Detecting potential customers of term deposit subscription

1 Introduction

Telemarketing is a very important strategy in business. Over the phone, salespeople can sell goods and services to potential customers directly. But telemarketing is not a random number of calls and to promote a few products by chance, such calls tend to antagonize customers and backfires. Through the analysis of customers, enterprises can find out target groups and improve transaction volume and customer satisfaction.

In this report, we are going to help a Portuguese retail bank to subscribe new users to a long-term deposit. Our goal is to build a model that can predict the result of telesales to sell a long-term deposit based on the information of customers. The data set 'Bank Marketing' contains 45211 observations and 20 features related to bank client data, the last contact of the current campaign, social and economic context and so on. According to the literature review and the information, we will use logistic regression and random forest to build our model.

2 Analysis Plan

2.1 Data Resource

We decide to use the full dataset because it contains more observations that lead to a more accurate model. Otherwise, it contains more variables and gives us more information about the customers. We will select features based on those variables.

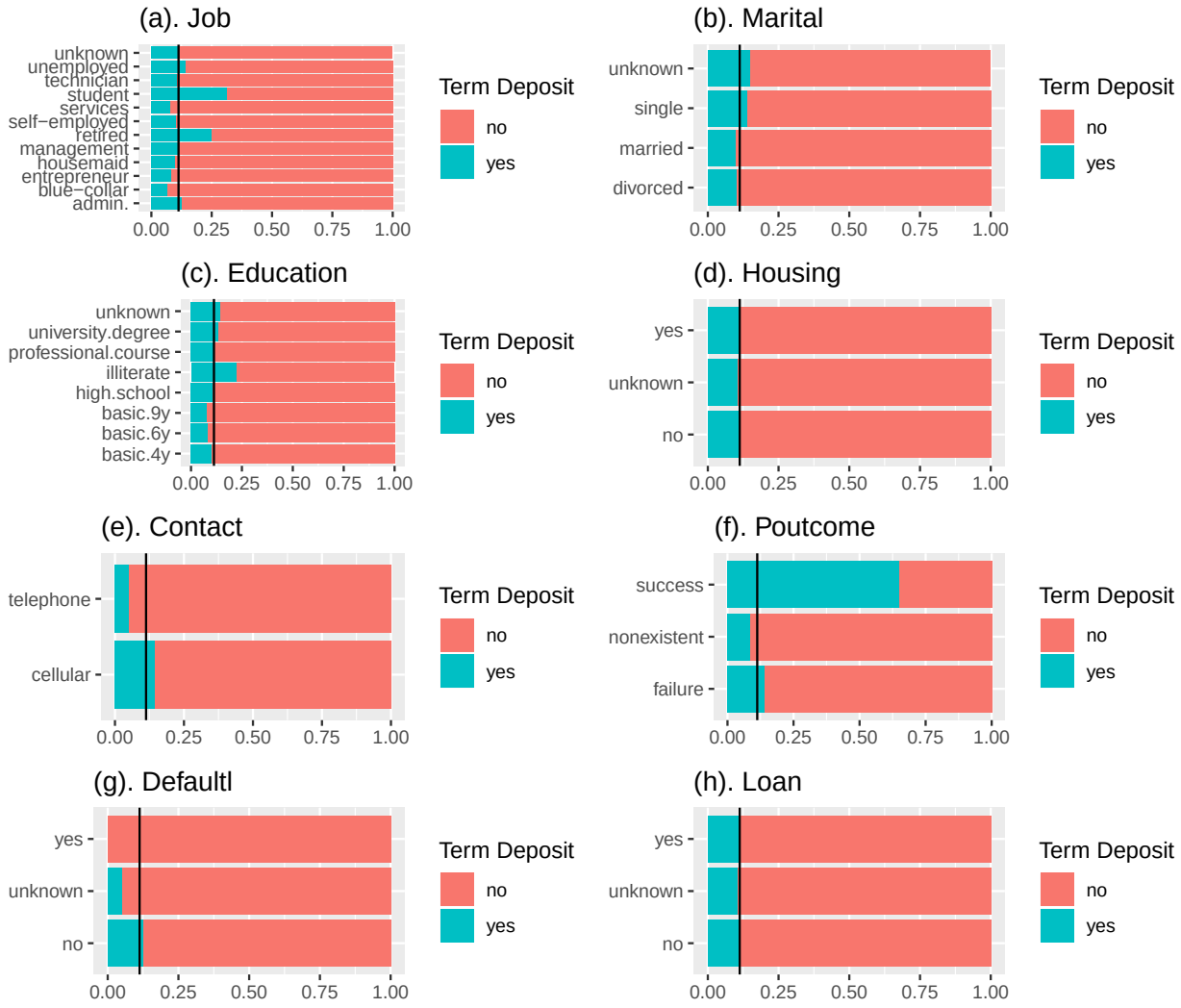
2.2 Descriptive Analysis

The original dataset contains 41188 observations with a term deposit ratio(numbers of clients subscribe to a term deposit divided by total clients) of 0.013. In this session, we aim to roughly explore the sources that would influence a client's decision on subscribing to a term deposit.

