UNIVERSITY OF
TORONTO

MIE1413 Final Project:
Predicting priorities of potential clients

Yunjie Xu, Zheyuan Fan

Department of Mechanical and Industrial Engineering, University of Toronto

MIE1413: Statistical Models in Empirical Research

Instructor: Prof. Birsen Donmez

April 14, 2022

Contents

1 Introduction

1.1 Background

The nature of the insurance pricing is that people with different background have different insurance prices. Thus, it might take a certain amount of time for the insurance company to determine the profit generated from each customer. Hence, determining the factors affecting the profitability of customers is crucial for insurance companies. That motivates us to do this project of predicting priorities of potential clients.

The setting of the problem is, we are two members of an actuary department at an insurance company called MEB insurance, and we were asked to assist the marketing department, who have been collecting past policyholders' data and scoring whether the profitability is high or low for them. They wished to predict the level of profitability for each prospective policyholder.

1.2 Method

Our research question is: how will the factors - age, education level, marital status, number of hours working per week, occupations, and the "score" developed by the marketing department, affect the company's decision on whether a potential customer is valuable or not. We will approach this problem in the following way. Firstly, we will prepare our data by exploring, cleaning, and splitting it into training and test data. Secondly, we are going to fit some Logistic models and implement model selection to find the best model. Next, we will interpret our results and discuss our findings for both the questions and the model. Finally, we will look for possible improvements of our project and discuss work we could do in the future.

2 Data Preparations

2.1 Variables and Dataset

The data are adapted online from the "Adult Data Set" contributed by R. Kohavi and B. Becker to the UCI Machine Learning Repository. (archive.ics.uci.edu) It contains 48,842 rows, indicating 48,842 observations, and 8 columns, which are the variables. The target variable is the $value\_flag$, which has two levels, 1 and 0, indicating whether the client was classified as valuable or not. There are 7 predictor variables in the original dataset; they are *age* which represents the age of the client; $education\_num$ is a categorical variable, which indicates the education level of the client; $marital\_status$, which shows the marital status of the client, a categorical variable with 7 levels(Divorced, Married-AF-spouse, Married-civ-spouse, Married-spouse-absent, Never-married, Separated, Widowed); *occupation*, a categorical variable with 6 levels, indicating the occupation groups of the clients. The dataset did not specify the description (or rank) of what kind of occupations each group represents; hours per week, indicating the number of hours worked per week, is a continuous variable; *score*, a continuous variable, representing the "score" developed by the company, it is an real number with two decimal places. Before moving to the next part, we removed all observations with ages below 25. Because we were informed that the model only would be used to predict customers aged 25 and older, younger policyholders were removed, leaving 40,410 observations to analyze.

2.2 Data Explorations

We explored our data by visualizing the distribution of 0 and 1s for the dependent variable (valuable or not) in each independent variable, as well as the distribution of each variable itself.

2.2.1 age

age is a continuous factor; all observations are between 25 to 90. Figure 1 and Figure 2 show the distribution of the response in age, as well as the distribution of age itself. Observe that older people are more likely to be classified as valuable in Figure 1. We clearly see from Figure 2 that the distribution

Figure 1: comparison age density

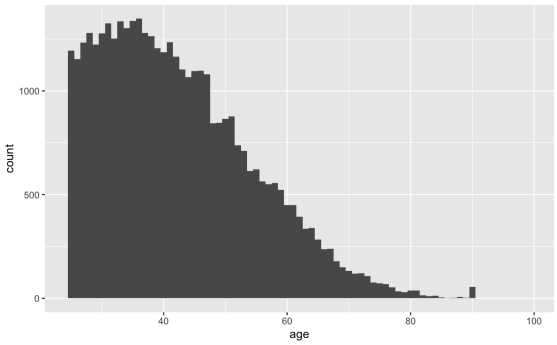

Figure 2: general age density

is smoothy and slightly right skewed. This can be further verified in Table 1. The average age of valuable people is 44.45; while the mean age for the less valuable is 41.5.

Table 1: Summary of age

	Min.	**1st Qu.**	**Median.**	**3rd Qu.**	**Max.**	**Mean**	**SD**
Overall	25.00	33.00	41.00	50.00	90.00	42.34	12.12
High-value	25.00	37.00	44.00	51.00	90.00	44.45	10.43
Low-value	25.00	31.00	39.00	49.00	90.00	41.50	12.64

2.2.2 education_num

Education_num is a numerical factor from 0 to 16, and all numbers are integer. The below graphs show the distribution of the dependent variable in *Education_num*. Obviously, as the level of education increases, the chance of being classified as valuable dramatically increases.

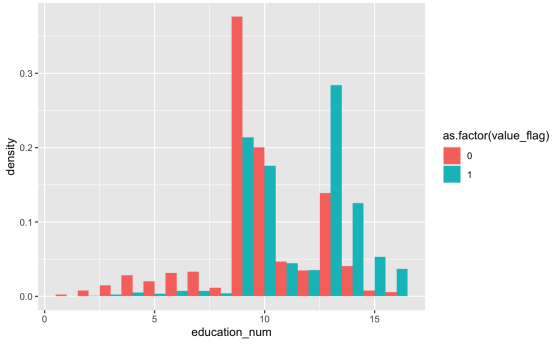

Figure 3: comparison education_num density

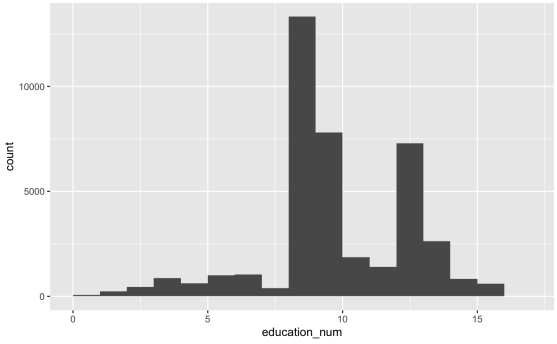

Figure 4: general education_num density

Table 2: Summary of education_num

	Min.	**1st Qu.**	**Median.**	**3rd Qu.**	**Max.**	**Mean**	**SD**
Overall	1.00	9.00	10.00	13.00	16.00	10.24	2.65
High-value	1.00	10.00	12.00	13.00	16.00	11.61	2.38
Low-value	1.00	9.00	9.00	11.00	16.00	9.69	2.55

2.2.3 hours_per_week

hours_per_week is a continuous variable that contains integers between 0 and 99. From *Table*3 : *Summaryofhoursperweek*, it is found that most people work around 40 hours per week. There are some individuals that work longer hours than normal, while some people even work 0 hours per week. The reason might be that some people might have several part-time jobs at the same time, while some people were currently unemployed or have already retired. Generally, as the number of hours working per week increases, the density of for the response variable tends to get higher as well.

