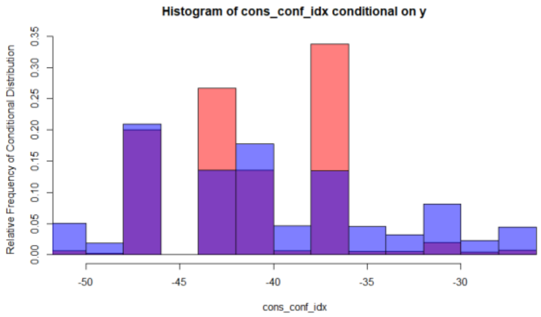
The bank aims to use direct marketing to target curtain customers to promote their product. So, I need to develop statistical models to offer the reference for the marketing. It requires the model could be used to predict if the customer will accept the offer. And this report will be comprised by four parts: data pre-process, the logistic regression model, the decision tree model and the conclusion.

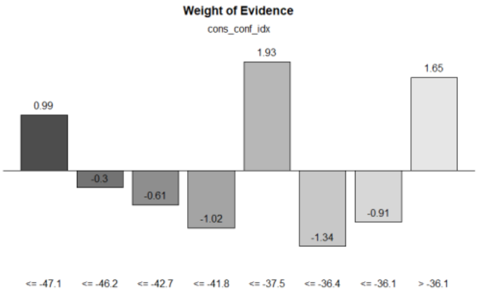
Data pre-process

From the list offered there are 19 independent variables and 1 dependent variable. The bank considers client attributes, marketing campaign attributes and the social and economic attributes as the aspects that might influence the dependent variable. There are 10000 data for every factor. Firstly, I detect no null value in this data set which means the completeness is very good. According to my pre-judge, the relevance of the data set is not very clear. For example, it's a little bit hard to understand the relationship between the expected outcome of individual customer and the number of employees. But I still need to test the eligibility of predicting in the following model selection. What's more, the comprehensibility of the data set is good. For every data entry, there are clear definition and the partition. But for some variables, there are still some unknown values such as marital, default, housing and loan.

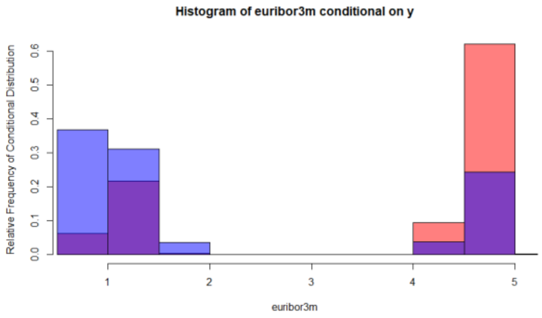
Now, let's move on to test the usefulness of every data entries by plot their condition distribution and plot of weight of evidence distribution. Furthermore, in order to get a better understanding of the data, I also need to compute the information values of all variables to see which one have the largest influence on the outcome. What's more, I need to clarify that for the continue variables, the suitable binning is necessary. Because the binning will affect the WOE and IV. And in order to get the proper outcome, I need to use some tools to get the optimal binning outcome. In this assignment, I use the smbinning() function to get the optimal binning.

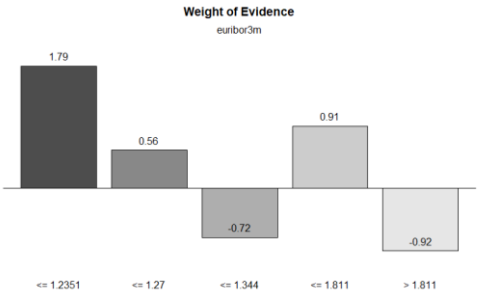
First of all, let's focus on the WOE situation of these variables and the conditional distributions to see how the variables affect the outcome. The first variable we try is consumer confidence index. The corresponding plot is shown below. As we can see in the picture, when this index is smaller than-47.1 and larger than -36.1, the customers are more likely to accept the offer. Moreover, there is also one positive period between -37.5 and -41.8 where the relative frequency of accepting the offer is higher than the other ones in the conditional distribution. It means in these periods where the WOE is positive, bank’s directly marketing will lead to a better situation.



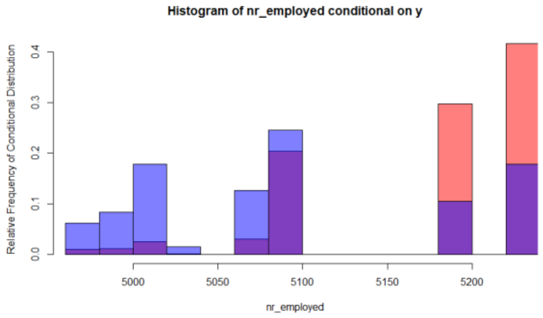


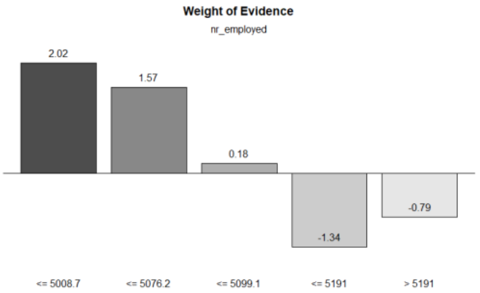
Then, the conditional distribution and WOE plots for Euribor 3-month rate are shown below. In the conditional distribution plot, the biggest relative difference happens in the period from 0 to 1.2351 which is consistent with the result of WOE. And more detailed information about how the variable affect the dependent variable could be obtained. Although WOE is negative when variable is greater than 1.811, because of the lack of the data from 2 to 4, the relative frequency of rejection starts to become larger than that of acceptation when the euribor3m is greater than 4.



