

**FACULTY OF COMPUTER AND MATHEMATICAL SCIENCES**

**STA761-STATISTICAL DATA MINING**

**CUSTOMER’S PROFILE PREDICTION FOR**

**GROUP MEMBERSHIP (TELCO)**

|  |  |  |
| --- | --- | --- |
| **PREPARED BY:** | NUR ALIA BINTI MOHAMMAD ZUBIR | 2020947963 |
|
| **DATE OF SUBMISSION:** | 30TH NOVEMBER 2020 | | |
| **TO BE REVIEWED BY:** | PROF. YAP BEE WAH | | |

Table of Contents

[STAGE 1 BUSINESS UNDERSTANDING 1](#_Toc57669419)

[STAGE 2 DATA UNDERSTANDING 1](#_Toc57669420)

[STAGE 3 DATA PREPARATION 3](#_Toc57669421)

[STAGE 4 DATA MODELING 6](#_Toc57669422)

[4.1 DATA VISUALIZATION 6](#_Toc57669423)

[4.2 LOGISTIC REGRESSION (ENTER, FORWARD AND BACKWARD) MODEL 7](#_Toc57669424)

[4.3 DECISION TREE (CART, CHAID AND C5.0) MODEL 14](#_Toc57669425)

[4.4 artificial neural network (mlp AND rbf) MODEL 16](#_Toc57669426)

[STAGE 5 MODEL EVALUATION 18](#_Toc57669427)

[5.1 LOGISTIC REGRESSION MODEL (ENTER, FORWARD AND BACKWARD) 18](#_Toc57669428)

[5.2 DECISION TREE MODEL COMPARISON (CART, CHAID AND C5.0) 20](#_Toc57669429)

[5.3 ARTIFICIAL NEURAL NETWORK COMPARISON (MLP AND RBF) 21](#_Toc57669430)

[5.4 MODEL COMPARISON – LOGISTIC REGRESSION VS DECISION TREE 22](#_Toc57669431)

[STAGE 6 MODEL DEPLOYMENT 23](#_Toc57669432)

[STAGE 7 CONCLUSION 24](#_Toc57669433)

[APPENDICES 25](#_Toc57669434)

[APPENDIX A LOGISTIC REGRESSION (ENTER) MODEL 25](#_Toc57669435)

[APPENDIX B LOGISTIC REGRESSION (FORWARD) MODEL 29](#_Toc57669436)

[APPENDIX C LOGISTIC REGRESSION (BACKWARD) MODEL 32](#_Toc57669437)

[APPENDIX D LOGISTIC REGRESSION (STEPWISE) MODEL 35](#_Toc57669438)

[APPENDIX E DECISION TREE (CART) MODEL 38](#_Toc57669439)

[APPENDIX E DECISION TREE (CHAID) MODEL 39](#_Toc57669440)

[APPENDIX F DECISION TREE (C5.0) MODEL 41](#_Toc57669441)

[APPENDIX G MLP NETWORK 52](#_Toc57669442)

[APPENDIX H RBF NETWORK 53](#_Toc57669443)

[APPENDIX I VARIABLE ANALYSIS AGAINST TARGET VARIABLES 54](#_Toc57669444)

**PREDICTIVE MODELING: PURCHASE OF ORGANIC PRODUCTS BY CUSTOMERS**

## STAGE 1 BUSINESS UNDERSTANDING

The objective is to determine group membership profiles of customers who are more likely to be in basic service, E-service, plus service and total service. The data mining goal is to build predictive models to predict what is customers group membership. The event of interest is Y=1 (if Customer’s group membership is Basic Service), Y=2 (if Customer’s group membership is E-Service), Y=3 (if Customer’s group membership is Plus Service) and Y=4 (if Customer’s group membership is Total Service).

## STAGE 2 DATA UNDERSTANDING

The target variable is DEFAULT which is a binary variable with two categories (1 = Default on bank loan, 0=otherwise). There are twelve (12) input variables. Table 1 provides the description of the variables.

Table 1 Description of Variables

| **Variable** | **Role** | **Measurement Level** | **Description** |
| --- | --- | --- | --- |
| CUSTCAT | Target | Nominal | Customer category  *1 = Basic Service*  *2 = E-Service*  *3 = Plus Service*  *4 = Total Service* |
| REGION | Input | Nominal | Geographic indicator  *1 = Zone 1*  *2 = Zone 2*  *3= Zone 3* |
| AGE | Input | Continuous | Age in years |
| MARITAL | Input | Flag | Marital status  *1 = Married*  *2 = Unmarried* |
| ADDRESS | Input | Continuous | Years at current address |
| INCOME | Input | Continuous | Household income in thousands |
| ED | Input | Ordinal | Level of education  *1 = College degree*  *2 = Did not complete high school*  *3 = High school degree*  *4 = Post-undergraduate degree*  *5 = Some college* |
| EMPLOY | Input | Continuous | Years with current employer |
| RETIRE | Input | Flag | Retired  *1 = No*  *2 = Yes* |
| GENDER | Input | Flag | Gender  *0 = Female*  *1 = Male* |
| RESIDE | Input | Continuous | Number of people in household |

## STAGE 3 DATA PREPARATION

The data was stored as a SPSS data file. The sample consists of 1,000 cases.

Figure 1 (a) and 1 (b) described the statistic from the Data Audit node and Data Audit results, respectively.

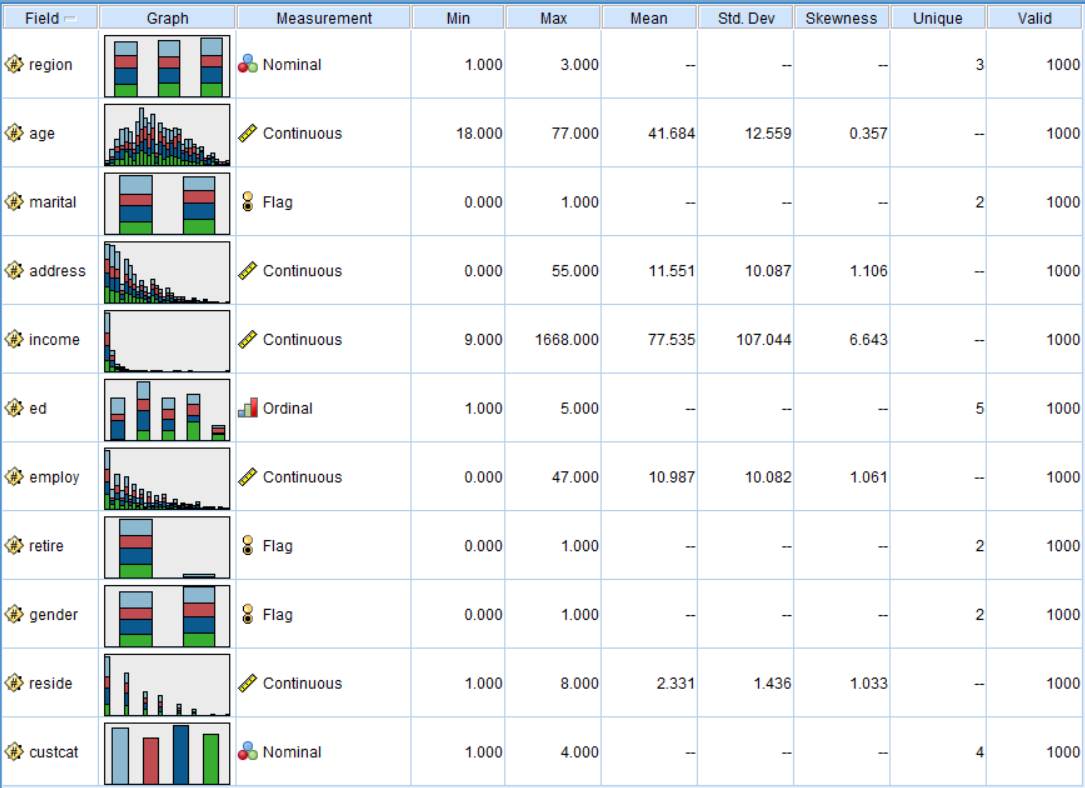


Figure 1(a) Descriptive Statistics from Data Audit node

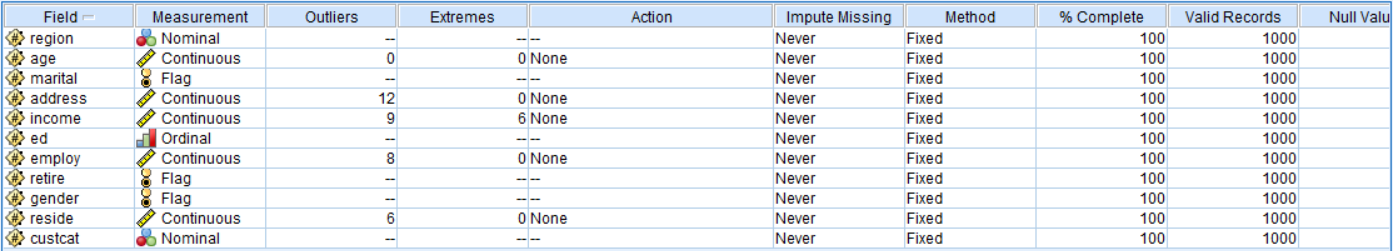


Figure 1(b) Data Audit results

From both Figure 1(a) and 1(b), the interpretation and summary of the variables is showed as per Table 2.

Table 2 Data Audit Summary and Interpretation

|  |  |
| --- | --- |
| **Variables** | **Summary and Interpretation of Variables and Sample Populations** |
| **CUSTCAT** | The target variable showing majority of the sample population indicates 28.1% will be in Plus Service membership follow by Basic Service (26.6%), Total Service (23.6%) and E-Service (21.7%). The total sample is 1000. |
| **REGION** | The **REGION** variable showing majority of the sample population indicates 34.4% will be in zone 3 follow by zone 2 (33.4%) and zone 1 (32.2%). The total sample is 1000. |
| **AGE** | The histogram is substantially symmetric. The skewness value is 0.357 described the distribution is significantly normally distributed. The **AGE** sample population majorly reside at the value of 33.672 with 25 of sample population found with this value. Data audit indicates that there were **0 outliers** (with 0 were extreme cases). |
| **MARITAL** | The **MARITAL** variable showing majority of the sample population indicates 50.5% is unmarried and 49.5% is married. The total sample is 1000. |
| **ADDRESS** | The histogram is substantially right skewed distribution. The skewness value is 1.106 described the distribution is significantly positive distributed. The **ADDRESS** sample population majorly reside at the value of 2.5781 with 45 of sample population found with this value. Data audit indicates that there were **12 outliers** (with 0 were extreme cases). Boxplot representation is available in [Appendix I](#_APPENDIX_G_VARIABLE). |
| **INCOME** | The histogram is substantially right skewed distribution. The skewness value is 6.643 described the distribution is significantly positive distributed. The **INCOME** sample population majorly reside at the value of 34.922 with 201 of sample population found with this value. Data audit indicates that there were **9 outliers** (with 6 were extreme cases). Boxplot representation is available in [Appendix I](#_APPENDIX_G_VARIABLE) |
| **ED** | The **ED** variable showing majority of the sample population indicates 28.7% has high school degree follow by college degree (23.4%), some college (20.9%), did not complete high school (20.4%) and post-undergraduate degree (6.6%). The total sample is 1000. |
| **EMPLOY** | The histogram is substantially right skewed distribution. The skewness value is 1.061 described the distribution is significantly positive skewed. The **EMPLOY** sample population majorly reside at the value of 0.7344 with 53 of sample population found with this value. Data audit indicates that there were **8 outliers** (with 0 were extreme cases). Boxplot representation is available in [Appendix I](#_APPENDIX_G_VARIABLE). |
| **RETIRE** | The **RETIRE** variable showing majority of the sample population indicates 95.3% has not retired yet and 4.7% is a percentage of retired. The total sample is 1000. |
| **GENDER** | The **GENDER** variable showing majority of the sample population indicates 51.7% is a male and 48.3% is a female. The total sample is 1000. |
| **RESIDE** | The histogram is substantially right skewed distribution. The skewness value is 1.033 described the distribution is significantly positive skewed. The **RESIDE** sample population majorly reside at the value of 1.1094 with 76 of sample population found with this value. Data audit indicates that there were **6 outliers** (with 0 were extreme cases). Boxplot representation is available in [Appendix I](#_APPENDIX_G_VARIABLE). |

Table 3 showed the Frequency Distribution for Categorical Variables.