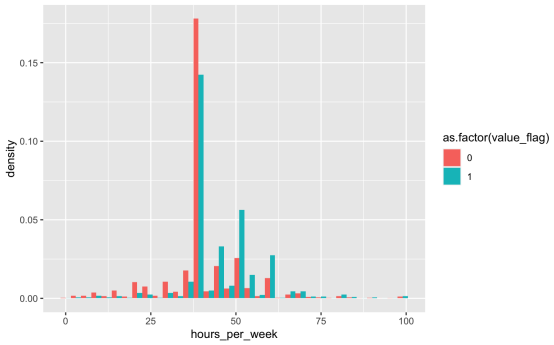

Figure 5: comparison working hour density

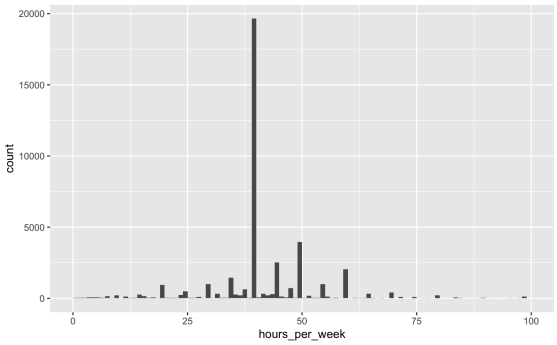

Figure 6: general working hour density

Table 3: Summary of hours per week

	Min.	**1st Qu.**	**Median.**	**3rd Qu.**	**Max.**	**Mean**	**SD**
Overall	1.00	40.00	40.00	45.00	99.00	41.99	11.80
High-value	1.00	40.00	40.00	50.00	99.00	45.46	11.09
Low-value	1.00	40.00	40.00	44.00	99.00	40.60	11.79

2.2.4 cap_gain

This is the specialist factor. The range is very large, from 0 to 99999, and more than 90 percent of observations are 0. Thus, it has a huge standard deviation compared to other variables. But we found that observations with nonzero cap gain are valuable in most instances.

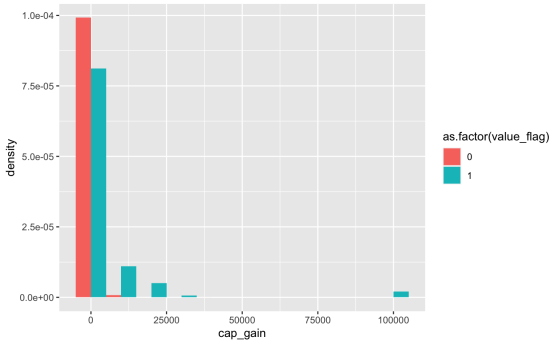

Figure 7: comparison cap_gain density

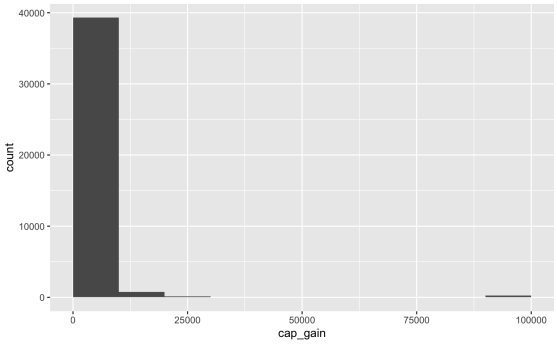

Figure 8: general cap_gain density

2.2.5 score

Figure 9 shows the distribution of 0 and 1s of the response in score. It is found that the "score" developed by the apartment does not seem to be an important variable here – the distributions of

Table 4: Summary of cap_gain

	Min.	**1st Qu.**	**Median.**	**3rd Qu.**	**Max.**	**Mean**	**SD**
Overall	0.00	0.00	0.00	0.00	99999.00	1271.76	8089.05
High-value	0.00	0.00	0.00	0.00	99999.00	4011.00	14674.78
Low-value	0.00	0.00	0.00	0.00	41310.00	169.70	941.74

0 and 1 do not generate a clear trend. Also, notice that the density of each bin is relatively low, compared to other variables, thus, a small difference in 0 and 1 densities is not significant, and we may not conclude a difference in the chance to be classified to valuable based on score.

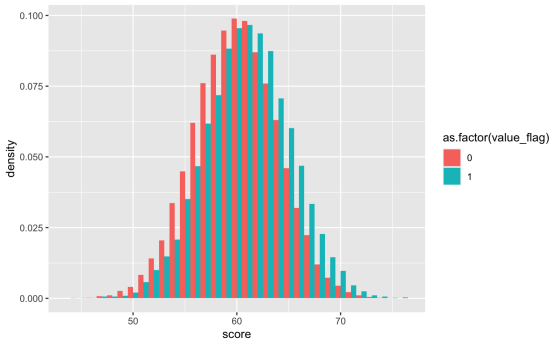

Figure 9: comparison score density

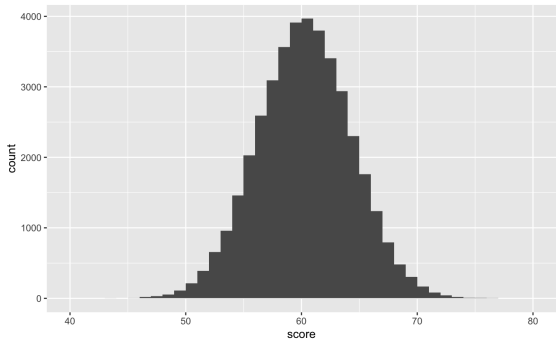

Figure 10: general score density

Table 5: Summary of score

	Min.	**1st Qu.**	**Median.**	**3rd Qu.**	**Max.**	**Mean**	**SD**
Overall	43.94	57.52	60.28	63.02	76.35	60.28	4.04
High-value	43.94	58.24	60.95	63.71	76.35	60.98	4.10
Low-value	45.36	57.29	60.02	60.02	75.86	59.99	3.99

2.2.6 marital_status

There are 7 levels for *marital_status*. From Figure 11 and Figure 12, we observe an incredible imbalance in both numbers of observations and proportions of valuable between marital groups. The largest group is Married-civ-spouse who has 21661 observations (53.06 percent); in contrast, the smallest group only has 31 observations (0.08 percent). Other groups range from a few hundred to a few thousand. From the proportion side, we observe that Married-civ-spouse and Married-AF-spouse have a higher proportion of high-value than other groups. We may conclude that married people are valued more than other people.

