STAT2604 Final Project

Abstract

This study aims to find out (1)the proportion of good customers that can be granted loans while ensuring that 5, 1, 0.5% of the bad customers are wrongly identified and (2)the top 3 most important explanatory variables that affect whether a customer is good or bad by studying the data set about past bank customers. Firstly, a summary table of the data set will be given. Then, an initial suggestion will provided. After that, a study of correlation of variables will be conducted in order to reduce the dimension of variable. The data quality issue will be solved before the last step. At last, a tuned logistic regression model will be provided to explain the importance of each variable and make prediction of good or bad customers.

Summary Statistics

The summary statistics is given by the following table. In this stage, the missing value is not being counted in the summary.

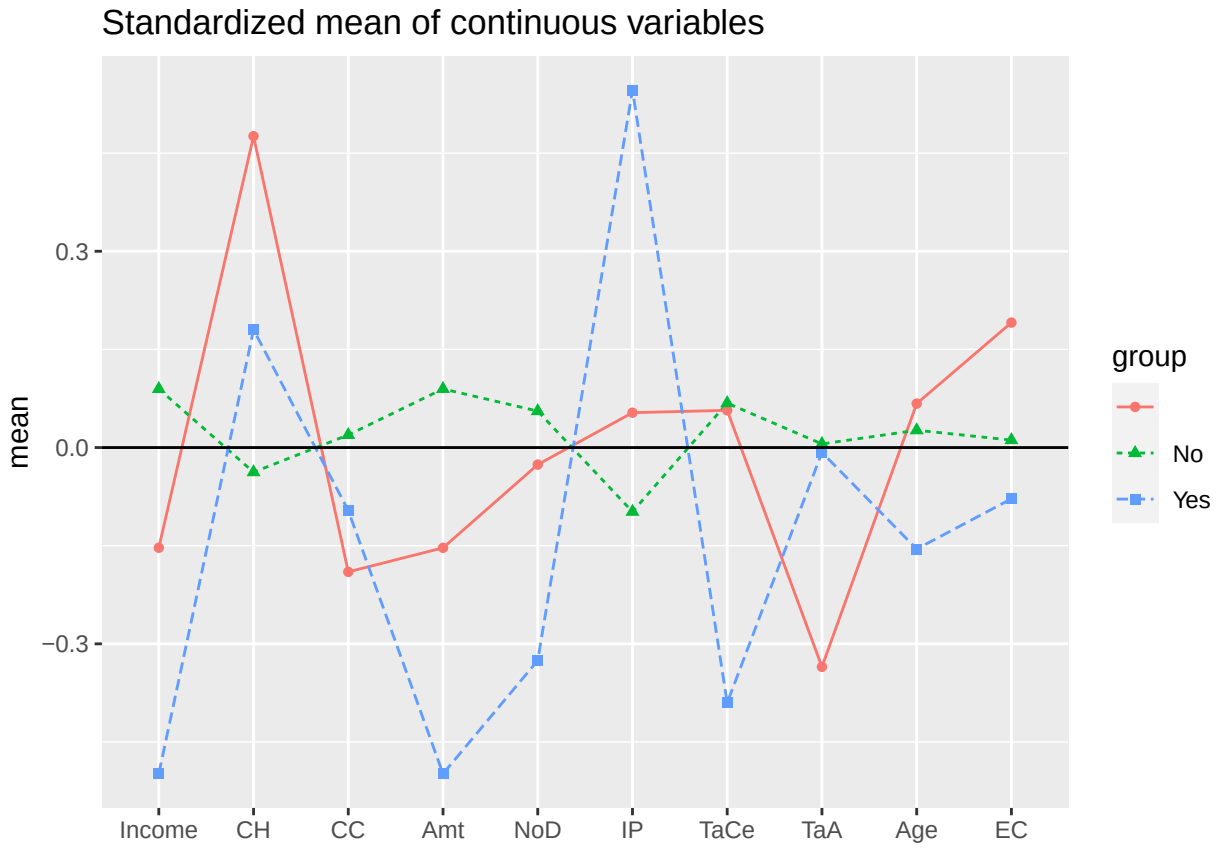
Summary Statistics	Not indicate(1%)	Bad Customer(84%)	Good Customer(15%)
Annual Gross Income in $			
min	4421.323454	1891.143246	1377.434352
max	86916.46917	252192.8146	165985.9497
mean (sd)	33,032.13 ± 22,963.10	40,909.79 ± 33,463.32	21,825.43 ± 20,093.79
Loan applications in past five years			
min	0	0	0
median	2	2	2
max	7	9	9
mean (sd)	2.81 ± 2.11	2.04 ± 1.47	2.36 ± 1.61
Credit cards currently held			
min	0	0	0
max	4	4	4
mean (sd)	1.62 ± 1.32	1.92 ± 1.41	1.75 ± 1.54
Loan amount			
min	23384.97036	15129.14597	11019.47481
max	270761.4075	766612.4437	507953.849
mean (sd)	109,085.78 ± 68,892.22	132,733.58 ± 100,388.90	75,467.76 ± 60,291.50
Installment Percentage			
min	15.58	15.2	15.3011
max	26.45	40	40
mean	17.6971428571429	17.3019229518565	18.9837902027027
Time at Current Employment(Years)			