Table 3 Frequency Distribution for Categorical Variables

| **Variable** | **Sub-Variable** | **Frequency (%)** |
| --- | --- | --- |
| CUSTCAT | Basic Service | 266 (26.6%) |
| E-Service | 217 (21.7%) |
| Plus Service | 281 (28.1%) |
| Total Service | 236 (23.6%) |
| REGION | Zone 1 | 322 (32.2%) |
| Zone 2 | 334 (33.4%) |
| Zone 3 | 344 (34.4%) |
| ED | College degree | 234 (23.4%) |
| Did not complete high school | 204 (20.4%) |
| High school degree | 287 (28.7%) |
| Post-undergraduate degree | 66 (6.6%) |
| Some college | 209 (20.9%) |
| MARITAL | Married | 495 (49.5%) |
| Unmarried | 505 (50.5%) |
| GENDER | Female | 517 (51.7%) |
| Male | 483 (48.3%) |
| RETIRE | No | 953 (95.3%) |
| Yes | 47 (4.7%) |

Based on the samples in the datasets whereby there were missing values, outliers and imbalanced data for target variable, model will include conditioning the datasets, by either to

* removal of outliers; or
* perform derive node to recategorized again.

## STAGE 4 DATA MODELING

### 4.1 DATA VISUALIZATION

Predictive modeling using logistic regression, decision tree and artificial neural network model were carried out. Figure 2 shows the predictive modeling flow. The data source was connected to Imputation node (impute REASON and JOB variables) followed by connection to Balance node (reduced of default variable) to the Type node then to the Data Partition node. The data was partitioned using 70% as training sample and 30% as testing (or evaluation) sample. The Logistic Regression nodes (ENTER, FORWARDS and BACKWARDS), Decision Tree nodes (C5, CART and CHAID) and Artificial Neural Network nodes (MLP and RBF) were connected to the Data Partition node. The performances of the three logistic regression models, three decision tree models and two artificial neural network models were then evaluated and compared using the Analysis node, Gain and ROC charts.

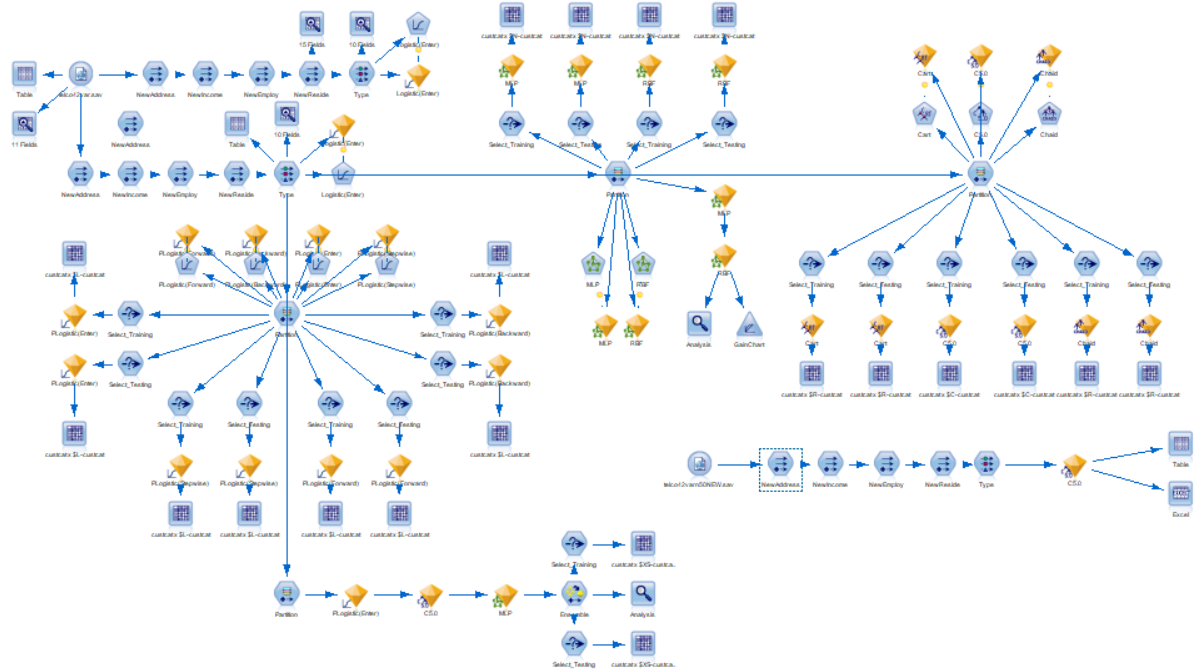


Figure 2 Predictive Modeling Flow

### 4.2 LOGISTIC REGRESSION (ENTER, FORWARD AND BACKWARD) MODEL

|  |
| --- |
| **ENTER** |
| **FORWARD** |
| **BACKWARD** |
| **STEPWISE** |

Figure 3 Predictor Importance - Logistic Regression Model

In Figure 3, the most important variable is Delinquencies (Number of Delinquent Trade Lines) for all three Logistic Regression models.

**Logistic Regression ENTER Model Summary**

Logistic Regression using ENTER selection method shows that the **model is significant** (Chi-Square= (42df) = 167.623, p<0.05). The Nagelkerke R-Square of the model is 0.227 and the classification accuracy is 40.7%. NewEmploy is a significant predictor since the likelihood ratio test table shown that the value of alpha for this variable is less than 0.05 and the value of Chi-square is 14.788.

Odds-Ratio interpretation is summarized as per Table 4.

Table 4 Odds-Ratio (OD) Interpretation

| **TARGET** | **VARIABLES** | **Odds-Ratio (OR)** | **Interpretation** |
| --- | --- | --- | --- |
| **E-SERVICE**  **vs**  **BASIC SERVICE** | ED(1) | 0.122 | Odds ratio for ED(1) = 0.122 indicates that customer with college degree are less likely in E-Service membership compared to Basic Service. |
| ED(2) | 0.264 | Odds ratio for ED(2) = 0.264 indicates that customer which did not complete high school are less likely in E-Service membership compared to Basic Service. |
| ED(3) | 0.258 | Odds ratio for ED(2) = 0.264 indicates that customer with high school degree are less likely in E-Service membership compared to Basic Service. |
| NEWADDRESS(1) | 0.303 | For every increase in years at current address, the odd ratio of custcat in E-Service compared to Basic Service decrease by 0.303 which is by 30.3%. |
| **PLUS SERVICE**  **VS**  **BASIC SERVICE** | NEWEMPLOY(1) | 0.291 | For every increase in years at current employ, the odd ratio of custcat in Plus Service compared to Basic Service decrease by 0.291 which is by 29.1%. |
| **TOTAL SERVICE**  **VS**  **BASIC SERVICE**  **TOTAL SERVICE**  **VS**  **BASIC SERVICE** | ED(1) | 0.024 | Odds ratio for ED(1) = 0.024 indicates that customer with college degree are less likely in Total Service membership compared to Basic Service. |
| ED(2) | 0.134 | Odds ratio for ED(2) = 0.134 indicates that customer which did not complete high school are less likely in Total Service membership compared to Basic Service. |
| ED(3) | 0.209 | Odds ratio for ED(3) = 0.209 indicates that customer with high school degree are less likely in Total Service membership compared to Basic Service. |
| NEWEMPLOY(1) | 0.379 | For every increase in years at current employ, the odd ratio of custcat in Total Service compared to Basic Service decrease by 0.379 which is by 37.9%. |

**Logistic Regression FORWARD, BACKWARD and STEPWISE Model Summary**

These three Logistic Regression selection method shows that the **model is significant** (Chi-Square= (21df) = 149.586, p<0.05). The Nagelkerke R-Square of the model is 0.205 and the classification accuracy is 40.3%. NewEmploy, NewReside and ED variables re a significant predictors since the likelihood ratio test table shown that the value of alpha for these three variables is less than 0.05 and the value of Chi-square is 113.644 (ED), 27.983 (NewEmploy) and 8.124 (NewReside).

Odds-Ratio interpretation is summarized as per Table 5.

Table 5 Odds-Ratio (OD) Interpretation

| **TARGET** | **VARIABLES** | **Odds-Ratio (OR)** | **Interpretation** |
| --- | --- | --- | --- |
| **E-SERVICE**  **vs**  **BASIC SERVICE** | ED(1) | 0.120 | Odds ratio for ED(1) = 0.120 indicates that customer with college degree are less likely in E-Service membership compared to Basic Service. |
| ED(2) | 0.252 | Odds ratio for ED(2) = 0.252 indicates that customer which did not complete high school are less likely in E-Service membership compared to Basic Service. |
| ED(3) | 0.247 | Odds ratio for ED(3) = 0.247 indicates that customer with high school degree are less likely in E-Service membership compared to Basic Service. |
| **PLUS SERVICE**  **vs**  **BASIC SERVICE** | NEWEMPLOY(1) | 0.189 | For every increase in years at current employ, the odd ratio of custcat in Plus Service compared to Basic Service decrease by 0.189 which is by 18.9%. |
| **TOTAL SERVICE**  **vs**  **BASIC SERVICE**  **TOTAL SERVICE**  **vs**  **BASIC SERVICE** | ED(1) | 0.024 | Odds ratio for ED(1) = 0.024 indicates that customer with college degree are less likely in Total Service membership compared to Basic Service. |
| ED(2) | 0.130 | Odds ratio for ED(2) = 0.130 indicates that customer which did not complete high school are less likely in Total Service membership compared to Basic Service. |
| ED(3) | 0.204 | Odds ratio for ED(3) = 0. 204 indicates that customer with high school degree are less likely in Total Service membership compared to Basic Service. |
| NEWEMPLOY(1) | 0.187 | For every increase in years at current employ, the odd ratio of custcat in Total Service compared to Basic Service decrease by 0.187 which is by 18.7%. |
| NEWRESIDE(1) | 0.367 | For every number of people in houshold, the odd ratio of custcat in Total Service compared to Basic Service decrease by 0.367 which is by 36.7%. |

A summary of the Logistic Regression Model is given in Table 6.