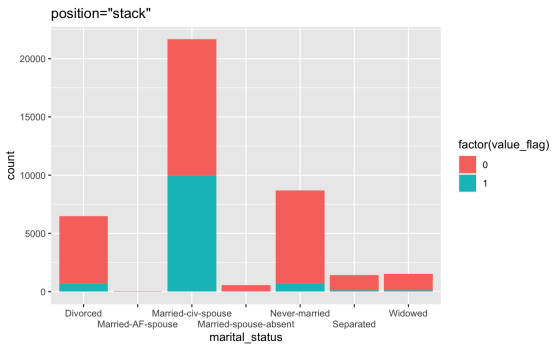

Figure 11: Count: marital_status

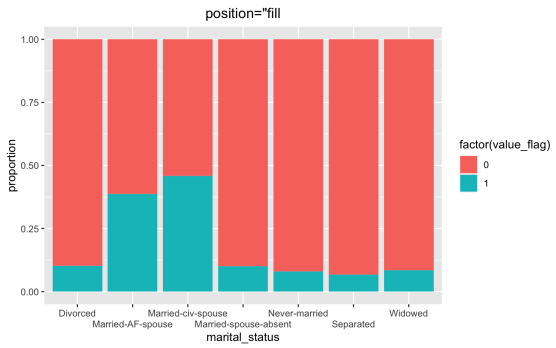

Figure 12: Proportion: marital_status

Table 6: Table of marital_status

	High value	**Low value**	**Total number**	**Valuable rate**	**Proportion**
Divorced	669	5829	6498	10.30%	16.08%
Married-AF-spouse	12	19	31	38.71%	0.08%
Married-civ-spouse	9934	11727	21661	45.86%	53.60%
Married-spouse-absent	58	515	573	10.12%	1.42%
Never-married	697	8000	8697	8.01%	21.52%
Separated	96	1342	1438	6.68%	3.56%
Widowed	128	1384	1512	8.47%	3.74%

2.2.7 occupation

occupation has 6 levels, one of which is a group where the occupation is unavailable. The situation in occupation is much more balanced than marital status. From Figure 14 and Table 7, we found that increasing occupation group numbers relate to an increasing proportion of high value.

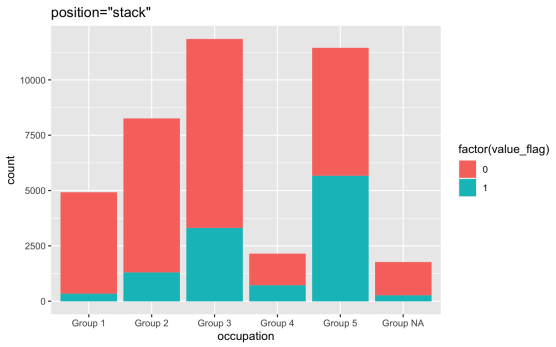

Figure 13: Count: occupation

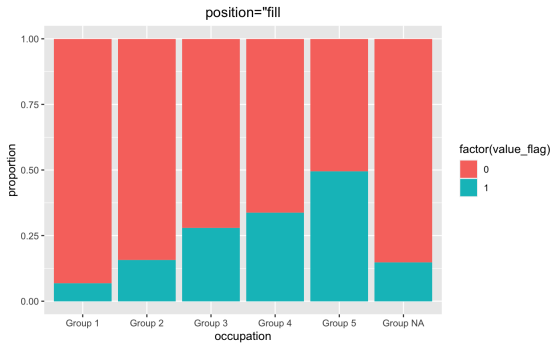

Figure 14: Proportion: occupation

Table 7: Table of occupation

	High value	**Low value**	**Total number**	**Valuable rate**	**Proportion**
Group 1	333	4594	4927	6.76%	12.19%
Group 2	1300	6951	8251	15.76%	20.42%
Group 3	3311	8545	11856	27.93%	29.34%
Group 4	725	1420	2145	33.80%	5.31%
Group 5	5663	5793	11456	49.43%	28.35%
Group NA	262	1513	1775	14.76%	4.39%

2.3 Data Cleaning

Before formally fitting the model, some data cleaning work needs to be done. Firstly, we removed observations with uncleared occupations. Secondly, integrate the number of levels for education_num from 16 to 4, by 25%, 50%, 75%, and 100% quantiles. Similarly, we implement this to marital_status - integrate levels of marital status from 8 levels to 2 levels, married and others (unmarried, divorced, widowed, etc.). The reason is that married clients have higher proportion of high-values than other groups, as discussed before. The purpose of this was to reduce the complexity of the regression model as much as possible. Finally, we split the dataset into training (80%) and test data (20%), which can help us evaluate different models objectively.

3 Modelling (Logistic Regression) Model Selection

3.1 model1: first try

3.1.1 model1 fit

We used Logistic Regression to solve this problem. We first fitted the model with full predictors and no interaction effects. We implemented the R function $glm()$ with a binomial link, the result is shown below. The equation of this model is:

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11}$$

where p is the probability of being classified to valuable. age (x_1) is a quantitative variable. $education_num$ is a categorical variable with 4 levels. It is represented by three indicator variables (x_2, x_3 *and* x_4), as follows:

$education\_num$	x_2	x_3	x_4
$education\_num\_Group0$	0	0	0
$education\_num\_Group1$	1	0	0
$education\_num\_Group2$	0	1	0
$education\_num\_Group3$	0	0	1

$marital\_status$ is a categorical variable with 2 levels, and it represented by only one variable x_5 , as follows:

$marital\_status$	x_5
$marital\_status\_Group0$	0
$marital\_status\_Group1$	1

occupation is a categorical variable with 5 levels. It is represented by three indicator variables (x_6, x_7, x_8 *and* x_9), as follows:

$marital\_status$	x_6	x_7	x_8	x_9
$occupation\_Group1$	0	0	0	0
$occupation\_Group2$	1	0	0	0
$occupation\_Group3$	0	1	0	0
$occupation\_Group4$	0	0	1	0
$occupation\_Group5$	0	0	0	1