Summary Statistics	Not indicate(1%)	Bad Customer(84%)	Good Customer(15%)
min	3	1	1
median	7	7	6
max	11	17	14
mean (sd)	7.05 ± 2.50	7.08 ± 2.63	5.87 ± 2.48
Time at Address(Years)			
min	1	0	1
max	9	14	11
mean (sd)	4.24 ± 2.19	5.01 ± 2.27	4.98 ± 2.22
Age(Years)			
min	24	19	19
max	53	75	74
mean (sd)	36.57 ± 9.04	36.10 ± 11.72	33.98 ± 11.27
Number of Dependants			
min	1	1	1
median	2	1	1
max	3	3	3
mean	1.63636363636364	1.70584284754869	1.38188976377953
Additional lines of credits			
min	1	1	1
max	3	4	4
mean (sd)	1.52 ± 0.60	1.42 ± 0.58	1.37 ± 0.56
Area_Indicator			
0	1 (4.76%)	20 (1.19%)	21 (7.09%)
1	3 (14.29%)	122 (7.25%)	80 (27.03%)
2	6 (28.57%)	479 (28.46%)	101 (34.12%)
3	9 (42.86%)	709 (42.13%)	73 (24.66%)
4	2 (9.52%)	353 (20.97%)	21 (7.09%)
Employment(Counts(%))			
Other	0 (0.00%)	0 (0.00%)	0 (0.00%)
Self Employment	1 (4.76%)	104 (6.18%)	23 (7.77%)
Part time	5 (23.81%)	545 (32.38%)	104 (35.14%)
Full time private sector	5 (23.81%)	326 (19.37%)	39 (13.18%)
Full time public sector	5 (23.81%)	465 (27.63%)	63 (21.28%)
Delayed or Missed Payments(Counts(%))			
No missed/delayed payments over last 3 years	18 (85.71%)	1,447 (85.98%)	220 (74.32%)
Delayed payments only over last 3 years	3 (14.29%)	215 (12.77%)	70 (23.65%)
Missed payments over last 3 years	0 (0.00%)	21 (1.25%)	6 (2.03%)
Residential_Status(Counts(%))			
Own	15 (71.43%)	1,238 (73.56%)	184 (62.16%)
Live with Family	2 (9.52%)	156 (9.27%)	43 (14.53%)
Rent	4 (19.05%)	289 (17.17%)	69 (23.31%)

The proportion of bad customer is much greater than good customer in our sample. From the table, we can find that the loan amount and area indicator of bad customer is greater than the good customer.

Initial suggestion

In order to understand the characteristics of good and bad customer, we are going to plot the mean of standardized data of the continuous variable of each group of customers. The variables that have a large different between groups maybe the important variables that can distinguish the customers.



Since there is some missing value of whether the customer is a good customer, there is a third group of not indicate. We will ignore those value at this stage.

From the graph, we can see that there is a huge difference in Annual Income(Income), Loan Amount(Amt), Installment Percentage(IP) between good customer and bad customer.

We find that the standardized mean of Annual Income(Income) and Loan Amount(Amt) are the same, this may indicate that the two variable can be explained together. We will try to find out there relationship in the next part.