Table 6 Summary of Logistic Regression model results

|  |  |  |  |
| --- | --- | --- | --- |
| **PARAMETER ESTIMATES** | **VARIABLE** | **ENTER** | **BW/FW/STPWISE** |
| E-service | age | 0.014 |  |
| [region=1.000] | 0.316 |  |
| [region=2.000] | 0.231 |  |
| [region=3.000] |  |  |
| [marital=.000] | -0.17 |  |
| [marital=1.000] |  |  |
| [ed=1.000] | -2.106 | -2.117 |
| [ed=2.000] | -1.332\* | -1.379\*\* |
| [ed=3.000] | -1.354\* | -1.397\*\* |
| [ed=4.000] | -0.825 | -0.838 |
| [ed=5.000] |  |  |
| [retire=.000] | 1.114 |  |
| [retire=1.000] |  |  |
| [gender=.000] | -0.199 |  |
| [gender=1.000] |  |  |
| [NewAddress=1] | -1.192\* |  |
| [NewAddress=2] |  |  |
| [NewEmploy=1] | -0.83 | -1.126 |
| [NewEmploy=2] | -0.356 | -0.372 |
| [NewEmploy=3] |  |  |
| [NewReside=1] | -0.201 | -0.203 |
| [NewReside=2] |  |  |
| age | 0.009 |  |
| Plus service | [region=1.000] | 0.366 |  |
| [region=2.000] | 0.172 |  |
| [region=3.000] |  |  |
| [marital=.000] | -0.174 |  |
| [marital=1.000] |  |  |
| [ed=1.000] | 0.5 | 0.478 |
| [ed=2.000] | 0.789 | 0.736 |
| [ed=3.000] | 0.488 | 0.427 |
| [ed=4.000] | 0.451 | 0.423 |
| [ed=5.000] |  |  |
| [retire=.000] | 0.448 |  |
| [retire=1.000] |  |  |
| [gender=.000] | -0.002 |  |
| [gender=1.000] |  |  |
| [NewAddress=1] | -0.739 |  |
| [NewAddress=2] |  |  |
| [NewEmploy=1] | -1.45\* | -1.664\*\* |
| [NewEmploy=2] | -0.561 | -0.606 |
| [NewEmploy=3] |  |  |
| [NewReside=1] | -0.207 | -0.264 |
| [NewReside=2] |  |  |
| age | 0.004 |  |
| [region=1.000] | 0.042 |  |
| Total service | [region=2.000] | 0.153 |  |
| [region=3.000] |  |  |
| [marital=.000] | -0.384 |  |
| [marital=1.000] |  |  |
| [ed=1.000] | -3.75 | -3.713 |
| [ed=2.000] | -2.009 | -2.037 |
| [ed=3.000] | -1.568\*\* | -1.59\*\* |
| [ed=4.000] | -0.783 | -0.759 |
| [ed=5.000] |  |  |
| [retire=.000] | 0.965 |  |
| [retire=1.000] |  |  |
| [gender=.000] | 0.124 |  |
| [gender=1.000] |  |  |
| [NewAddress=1] | -0.969 |  |
| [NewAddress=2] |  |  |
| [NewEmploy=1] | -1.693\* | -1.679\*\* |
| [NewEmploy=2] | -1.015 | -0.882 |
| [NewEmploy=3] |  |  |
| [NewReside=1] | -0.838\* | -1.002\* |
| [NewReside=2] |  |  |
|  |  |  |  |
| Nagelkerke R-Square |  | 0.227 | 0.205 |
| Accuracy |  | 40.7% | 40.3% |

\*p<0.05;\*\*p<0.01

### 4.3 DECISION TREE (CART, CHAID AND C5.0) MODEL

|  |
| --- |
| **CART** |
| **CHAID** |
| **C5.0** |

Figure 4 Predictor Importance - Decision Tree Model

In Figure 4, the most important variable is ED (level of education) for three (3) Decision Tree Models which are CART, CHAID and C5.0 model.

The details of Decision Tree results comprise of the models, the decision rules and interpretation of the decision rules are available in the following Appendices:

* [Appendix D](#_APPENDIX_D_DECISION) – Decision Tree CART
* [Appendix E](#_APPENDIX_E_DECISION) – Decision Tree CHAID
* [Appendix F](#_APPENDIX_F_DECISION) – Decision Tree C5.0

Decision Tree CART have 2 rules, whilst Decision Tree CHAID and Decision Tree C5.0 have 7 and 151 rules, respectively.

### 4.4 artificial neural network (mlp AND rbf) MODEL

|  |  |
| --- | --- |
| **MLP** | **RBF** |
|  |  |
|  |  |
|  |  |

Figure 5 Result of MLP and RBF model

In Figure 5, the most important variable for MLP model is NewAddress while for RBF model is age in years. Model MLP have 1 hidden layer and 8 neurons. For RBF model, it consists of 0 hidden layer and 3 neurons.

The accuracy for both models are 56.4% for MLP model and 32.7% for RBF model. This indicator shown that the MLP model is more accurate than RBF model.

The details of Artificial Neural Network’s network results of the models, are available in the following Appendices:

* [Appendix G](#_APPENDIX_D_DECISION) – Network of MLP model
* [Appendix](#_APPENDIX_E_DECISION) H – Network of RBF model

## STAGE 5 MODEL EVALUATION

### 5.1 LOGISTIC REGRESSION MODEL (ENTER, FORWARD AND BACKWARD)

The model evaluation results are shown in Table 6. The performance of the three models are quite similar. The ENTER model has slightly higher training accuracy (40.71%) compared to other models (BACKWARD, FORWARD and STEPWISE) that have a same value of accuracy which is 40.29%. The ENTER model and other three (BACKWARD, FORWARD and STEPWISE) models have different classification performance as they selected the different predictor variables.

Table 6 Logistic Regression Model Evaluation results

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Performance Measure** | **Training** | **Testing** |
| ENTER | Accuracy | 40.71% | 40.67% |
| Nagelkerke | 0.227 | |
| FWD | Accuracy | 40.29% | 40.67% |
| Nagelkerke | 0.205 | |
| BWD | Accuracy | 40.29% | 40.67% |
| Nagelkerke | 0.205 | |
| STEPWISE | Accuracy | 40.29% | 40.67% |
| Nagelkerke | 0.205 | |

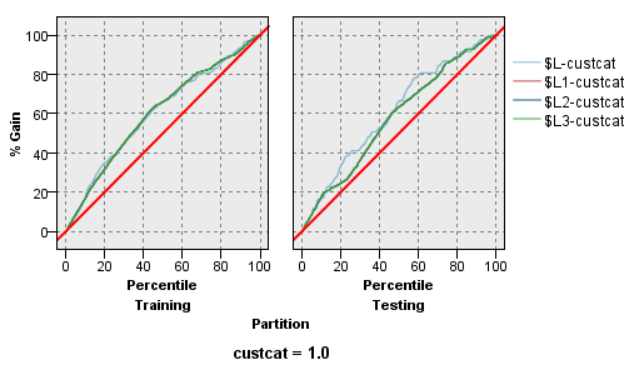


Figure 6 Logistic Regression Gain Chart

The Gain chart in Figure 6 shows the area Under the Curve (AUC) is quite similar for all training Logistic Regression model but there is a different in testing. Comparing Logistic Regression of ENTER, STEPWISE, FORWARD and BACKWARD, the model performance in Logistic ENTER performed better than Logistic Regression STEPWISE, FORWARD and BACKWARD since the model are capable to predict more samples correctly. Area Under the Curve (AUC) in training for Logistic ENTER ($L-Custcat) are larger than Logistic Regression STEPWISE, FORWARD and BACKWARD.

Selection of best model is based on the testing sample results as well as evaluating the performance as depicted by Gain Chart. Thus, for this project we recommend Logistics Regression ENTER model to be compared further with Decision Tree models and Artificial Neural Network models.

### 5.2 DECISION TREE MODEL COMPARISON (CART, CHAID AND C5.0)

The model evaluation results are shown in Table 7. The performance of the three models are quite similar. The Decision Tree CHAID model has slightly higher Testing accuracy (40.00%) follow by CART model with (35.00%) of accuracy and C5 model with (32.00%) of accuracy.

Table 7 Decision Tree Model Evaluation results

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Performance Measure** | **Training** | **Testing** |
| CART | Accuracy | 37.00% | 35.00% |
| CHAID | Accuracy | 39.14% | 40.00% |
| C5.0 | Accuracy | 63.57% | 32.00% |

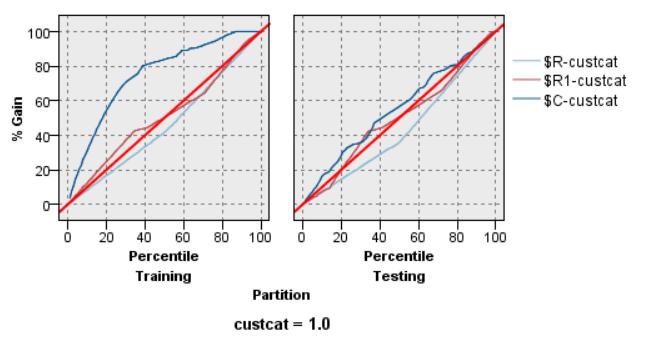


Figure 7 Decision Tree Gain Chart

The Gain chart in Figure 7 shows the Decision Tree C5 model able to predict more samples correctly in training as compared to both Decision Tree CART and CHAID where both performances are almost identical. Area Under the Curve (AUC) of Decision Tree C5 is larger than AUC of both Decision Tree CART and CHAID. But in testing, AUC for C5 have large different compared to training AUC for C5. It will cause the model overfit and it is not a preferable model. Because of that, selection of best model is based on the testing sample results as well as evaluating the performance as depicted by Gain Chart. Thus, for this project we recommend Decision Tree CHAID be further compared with Logistic Regression ENTER model.

### 5.3 ARTIFICIAL NEURAL NETWORK COMPARISON (MLP AND RBF)

The model evaluation results are shown in Table 8. The performance of the two models are different. The Artificial Neural Network RBF model has slightly higher Testing accuracy (32.00%) compared to MLP model with (30.33%) of accuracy.

Table 8 Artificial Neural Network Model Evaluation results

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Performance Measure** | **Training** | **Testing** |
| MLP | Accuracy | 52.86% | 30.33% |
| RBF | Accuracy | 35.00% | 32.00% |

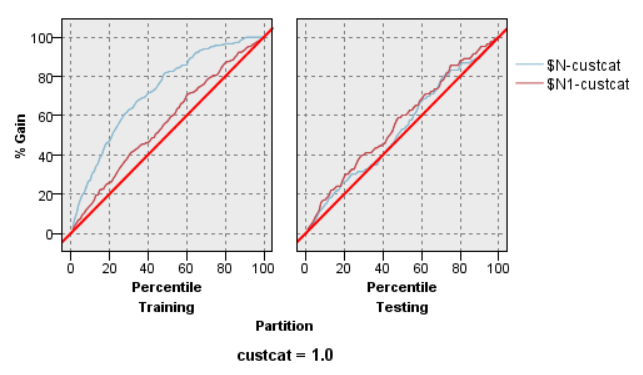


Figure 8 Artificial Neural Network Gain Chart

The Gain chart Figure 8 shows the Artificial Neural Network MLP model able to predict more samples correctly as compared to Artificial Neural Network RBF where its performances are different for training cases and the Area Under the Curve (AUC) of Artificial Neural Network MLP is larger than AUC of Artificial Neural Network RBF. In this matter, the MLP model is not preferable model because the model is overfitting due to large different between training and testing performance. Selection of best model is based on the testing sample results as well as evaluating the performance as depicted by Gain Chart. Thus, for this project we recommend Artificial Neural Network RBF to be further compared with Decision Tree CHAID and Logistic Regression ENTER model.