$hours\_per\_week$ (x_{10}) and $score$ (x_{11}) are quantitative variables. β_1, $\beta_2, \ldots, \beta_{11}$ are the coefficients for $x_1, x_2, \ldots, x_{11}$ respectively, as the equation shown above.The AIC of this model is at 25474, we will compare it to the other models. The ratio of deviance and Pearson chi-square statistics around 1, meaning that Binomial here is a good fit for Generalized Linear Model. Below figures shows the R output:

Figure 15: Confusion Matrix for Model 1

```
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)          -10.045744   0.267394 -37.569   <2e-16 ***
age                    0.022920   0.001400  16.368   <2e-16 ***
education_num_Group1   0.651576   0.043852  14.859   <2e-16 ***
education_num_Group2   0.752000   0.059258  12.690   <2e-16 ***
education_num_Group3   1.527085   0.042649  35.806   <2e-16 ***
Marital_Group1         2.264525   0.038011  59.575   <2e-16 ***
occupationGroup 2      0.678890   0.079117   8.581   <2e-16 ***
occupationGroup 3      1.013926   0.074314  13.644   <2e-16 ***
occupationGroup 4      1.448337   0.092532  15.652   <2e-16 ***
occupationGroup 5      1.677683   0.076486  21.935   <2e-16 ***
hours_per_week         0.027833   0.001390  20.019   <2e-16 ***
score                  0.060262   0.003898  15.458   <2e-16 ***
```

Figure 16: The R output of model 1.

3.1.2 model1 performance

We also built the confusion matrix to evaluate each model's performance. The threshold used by the confusion matrix was the one that would achieve the highest accuracy. Since the research question focuses on finding valuable people, we are also interested in recall rate here. From below results, model1 only correctly classified 1382 high-value observations, resulting a large type I error.

Table 8: Confusion Matrix for Model 1

	High-value	**Low-value**
High-value	1382	634
Low-value	1193	5376

$$Accuracy = \frac{TP + TN}{T + N} = \frac{1382 + 5376}{1382 + 5376 + 1193 + 634} = 78.7187\%$$

$$Recall = \frac{TP}{T} = \frac{1382}{1382 + 1193} = 53.6699\%$$

3.2 model2: second try

3.2.1 model2 fit

We then fitted the model with full predictors and all interaction terms. The model became very complicated and had 320 different parameters now. But very few variables appear to be significant, and AIC is about the same as the model1. The ratio of deviance and Pearson chi-square statistics is around 1, meaning that Binomial here is a good fit for Generalized Linear Model. Because of the limited space, we decided not to write the model down and show the R output here. Please see the R outputs of this model in the appendix.

3.2.2 model2 performance

The accuracy of model 2 is 78.6371%. The accuracy is decreased a little bit compared with model 1. The recall rate has improved to 55.3786% (1426 correct high-value classifications this time). However, the type I error is still extensive.

3.3 model3: stepwise selection

3.3.1 model3 fit

The full model with all interactions, model2, is not overall a good model, considering the fact that it has very few significant variables. Thus, we decided to use stepwise selection method, with minimizing AIC, to select the better model. The equation of this model is:

$$
\begin{aligned}
log\left(\frac{p}{1-p}\right) = {} & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \\
& \beta_{11} x_{11} + \beta_{12} x_1 x_5 + \beta_{13} x_2 x_5 + \beta_{14} x_3 x_5 + \beta_{15} x_4 x_5 + \beta_{16} x_1 x_6 + \beta_{17} x_1 x_7 + \\
& \beta_{18} x_1 x_8 + \beta_{19} x_5 x_9 + \beta_{20} x_5 x_6 + \beta_{21} x_5 x_7 + \beta_{22} x_5 x_8 + \beta_{23} x_5 x_9 + \beta_{24} x_1 x_{10} + \\
& \beta_{25} x_2 x_{10} + \beta_{26} x_3 x_{10} + \beta_{27} x_4 x_{10} + \beta_{28} x_5 x_{10} + \beta_{29} x_6 x_{10} + \beta_{30} x_7 x_{10} + \beta_{31} x_8 x_{10} + \\
& \beta_{32} x_9 x_{10} + \beta_{33} x_1 x_{11} + \beta_{34} x_5 x_{11} + \beta_{35} x_6 x_{11} + \beta_{36} x_7 x_{11} + \beta_{37} x_8 x_{11} + \beta_{38} x_9 x_{11} + \\
& \beta_{39} x_{10} x_{11} + \beta_{40} x_1 x_5 x_{10} + \beta_{41} x_2 x_5 x_{10} + \beta_{42} x_3 x_5 x_{10} + \beta_{43} x_4 x_5 x_{10} + \beta_{44} x_1 x_6 x_{10} + \\
& \beta_{45} x_1 x_7 x_{10} + \beta_{46} x_1 x_8 x_{10} + \beta_{47} x_1 x_9 x_{10} + \beta_{48} x_1 x_5 x_{11} + \beta_{49} x_5 x_6 x_{11} + \beta_{50} x_5 x_6 x_{11} + \\
& \beta_{51} x_5 x_7 x_{11} + \beta_{52} x_5 x_8 x_{11} + \beta_{53} x_5 x_9 x_{11} + \beta_{54} x_1 x_{10} x_{11} + \beta_{55} x_5 x_{10} x_{11} + \beta_{56} x_1 x_5 x_{10} x_{11}
\end{aligned}
$$