For the other non-continuous variables, we will have chi-square test between them and the good-customer variables to find whether there are relationship.

```
chisq.test(x$Good_Customer, x$Area, correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  x$Good_Customer and x$Area
## X-squared = 187.93, df = 8, p-value < 2.2e-16
```

```
chisq.test(x$Good_Customer, x$character_Employment, correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  x$Good_Customer and x$character_Employment
## X-squared = 23.206, df = 8, p-value = 0.00311
```

```
chisq.test(x$Good_Customer, x$character_Delayed_Missed_Payments, correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  x$Good_Customer and x$character_Delayed_Missed_Payments
## X-squared = 26.151, df = 4, p-value = 2.95e-05
```

```
chisq.test(x$Good_Customer, x$Residential_Status, correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  x$Good_Customer and x$Residential_Status
## X-squared = 16.734, df = 4, p-value = 0.002177
```

For the non continuous variable, it can be found that Good customer and area indicator may have a strong relationship because the chi-square value is very high and the p-value is very low.

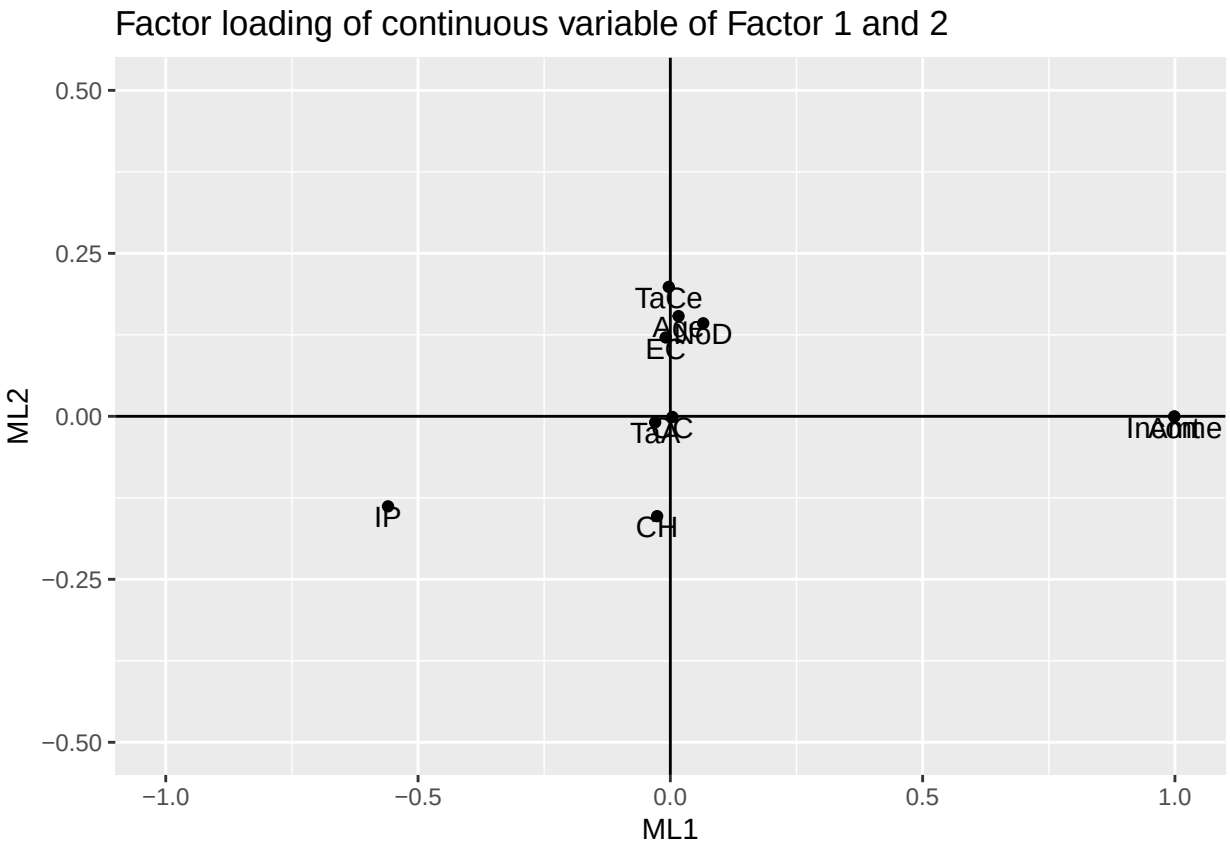
As an initial suggestion, Annual Income(Income), Installment Percentage(IP) and Area Indicator may be the top three most important variable.