### 5.4 MODEL COMPARISON – LOGISTIC REGRESSION VS DECISION TREE

Based on result from Logistic Regression, Decision Tree and Artificial Neural Network, Logistic Regression BACKWARD model, Decision Tree of C5 model and Artificial Neural Network of MLP model were further compared. The results are tabulated as per Table 6.

Table 8 Model Evaluation results (BACKWARD, C5 and MLP)

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **PERFORMANCE MEASURE** | **Training** | **Testing** |
| Logistic Regression ENTER model | Accuracy | 40.71% | 40.67% |
| Decision Tree CHAID model | Accuracy | 39.14% | 40.00% |
| Artificial Neural Network RBF | Accuracy | 35.00% | 32.00% |

From Table 8, the Logistic Regression ENTER model is found to offer better prediction at Testing accuracy of 40.67% as compared to Decision Tree CHAID model which is 40.00% and Artificial Neural Network RBF model which is 32.00%. From the result, Logistic Regression ENTER model was selected for model deployment.

## STAGE 6 MODEL DEPLOYMENT

The Decision Tree CHAID model was deployed to the 50 new TELCO datasets. This new TELCO datasets, was complete with all the variables having a no missing value. 1 outlier was found in ADDRESS, EMPLOY and RESIDE variables and 1 extreme value was found in INCOME variable. The derive nodes are used to recategorized the variables that have an outlier and extreme value, and model deployment was carried out to the new datasets. From the results as shown in Figure 9, only 12 customers out of 50 were predicted as more likely to be Basic Service group membership, 2 customers to be in E-Service group membership, 22 customers in Plus Service and 14 customers in Total Service.

 Figure 9 Sample of the Predicted Default Category and Probabilities for the New Customers

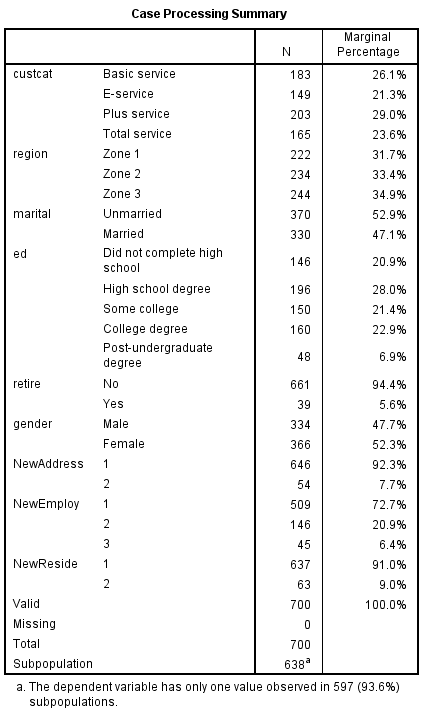
## STAGE 7 CONCLUSION

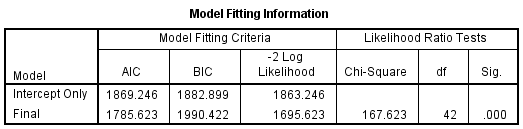
Based on the result obtain from the overall analysis, it can be concluded as follows:

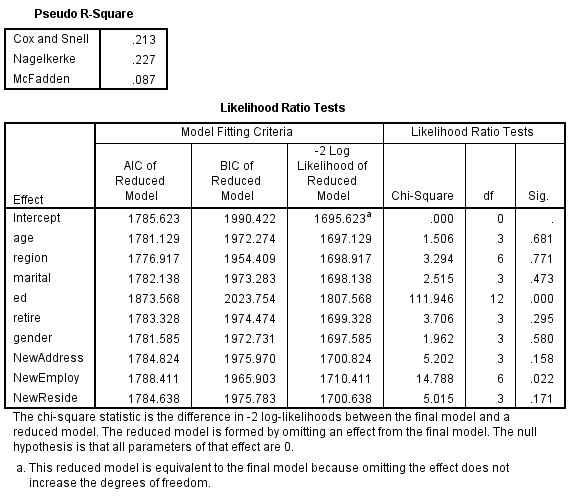
* Model performance – overall performance is considered as average performance. With better data quality, the model shall be retrained to achieve better model performance.
* Require more data to increase the accuracy of model.

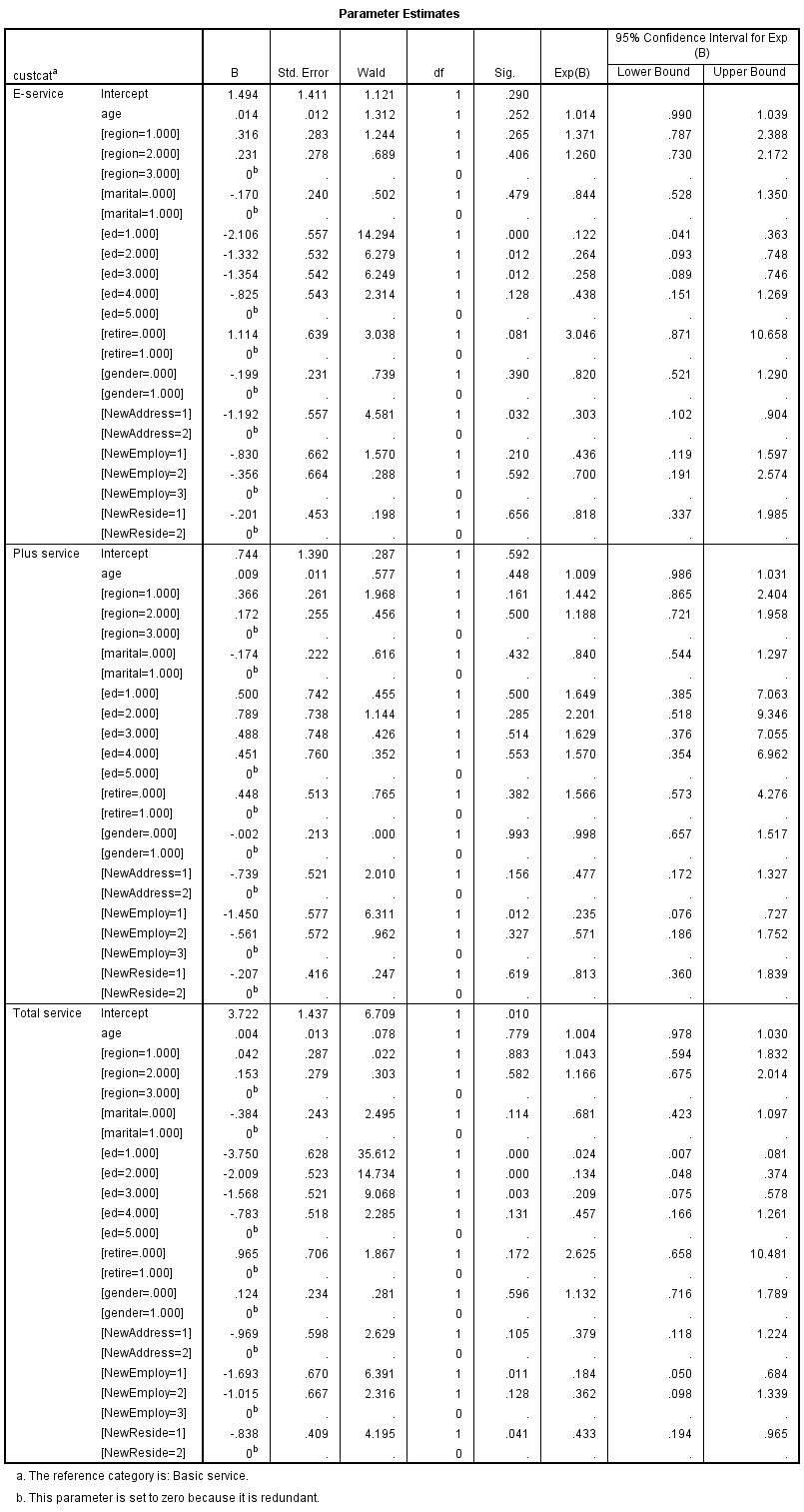
## APPENDICES

### APPENDIX A LOGISTIC REGRESSION (ENTER) MODEL



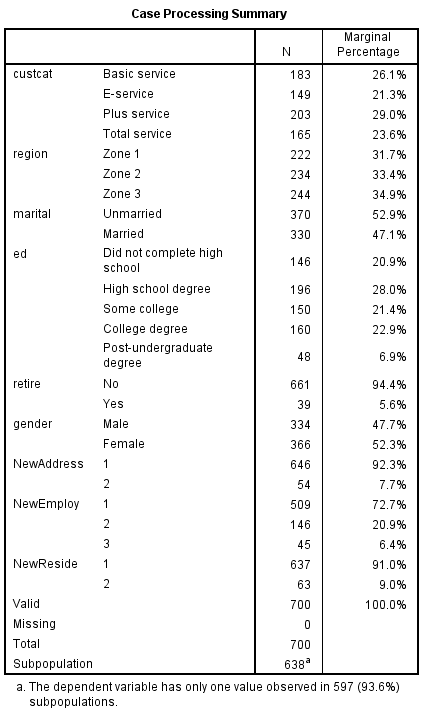


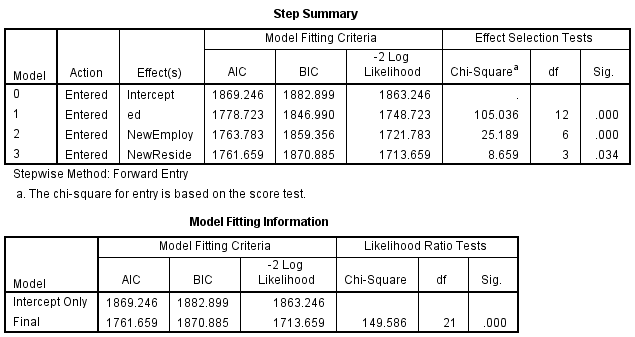


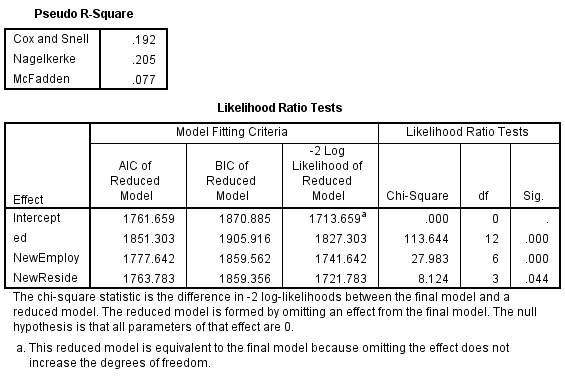


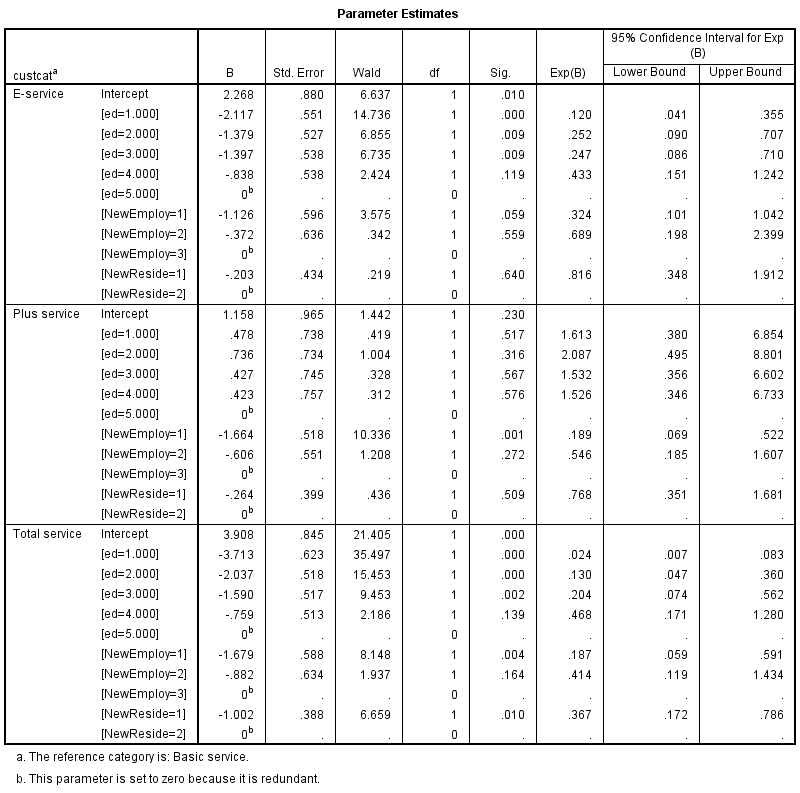
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classification** | | | | | |
| Observed | Predicted | | | | |
| Basic service | E-service | Plus service | Total service | Percent Correct |
| Basic service | 88 | 2 | 49 | 44 | 48.1% |
| E-service | 39 | 12 | 44 | 54 | 8.1% |
| Plus service | 71 | 0 | 93 | 39 | 45.8% |
| Total service | 36 | 7 | 30 | 92 | 55.8% |
| Overall Percentage | 33.4% | 3.0% | 30.9% | 32.7% | 40.7% |