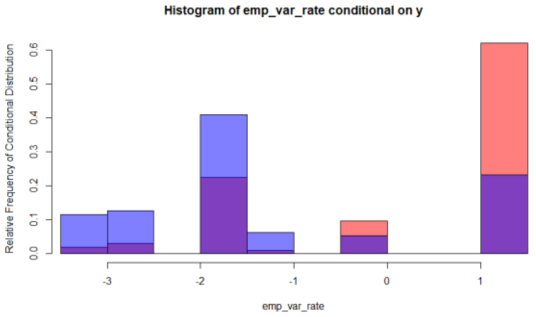


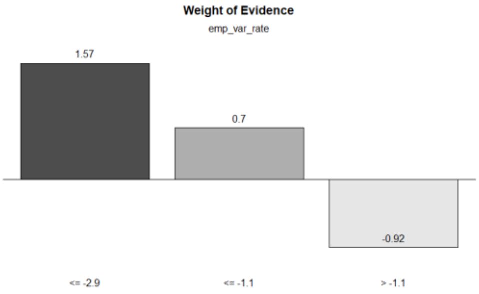
For the number of employees, the trend is very clear with the increase of nr_employed, the tendency of rejecting the offer for the guest becomes stronger. It means that when the bank considers to hold the promote campaign, they are supposed to avoid the time when the number of employees is greater than 5191. But, because of the data lack between 5100 to 5175, more data during this period is required to help to locate the approximate threshold.



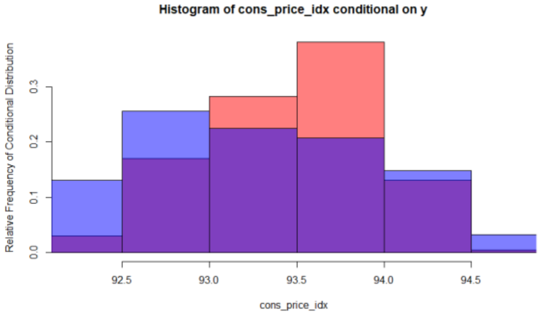


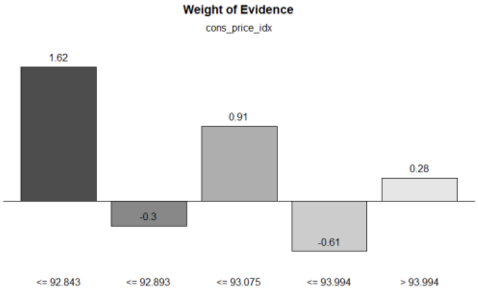
Then, let's turn to employment variation rate. This variable could be divided into 3 parts. The first one is smaller than -2.9 where customers are most likely to accept the offer. As well, the WOE of the interval from -1.1 to -2.9 is relatively low but still positive where the bank could still come off the campaign. But when the variable is bigger than -1.1, the WOE becomes negative. Furthermore, when the variable is greater than 1, the relative frequency of rejecting the offer is much higher than that of accepting. There are also some missing data intervals which make the judge vague.



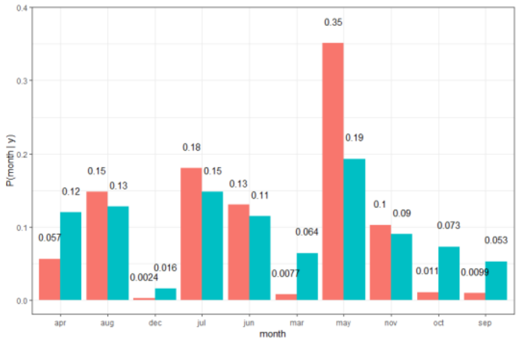


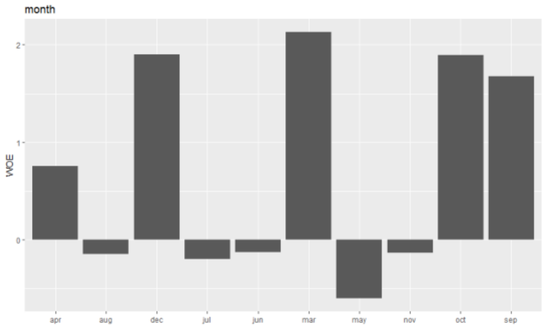
The last social and economic attribute is consumer price index. Similarly, the graph of conditional distribution plot and WOE plot fit well. But because of the value of conditional histogram's break, there is one negative interval from 92.843 to 92.893 absenting in the histogram. These two pictures illustrate that the bank is supposed to do the campaign when the index is smaller than 92.843 where the relative rate of accepting is the highest.



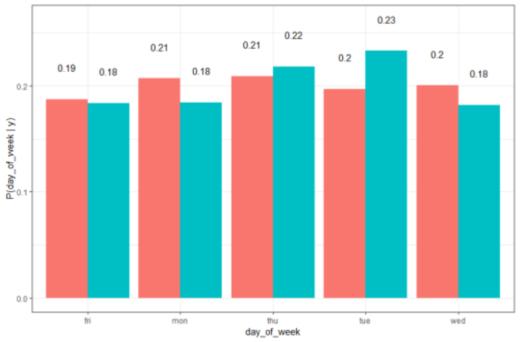


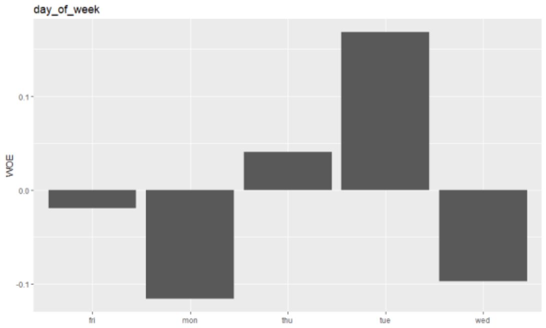
Next, we are going to analyse the marketing campaign attributes. The first variable I start with is month. From the graphs shown below, the WOE of April, December, March, October and September are positive and in March, the relative ratio of accepting the offer is the biggest. But compared with other months, the data in March is small. There might be some statistic error because of the volume of sample statistics. In addition, in order to get a better result of campaign, the bank should avoid the hold the promotion in the other months, especially in May.