Correlation of variable

We are going to find out whether there is correlation between different continuous variables in order to avoid over-estimate the importance of the variables. If there is any variables have a strong Correlation, we will group them together as factor by factor analysis. The result is the following:

```
## Factor Analysis using method =  ml
## Call: fa(r = scale(x[, c("Income", "CH", "CC", "Amt", "NoD", "IP",
##     "TaCe", "TaA", "Age", "EC")]), nfactors = 10, rotate = "none",
##     fm = "ml")
## Standardized loadings (pattern matrix) based upon correlation matrix
##          ML1   ML2   ML3   ML4   ML5 ML6 ML7 ML8 ML9 ML10    h2     u2 com
## Income  1.00  0.00  0.00  0.00  0.00   0   0   0   0    0 0.997 0.0025 1.0
## CH     -0.03 -0.15 -0.03 -0.08  0.06   0   0   0   0    0 0.035 0.9649 2.0
## CC      0.00  0.00  0.02  0.16  0.06   0   0   0   0    0 0.029 0.9709 1.3
## Amt     1.00  0.00  0.00  0.00  0.00   0   0   0   0    0 0.998 0.0025 1.0
## NoD     0.07  0.14 -0.13 -0.11  0.01   0   0   0   0    0 0.053 0.9466 3.3
## IP     -0.56 -0.14  0.16 -0.02 -0.03   0   0   0   0    0 0.359 0.6413 1.3
## TaCe    0.00  0.20  0.03  0.06  0.00   0   0   0   0    0 0.044 0.9556 1.3
## TaA    -0.03 -0.01 -0.07  0.08 -0.09   0   0   0   0    0 0.022 0.9781 3.2
```

```
## Age     0.02  0.15  0.16 -0.10 -0.02   0   0   0   0    0 0.058 0.9419 2.7
## EC     -0.01  0.12  0.14  0.03  0.05   0   0   0   0    0 0.036 0.9642 2.3
##
##                        ML1  ML2  ML3  ML4  ML5  ML6  ML7  ML8  ML9 ML10
## SS loadings           2.31 0.14 0.09 0.06 0.02 0.00 0.00 0.00 0.00 0.00
## Proportion Var        0.23 0.01 0.01 0.01 0.00 0.00 0.00 0.00 0.00 0.00
## Cumulative Var        0.23 0.25 0.25 0.26 0.26 0.26 0.26 0.26 0.26 0.26
## Proportion Explained  0.88 0.05 0.04 0.02 0.01 0.00 0.00 0.00 0.00 0.00
## Cumulative Proportion 0.88 0.93 0.97 0.99 1.00 1.00 1.00 1.00 1.00 1.00
##
## Mean item complexity =  1.9
## Test of the hypothesis that 10 factors are sufficient.
##
## The degrees of freedom for the null model are  45  and the objective function was  13.77 with Chi Sq
## The degrees of freedom for the model are -10  and the objective function was  7.07
##
## The root mean square of the residuals (RMSR) is  0.01
## The df corrected root mean square of the residuals is  NA
##
## The harmonic number of observations is  1944 with the empirical chi square  20.7  with prob <  NA
## The total number of observations was  2000  with Likelihood Chi Square =  14063.85  with prob <  NA
##
## Tucker Lewis Index of factoring reliability =  3.316
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##                                                 ML1   ML2   ML3   ML4   ML5
## Correlation of (regression) scores with factors   1  0.36  0.31  0.25  0.14
## Multiple R square of scores with factors          1  0.13  0.10  0.06  0.02
## Minimum correlation of possible factor scores     1 -0.74 -0.81 -0.88 -0.96
##                                                 ML6 ML7 ML8 ML9 ML10
## Correlation of (regression) scores with factors   0   0   0   0    0
## Multiple R square of scores with factors          0   0   0   0    0
## Minimum correlation of possible factor scores    -1  -1  -1  -1   -1
```



From the diagram, we find that the factor loading of Annual Income(Income) and Loan amount(Amt) in factor 1 is the same. The two variables have a strong correlation. Therefore, we are going to use the factor score of factor 1 to replace the Annual Income(Income) and Loan amount(Amt) variable in the modeling part. For the other variables, we decide to keep them as their correlation are not as strong as Annual Income(Income) and Loan amount(Amt).

