Bank Customer Response Based on Data Mining

Yinbin Liu[a], Mimi Yu[b, *], and Ziyi Hu[c]

School of Management, Shanghai University, Shanghai, China

[a]yinbinliu@126.com, [b]doublemi_yu@163.com, [c]512608474@qq.com

*Corresponding author

Keywords: date, banking house, customer response.

Abstract: The development of the era of big data has made China's banking industry face fierce external competition and huge internal limitations. Moreover, the current research on the banking industry mostly stays in the theoretical analysis, and rarely continues to quantitatively analyze the bank customer response problem. First based on the viewpoint of precision marketing, this paper reviews the research status of using data mining technology in precision marketing. Second, we use the telemarketing data of a bank in Europe in 2014, then the collected bank data is cleaned and converted, and the appropriate bank customer response model: logistic regression model is selected. Then the model is evaluated and predicted. At last, the modeling results show that the precise marketing based on data mining technology plays an important role in customer response of banking industry, and discusses the important role of precise marketing in improving marketing efficiency and reducing bank marketing costs.

1. Introduction

With the rise of Internet technology, we have entered a big data era. The behavioral characteristics of people's daily lives can be recorded and saved through data. And by analyzing and processing these huge amounts of data, companies can identify customer needs and identify market directions to enhance their competitiveness. Generally, big data has the characteristics of huge volume, various types, low value density and fast processing speed, and with the rapid development of bank Informa ionization, the large amount of business data generated is in line with big data characteristics [1]. A large amount of business data, including basic information, transaction information and other raw data, as well as consumer behavior analysis derived from raw data, demand preferences and other in-depth data, extract valuable information from these massive data, and serve the bank's marketing decision-making is an important application of data mining [2].

One of the most frequently used predictive models for data mining is the response model, which uses a response model to predict which customers are most likely to respond to an activity. The response models in data mining methods are described in many documents and are widely used in practical commercial applications, but these methods are not fully in line with business objectives. And for companies, the main task of building a response model is to increase the return on investment by identifying the customers who are most likely to respond. Also, for those customers who are most likely to respond, the company will take responsive actions, and customers who are not responsive do not need to act on them, which can effectively reduce the cost of activities and thus improve the rate of return on investment. According to whether customers respond to marketing activities, they are divided into two categories: responding to customers and not responding to customers.

In recent years, the application of data mining technology to bank precision marketing has become a research hotspot. As early as 1995, more than 50 of the top 100 commercial banks in the United States produced data mining projects. With the development of data mining technology, many large foreign banks have applied data mining technology to all aspects of operation, management and decision-making in order to obtain useful information from huge data [3]. It can be said that data mining technology has become a key technology for foreign banks to win in the fierce market competition. The research and application of precision marketing in China's banks is only a few years, and the practical application of data mining in domestic commercial banks is still rare. This paper collects the data mining and analysis of a foreign bank data combined with R language [4], and uses the quantitative analysis method to conduct bank customer response research, which is beneficial to combine practical application with theory to make up for the domestic research gap in this field.

2. Data pre-processing

Data quality is the key to modeling. Data pre-processing consumes the longest time in the whole data mining process. And it is a very important stage in the data mining process. The purpose is to prepare a "clean" data set for modeling, improve the quality of data, and further improve the efficiency and accuracy of data mining and the effectiveness of the final mining pattern. Figure 1 shows the process of data pre-processing.