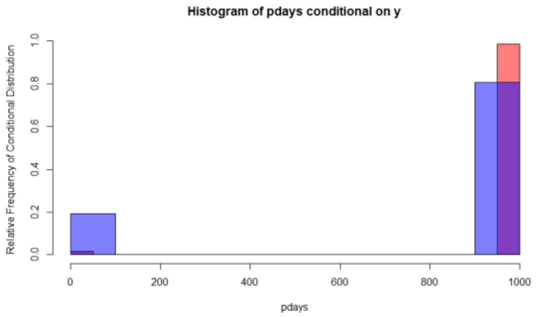


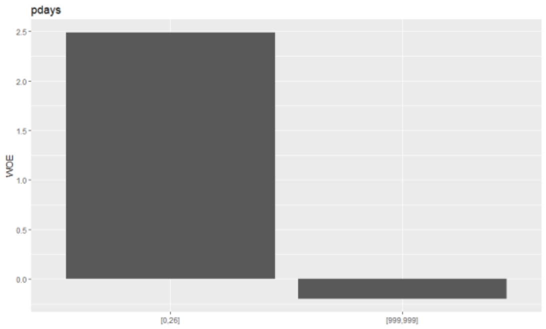
After the month, we are also interested in the influence of day of week. According to the corresponding graph, the difference between accepting and rejecting is very small. Furthermore, the ratios in the five days are approximately the same which means the chose of day of week applies little influence on if the customer will sign the contract. But if the bank needs to take a chose of day of week, the Thursday is the best choice and Monday is not a good option.



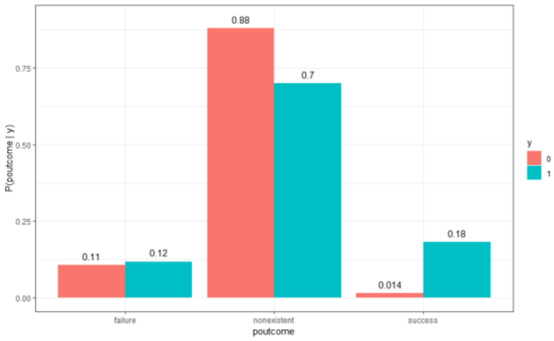


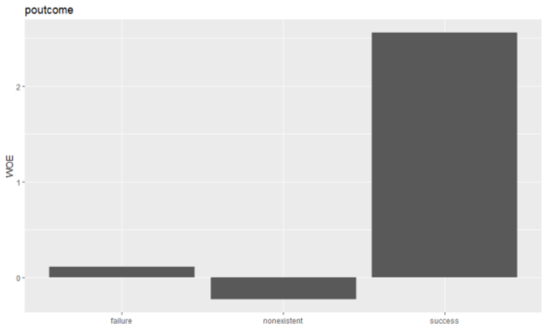
Also, the days after the last contact needs to be get into consideration. According to the pictures, for the old customer, the customers are more likely to sign the contract after some time since the campaign. To some extent, it means the people need to spend some time to get a well understanding of the product and fit their own needs. And the customer contacted recently might need more time to consider. Moreover, the high accepting ratio from 0 to about 100 suggests that the product is very attracting. For the guests who never be contacted before, they might don't know it well and they are more likely to reject. It means the bank is supposed to strength the propaganda to help people understand it better.



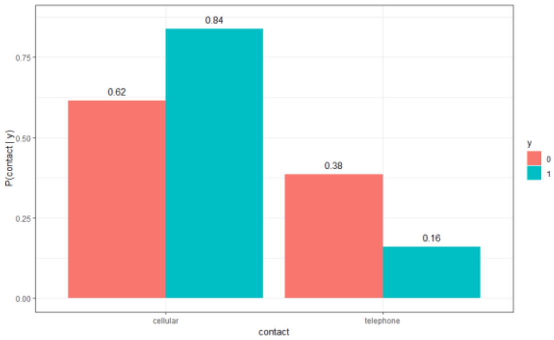


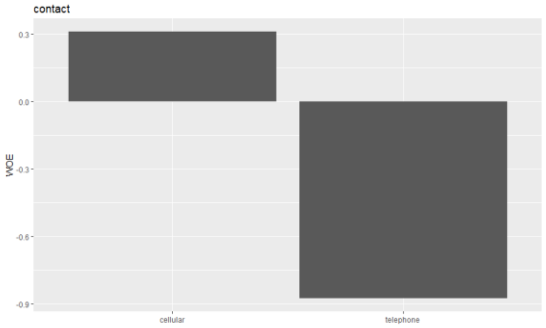
If we consider the demand of the contract, the outcome of previous campaign should also be included. We could conclude that for the guests who signed the contract at last campaign, most of them are likely to sign it again and the willing of it is very strong which implies the product is very attracting. Another point that is worthy to notice is that for the one who rejects the offer before, the ratio of accepting and rejecting is greater than 1. The reason might be that during this time period, they learned some new information about the product and changed their opinion. However, the customer who never attended the campaign before also express the trend of rejecting. So, when doing the direct marketing, the bank could focus on the ones who attended the campaign before and attract more people to join the campaign. Because it will help them to understand the product and might sign the contract in the next campaign.