data quality

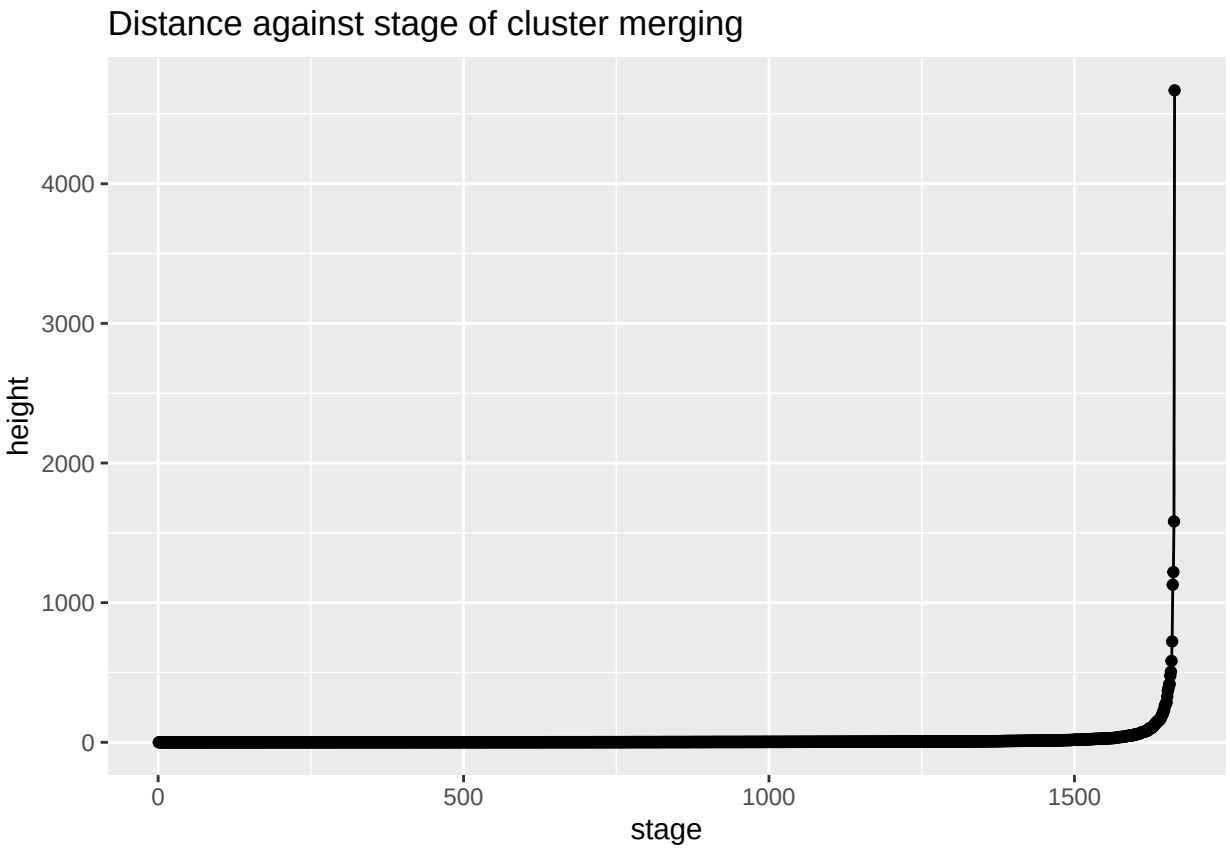
From the data set, we find that there are a lot of outliers and missing data. We are going to handle them in different ways.

outliers

For the outliers, we are going to remove the customers data that contain at least one outliers data of the continuous variables. After removing, there are still 1665 observation.

missing data

we find that there are missing data in Number of Dependents, Installment Percentage and Good Customer. We are going to replace the missing value by the value of similar data. In order to find similar data, we decided to cluster the data set by wards method. Then the missing data will be replace by the cluster mean or median or proportion based on their nature.



From the height - stage plot, we find that there is a huge increase of distance between 3 cluster solution to a 2 cluster solution, so a 3 cluster solution is suggested. There are 3 set of similar data. For cluster 1, the mean of Installment Percentage is 17.64414, median of Number of Dependents is 1 and 79% is a bad customers. For cluster 2, the mean of Installment Percentage is 15.82150, median of Number of Dependents is 1 and 93% is a bad customers. For cluster 3, the mean of Installment Percentage is 17.59769, median of Number of Dependents is 1 and 85% is a bad customers. Therefore the missing data will replace by the above value if any.

