6+v5,
+       data = trainingData,
+       preProcess = c("center", "scale"),
+       method = "knn"))
k-Nearest Neighbors 
372 samples
 11 predictor
  2 classes: '0', '1' 
Pre-processing: centered (11), scaled (11) 
Resampling: Bootstrapped (25 reps) 
Summary of sample sizes: 372, 372, 372, 372, 372, 372, ... 
Resampling results across tuning parameters:
  k  Accuracy   Kappa    
  5  0.8190253  0.6378894
  7  0.8014417  0.6024340
  9  0.7746455  0.5483935
Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 5.
```

```
> (train(v4~a1+a2+a3+a4+a5+a6+v1+v2+v3+v6+v5,
+       data = trainingData,
+       preProcess = c("center", "scale"),
+       method = "glm"))
Generalized Linear Model
372 samples
 11 predictor
2 classes: '0', '1'

Pre-processing: centered (11), scaled (11)
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 372, 372, 372, 372, 372, 372, ...
Resampling results:
  Accuracy  Kappa
  0.816514  0.6318915
> (train(v4~a1+a2+a3+a4+a5+a6+v1+v2+v3+v6+v5,
+       data = trainingData,
+       preProcess = c("center", "scale"),
+       method = "ranger"))
Random Forest
372 samples
 11 predictor
  2 classes: '0', '1'
Pre-processing: centered (11), scaled (11)
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 372, 372, 372, 372, 372, 372, ...
Resampling results across tuning parameters:
  mtry  splitrule   Accuracy   Kappa
   2    gini        0.8473465  0.6948027
   2    extratrees  0.8484587  0.6970588
   6    gini        0.8599116  0.7199125
   6    extratrees  0.8593760  0.7188610
  11    gini        0.8548194  0.7095666
  11    extratrees  0.8542661  0.7083438
Tuning parameter 'min.node.size' was held constant at a value of 1
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were mtry = 6, splitrule = gini
 and min.node.size = 1.
> (gbm <- train(v4~a1+a2+a3+a4+a5+a6+v1+v2+v3+v6+v5,
+       data = trainingData,
+       method = "gbm",
+       preProcess = c("center", "scale"),
+       verbose = F))
Stochastic Gradient Boosting
372 samples
 11 predictor
  2 classes: '0', '1'
Pre-processing: centered (11), scaled (11)
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 372, 372, 372, 372, 372, 372, ...
Resampling results across tuning parameters:
  interaction.depth  n.trees  Accuracy   Kappa
  1                   50      0.7526577  0.5072293
  1                  100      0.7901919  0.5804269
  1                  150      0.7967340  0.5932775
  2                   50      0.8234201  0.6460869
  2                  100      0.8457452  0.6906979
  2                  150      0.8501107  0.6995848
  3                   50      0.8376948  0.6746539
  3                  100      0.8499825  0.6993852
  3                 150      0.8463252  0.6921243