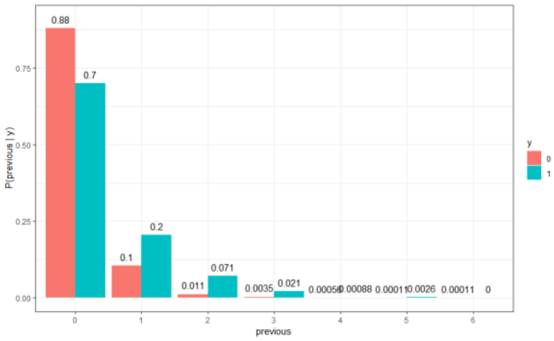


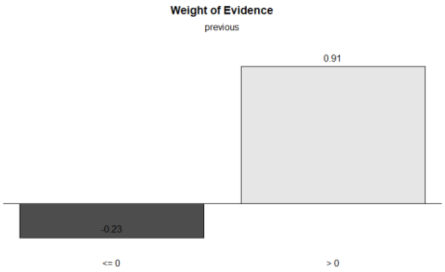
Another factor influencing the outcome is the contact method. According to the graphs, it is clearly the success possibility of contacting by cellular is greater than that of contacting by telephone. Considering the age of the user of cellular and telephone, we might have an assumption that because a big part of user of cellular are relatively younger and this product might be more suit for the younger, so they are more tendency to accept this product. In order to reach a better result and save the cost of the campaign, bank could consider the cellular as the main contact way to the guest.



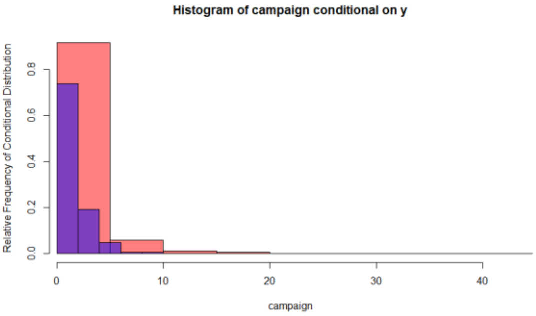


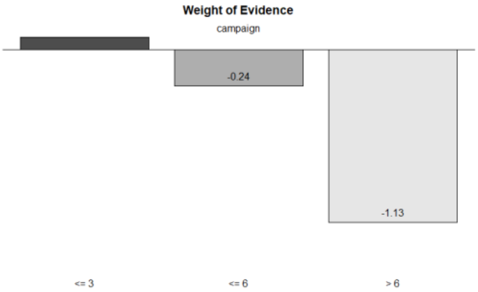
The previous number of contacts performed for the client should also be considered as a factor. The number of contacts performed before the campaign illustrate the times of being introduce the produce to for a guest. Because as we analysis before, the product is very attracting if you know it well. The similar conclusion also could be gotten from the graphs below. For the ones who never get the contact before, they are more likely to reject the proposal. But for the ones who know some information about the product. They are more likely to be happy to try during this campaign. In addition, the proportion of the first contact guest is very large which means most of the sample people have never been contacted before.



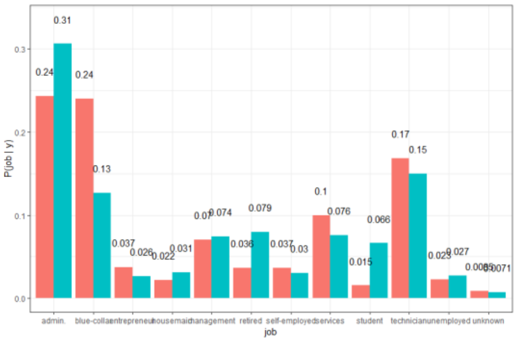


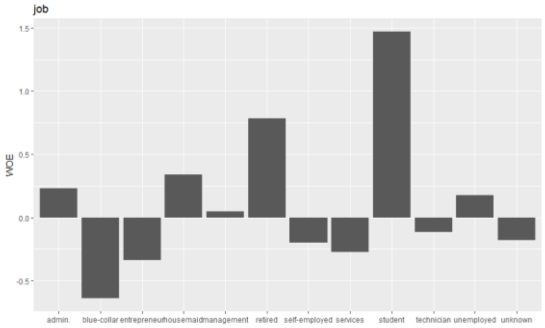
After considering the previous campaign, the number of contacts performed in this campaign is also important. According to the WOE plot, the first 3 contacts could persuade most the prospect clients to accept it and the remainning people might not have this kind of demand. With the increase of the times, the WOE of acception decreases dramatically which means after several contacts the rest people don't have enough interest in it.



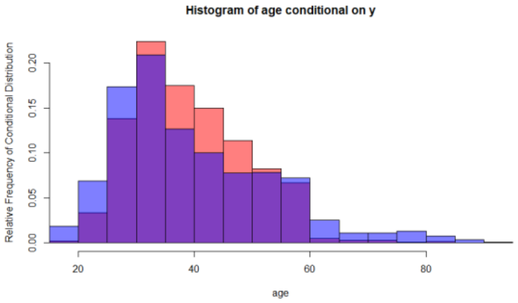


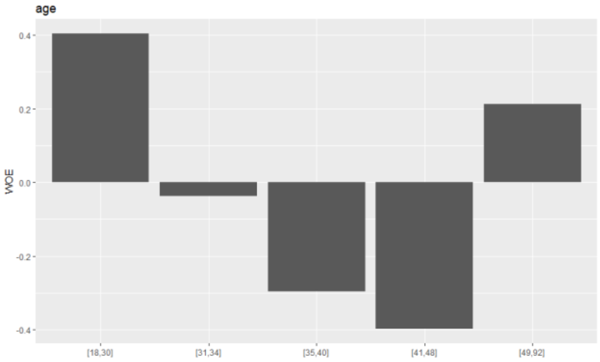
After the marketing campaign attribute, the client attributes also need to be included, such as job. The graphs illustrate that the bank is supposed to regard the admin, housemaid, retired people, studens and the unemployed as the target group. While the marketing department could put less emphasis on the blue-collar, entrepreneur, self-employed, sevices worker, technician and the others. It will help for the bank to hold the campaign in the campus or the supermarket where there are a lot of target people.