As we can see from the above equation, minimizing AIC does not greatly reduce the complexity. The model involves third- and fourth-order interaction terms between variables. AIC only decreased from 25475 in model1 to 25244, which is not a large decrement. More importantly, like model2, this model does not generate a proper number of significant variables.

```
Coefficients:
                                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                                         1.580e-01  7.012e+00   0.023   0.9820
age                                                -1.327e-01  1.355e-01  -0.979   0.3275
education_num_Group1                                8.541e-01  3.939e-01   2.168   0.0301 *
education_num_Group2                                1.966e-01  5.494e-01   0.358   0.7204
education_num_Group3                                1.748e+00  3.252e-01   5.374 7.72e-08 ***
Marital_Group1                                     -1.235e+01  8.141e+00  -1.517   0.1293
occupationGroup 2                                  -1.586e+00  3.098e+00  -0.512   0.6088
occupationGroup 3                                  -2.513e+00  2.905e+00  -0.865   0.3871
occupationGroup 4                                  -6.259e+00  3.429e+00  -1.825   0.0680 .
occupationGroup 5                                  -2.686e+00  2.814e+00  -0.955   0.3397
hours_per_week                                     -6.241e-02  1.413e-01  -0.442   0.6587
score                                              -1.600e-01  1.158e-01  -1.382   0.1669
age:Marital_Group1                                  2.255e-01  1.588e-01   1.420   0.1556
education_num_Group1:Marital_Group1                 1.787e-01  4.372e-01   0.409   0.6828
education_num_Group2:Marital_Group1                 1.310e+00  6.101e-01   2.147   0.0318 *
education_num_Group3:Marital_Group1                -4.113e-01  3.625e-01  -1.135   0.2565
age:occupationGroup 2                              -5.216e-03  1.935e-02  -0.270   0.7875
age:occupationGroup 3                               4.153e-03  1.835e-02   0.226   0.8209
age:occupationGroup 4                              -6.667e-03  2.565e-02  -0.260   0.7949
age:occupationGroup 5                              -2.664e-02  1.810e-02  -1.472   0.1410
Marital_Group1:occupationGroup 2                    3.273e+00  3.261e+00   1.004   0.3155
Marital_Group1:occupationGroup 3                    3.148e+00  3.047e+00   1.033   0.3016
Marital_Group1:occupationGroup 4                    6.350e+00  3.599e+00   1.764   0.0777 .
Marital_Group1:occupationGroup 5                    6.781e+00  2.969e+00   2.284   0.0224 *
age:hours_per_week                                  2.995e-03  3.050e-03   0.982   0.3263
education_num_Group1:hours_per_week                -7.789e-03  8.475e-03  -0.919   0.3580
education_num_Group2:hours_per_week                 1.151e-02  1.174e-02   0.981   0.3266
education_num_Group3:hours_per_week                -2.374e-03  6.844e-03  -0.347   0.7287
Marital_Group1:hours_per_week                       1.844e-01  1.685e-01   1.094   0.2738
occupationGroup 2:hours_per_week                   -2.164e-02  2.204e-02  -0.982   0.3261
occupationGroup 3:hours_per_week                    1.156e-03  2.062e-02   0.056   0.9553
occupationGroup 4:hours_per_week                   -1.986e-02  2.810e-02  -0.707   0.4797
occupationGroup 5:hours_per_week                   -3.778e-02  2.043e-02  -1.849   0.0645 .
age:score                                           3.284e-03  2.226e-03   1.475   0.1402


age:Marital_Group1                                  2.255e-01  1.588e-01   1.420   0.1556
education_num_Group1:Marital_Group1                 1.787e-01  4.372e-01   0.409   0.6828
education_num_Group2:Marital_Group1                 1.310e+00  6.101e-01   2.147   0.0318 *
education_num_Group3:Marital_Group1                -4.113e-01  3.625e-01  -1.135   0.2565
age:occupationGroup 2                              -5.216e-03  1.935e-02  -0.270   0.7875
age:occupationGroup 3                               4.153e-03  1.835e-02   0.226   0.8209
age:occupationGroup 4                              -6.667e-03  2.565e-02  -0.260   0.7949
age:occupationGroup 5                              -2.664e-02  1.810e-02  -1.472   0.1410
Marital_Group1:occupationGroup 2                    3.273e+00  3.261e+00   1.004   0.3155
Marital_Group1:occupationGroup 3                    3.148e+00  3.047e+00   1.033   0.3016
Marital_Group1:occupationGroup 4                    6.350e+00  3.599e+00   1.764   0.0777 .
Marital_Group1:occupationGroup 5                    6.781e+00  2.969e+00   2.284   0.0224 *
age:hours_per_week                                  2.995e-03  3.050e-03   0.982   0.3263
education_num_Group1:hours_per_week                -7.789e-03  8.475e-03  -0.919   0.3580
education_num_Group2:hours_per_week                 1.151e-02  1.174e-02   0.981   0.3266
education_num_Group3:hours_per_week                -2.374e-03  6.844e-03  -0.347   0.7287
Marital_Group1:hours_per_week                       1.844e-01  1.685e-01   1.094   0.2738
occupationGroup 2:hours_per_week                   -2.164e-02  2.204e-02  -0.982   0.3261
occupationGroup 3:hours_per_week                    1.156e-03  2.062e-02   0.056   0.9553
occupationGroup 4:hours_per_week                   -1.986e-02  2.810e-02  -0.707   0.4797
occupationGroup 5:hours_per_week                   -3.778e-02  2.043e-02  -1.849   0.0645 .
age:score                                           3.284e-03  2.226e-03   1.475   0.1402
Marital_Group1:score                                3.180e-01  1.345e-01   2.364   0.0181 *
occupationGroup 2:score                             5.080e-02  4.952e-02   1.026   0.3049
occupationGroup 3:score                             6.150e-02  4.634e-02   1.327   0.1844
occupationGroup 4:score                             1.364e-01  5.332e-02   2.558   0.0105 *
occupationGroup 5:score                             9.657e-02  4.482e-02   2.155   0.0312 *
hours_per_week:score                                2.312e-03  2.329e-03   0.993   0.3208
age:Marital_Group1:hours_per_week                  -4.716e-03  3.611e-03  -1.306   0.1915
education_num_Group1:Marital_Group1:hours_per_week -3.547e-04  9.440e-03  -0.038   0.9700
education_num_Group2:Marital_Group1:hours_per_week -2.847e-02  1.311e-02  -2.172   0.0299 *
education_num_Group3:Marital_Group1:hours_per_week  5.638e-03  7.691e-03   0.733   0.4635
age:occupationGroup 2:hours_per_week                1.607e-04  4.614e-04   0.348   0.7276
age:occupationGroup 3:hours_per_week                8.385e-05  4.352e-04   0.193   0.8472
age:occupationGroup 4:hours_per_week                4.571e-04  6.211e-04   0.736   0.4618
age:occupationGroup 5:hours_per_week                8.416e-04  4.303e-04   1.956   0.0505 .
age:Marital_Group1:score                           -4.921e-03  2.616e-03  -1.881   0.0600 .
Marital_Group1:occupationGroup 2:score             -5.255e-02  5.428e-02  -0.968   0.3330
Marital_Group1:occupationGroup 3:score             -6.324e-02  5.080e-02  -1.245   0.2131
Marital_Group1:occupationGroup 4:score             -1.092e-01  5.962e-02  -1.832   0.0669 .
Marital_Group1:occupationGroup 5:score             -1.167e-01  4.952e-02  -2.357   0.0184 *
age:hours_per_week:score                           -6.020e-05  5.020e-05  -1.199   0.2304
Marital_Group1:hours_per_week:score                -4.200e-03  2.781e-03  -1.510   0.1310
age:Marital_Group1:hours_per_week:score             9.476e-05  5.955e-05   1.591   0.1116
```

Figure 17: The R output of model 3.