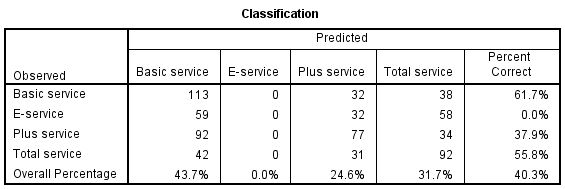
### APPENDIX B LOGISTIC REGRESSION (FORWARD) MODEL



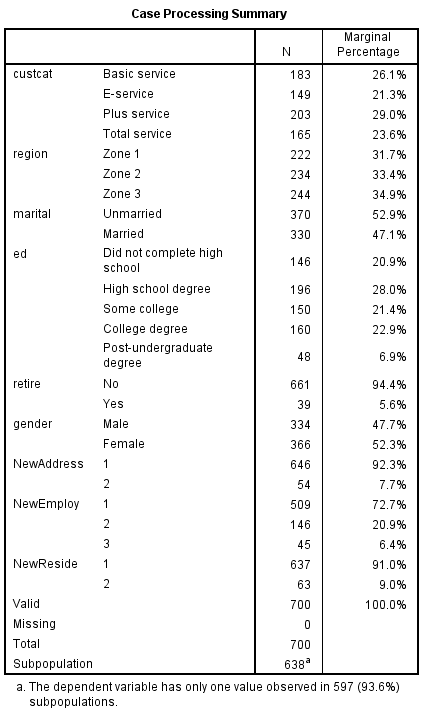


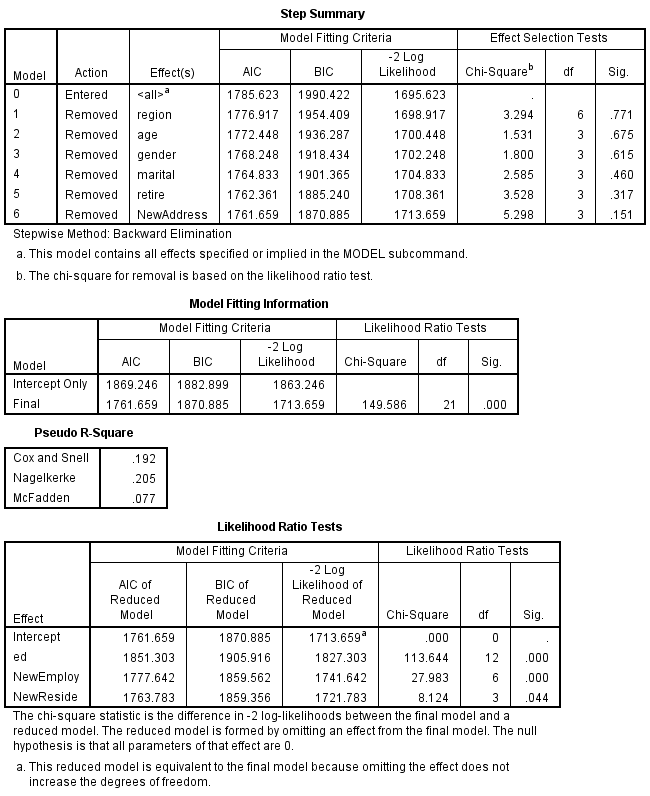


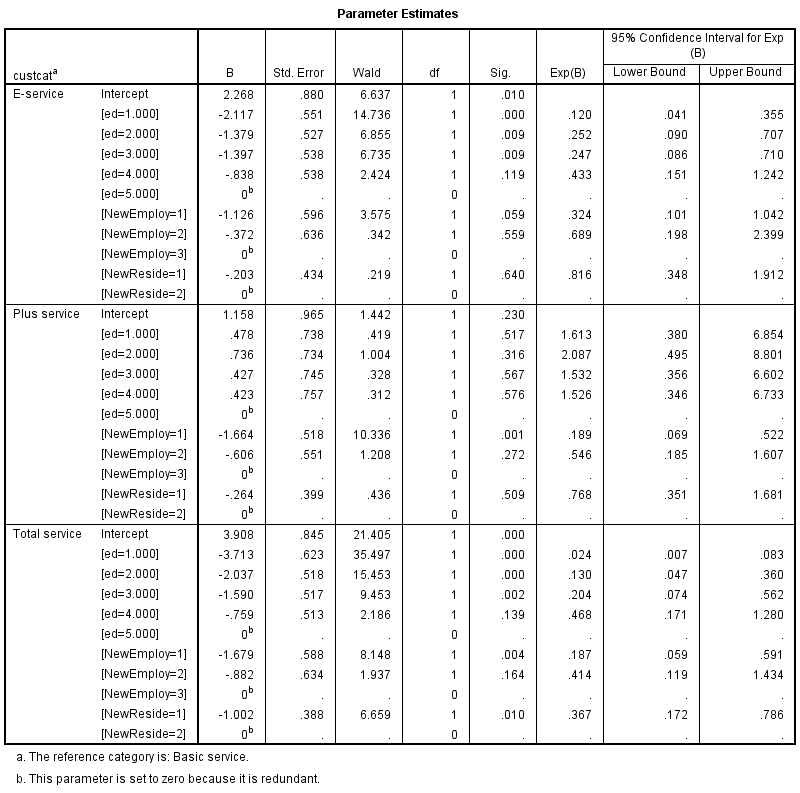


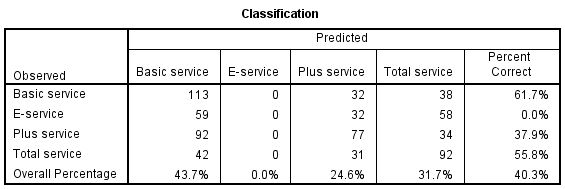


### APPENDIX C LOGISTIC REGRESSION (BACKWARD) MODEL

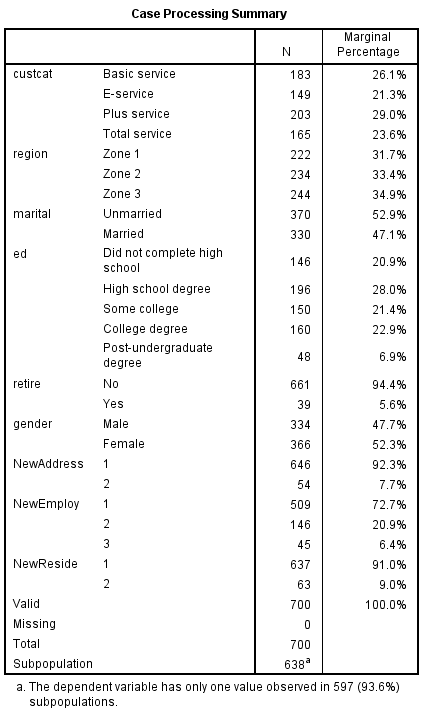


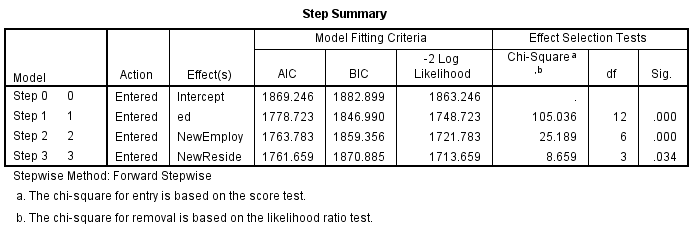


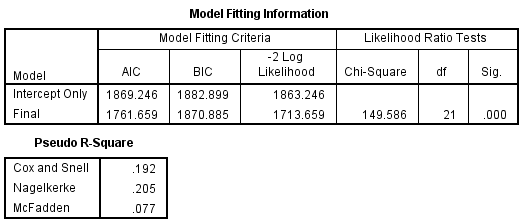


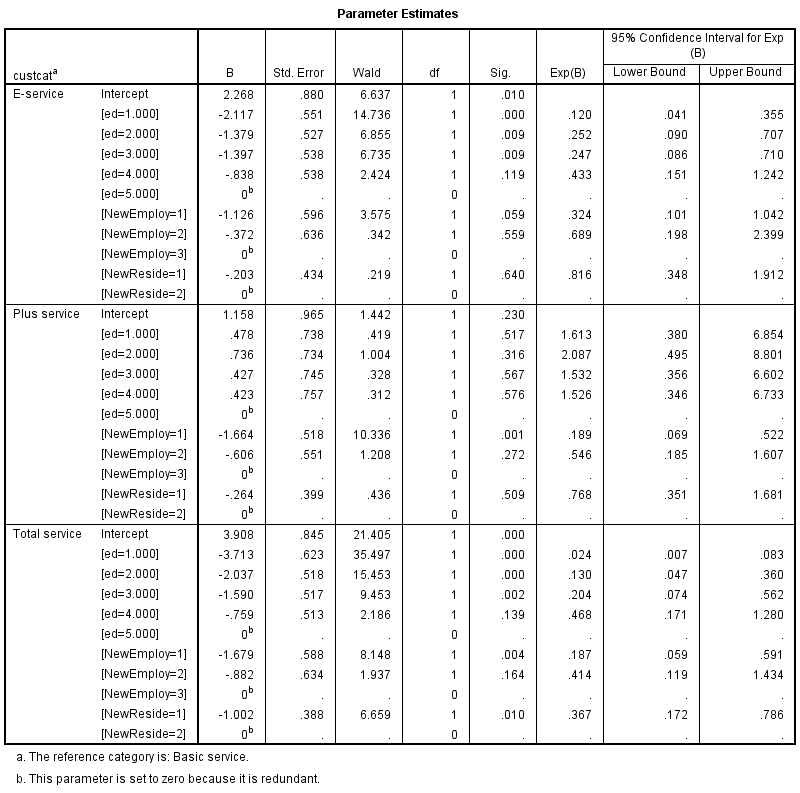


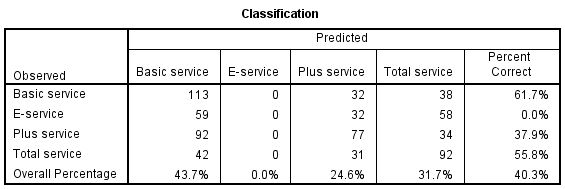
### APPENDIX D LOGISTIC REGRESSION (STEPWISE) MODEL

****

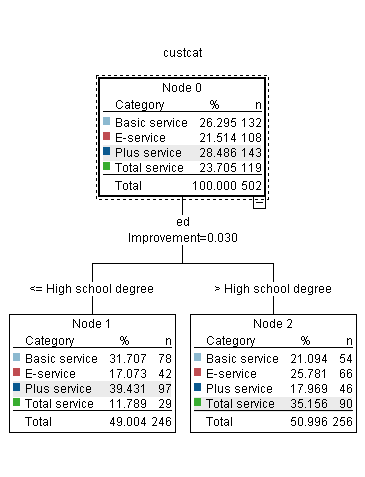
****

****

****

****

### APPENDIX E DECISION TREE (CART) MODEL

****

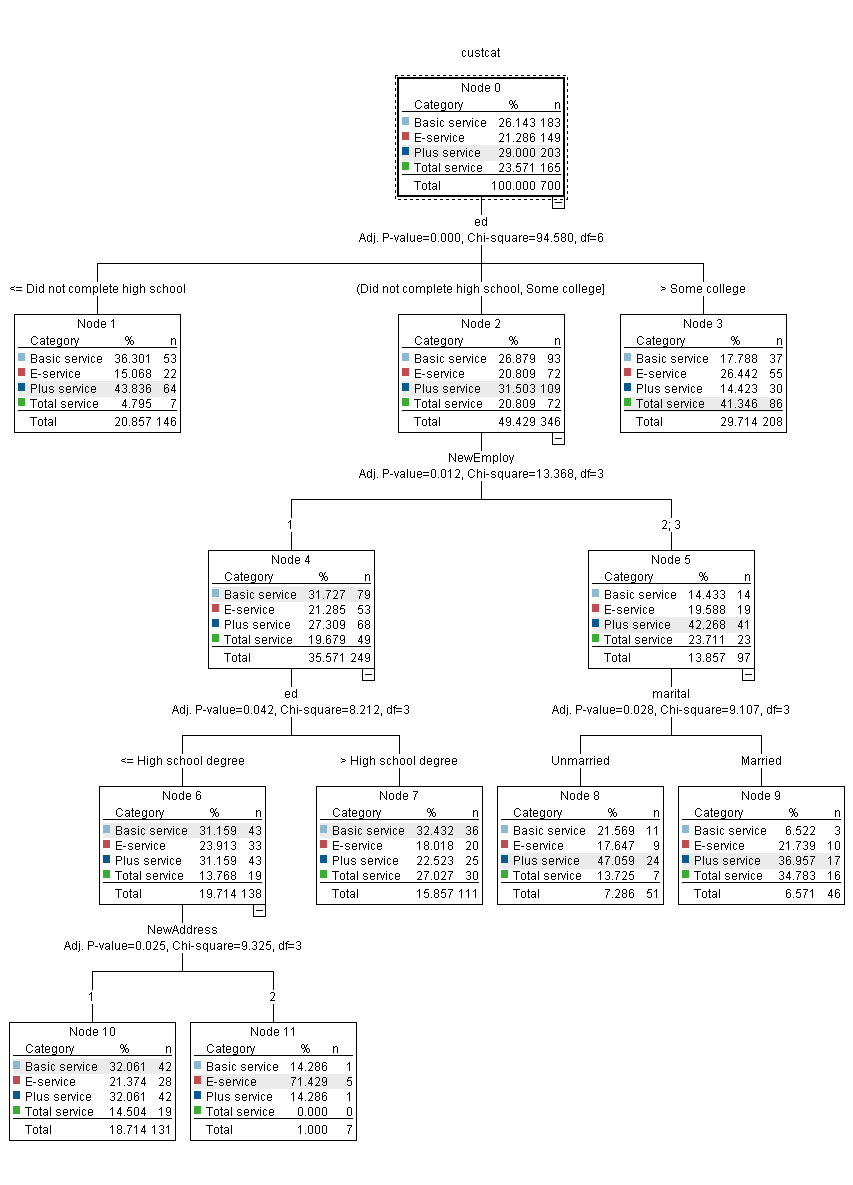
**CART DECISION RULES**



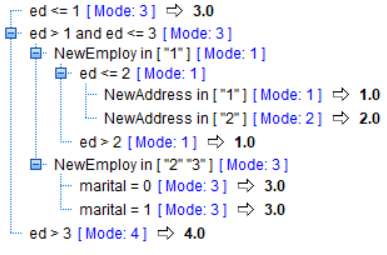
**Interpretation of CART Decision Rules**

1. If ED is college degree or did not complete high school, then CUSTCAT will be in Plus Service membership.
2. If ED is high school degree or post undergraduate degree and some college, then CUSTCAT will be in Total Service membership.

### APPENDIX E DECISION TREE (CHAID) MODEL

****

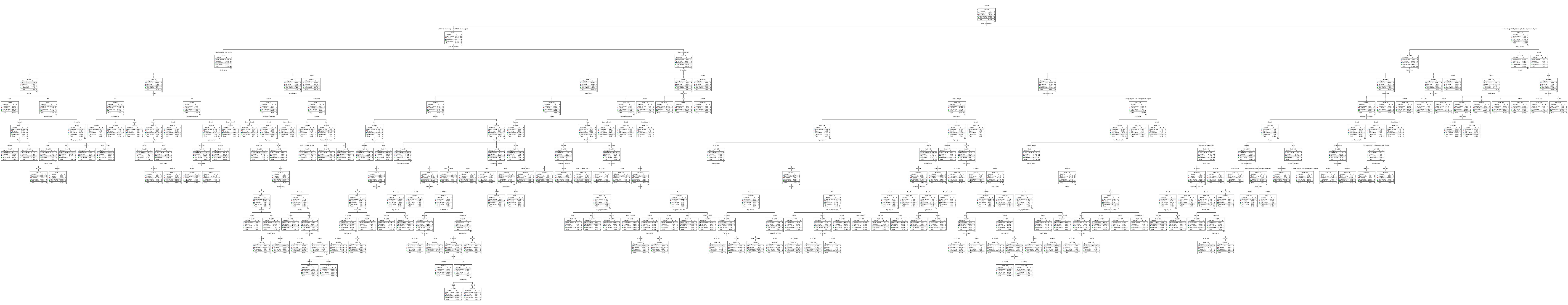
**CHAID DECISION RULES**



**Interpretation of CHAID Decision Rules**

1. If ED less or equal to 1, then CUSTCAT is Plus Service.
2. If ED more than 1 but less or equal to 3, and NEWEMPLOY in 1, and ED less or equal 2, and NEWADDRESS in 1, then CUSTCAT is Basic Service.
3. If ED more than 1 but less or equal to 3, and NEWEMPLOY in 1, and ED less or equal 2, and NEWADDRESS in 2, then CUSTCAT is E Service.
4. If ED more than 1 but less or equal to 3, and NEWEMPLOY in 1, and ED more than 2, then CUSTCAT is Basic Service.
5. If ED more than 1 but less or equal to 3, and NEWEMPLOY in 2 or 3, and marital equal to 0, then CUSTCAT is Plus Service.
6. If ED more than 1 but less or equal to 3, and NEWEMPLOY in 2 or 3, and marital equal to 1, then CUSTCAT is Plus Service.