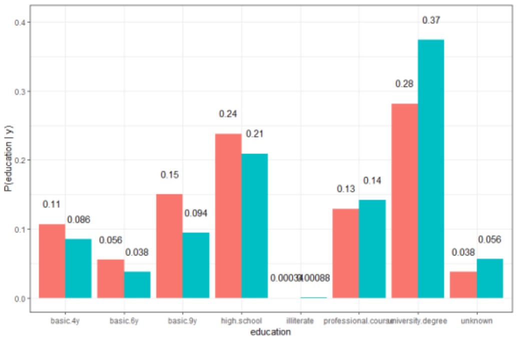


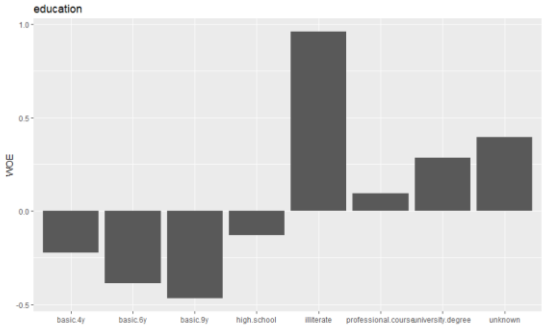
Age is also one relative variable for the promotion. Because different age present different situation. According to the tables, the young under 30 are the majority of the customer. For the people between 31 and 48, they have a higher trend to reject the offer. However, after the 48, people show some interest in accepting it. For the bank, they should put the focus of the campaign to the younger and older people instead of the middle-age.



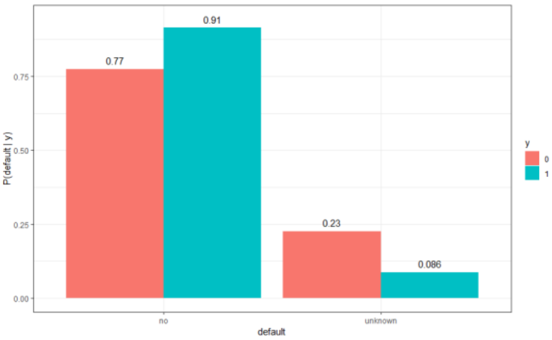


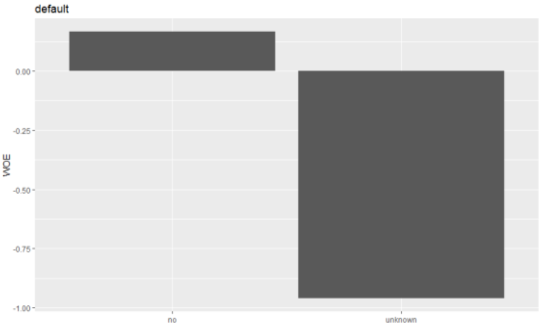
In marketing, the education could also be an influence factor. In the graph, the ones received basic course and the high school course shows less interest in it than others. We noticed that for the illiterate, they have the highest WOE but because the proportion of it is too small, so there might be some error in the data. As a conclusion, the bank could regard the ones received high degree education as the main target crowd.



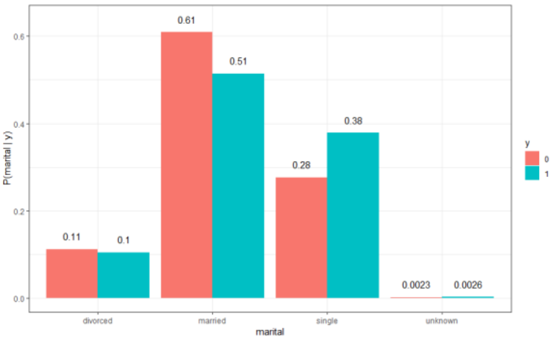


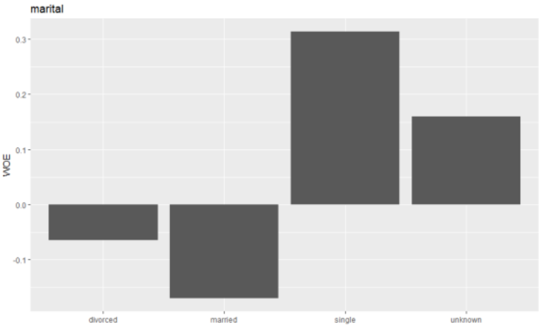
For a financial product, evaluations about the clients should be regarded as a necessary process. The first thing about the financial is the credit. It can be seen from the chart that for people who do not have a default, their tendency to accept this offer is positive. Because people with defaults are often reluctant to disclose their credit status, we can think of some default people whose credit situation is unknown. The product is an attractive product according to the previous analysis. Therefore, the WOE of the population who does not have a default situation is positive. This shows that people with default situations have a strong tendency to reject this offer.