3.3.2 model3 performance

The accuracy of model 3 is 79.0681%. The recall rate is 54.6796% (1167 correct classifications this time). The performance did not significantly improve compared to previous models.Therefore, we decided to use the model1 as our final logistic regression model, in which there are no interaction terms.

3.4 Likelihood Ratio Test

We conducted the likelihood ratio test to compare our model with the simplest model, which has no predictor. We wanted to test if our model (model1) is better at distinguishing the high-values and the low values from the population.

$$H_0 : \textit{The simplest model with no predictors is adequate at predicting outcomes. vs.}$$

$$H_a : \textit{model1 has more predictive power than simplest model}$$

.

$$\textit{Test statistics} : \quad X^2 \textit{ statistic } = 2 \times |l(b_f) - l(b_r)|$$

$$\textit{Critical value} : \quad X^2_{\alpha, p_f - p_r}$$

```
Likelihood ratio test

Model 1: value_flag ~ 1
Model 2: value_flag ~ age + education_num_Group + Marital_Group + occupation +
    hours_per_week + score
  #Df LogLik Df Chisq Pr(>Chisq)
1   1 -18133
2  12 -12725 11 10815  < 2.2e-16 ***
```

Figure 18: The R output of Likelihood Ratio Test.

From above R output shows, p-value approaches to 0. Thus, the H_0 should be rejected. Thus, we can conclude that the full model is better than the null model. And we should use all parameters in the model.

3.5 Logistic Regression Diagnostics

3.5.1 Assumption checking: Multicollinearity

We first checked if multicollinearities between predictor variables exist. We used VIF (variance-inflation factors) to check the multicollinearity. Notice that since categorical variables are involved in this question, we need to use generalized VIF to check the collinearity. We used the VIF function in R package 'car' to implement this. The result is shown below. We observed GVIFs slightly above 1 for the predictor variables, way less than the rule-of-thumb threshold of 5, indicating that there are not many strong correlations between these variables.

```
> # Multicollinearity
> car::vif(fit1_train)
                        GVIF Df GVIF^(1/(2*Df))
age                 1.045044  1        1.022274
education_num_Group 1.355563  3        1.052010
Marital_Group       1.095281  1        1.046557
occupation          1.333444  4        1.036625
hours_per_week      1.026697  1        1.013261
score               1.002798  1        1.001398
```

Figure 19: The R output of multicollinearity.

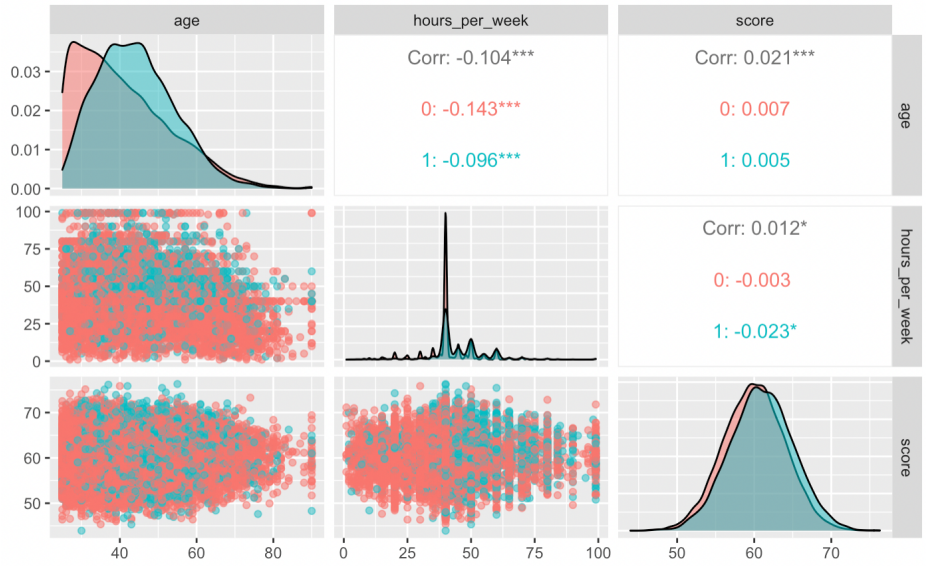

Figure 20:

We also gave a visualization of the correlations between three continuous variables. The scatterplots showed no obvious correlation between the variables. The actual correlations are also shown in the plots, which we can see are relatively small. Therefore, we conclude that multicollinearities does not exist in our model.

3.5.2 Assumption checking: Linearity

The linearity assumption between continuous predictor variables and the logit of the outcome is also an important assumption. We also checked this assumption by creating the scatterplot between each predictor and the logit values (Figure 21). It is observed that the blue lines for age and score are smoothed, while the scatter plot for hours_per_week does not seem to be "smooth" enough. This indicates that age and score are all quite linearly associated with the diabetes outcome in logit scale, and hours_per_week may not be linear and might need some transformations such as including 2 or 3-power terms.