7. If ED more than 3, then CUSTCAT is Total Service.

### APPENDIX F DECISION TREE (C5.0) MODEL

****

**C5.0 DECISION RULES**

ed in [ 1.000 2.000 ] [ Mode: 3 ]

ed in [ 1.000 ] [ Mode: 3 ]

NewEmploy = 1 [ Mode: 1 ]

retire = 1.000 [ Mode: 3 ] => 3.0

retire = 0.000 [ Mode: 1 ]

marital = 1.000 [ Mode: 1 ]

gender = 1.000 [ Mode: 1 ] => 1.0

gender = 0.000 [ Mode: 3 ] => 3.0

marital = 0.000 [ Mode: 1 ]

region in [ 1.000 ] [ Mode: 3 ]

age <= 42 [ Mode: 1 ] => 1.0

age > 42 [ Mode: 3 ] => 3.0

region in [ 2.000 ] [ Mode: 1 ] => 1.0

region in [ 3.000 ] [ Mode: 3 ] => 3.0

region in [ 4.000 5.000 ] [ Mode: 1 ] => 1.0

NewEmploy = 2 [ Mode: 3 ]

retire = 1.000 [ Mode: 1 ]

NewAddress = 1 [ Mode: 1 ] => 1.0

NewAddress = 2 [ Mode: 3 ] => 3.0

NewAddress = default [ Mode: 1 ] => 1.0

retire = 0.000 [ Mode: 3 ]

region in [ 1.000 ] [ Mode: 3 ]

gender = 1.000 [ Mode: 1 ] => 1.0

gender = 0.000 [ Mode: 3 ]

age <= 49 [ Mode: 2 ] => 2.0

age > 49 [ Mode: 3 ] => 3.0

region in [ 2.000 ] [ Mode: 3 ] => 3.0

region in [ 3.000 ] [ Mode: 3 ]

age <= 52 [ Mode: 1 ]

marital = 1.000 [ Mode: 1 ] => 1.0

marital = 0.000 [ Mode: 2 ] => 2.0

age > 52 [ Mode: 3 ] => 3.0

region in [ 4.000 5.000 ] [ Mode: 3 ] => 3.0

NewEmploy = 3 [ Mode: 3 ]

marital = 1.000 [ Mode: 3 ]

region in [ 1.000 3.000 ] [ Mode: 3 ] => 3.0

region in [ 2.000 ] [ Mode: 4 ]

age <= 62 [ Mode: 1 ] => 1.0

age > 62 [ Mode: 4 ] => 4.0

region in [ 4.000 5.000 ] [ Mode: 3 ] => 3.0

marital = 0.000 [ Mode: 2 ]

retire = 1.000 [ Mode: 1 ] => 1.0

retire = 0.000 [ Mode: 2 ]

region in [ 1.000 4.000 5.000 ] [ Mode: 2 ] => 2.0

region in [ 2.000 ] [ Mode: 3 ] => 3.0

region in [ 3.000 ] [ Mode: 2 ] => 2.0

NewEmploy = default [ Mode: 3 ] => 3.0

ed in [ 2.000 ] [ Mode: 3 ]

NewAddress = 1 [ Mode: 3 ]

NewEmploy = 1 [ Mode: 3 ]

retire = 1.000 [ Mode: 3 ]

gender = 1.000 [ Mode: 3 ] => 3.0

gender = 0.000 [ Mode: 1 ] => 1.0

retire = 0.000 [ Mode: 3 ]