Tuning parameter 'shrinkage' was held constant at a value of 0.1
Tuning parameter 'n.minobsinnode' was held constant at a value of 10
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were n.trees = 150, interaction.depth =
 2, shrinkage = 0.1 and n.minobsinnode = 10.
```

Since this output, we see that Random Forest and Stochastic Gradient Boosting provides the most accurate estimate in estimating the modelv4.

4. Conclusions

- The importance of the predictors *var2*, *a3*, *a6* are neligible. These variables have not much effect on the decision of choosing to study at UNI. These could be expected since the fact that the decision of customers in choosing where to study their undergraduate level does not depend on the distance from where they are linving (*var2*), either if there are their family members working in the UNI (variable *a3*). The social media (variable *a6*) also has not much effect on their choice.
- The most essential responses are *v4* and *v6*: The two models constructed which are quite nice with high Goodness-of-Fit, *modelv6* and *modelv4* (*modelv40*). Two key responses for our real world problem here are *a4,* the decision-making based on demand of the labor market which is the most important function of a university or a training company, and *a6*, the decision-making based on ability of understading and applying the social and scientific knowledge of the customer in their high school level education.
- The most influent predictors to the response *v4* are *v1*, *v2*, *v6* and *a4*, *a5*. From these influences, the coefficients of a4, a5 in modelv4 are positive, 1.42387 and 1.37642 respectively. This needs to odd ratio to be quite big $e^{1.43287} \approx 4.1532$ and $e^{1.37642} \approx 3.961$, respectively. The fact here shows that local customers who are living in Thai Nguyen province are dominant to those from other provinces studing in university. This also tells us that the Brochure of UNI distributing to customer is a really important information resource provided to the students/customers. This effect is positive on their choice.
- The most influent predictors to the response *v6* are *v4, v5*, *v2*, *a4*, *a5*. This is showed in *modelv6*. Similar to model with the response *v4*, *modelv6* have positive effect caused by predictors *a4* and *a5*.

Appendix

SURVEY FORM

This survey is distributed to 579 customers/students who is currently studing in UNI over 12 universities and colleges, schools listed in the introduction. We are very grateful for this help in answering the question in this survey. The survey is in Vietnamese. Here we translate it and put on the English version.

SURVEY ON CHOOSING YOUR CURRENT UNIVERSITY

We are thankful to you in helping us getting the true information to serve for our study.

Please provide us the following answers, put a tick in the box on your choice:

1. At which year you are studying in your university: K...........?

2. What is the first information resource about your university (UNI) provided to you? :

□ *You have relatives (or friends) finished their study in UNI.*

□ *Having relatives or friends who are studing in UNI.*

□ *Having relatives or friends who are working in UNI.*

□ *You are living in Thai Nguyen province, so you had known UNI since you were studing in a high school.*

□ *Since the information provided in Brochure about UNI when you tried to make a choice on your future study/career at the end of your high school grades.*

□ *Seeing the advertisement of UNI on the social media by chance.*

3. How far is it from your hometown to UNI:

□ *Below 15Km.* □ *From 15KM- 50KM.* □ *Above 50KM.*

4. What is the most important reason of your choice to study in UNI (**multiple choices should be order with respect to the decreasing important level**):

□ *Since your blood relatives or friends who had been trained in UNI advised/persuaded you*.

□ *Since your blood relatives or friends who are being training in UNI advised/affected you..*

□ *Since your blood relatives who had not been or are not being trained in UNI advised/persuaded you..*

□ *Since your self-decision making based on your knowledge of career opportunities and your passion..*

□ *Since your prior calculation to the expense on which you might spend for your future study and living in UNI was less than that in other univeristies in the country which have the same program*.

□ *Since the admission requirements of UNI are suitable for your GPA in high school*.

5. Would you change your decision if you had a chance to choose the university again?

□ Yes □ No

Acknowledgments

This work is a part of a scientific project funded by Thai Nguyen University of Technology. We are firmly appreciate this valuable support for us to complete this project.

References

[1] A Douglas C. Montgomery, George C. Runger, Applied Statistics and Probability for Engineers, John Wiley & Sons, 2014.

[2] Ronald E. Walpole, Raymond H. Meyers, Sharon L. Meyers, Keying Ye, Probability & Statistics for Engineers & Scientists, Prentice Hall, Person, 9th editition, 2012.

[3] Sheldon M. Ross, Introduction to Probability Models, Elsevier, 10th edition, 2010.

[4] Daniel McFadden, (1977). "Quantitative Methods for Analyzing Travel Behaviour of Individuals: Some Recent Developments", Cowles Foundation for Research in Economics, Yale University, 1977, pp. 474.

[5] N.V. Dung, Morden Mass Communication Journal from Academic to Real Life, Viet Nam National University Publishing, 2011.

[6] T.H. Quang, Socialist about Mass Communication, Ho Chi Minh City Open University Publishing, 2009.

First Author: M.Sc Tran Thi Hue, full-time lecture in Faculty of International Training, Thai Nguyen University of Technology (TNUT); Bachelor of Mechanical Mathematics in Viet Nam National University, Ha Noi, University of Natural Science (graduated in 2002), Master of Mechanical Engineering issed by Thai Nguyen University of Technology, Thai Nguyen University in 2004; Working as a full-time lecture in Thai Nguyen University since 2002 until now; Being Head of Division Nature Science in Faculty of International Training, TNUT; Having 3 scienctific papers published in national journals of science and technology; Being one of three co-authors of a textbook published in 2020 which is decided to be used as a material to teach applied math for student in TNUT; Was one the director of a scientific project granted by Thai Nguyen University of Technology and inspected with excellent quality; Current research area interest: Differential equations, Statistics.

Second Author M.Sc Ngo Thi Quynh Nhung, full-time lecture in Faculty of Postal Profession Training, Thai Nguyen College of Economics and Finance; graduated from Thai Nguyen University of Economics and Business Administration, Thai Nguyen University, at Bachelor degree of Accounting in 2009; graduated from Thai Nguyen University of Economics and Business Administration at Master of Business Administration in 2012; Working as a full-time lecture in Thai Nguyen College of Economics and Finance since 2009 until now; Having a textbook published by a national publisher; Awarded a Excellent Teacher Price of Thai Nguyen Province.