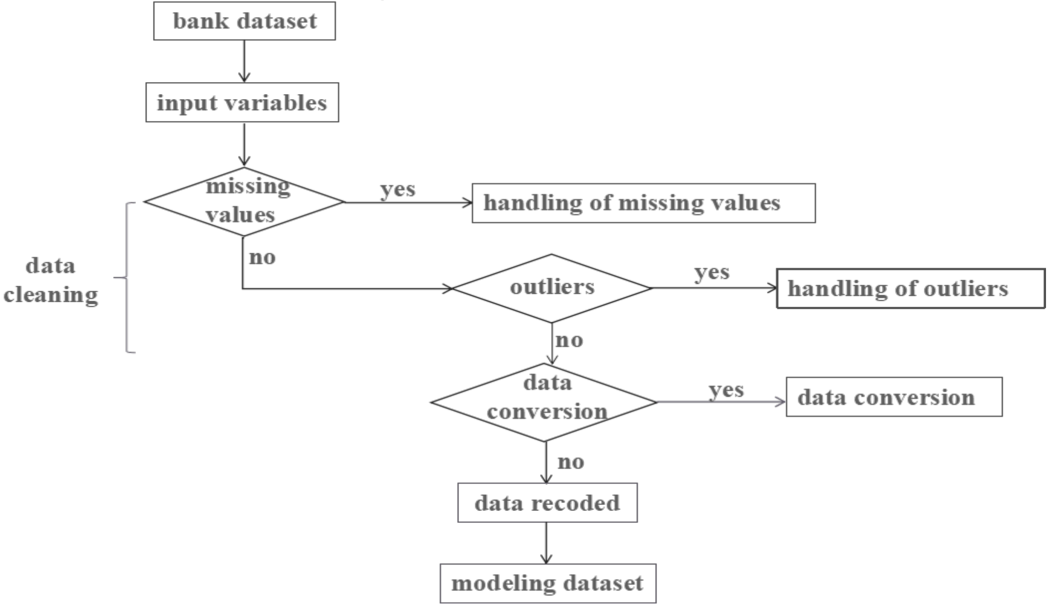

Figure 1. The process of data preprocessing

When receiving the paper, we assume that the corresponding authors grant us the copyright to use the paper for the book or journal in question. Should authors use tables or figures from other Publications, they must ask the corresponding publishers to grant them the right to publish this material in their paper.

2.1 Bank dataset selection

Data selection refers to retrieving and analyzing data related to tasks from the database, including the following three aspects:

(1) Modeling data

The telephone marketing data of a bank in Europe in 2014 was selected. After screening and cleaning, a total of 41,188 data were collected.

(2) Target variable (Response variable)

The target variable is the dependent variable or response variable, which represents the goal of data mining. And the choice of target variables is closely related to the definition of business issues and requires a clear answer from the business staff. The target variable for customer response modeling in this paper is whether the customer purchases a product recommended by the bank, denoted by y.

(3) Input variables (Predictors)

Input variables are also called independent variables or predictors, which are used to interpret the variables of customer responses (i.e. dependent variables). Creating these predictors is the key to building an effective model and there are 20 input variables as shown in Table 1 in this paper.

Table 1. 20 input variables (predictors)

Number	Variables	Description
1	age	Age, Numerical type
2	job	Job: admin, blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown;
3	marital	Marital status: divorced, married, single, unknown;
4	education	Education level: basic.4y, basic.6y, basic.9y, high school, illiterate, professional course, university degree;
5	default	Credit card default: no, yes, unknown;
6	housing	Mortgage: no, yes, unknown;
7	loan	Personal loan: no, yes, unknown;
8	contact	Contact method: cellular, telephone;
9	month	Last contact month: Jan, Feb, Mar, ..., Nov, Dec;
10	day_of_week	Last time was contacted by the day of the week: mon, tue, wed, thu, fri;
11	duration	The duration of the last call, in seconds.
12	campaign	How many times did the promotion contact?
13	pdays	How many days have you been contacted by the last time you last contacted the promotion?
14	previous	How many times have the bank contacted the customer during the last promotion?
15	poutcome	Results of the last marketing campaign: failure, non-existent, success
16	emp.var.rate	Employment change rate, quarterly indicator
17	cons.price.idx	Consumer price index, monthly indicator
18	cons.conf.idx	Consumer confidence index, monthly indicator
19	euribor3m	European bank 3-month lending rate, daily indicator
20	nr.employed	Number of people employed

2.2 Data cleaning

In the original data, there are inevitably some "dirty" data such as vacancy values, noise (error or abnormal data). And these "dirty data" will mislead the data mining search process, so data cleaning is needed to improve the quality of the data. The data cleaning process usually involves identifying missing values and identifying outliers.