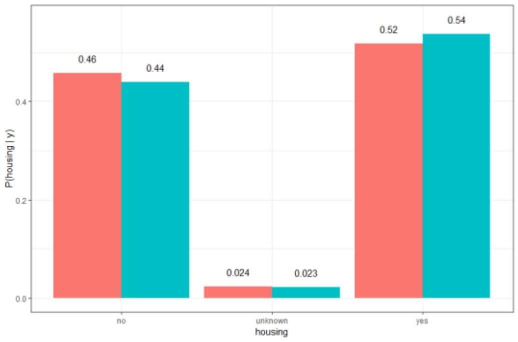


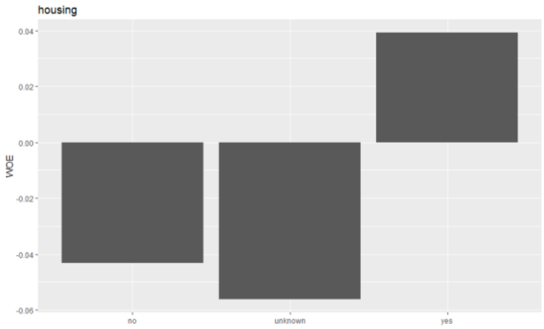
Marriage is also an important factor that can affect a person's financial situation. According to the chart, single people are more accepting of this product. Married people have a lower demand for this product. The divorced people did not show a strong tendency. Because married people have more liquidity needs in finance. At the same time, the divorced people also have the liquidity needs but weaker than that of the couple. Single people don’t have that many demands. So, they would like to put their money into the long-term deposit. This product may be more attractive to people with less liquidity needs. Banks can consider focusing on single-person crowds for marketing.



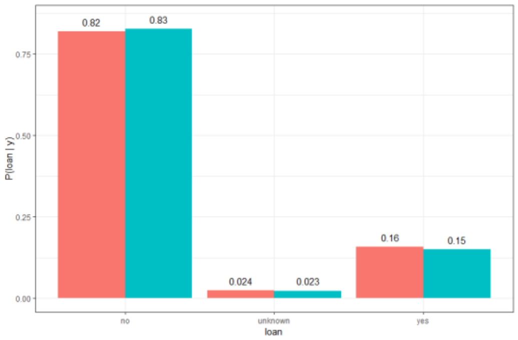


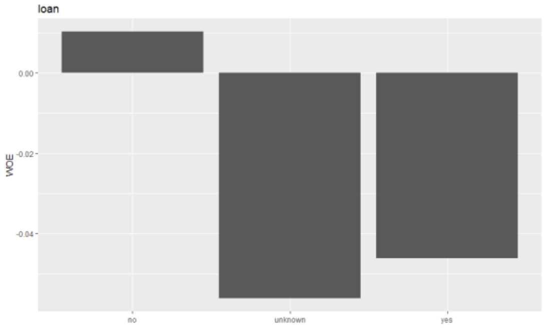
Real estate as an important asset for people also needs to be judged. As can be seen from the figure below, in the WOE diagram, the WOE value of a person with a house is positive. However, the population with no room and unknown situation has a negative WOE value. But this does not lead to the conclusion that housing can be an important factor. Because according to the conditional possibility plot, the ratio of rejection and acceptance in the three groups is roughly the same, the difference is small, which indicates that housing does not significantly affect the results.



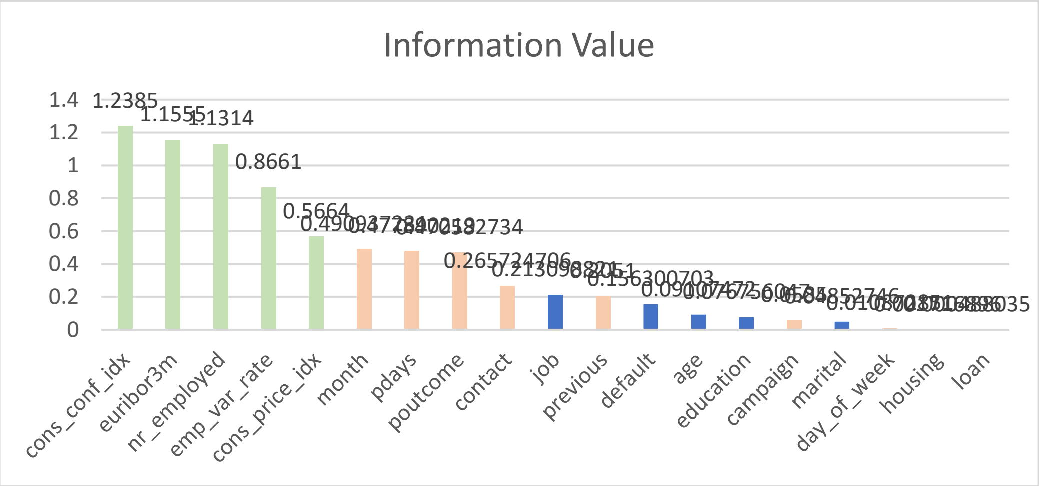


The case of loan is roughly the same as that of housing, although the WOE of each value has a positive or negative relationship. However, the difference between acceptance and rejection in each conditional probability is small. Explain that this variable does not have a large impact on the results.





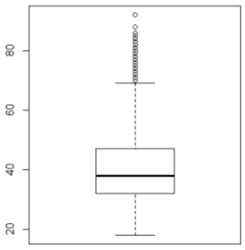
In summary, after optimally binning and calculating the IV value of the corresponding variable, we can sort the variables according to the IV value of each variable. The figure below shows the order of importance of each variable in interpreting the dependent variable and the corresponding IV value. According to the chart, we can see that the social and economic attribute is the one that has the greatest impact on the outcome. In this group, only the employee variation rate and the expected outcome have a about monotonous inverse relationship. The rest of the macro variables are generated in the specific interval for the customer to make the customer accept the offer.

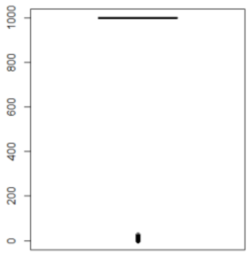


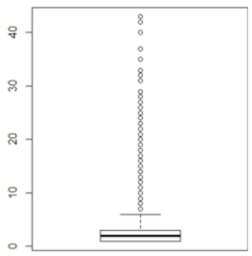
Next, I found the correlation coefficient matrix for all continuous numeric variables. In the coefficient matrix, only the partial correlation coefficient between the four variables emp_var_rate, euribor3m, nr_employed, and cons_price_idx is greater than 0.7. Therefore, I obtained the coefficient matrix for these four variables as follows. The correlation coefficients between euribor3m and emp_var_rate and nr_employed are also above 0.9. There may be some causality between the three. At the same time, there is a correlation between cons_price_idx and the three of them, but the correlation coefficient does not reach 0.9. It can be considered as an independent variable. The choices about them also need to be further judged in the regression.