Modeling

In order to explain the relative importance of the continuous variables and levels of factor, we decided to apply logistics regression model on the data and then calculate the relative importance of the variables by the part-worth utilities.

We first split the the data into 80% of training set and 20% of testing set.

```
x = x[,c("CH", "CC", "ML1", "NoD", "IP", "TaCe", "TaA", "Age", "EC", "Area", "character_Employment", "cl
x$Good_Customer = factor(x$Good_Customer)
x$character_Employment = factor(x$character_Employment)
x$character_Delayed_Missed_Payments = factor(x$character_Delayed_Missed_Payments)
x$Area = factor(x$Area)
x$Residential_Status = factor(x$Residential_Status)
set.seed(2604)
trainIndex <- createDataPartition(x$Good_Customer, p = .8, list = FALSE)
train <- x[ trainIndex,]
```

```
test  <- x[-trainIndex,]
head(train)
```

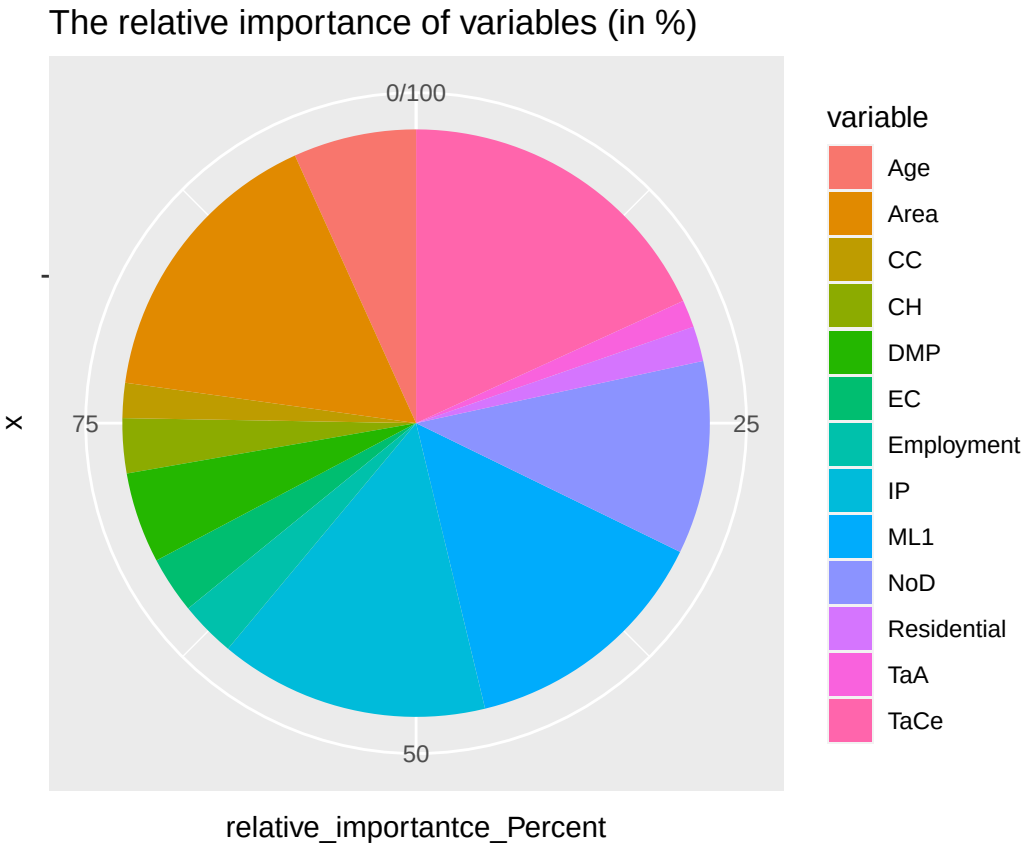
```
##    CH CC        ML1 NoD      IP TaCe TaA Age EC Area character_Employment
## 5   3  0  2.6261056   1 15.5909    8   5  19  1    3                     3
## 6   4  3 -1.0311713   2 19.3823    6   6  32  1    2                     3
## 7   3  0  0.8670612   1 15.9944    3   5  25  2    3                     3
## 9   3  2 -0.6750263   1 17.6382    7   2  33  2    3                     3
## 10  1  3 -0.4954476   1 17.2160    6   3  30  1    3                     3
## 11  4  0 -0.5275475   2 17.3237    4   8  22  2    3                     3
##    character_Delayed_Missed_Payments Good_Customer Residential_Status
## 5                                  0            No                Own
## 6                                  0           Yes                Own
## 7                                  1           Yes                Own
## 9                                  1            No                Own
## 10                                 0            No                Own
## 11                                 0            No   Live with Family
```