Handling of missing values: since the “unknown” class in Table 1 does not provide us with information, this article treats this class as a missing value and uses random forest interpolation to deal with missing values.

Handling of outliers: outliers generally consider outliers and outliers that violate common sense. And the way to deal with them is to treat the outliers as missing values or delete or interpolate. There are two specific ways to handle outliers. One way of processing is according to the model used in the article. For example, the decision tree model allows missing values, but there is no missing value in the logistic regression model. The other is based on the proportion of outliers. If the amount of abnormal data is relatively small, it can be deleted directly. However, if there is a large amount of abnormal data, it is interpolated as a missing value, because deletion will reduce useful information.

2.3 Data conversion

Data conversion is mainly to normalize the data and transform the data into a form suitable for mining. In order to avoid the influence of different value ranges, the data are standardized before analysis so that they are all in a similar interval.

2.4 Data recoded

Due to the excessive classification of some data in the input variables, it is not conducive to modeling and analysis. So, the classifications other than “unknown” in the job, month and education variables are recoded.

1) There are 12 original classifications of the job variables, which are simplified into four categories: management, services, entrepreneur and unemployed, as shown in Table 2.

Table 2. Data recoded of variable job

original classification	after processing	original classification	after processing
admin.	management	self-employed	entrepreneur
blue-collar	services	services	services
entrepreneur	entrepreneur	student	unemployed
housemaid	entrepreneur	technician	services
management	management	unemployed	unemployed
retired	unemployed		

2) There are 12 original classifications of month, which are simplified into four quarterly: Q1, Q2, Q3 and Q4, as shown in Table 3.

Table 3. Data recoded of variable month

original	after	original	after	original	after	original	after
Jan	Q1	Apr	Q2	Jul	Q3	Oct	Q4
Feb	Q1	Mar	Q2	Aug	Q3	Nov	Q4
Mar	Q1	Jun	Q2	Sep	Q3	Dec	Q4

3) There are 7 original classifications of education, which are simplified into three categories: primary, secondary and tertiary, as shown in Table 4.

Table 4. Data recoded of variable education

original classification	after processing	original classification	after processing
illiterate	primary	high school	secondary
basic.4y	primary	professional course	tertiary
basic.6y	primary	university degree	tertiary
basic.9y	primary		

The data mining modeling preparation based on the above data refers to a series of data preparation work for the data source for implementing various data mining methods, mainly including missing value and outlier processing, variable conversion and data recoded. In this study, the logistic regression method is used to establish a bank customer response model, which needs to model and analyze the data after the missing value processing.

3. Customer Response Model

3.1 Selection of customer response models

With the deepening of research, more and more methods are used to build customer response prediction model, and each model has its own advantages and disadvantages. At present, the current models mainly include multiple linear regression models, multiple logistic regression models (Logistic), and Probit models.

For customer response prediction problems, it is found that the influencing factors may be quantitative or qualitative. The response variable (also called attribute variables) presents two qualitative variables with different reaction outcomes. For the characteristic represented by the attribute variable Y, that is, whether the customer purchases, the statistical characteristic describing whether this characteristic occurs or not is the probability of its occurrence. Customer response model predicts whether a customer will buy a product promoted by a bank. So, the target variable belongs to a discrete variable. And the variable only has 0 and 1 values indicating whether it is yes or no (0 means no purchase, 1 means purchase), this is related to Logistic. The characteristics of the regression model are consistent. From the bank data type, it can be seen that the problem of the dependent variable and the independent variable is fully satisfied, and the logistic regression analysis method is applied.