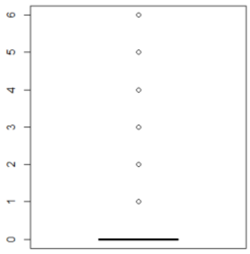
	emp_var_rate	euribor3m	nr_employed	cons_price_id
emp_var_rate	1			
euribor3m	0.9716538	1		
nr_employed	0.9052118	0.9447779	1	
cons_price_idx	0.7745721	0.6847596	0.5157321	1

In terms of data quality, the data is generally of good quality. Except for the evaluations mentioned at the beginning of the report. As shown in the figure, there are more outliers in the four variables “age”, “campaign”, “pdays” and “previous”, which will cause the estimated model to be biased. Due to the large concentration of data in the “pdays” variable at 999, the fitting results of other data may deviate from the actual one. A large amount of data in the “campaign” also cluster at 0, and the sample distribution is uneven. The same problem also occurs with the “previous” variable. The “age” variable has more outliers on the larger digits.





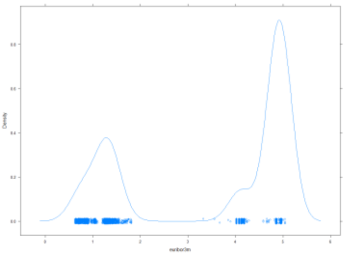


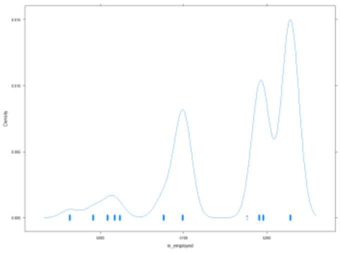


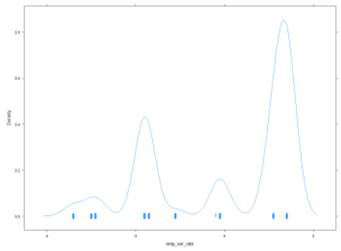
Age campaign pdays previous

In addition, according to the density map of continuous variables, the three variables "euribor3m", "nr_employed" and "emp_var_rate" are not ideally distributed. In some intervals, these variables have more missing values, which will lead to errors in the construction of the classification model. In terms of continuity, the quality of these three data sets is not ideal.

euribor3m nr.employed emp.var.rate







Logistic regression

Firstly, we perform a logistic regression on all the variables, and the resulting formula is very complicated. And the p-value of the coefficient in the formula does not meet the requirements. So, you need to delete the variables. In order to get the best logistic regression model, we can obtain four regression equations by stepwise regression in the forward and backward directions by AIC criterion and BIC criterion. For the AIC criterion, it takes into account the statistical fit of the model and the number of parameters used to fit. The smaller the AIC value, the better the model, which means that the model achieves sufficient fit with fewer parameters. During the running of the function, the model adds (forward regression) or deletes (backward regression) a variable until a certain judgment condition is met. BIC=log(n)k − 2 log(L), while AIC=2k − 2 log(L). So we set k=log(n), where n is the number of data. The stepwise regression of BIC is also obtained by the same method. The 4 outcomes are shown below.

model	type	formula
M1	Forward AIC	`y ~ nr_employed + month + poutcome + contact + default + day_of_week + cons_conf_idx + pdays + campaign + previous + marital`
M2	Backward AIC	`y ~ marital + default + contact + month + day_of_week+ campaign + pdays + previous + poutcome + emp_var_rate + cons_price_idx + cons_conf_idx + euribor3m`
M3	Backward BIC	`y ~ default + contact + month + pdays + previous + emp_var_rate + cons_price_idx + euribor3m`
M4	Forward BIC	`y ~ nr_employed+ poutcome + month + contact + default`

In order to select a better model among the four models, we need to perform Anova's chi-square distribution test on these four models. Anova is used to study whether different levels of a control variable have a significant impact on observed variables. In the test of m1, the variables "previous" and "martial" did not pass the test. In the test of m2, only the variable "previous" did not pass the test. In m3, the variable "previous" also failed the test. Only in m4, all variables passed the test. The test results corresponding to m4 are as follows.

Coefficients	y ~ nr_employed + poutcome + month + contact + default		Pr(>Chi)	
intercept	55.90459228			
nr_employed	-0.01133646		< 2.2e-16	***
Poutcome	nonexistent	0.67242687	< 2.2e-16	***

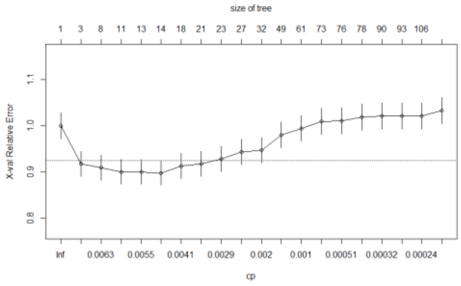
	success	1.89561317		
Month	Aug	-0.09072748	< 2.2e-16	***
	Jul	0.20581592		
	Mar	0.89021324		
	Nov	-0.31265442		
	Sep	-0.48931445		
	Dec	0.15310394		
	Jun	0.24011402		
	May	-0.69869136		
	Oct	0.14980945		
Contact	Telephone	-0.52609629	3.977e-08	***
default	unknown	-0.40574868	0.0002841	***

Next, the sensitivity is used to represent the ratio of TP to TP+FN. In order to determine the threshold when the sensitivity is 80%, we first need to obtain the predicted probabilities of the logical regression model of m4. Then I need to ask for the point where the sensitivity is equal to 0.8, but the points that are not determined can be accurately matched. Through the interpolation method, the sensitivity of the model can be calculated to be up to about 80% where the threshold is 0.0663888. The truth table of it is shown as below.