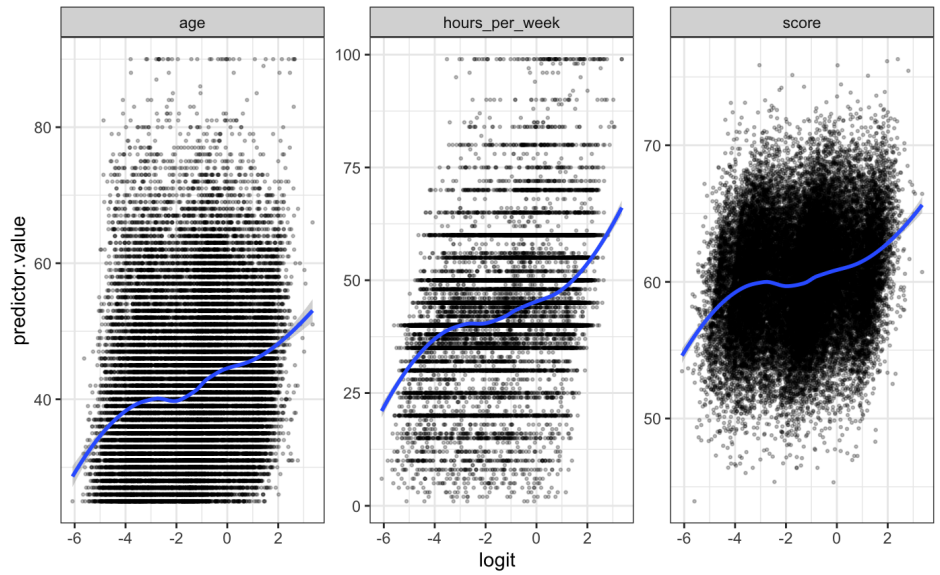

Figure 21: The R output of linearity checking .

3.5.3 Assumption checking: Influential values

Influential values are extreme individual data points that greatly affect the quality of the logistic regression model. To check the extreme values, we first plotted the Cook's distance plot, and labelled the top 5 points with the highest Cook's distance (Figure 22), then visualized the standardized residuals (Figure 23). We filtered potential influential data points with an absolute standardized residuals above 3 and found 4 of them, (shown below), they will be deleted. Interestingly, both Cook's distance and standardized residuals suggest that data point 34462 is an influential value. Thus, it should be deleted definitely.

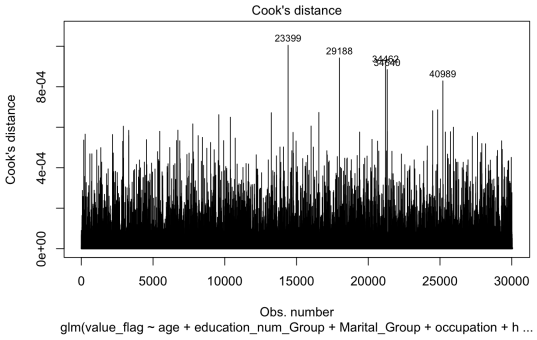

Figure 22: The R output of Cook's distance

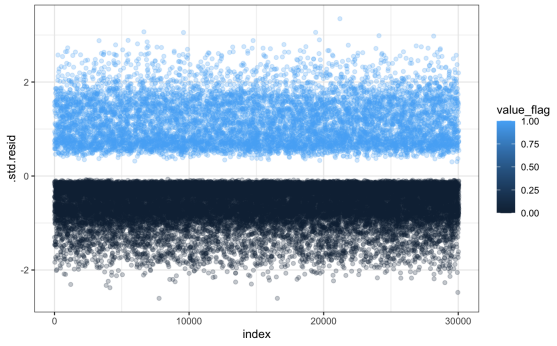

Figure 23: The R output of standardized residuals

.rownames <chr>	value_flag <dbl>	age <int>	education_num_Group <fctr>	Marital_Group <fctr>	occupation <fctr>	hours_per_week <int>	score <dbl>	.fitted <dbl>
10814	1	29	0	0	Group 1	40	59.26	-4.696609
15568	1	61	0	0	Group 1	32	51.51	-4.652872
31513	1	27	0	0	Group 1	50	55.98	-4.661779
34462	1	30	0	0	Group 1	40	43.94	-5.596907

Figure 24: Screenshot of absolute standardized residuals >3.

4 Results Interpretation

4.1 Continuous factor effect

4.1.1 age

We will use age effect as an example of continuous factors and carefully write steps down on how to find it.

$$Log{-}odds(age\ =\ z) = \beta_0{+}\beta_1 Z{+}\beta_2 x_2{+}\beta_3 x_3{+}\beta_4 x_4{+}\beta_5 x_5{+}\beta_6 x_6{+}\beta_7 x_7{+}\beta_8 x_8{+}\beta_9 x_9{+}\beta_{10} x_{10}{+}\beta_{11} x_{11}$$

$$Log{-}odds(age\ =\ z{+}5) = \beta_0{+}\beta_1(Z{+}5){+}\beta_2 x_2{+}\beta_3 x_3{+}\beta_4 x_4{+}\beta_5 x_5{+}\beta_6 x_6{+}\beta_7 x_7{+}\beta_8 x_8{+}\beta_9 x_9{+}\beta_{10} x_{10}{+}\beta_{11} x_{11}$$

$$Odds(age\ =\ z) = exp(\beta_0{+}\beta_1 Z{+}\beta_2 x_2{+}\beta_3 x_3{+}\beta_4 x_4{+}\beta_5 x_5{+}\beta_6 x_6{+}\beta_7 x_7{+}\beta_8 x_8{+}\beta_9 x_9{+}\beta_{10} x_{10}{+}\beta_{11} x_{11})$$

$$Odds(age\ =\ z{+}5) = exp(\beta_0{+}\beta_1(Z{+}5){+}\beta_2 x_2{+}\beta_3 x_3{+}\beta_4 x_4{+}\beta_5 x_5{+}\beta_6 x_6{+}\beta_7 x_7{+}\beta_8 x_8{+}\beta_9 x_9{+}\beta_{10} x_{10}{+}\beta_{11} x_{11})$$

$$Odds\ ratio\ ((z+5)\ vs.\ \ z)\ \ =\ \ exp(5\times\beta_1) = exp(5\times 0.022920) = 1.12142$$

$$Upper\ bound:\ \ exp(\ 5\times(0.22920 + se(age)\times 1.96\ ))\ \ =\ \ 1.1369$$

$$Lower\ bound:\ \ exp(\ 5\times(0.22920 - se(age)\times 1.96\ ))\ \ =\ \ 1.1061$$

Thus,controlling for other covariates, the odds of classifying the customer to valuable will increase 12.14% (95% CI:[10.61%,13.69%]) for every 5 years of increase.

4.1.2 hours_per_week:

The coefficient for hours_per_week is 0.027833, meaning that controlling for other covariates, the odds of classifying the customer to valuable will increase 2.82%(95% CI:[2.54%,3.11%]) for every unit increase in hours worked per week.

4.1.3 score

The coefficient for score is 0.06262, meaning that controlling for other covariates, the odds of classifying the customer to valuable will increase 6.21%(95% CI:[5.40%,7.02%]) for every unit increase in hours worked per week.