In addition, it is also considered that the Logistic regression has no assumptions about the distribution of independent variables. In the selection of variables, the multivariate linear model requires the variable data to be normally distributed, which undoubtedly limits the optional range of the model. The Logistic regression model is designed to calculate the conditional probability of the research object. Since it is based on the cumulative probability function, the condition that the independent variable obeys the multivariate normal distribution is not needed.

Although the logistic regression model requires no multiple linear correlation between variables, according to the correlation analysis of each index variable, some indicators with multiple linear correlations can be eliminated, which can be consistent with the model to the variable indicators. There are multiple linear correlation requirements, which increase the possibility of universal application of the model and improve the accuracy of the model prediction.

In summary, the use of logistic regression method to establish a customer response model has a higher recognition rate, also is more feasible in the selection of samples, and is more extensive in the selection of indicators. Therefore, this paper uses Logistic regression model to study, in order to achieve the purpose of responding to bank customer response.

3.2 Logistic regression model

Logistic regression [5] is a nonlinear regression model designed for binary dependent variables (dichotomous variables). It is mainly used to predict binary response variables (such as yes or no) and is an effective way to solve the 0-1 regression problem.

Let p be the probability of occurrence of an event Y, ranging from 0 to 1, then (1-p) is the probability that the event does not occur. The probability p in which the response variable Y appears is related to the independent variable (x1, x2, …, xn) and can be written as a function form in mathematics, that is, p=f (x1, x2, …, xn). The defined probability p can only take a value between 0 and 1, and the value of the applied variable of the function is in the real number set, beyond the range of the probability p, which is unlikely to occur outside (0, 1). As a result, it is necessary to mathematically transform the probability p to appear for Y under the action of (x1, x2, ..., xn).

Let p/ (1-p) take the natural logarithm of ln (p/ (1-p)), that is, logit transform for p. Under the action of the independent variable, the probability model for establishing Y=1 is:

$$\operatorname{logit}(p) = \ln(p/(1-p)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \qquad (1)$$

(Where (x1, x2, …, xn) is the predictor; β0 is a constant term, indicating the natural logarithmic parameter of the ratio when the value of the independent variable is all zero; (β0, β1, ..., βn) is called the regression coefficient indicates that when the value of the other independent variable remains unchanged, the value of the independent variable is increased by one unit to cause the change of the natural logarithm of p/ (1-p). After deriving the Logistic regression coefficient, a Logistic regression model can be established and used for prediction.

Deformation of equation (1) to obtain equation (2), that is $\mathrm{p} = \frac{exp(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)}{1+exp(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)}$. After obtaining the specific structure and parameters of the model, the parameters are substituted into equation (2) to estimate the probability of each individual response. In practice, a minimum level of acceptable customer response probability (cutoff point) is selected as the dividing line. Generally, the model chooses 0.5 as the segmentation point. If the predicted probability is greater than 0.5, the predicted customer will purchase the product promoted by the bank, otherwise they will not buy.

In the R language environment, the output of Table 5 is obtained: it can be seen from the results that some variables are not significantly related, and the variable filtering operation is required.

3.3 Variable screening

In a regression model, not all independent variables have a significant relationship with the dependent variables, and sometimes the effects of some independent variables can be ignored. Therefore, it is necessary to choose the truly statistically significant parts, and to select some of the variables that have significant influence on the dependent variables, and establish a better model.

In the logistic regression modeling, stepwise regression is used to screen the variables. The stepwise regression method is divided into forward stepwise regression, backward stepwise regression and two-way stepwise regression. This paper uses two-step stepwise regression for variable screening. The results after the variable screening are shown in Equation (3) and Table 5.

glm (formula = y~ job + education + month + day_of_week + duration + campaign + pdays+poutcome + emp.var.rate + cons.price.idx + cons.conf.idx + nr.employed