NewReside = 1 [ Mode: 1 ]

region in [ 1.000 ] [ Mode: 3 ]

marital = 1.000 [ Mode: 1 ]

gender = 1.000 [ Mode: 1 ] => 1.0

gender = 0.000 [ Mode: 1 ]

age <= 38 [ Mode: 2 ] => 2.0

age > 38 [ Mode: 1 ] => 1.0

marital = 0.000 [ Mode: 3 ]

gender = 1.000 [ Mode: 3 ] => 3.0

gender = 0.000 [ Mode: 3 ]

age <= 31 [ Mode: 3 ] => 3.0

age > 31 [ Mode: 2 ]

age <= 54 [ Mode: 2 ] => 2.0

age > 54 [ Mode: 1 ] => 1.0

region in [ 2.000 ] [ Mode: 1 ]

marital = 1.000 [ Mode: 1 ]

age <= 49 [ Mode: 2 ]

age <= 45 [ Mode: 2 ] => 2.0

age > 45 [ Mode: 3 ] => 3.0

age > 49 [ Mode: 1 ] => 1.0

marital = 0.000 [ Mode: 1 ]

age <= 34 [ Mode: 1 ] => 1.0

age > 34 [ Mode: 2 ] => 2.0

region in [ 3.000 ] [ Mode: 3 ]

age <= 29 [ Mode: 1 ] => 1.0

age > 29 [ Mode: 3 ]

marital = 1.000 [ Mode: 3 ]

age <= 33 [ Mode: 2 ] => 2.0

age > 33 [ Mode: 3 ] => 3.0

marital = 0.000 [ Mode: 3 ]

age <= 48 [ Mode: 3 ]

gender = 1.000 [ Mode: 3 ] => 3.0

gender = 0.000 [ Mode: 1 ]

age <= 32 [ Mode: 3 ] => 3.0

age > 32 [ Mode: 1 ] => 1.0

age > 48 [ Mode: 2 ] => 2.0

region in [ 4.000 5.000 ] [ Mode: 1 ] => 1.0

NewReside = 2 [ Mode: 3 ]

region in [ 1.000 ] [ Mode: 3 ] => 3.0

region in [ 2.000 ] [ Mode: 4 ] => 4.0

region in [ 3.000 ] [ Mode: 1 ]

age <= 29 [ Mode: 2 ] => 2.0

age > 29 [ Mode: 1 ] => 1.0

region in [ 4.000 5.000 ] [ Mode: 3 ] => 3.0

NewReside = default [ Mode: 1 ] => 1.0

NewEmploy = 2 [ Mode: 3 ]

gender = 1.000 [ Mode: 3 ] => 3.0

gender = 0.000 [ Mode: 4 ]

marital = 1.000 [ Mode: 3 ]

region in [ 1.000 ] [ Mode: 3 ] => 3.0

region in [ 2.000 ] [ Mode: 2 ] => 2.0

region in [ 3.000 4.000 5.000 ] [ Mode: 3 ] => 3.0

marital = 0.000 [ Mode: 1 ]

age <= 47 [ Mode: 4 ] => 4.0

age > 47 [ Mode: 1 ] => 1.0

NewEmploy = 3 [ Mode: 3 ]

region in [ 1.000 3.000 ] [ Mode: 4 ] => 4.0

region in [ 2.000 ] [ Mode: 3 ] => 3.0

region in [ 4.000 5.000 ] [ Mode: 3 ] => 3.0

NewEmploy = default [ Mode: 3 ] => 3.0

NewAddress = 2 [ Mode: 3 ]

NewEmploy in [ "1" ] [ Mode: 2 ] => 2.0

NewEmploy in [ "2" "3" ] [ Mode: 3 ] => 3.0

NewEmploy in [ "default" ] [ Mode: 3 ] => 3.0

NewAddress = default [ Mode: 3 ] => 3.0

ed in [ 3.000 4.000 5.000 ] [ Mode: 4 ]

NewAddress = 1 [ Mode: 4 ]

NewEmploy = 1 [ Mode: 4 ]

ed in [ 3.000 ] [ Mode: 1 ]

NewReside = 1 [ Mode: 1 ]

age <= 38 [ Mode: 1 ]

marital = 1.000 [ Mode: 1 ]

gender = 1.000 [ Mode: 4 ]

region in [ 1.000 ] [ Mode: 4 ] => 4.0

region in [ 2.000 ] [ Mode: 2 ] => 2.0

region in [ 3.000 ] [ Mode: 1 ] => 1.0

region in [ 4.000 5.000 ] [ Mode: 4 ] => 4.0

gender = 0.000 [ Mode: 1 ]

region in [ 1.000 ] [ Mode: 1 ]

age <= 32 [ Mode: 1 ] => 1.0

age > 32 [ Mode: 2 ] => 2.0

region in [ 2.000 ] [ Mode: 1 ] => 1.0

region in [ 3.000 ] [ Mode: 2 ] => 2.0

region in [ 4.000 5.000 ] [ Mode: 1 ] => 1.0

marital = 0.000 [ Mode: 1 ]

gender = 1.000 [ Mode: 2 ]

age <= 34 [ Mode: 2 ]

age <= 32 [ Mode: 1 ] => 1.0

age > 32 [ Mode: 2 ] => 2.0

age > 34 [ Mode: 3 ]

region in [ 1.000 3.000 ] [ Mode: 3 ] => 3.0

region in [ 2.000 ] [ Mode: 2 ] => 2.0

region in [ 4.000 5.000 ] [ Mode: 3 ] => 3.0

gender = 0.000 [ Mode: 1 ]

region in [ 1.000 ] [ Mode: 4 ] => 4.0

region in [ 2.000 ] [ Mode: 3 ]

age <= 34 [ Mode: 3 ] => 3.0

age > 34 [ Mode: 4 ] => 4.0

region in [ 3.000 ] [ Mode: 1 ] => 1.0

region in [ 4.000 5.000 ] [ Mode: 1 ] => 1.0

age > 38 [ Mode: 1 ]

marital = 1.000 [ Mode: 3 ]

region in [ 1.000 ] [ Mode: 3 ]

age <= 41 [ Mode: 3 ] => 3.0

age > 41 [ Mode: 1 ] => 1.0

region in [ 2.000 ] [ Mode: 3 ] => 3.0

region in [ 3.000 ] [ Mode: 4 ]

age <= 44 [ Mode: 3 ] => 3.0

age > 44 [ Mode: 4 ] => 4.0

region in [ 4.000 5.000 ] [ Mode: 3 ] => 3.0

marital = 0.000 [ Mode: 1 ] => 1.0

NewReside = 2 [ Mode: 3 ]

age <= 23 [ Mode: 4 ] => 4.0

age > 23 [ Mode: 3 ]

age <= 31 [ Mode: 2 ] => 2.0

age > 31 [ Mode: 3 ] => 3.0

NewReside = default [ Mode: 1 ] => 1.0

ed in [ 4.000 5.000 ] [ Mode: 4 ]

NewReside = 1 [ Mode: 4 ]

ed in [ 4.000 ] [ Mode: 4 ]

marital = 1.000 [ Mode: 4 ]

age <= 26 [ Mode: 1 ] => 1.0

age > 26 [ Mode: 4 ] => 4.0

marital = 0.000 [ Mode: 4 ]

gender = 1.000 [ Mode: 1 ]

region in [ 1.000 ] [ Mode: 2 ]

age <= 37 [ Mode: 2 ] => 2.0

age > 37 [ Mode: 4 ] => 4.0

region in [ 2.000 ] [ Mode: 1 ]

age <= 28 [ Mode: 4 ] => 4.0

age > 28 [ Mode: 1 ]

age <= 43 [ Mode: 1 ] => 1.0

age > 43 [ Mode: 2 ] => 2.0

region in [ 3.000 ] [ Mode: 1 ]

age <= 39 [ Mode: 2 ] => 2.0

age > 39 [ Mode: 1 ] => 1.0

region in [ 4.000 5.000 ] [ Mode: 1 ] => 1.0

gender = 0.000 [ Mode: 4 ]

region in [ 1.000 ] [ Mode: 4 ]

age <= 39 [ Mode: 4 ] => 4.0

age > 39 [ Mode: 1 ] => 1.0

region in [ 2.000 ] [ Mode: 2 ] => 2.0

region in [ 3.000 ] [ Mode: 4 ] => 4.0

region in [ 4.000 5.000 ] [ Mode: 4 ] => 4.0

ed in [ 5.000 ] [ Mode: 4 ]

age <= 38 [ Mode: 4 ]

region in [ 1.000 ] [ Mode: 1 ]

age <= 31 [ Mode: 2 ] => 2.0

age > 31 [ Mode: 1 ] => 1.0

region in [ 2.000 ] [ Mode: 4 ] => 4.0

region in [ 3.000 ] [ Mode: 4 ]

marital = 1.000 [ Mode: 4 ] => 4.0

marital = 0.000 [ Mode: 4 ]

age <= 32 [ Mode: 1 ] => 1.0

age > 32 [ Mode: 4 ] => 4.0

region in [ 4.000 5.000 ] [ Mode: 4 ] => 4.0

age > 38 [ Mode: 2 ] => 2.0

NewReside = 2 [ Mode: 4 ] => 4.0

NewReside = default [ Mode: 4 ] => 4.0

NewEmploy = 2 [ Mode: 3 ]

NewReside = 1 [ Mode: 3 ]

region in [ 1.000 ] [ Mode: 3 ]

gender = 1.000 [ Mode: 3 ]

ed in [ 3.000 ] [ Mode: 2 ] => 2.0

ed in [ 4.000 5.000 ] [ Mode: 3 ]

age <= 52 [ Mode: 1 ] => 1.0

age > 52 [ Mode: 4 ] => 4.0

gender = 0.000 [ Mode: 3 ]

ed in [ 3.000 ] [ Mode: 3 ] => 3.0

ed in [ 4.000 5.000 ] [ Mode: 2 ] => 2.0

region in [ 2.000 ] [ Mode: 3 ]

ed in [ 3.000 ] [ Mode: 3 ]

marital = 1.000 [ Mode: 2 ] => 2.0

marital = 0.000 [ Mode: 3 ] => 3.0

ed in [ 4.000 5.000 ] [ Mode: 4 ]

age <= 51 [ Mode: 4 ] => 4.0

age > 51 [ Mode: 2 ] => 2.0

region in [ 3.000 ] [ Mode: 1 ] => 1.0

region in [ 4.000 5.000 ] [ Mode: 3 ] => 3.0

NewReside = 2 [ Mode: 4 ] => 4.0

NewReside = default [ Mode: 3 ] => 3.0

NewEmploy = 3 [ Mode: 4 ]

age <= 55 [ Mode: 3 ] => 3.0

age > 55 [ Mode: 4 ] => 4.0

NewEmploy = default [ Mode: 4 ] => 4.0

NewAddress = 2 [ Mode: 4 ]

gender = 1.000 [ Mode: 2 ]