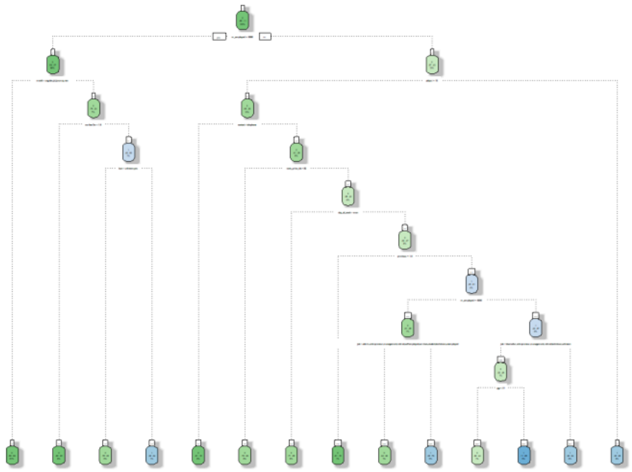
	Predicted class		
True class		1	0
	1	912	222
	0	3916	4950

Decision tree

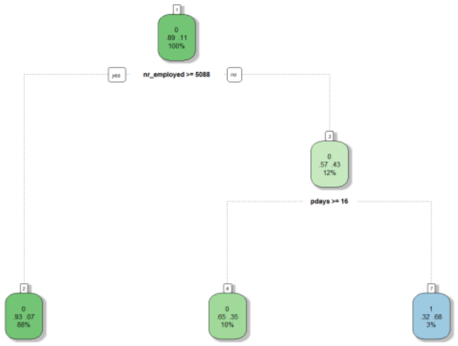
The basic idea of the decision tree model is to classify the existing data as a whole by greedy algorithms. First, we set the cp value to 0 and generate an original decision tree. Next, the original data and the generated model are used for prediction to obtain the corresponding relative error and 10-fold Cross-Validation estimate of relative error. Because CV error is known, we can query the CP value corresponding to the minimum value of CV error by function, and use this CP value as the reference value of the original decision tree to get DT with minimum CV error. Another DT model is obtained by trimming the original decision tree through the 1-SD principle. First, we first find the point with the least relative error and add a standard deviation to the relative error of the point. Then we find the first relative error value on the curve that is smaller than this value. The number of branches corresponding to this error value is the best decision tree selected by the 1-SD rule. The corresponding image is as shown below.



According to the above method, we obtain the cp value of DT selected by minimizing CV error is 0.004850088, and the number of decision tree splits generated is 13. The cp value of DT obtained by the 1-SD rule is 0.00670194, and the number of corresponding splits is two.



CV error



1- SD rule

Using the model and the original data to make predictions, the ROC curves corresponding to the two models are obtained respectively. The auc of the 1-SD model is 0.3012. The auc of the model obtained by CV error is 0.7589. So, we choose the CV error one. The table below describes the variables and importance used by the DT we chose.

nr_employed	euribor3m	cons_conf_idx	emp_var_rate	cons_price_idx	month
290.94	270.29	210.63	168.90	152.91	148.78
pdays	poutcome	previous	contact	day_of_week	job
46.85	41.88	15.10	10.02	7.95	7.29
age	loan	education	marital	housing	campaign
6.08	4.93	1.69	0.41	0.27	0.22

Similarly, we use the method used to find the threshold used in logistic regression. It can be found that the threshold with a sensitivity of 0.8 is close to 0.481. When the threshold is equal to 0.481, its corresponding truth table is as follows.

	Predicted class		
True class		1	0
	1	351	783
	0	173	8693

conclusion

Among the three sets of independent variables given, the economic condition group has the greatest impact on the customer's response. When these macro variables are in certain specific intervals, they will have a greater impact on the customer's response. Banks can use these indicators to make choices about the timing of the promotion. In terms of ensuring the practicability of the model, I chose to determine the corresponding threshold to ensure that the model's sensitivity remains at 0.8.

In the data processing, I found some quality problems and deviation points of the given data. And for macroscopic continuity variables, I found a large correlation between them. This correlation may cause deviations in the coefficients during logistic regression, but with the step function, the resulting model leaves only one macro variable, so the correlation has little effect in the fitting.

For the logistic regression model, all the variables I chose for the best model passed the chi-square test. According to the best equation we get, we can know that the bank can conduct targeted marketing through the number of employees, the results of the last campaign, the contact method, the month and whether the customer itself has a default. The practice for specific variables has been explained above. For the decision tree model, the model I selected has 18 variables, but the classification results are more accurate. Through the results, we can see that the macroeconomic variables have the greatest impact on the campaign, the influence of the market campaign is second, and the personal influence of the customer is the least. This means that the bank can first consider the macro factors to determine the time when holding the campaign, and then screen the target population through the market campaign factor. If you want to further segment the target population, you can use the customer's personal characteristics for analysis.

Logistic regression classifiers establish equations to determine the relationship between variables and final customer decisions. The result of this method is a probability between 0 and 1, and can be applied to continuous variables and categorical variables. But at the same time, it is sensitive to the interrelationship of the independent variables in the model. Moreover, the transition from odds to probability in the prediction results is nonlinear, resulting in a lot of interval variable changes without discrimination.

For decision trees, the decision tree creation method is simpler and the classifier created is more accurate. This method is applicable to high-dimensional data, continuous class data, and sub-type data. But the disadvantage is that it is prone to overfitting and it is easy to ignore the correlation between attributes.