Table 5. Output results after variable screening

Coefficients:	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-2.048e-02	1.423e+01	-14.387	< 2e-16
jobservices	-1.243e-01	4.994e-02	-2.488	0.0126
jobentrepreneur	-1.294e-01	7.766e-02	-1.667	0.00956
jobunemployed	2.672e-01	6.519e-02	4.099	4.15e-05
educationsecondary	9.694e-02	5.833e-02	1.662	0.0956
educationtertiary	2.906e-02	5.263e-02	5.523	3.34e-08
contacttelephone	-6.109e-01	7.067e-02	-8.645	< 2e-16
monthQ2	-2.245e+00	1.108e-01	-20.251	< 2e-16
monthQ3	-1.479e+00	1.147e-01	-12.897	< 2e-16
monthQ4	-1.993e+00	1.183e-01	-16.840	< 2e-16
duration	4.655e-01	7.367e-05	63.180	< 2e-16
campaign	-4.528e-02	1.155e-02	-3.920	8.85e-09
pdays	-8.902e-04	2.022e-04	-4.402	1.07e-05
poutcomenonexistent	5.220e-01	6.354e-02	8.228	< 2e-16
poutcomesuccess	9.934e-01	2.026e-01	4.902	9.48e-07
emp.var.rate	-1.393e+00	6.870e-02	-20.283	< 2e-16
cons.price.idx	1.864e+00	1.038e-01	17.952	< 2e-16
cons.conf.idx	3.696e-02	4.836e-03	7.642	2.14e-14
nr.employed	5.839e-03	9.754e-04	5.986	2.15e-09

3.4 The meaning of Logistic regression coefficient

The odds, refers to the ratio of the probability of an event to the probability of not occurring. Odds = p/ (1-p) = -1+ 1/ (1-p) indicates the relationship between purchase and disregard. Odds is a number from 0 to infinity. And the larger the value of odds, the more likely the event will occur.

Table 6 gives the results of the odds corresponding to each variable. For example, the benchmark level of the job variable is management. From the above results, the odds of the service industry and the autonomous laborer purchasing bank products are 0.88 times that of the management employees. And the probability of unemployed people purchasing bank products is 1.25 times that of management personnel. Therefore, it is possible to promote the marketing of banking products and increase the success rate.

Table 6. Results of the odds corresponding to each variable

(Intercept)	Age	Jobservices	Jobentrepreneur	Jobemployed	Maritalmarried
1.0820e-85	1.0021e+00	8.7534e-00	8.7770e-01	1.2523e+00	9.9370e-01
educationtertiary	defaultyes	housingyes	contactelephone	loanyes	monthQ2
1.3184e+00	7.9858e-04	9.9218e-01	5.3961e-01	9.4691e-01	1.0807e-01
day_of_weekmon	day_of_weekthu	day_of_weektue	day_of_weekwed	duration	campaign
8.8772e-01	1.0699e+00	1.0724e+00	1.1625+00	1.0046+00	9.5606e-01
poutcomenonexistent	poutcomesuccess	emp.var.rate	cons.price.idx	cons.conf.idx	Euribor3m
1.5737e+00	2.5423e+00	2.4506e-01	6.1883e+00	1.0344e+00	1.0703e+00

4. Model evaluation

The model obtained in data mining may be of no practical significance or use value. The effectiveness evaluation can determine whether an effective and useful model is obtained. And the model should be evaluated using data that is not involved in model to get accurate results. Therefore, after modeling, the new sample data is used as verification set [6], which is used for verification and correction of the model to verify the model. The method of verification is to use the model to predict the data of known customer status, then compare the predicted value with the actual customer status, and apply the model when the prediction accuracy reaches the standard.