Result

The following are the results of the tuned model:

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8523  -0.5113  -0.2983  -0.1537   2.9952
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                        1.643322   1.636810   1.004 0.315389
## CH                                 0.096672   0.065916   1.467 0.142488
## CC                                -0.094738   0.064061  -1.479 0.139174
## ML1                               -0.597872   0.167810  -3.563 0.000367 ***
## NoD                               -0.510050   0.131334  -3.884 0.000103 ***
## IP                                 0.129646   0.078931   1.643 0.100483
## TaCe                              -0.232852   0.037629  -6.188 6.09e-10 ***
## TaA                               -0.029928   0.043530  -0.688 0.491746
## Age                               -0.021802   0.009089  -2.399 0.016454 *
## EC                                -0.304947   0.180185  -1.692 0.090568 .
## Area1                             -0.602112   0.554958  -1.085 0.277935
## Area2                             -1.937857   0.540593  -3.585 0.000337 ***
## Area3                             -2.754369   0.545960  -5.045 4.54e-07 ***
## Area4                             -3.083709   0.591955  -5.209 1.89e-07 ***
## character_Employment2              0.246836   0.397086   0.622 0.534192
## character_Employment3              0.036007   0.375761   0.096 0.923661
## character_Employment4             -0.597933   0.419374  -1.426 0.153933
## character_Employment5             -0.307543   0.401063  -0.767 0.443188
## character_Delayed_Missed_Payments1  0.969790   0.226727   4.277 1.89e-05 ***
```

```
## character_Delayed_Missed_Payments2  0.525765   0.738002   0.712 0.476207
## Residential_StatusOwn              -0.378317   0.295726  -1.279 0.200798
## Residential_StatusRent             -0.189140   0.341888  -0.553 0.580110
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1095.42  on 1332  degrees of freedom
## Residual deviance:  818.45  on 1311  degrees of freedom
## AIC: 862.45
##
## Number of Fisher Scoring iterations: 6
```




```
##       variable relative_importantce_Percent
## 1           CH                     3.019261
## 2           CC                     1.926080
## 3          ML1                    14.003123
## 4          NoD                    10.619469
## 5           IP                    14.836023
## 6         TaCe                    18.167621
## 7          TaA                     1.509630
## 8          Age                     6.767309
## 9           EC                     3.123373
## 10        Area                    16.033316
```

```
## 11  Employment                    3.071317
## 12         DMP                    4.997397
## 13 Residential                    1.926080
```

Objective One

To let the wrongly prediction of bad customer be a certain level, we decide to study the certain quantile of the predict value of the testing data of the bad customer, if the threshold is set below the quantile, certain % of prediction of bad customer will be make. Therefore, we can know the threshold value in different false negative level. The proportion of good customer is the proportion of the predict value that smaller than the threshold value.

```
pred_test.logit1 = pred_test.logit[pred_test.logit$obs == "No", ]

sum(pred_test.logit$No < quantile(pred_test.logit1$No, 0.05)) /332
```

```
## [1] 0.1054217
```

```
sum(pred_test.logit$No < quantile(pred_test.logit1$No, 0.01)) /332
```

```
## [1] 0.03012048
```

```
sum(pred_test.logit$No < quantile(pred_test.logit1$No, 0.005)) /332
```

```
## [1] 0.01807229
```

At 5 % level, the proportion of good customer is 10.5% At 1 % level, the proportion of good customer is 3% At 0.5 % level, the proportion of good customer is 1.8%

Conclusion

Based on the part-worth utility of the levels of factor and different continuous variable, pie chart and tables of relative importance is produced. From the result, the top three most importance variables are Time at Current Employment, Area indicator and Installment Percentage. Their contribution are 18.2%, 16% and 14.8% Representative.The result is not fully consistent to the previous suggestions. We believe it is because of the outliers removing and missing data handling.