NewEmploy = 1 [ Mode: 4 ]

age <= 53 [ Mode: 2 ] => 2.0

age > 53 [ Mode: 4 ] => 4.0

NewEmploy = 2 [ Mode: 2 ] => 2.0

NewEmploy = 3 [ Mode: 3 ] => 3.0

NewEmploy = default [ Mode: 2 ] => 2.0

gender = 0.000 [ Mode: 4 ]

age <= 62 [ Mode: 4 ] => 4.0

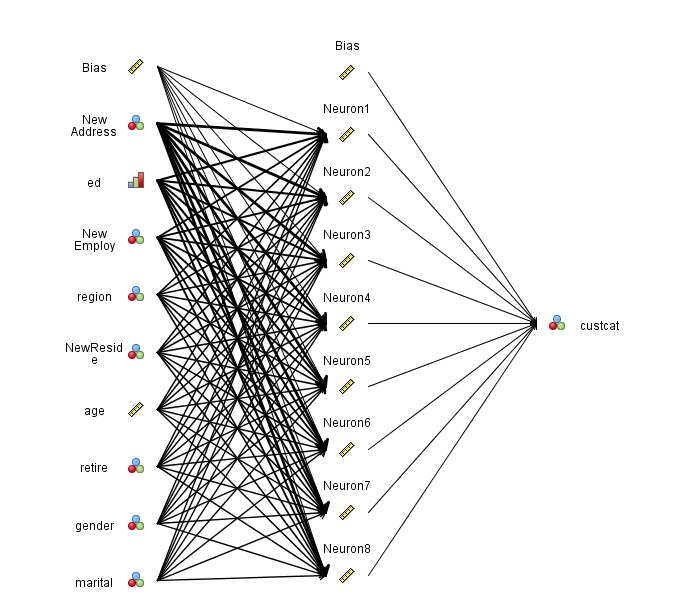
age > 62 [ Mode: 3 ] => 3.0

NewAddress = default [ Mode: 4 ] => 4.0

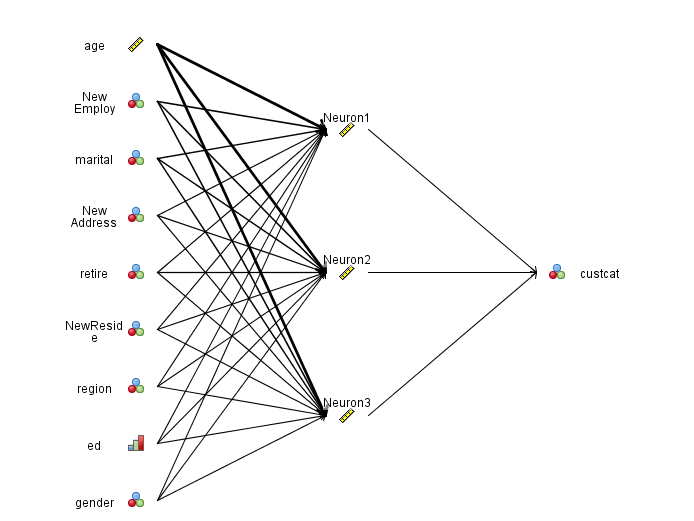
**Some of Interpretation of C5.0 Decision Rules**

1. If ED is 1, and NEWEMPLOY equal to 1, and RETIRE equal to 1, then CUSTCAT is Plus Service.
2. If ED is 1, and NEWEMPLOY equal to 1, and RETIRE equal to 0, and MARITAL equal to 1, and GENDER equal to 1, then CUSTCAT is Basic Service.
3. If ED is 1, and NEWEMPLOY equal to 1, and RETIRE equal to 0, and MARITAL equal to 1, and GENDER equal to 0, then CUSTCAT is Plus Service.
4. If ED is 1, and NEWEMPLOY equal to 1, and RETIRE equal to 0, and MARITAL equal to 0, and region in 1, and age less or equal to 42, then CUSTCAT is Basic Service.
5. If ED is 1, and NEWEMPLOY equal to 1, and RETIRE equal to 0, and MARITAL equal to 0, and region in 1, and age more than 42, then CUSTCAT is Plus Service.
6. If ED is 1, and NEWEMPLOY equal to 1, and RETIRE equal to 0, and MARITAL equal to 0, and region in 2, then CUSTCAT is Basic Service.
7. If ED is 1, and NEWEMPLOY equal to 1, and RETIRE equal to 0, and MARITAL equal to 0, and region in 3, then CUSTCAT is Plus Service.
8. If ED is 1, and NEWEMPLOY equal to 1, and RETIRE equal to 0, and MARITAL equal to 0, and region in 4 or 5, then CUSTCAT is Basic Service.
9. If ED is 1, and NEWEMPLOY equal to 2, and RETIRE equal to 1, and NEWADDRESS equal to 1, then CUSTCAT is Basic Service.
10. If ED is 1, and NEWEMPLOY equal to 2, and RETIRE equal to 1, and NEWADDRESS equal to 2, then CUSTCAT is Plus Service.
11. If ED is 1, and NEWEMPLOY equal to 2, and RETIRE equal to 1, and NEWADDRESS equal to default, then CUSTCAT is Basic Service.
12. If ED is 1, and NEWEMPLOY equal to 2, and RETIRE equal to 0, and REGION in 1, and GENDER equal to 1 then CUSTCAT is Basic Service.
13. If ED is 1, and NEWEMPLOY equal to 2, and RETIRE equal to 0, and REGION in 1, and GENDER equal to 0, and AGE less or equal to 49 then CUSTCAT is E Service.
14. If ED is 1, and NEWEMPLOY equal to 2, and RETIRE equal to 0, and REGION in 1, and GENDER equal to 0, and AGE more than 49 then CUSTCAT is Plus Service.
15. If ED is 1, and NEWEMPLOY equal to 2, and RETIRE equal to 0, and REGION in 2, then CUSTCAT is Plus Service.
16. If ED is 1, and NEWEMPLOY equal to 2, and RETIRE equal to 0, and REGION in 3, and AGE less or equal to 52, and MARITAL equal to 1, then CUSTCAT is Basic Service.
17. If ED is 1, and NEWEMPLOY equal to 2, and RETIRE equal to 0, and REGION in 3, and AGE less or equal to 52, and MARITAL equal to 0, then CUSTCAT is E Service.
18. If ED is 1, and NEWEMPLOY equal to 2, and RETIRE equal to 0, and REGION in 3, and AGE more than 52, then CUSTCAT is Plus Service.
19. If ED is 1, and NEWEMPLOY equal to 2, and RETIRE equal to 0, and REGION in 4 or 5, then CUSTCAT is Plus Service.
20. If ED is 1, and NEWEMPLOY equal to 3, and MARITAL equal to 1, and REGION in 1 or 3, then CUSTCAT is Plus Service.
21. If ED is 1, and NEWEMPLOY equal to 3, and MARITAL equal to 1, and REGION in 2, and AGE less or equal to 62 then CUSTCAT is Basic Service.
22. If ED is 1, and NEWEMPLOY equal to 3, and MARITAL equal to 1, and REGION in 2, and AGE more than 62 then CUSTCAT is Total Service.
23. If ED is 1, and NEWEMPLOY equal to 3, and MARITAL equal to 1, and REGION in 4 or 5, then CUSTCAT is Plus Service.
24. If ED is 1, and NEWEMPLOY equal to 3, and MARITAL equal to 0, and RETIRE in 1, then CUSTCAT is Basic Service.
25. If ED is 1, and NEWEMPLOY equal to 3, and MARITAL equal to 0, and RETIRE in 0, and REGION in 1 or 4 or 5, then CUSTCAT is E Service.
26. If ED is 1, and NEWEMPLOY equal to 3, and MARITAL equal to 0, and RETIRE in 0, and REGION in 2, then CUSTCAT is Plus Service.
27. If ED is 1, and NEWEMPLOY equal to 3, and MARITAL equal to 0, and RETIRE in 0, and REGION in 3, then CUSTCAT is E Service.
28. If ED is 1, and NEWEMPLOY is default, then CUSTCAT is Plus Service.

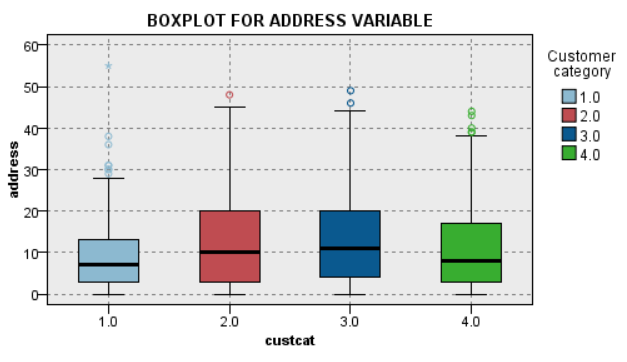
### APPENDIX g mlp network

****

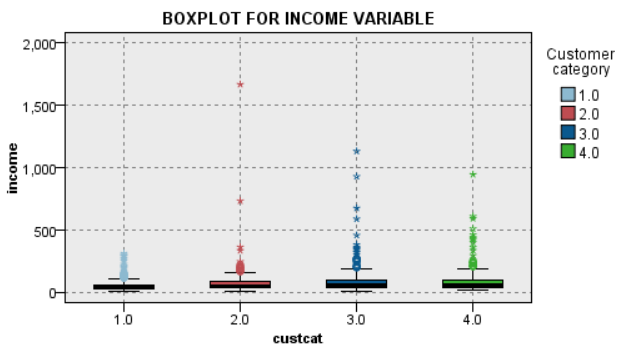
### APPENDIX h rbf network

****

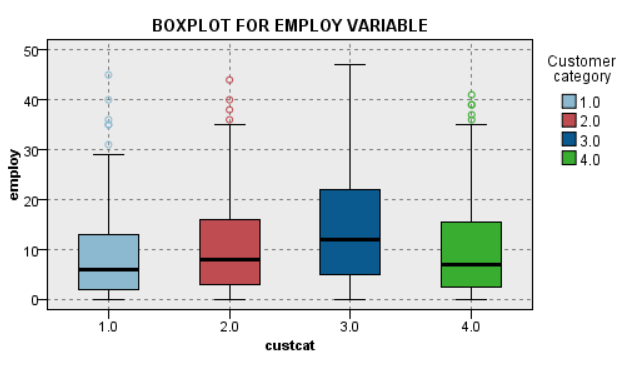
### APPENDIX i VARIABLE ANALYSIS AGAINST TARGET VARIABLES



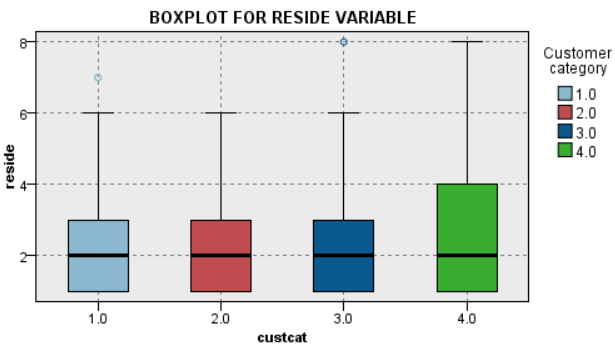
Boxplot in ADDRESS variable shown that the data is slightly skew to the right (positive skew). The dots on the upper whisker represent an outlier. Outliers was observed in ADDRESS variables above maximum (Q3+1.5\*IQR where IQR is Interquartile Range).



Boxplot in INCOME variable shown that the data is slightly skew to the right (positive skew). The dots on the upper whisker represent an outlier. Outliers was observed in INCOME variables above maximum (Q3+1.5\*IQR where IQR is Interquartile Range).



Boxplot in EMPLOY variable shown that the median line of a box plot lies inside of the box of comparison box plot, then there is likely to be a similarity between the two groups. Both boxes have different length and that is indicate there are difference in the data dispersion. The dots on the upper whisker represent an outlier. Outliers was observed in VALUE variables above maximum (Q3+1.5\*IQR where IQR is Interquartile Range). For the skewness, it can be concluded that the data is a positive skewed.



Boxplot in RESIDE variable shown that the median line of a box plot lies inside of the box of comparison box plot, then there is likely to be a similarity between the two groups. Both boxes have different length and that is indicate there are difference in the data dispersion. The dots on the upper whisker represent an outlier. Outliers was observed in RESIDE variables above maximum (Q3+1.5\*IQR where IQR is Interquartile Range). For the skewness, it can be concluded that the data is a positive skewed.