Use the ROC curve [7, 8] to evaluate the model, that is, use the data to predict the model. The ROC curve is based on a series of different two-category methods (demarcation value or decision threshold), with the true positive rate (sensitivity) as the ordinate, and the false positive rate (Special effect) as the abscissa. Sensitivity [7] also known as the true case rate, that is, the correct identification of the percentage of positive elements and sensitivity = TP / (TP + FN). Specificity [7] also known as the true negative rate, that is, the correct recognition of negative the percentage of tuples, Specificity = TN / (TN + FP). True Positive (TP) is a positive sample that is predicted to be positive by the model; False Negative (FN) is a positive sample that is predicted to be negative by the model; False Positive (FP) is a negative sample that is predicted to be positive by the model. True Negative (TN) is a negative sample that is predicted to be negative by the model. The more convex the ROC curve is, the closer the upper left corner is to indicate that the value of the model is larger. That is, the area under the curve (AUC) [9] can evaluate the accuracy of the customer response. In other words, if the area under the curve is larger, the accuracy of the customer response model is higher.

Figure 2 is the ROC curve obtained by evaluating the model. And it can be seen from the figure that the probability threshold of the curve is 0.084. When the probability threshold is more than the 0.084, the customer will purchase the products promoted by the bank, otherwise the customer will not purchase the products promoted by the bank. Moreover, the value of AUC at this time is 0.934, which is close to the maximum value of 1, indicating that the accuracy of the prediction of the model is still very high.

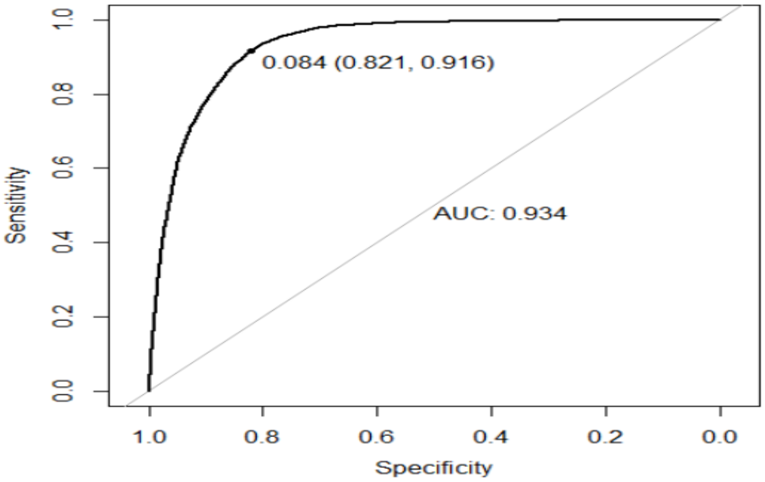

Figure 2. ROC curve

In this paper, the accuracy and sensitivity of the probability thresholds of 0.084 and 0.05 are calculated and compared, as shown in Table 7. Although the accuracy of the probability threshold of 0.084 is lower than the accuracy of 0.05, the sensitivity of the probability threshold of 0.084, that is, the actual response of the customer is recognized almost twice the probability threshold of 0.05.

Table 7. Comparison of probability thresholds

p>0.084	0	1	Accura-cy	Sensiti-vity	**p>0.05**	0	1	Accuracy	Sensitivity
0	30000	6548	0.83	0.91	0	35585	963	0.91	0.423
1	389	4251			1	2677	1963		

5. Conclusion

This paper collects data, then cleans data, transforms and re-encodes data, builds a customer response model based on data mining technology, and effectively combines R language technology with precise marketing concepts. Table 8 gives a comparison of the customer response model before and after. The main goal of establishing a response model for accurate marketing of banks is to increase the return on investment of marketing activities, identify the customers who are most likely to respond, and take appropriate promotions for those customers who are most likely to respond. Customers with low responsiveness do not need to market them, which can improve the return on investment, improve the success rate of marketing activities, improve the efficiency of marketing activities, reduce the cost of marketing activities, and have practical value.

Table 8. Comparison of effects before and after modeling

Customer response model	Sales basis	Success rate	Cost
yes	blind sales	low	high
no	have a basis for sales	high	low

This paper analyzes whether the customer purchases the products promoted by the bank and establishes a model for predictive evaluation. The research of this paper can be promoted and applied in our daily life. For example, the issuance of coupons can be targeted and cost-effective. Future research can consider the actual cost and carry out predictive analysis on the basis of cost. For this research, the books and literatures published so far have not been considered, and it is a very valuable research direction.