4.2 Categorical factor effect

4.2.1 education_num

We only show one example of a categorical variable group effect and carefully write steps down on how to find it. Other effects will be delivery by tables.
For education_num_Group1:

$$log - odds(education\_num\_Group0) = \beta_0 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11}$$

$$log - odds(education\_num\_Group1) = \beta_0 + \beta_1 x_1 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11}$$

$$Odds(education\_num\_Group0) = exp(\beta_0 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11})$$

$$Odds(education\_num\_Group1) = exp(\beta_0 + \beta_1 x_1 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11})$$

$$Odds(education\_num\_Group0\ vs.\ education\_num\_Group1) = \ exp(\beta_1) = exp(0.651676) = 1.9188$$

$$Upper\ bound:\ exp(\ 0.651676 + se(education\_num\_Group1) \times 1.96\ )\ =\ 2.0910$$

$$Lower\ bound:\ exp(\ 0.651676\ - se(education\_num\_Group1) \times 1.96\ ))\ =\ 1.7607$$

Controlling all other covariates to constant, comparing to the group of lowest education level(baseline), the odds of classifying the customer to valuable will increase 91.86% (95% CI:[76.05%,109.08%]). *education_num* has four levels and the reference level is *education_num_Group*0. Thus, other group effect compared to the reference level is:

Table 9: education_num effect

	Increase/decrease	95%CI**:LB**	95%CI**:UB**
*education_num_Group*1	91.86%	76.05%	109.08%
*education_num_Group*2	112.12%	88.86%	138.25%
*education_num_Group*3	360.47%	323.55%	400.02%

4.2.2 marital_status

marital_status has two levels, and the reference level is *marital_status_Group*0. Thus, *marital_status_Group*1 compared to the reference level is:

Table 10: marital_status effect

	Increase/decrease	95%CI**:LB**	95%CI**:UB**
*marital_status_Group*1	862.66%	793.54%	937.11%

4.2.3 occupation

occupation has five levels, and the reference level is *occupation_Group*1. Thus, other groups compared to the reference level is:

Table 11: occupation effect

	Increase/decrease	95%CI**:LB**	95%CI**:UB**
$occupation_Group2$	97.17%	68.85%	130.24%
$occupation_Group3$	138.28%	138.28%	218.86%
$occupation_Group4$	325.60%	255.01%	410.23%
$occupation_Group5$	435.31%	360.79%	521.89%

5 Future works and Conclusion

5.1 Conclusion

Based on our model, it is found that education level, marital status, occupation group, are the most influential factors on determining if a customer is valuable or not. If a person is married, then he/she has a few times more chances of being classified to valuable than the unmarried people. Similarly, if a person's occupation is grouped into the "higher group" (group number is high), then he/she is likely to be classified as valuable. On the other hand, other factors, hours working per week, and insurance score have much less effects on determining valuableness. The odds of being classified to "high-value" will only increase about two to six percentages for every unit increase in working hours per week, and company-developed score. Age has a moderate effect on valuableness – a 5-year increase in age would result in 12% increase in the chance of being classified to valuable.

5.2 Future works: Decision Tree

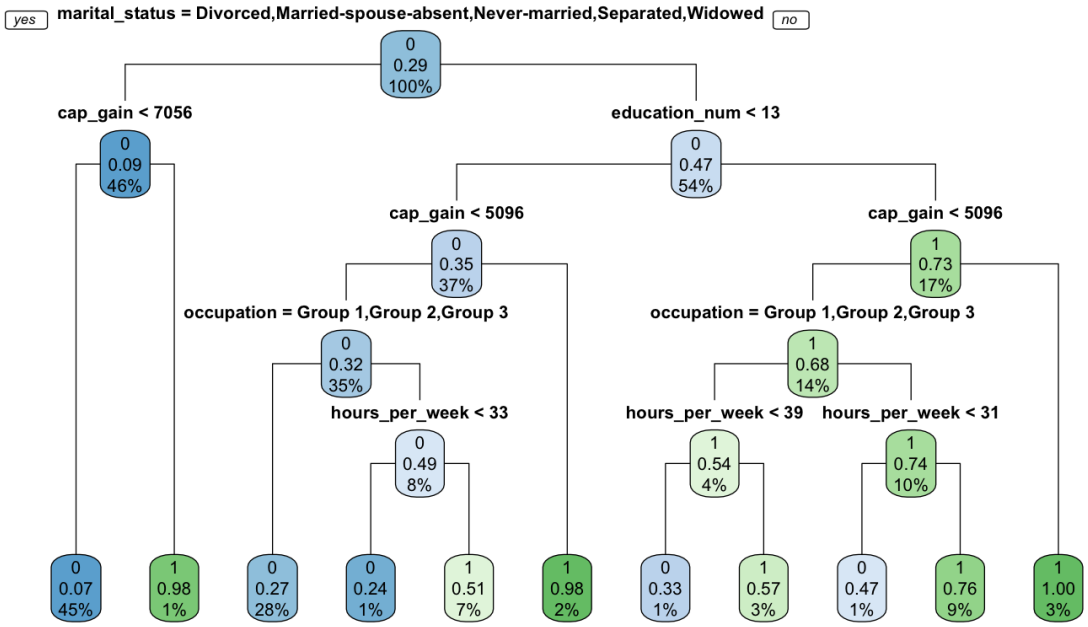

Figure 25: The R output of decision tree

Decision tree is a non-parametric supervised learning method used for classification. Compared with logistic regression, decision tree algorithm is better at managing categorical data types, outliers, and skewed data. Thus, in our case, we can keep cap_gain and treat other variables the same as before, and split the dataset into training (80%) and test data (20%) again. Below, the figure shows the results of decision tree algorithm.

Table 12: Confusion Matrix for Decision Tree

	High-value	**Low-value**
High-value	1732	1079
Low-value	725	6122

$$Accuracy = \frac{TP + TN}{T + N} = \frac{1732 + 6122}{1732 + 6122 + 1079 + 752} = 81.0945\%$$

$$Recall = \frac{TP}{T} = \frac{1732}{1732 + 725} = 70.4925\%$$

We built the confusion matrix again to evaluate the decision tree's performance. The threshold used by the confusion matrix was the one that would achieve the highest accuracy, which is also the same as the logistics regression used above. Compared with model1, accuracy increased from 78.7% to 81.1%, and the recall rate increased from 53.7% to 70.5%. It is obvious that the decision tree performance is better than model1.