References

[1] Gandomi A, Haider M. Beyond the hype: big data concepts, methods, and analytics [J]. International Journal of Information Management, 2015, 35(2): 137-144.

[2] Elsalamony H A. Bank direct marketing analysis of data mining techniques [J]. International Journal of Computer Applications, 2014, 85(7):12-22.

[3] Ahmad R, Buttle F. Retaining telephone banking customers at frontier bank [J]. International Journal of Bank Marketing, 2002, 20(1): 5-16.

[4] Faraway J J. Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models [M]. CRC Press, 2016.

[5] Harrell F E. Ordinal logistic regression [M]//Regression modeling strategies. Springer, Cham, 2015: 311-325.

[6] Moro S, Cortez P, Rita P. A data-driven approach to predict the success of bank telemarketing [J]. Decision Support Systems, 2014, 62: 22-31.

[7] Carter J V, Pan J, Rai S N, et al. ROC-ing along: evaluation and interpretation of receiver operating characteristic curves [J]. Surgery, 2016, 159(6): 1638-1645.

[8] Althouse A D. Statistical graphics in action: making better sense of the ROC curve [J]. International Journal of Cardiology, 2016, 215: 9-10.

[9] Parlar T, ACARAVCI S K. Using data mining techniques for detecting the important features of the bank direct marketing data [J]. International Journal of Economics and Financial Issues, 2017, 7(2): 692-696

Table 5. Output of Logistic regression

Coefficients:	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-1.956e+02	1.931e+01	-10.133	< 2e-16
age	2.153e-03	1.977e-03	1.089	0.276224
jobservices	-1.331e-01	4.935e-02	-2.698	0.006978
jobentrepreneur	-1.305e-01	7.787e-02	-1.675	0.093882
jobunemployed	2.250e-01	6.692e-02	3.362	0.000774
maritalmarried	-6.314e-03	6.793e-02	-0.093	0.925949
maritalsingle	7.028e-02	7.731e-02	0.909	0.363299
educationsecondary	1.000e-01	6.014e-02	1.663	0.096341
educationtertiary	2.847e-01	5.386e-02	5.285	1.26e-07
defaultyes	-7.140e+00	1.134e+02	-0.063	0.949784
housingyes	6.835e-03	4.072e-02	0.168	0.866718
loanyes	-6.325e-02	5.688e-02	-1.112	0.266097
contacttelephone	-6.160e-01	7.122e-02	-8.649	< 2e-16
monthQ2	2.223e+00	1.125e-01	19.751	< 2e-16
monthQ3	7.573e-01	6.594e-02	11.484	< 2e-16
monthQ4	2.185e-01	6.594e-02	2.501	0.012368
day_of_weekmon	-1.218e-01	6.577e-02	-1.852	-1.852
day_of_weektue	6.812e-02	6.533e-02	1.043	0.297047
day_of_weekwed	1.484e-01	6.518e-02	2.276	0.022830
day_of_weekthu	6.597e-02	6.373e-02	1.035	0.300616
duration	4.655e-03	7.369e-05	63.175	< 2e-16
campaign	-4.486e-02	1.155e-02	-3.883	0.000103
pdays	-9.725e-04	2.161e-04	-4.500	6.78e-06
previous	-5.658e-02	5.872e-02	-0.964	0.335289
poutcomenonexistent	4.540e-01	9.363e-02	4.849	1.24e-06
poutcomesuccess	9.312e-01	2.104e-01	4.426	9.61e-06
emp.var.rate	-1.407e+00	7.683e-02	-18.318	< 2e-16
cons.price.idx	1.829e+00	1.190e-01	15.363	< 2e-16
cons.conf.idx	3.410e-02	6.662e-03	5.119	3.07e-07
euribor3m	6.589e-02	1.124e-01	0.586	0.557874
nr.employed	4.816e-03	1.870e-03	2.575	0.010021