2.2.1 Influential Resources From Qualitative Variables

We draw grouped percentage bar plots for the qualitative variables(Figure 1). Considering a specific categorical variable 'job', if the student group in this variable has no influence on the term-deposit, then the percentage of clients that subscribe to a term deposit in this group will be equivalent to term-deposit ratio 0.013. Correspondingly, if clients in the student group are more likely to subscribe to the term deposit, then the percentage will be higher than the ratio, leading to the blue bar in the plot exceeding the ratio line. We highlight the key information from the bar plot above.
* Clients with jobs as 'student' and 'retired' tend to have term deposits, clients with jobs as 'blue-collar' tend to reject term deposits.
* Clients with illiteracy education are likely to subscribe to term deposits.
* If a client has experience with subscribing to a term deposit or is contacted with cellular, he tends to have term deposit. If a client has credit, it's of little chance for him to subscribe to a term deposit.

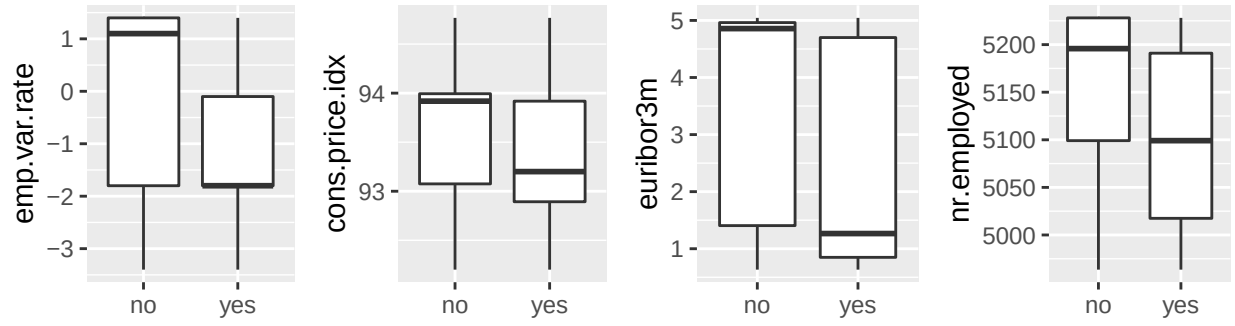
Figure 1



2.2.2 Influential Resources From Social and Economic Context

Economic factors affect the behavior of subscribing to a term deposit. Higher consumption confidence and higher substitution deposit rate drive people to abandon term deposit. Figure 2 shows an increase in employee rate, consumption confidence, three month rate lead to an decreasing in subscribing to term deposits.

Figure 2



2.3 Model

In this report, we are going to construct a logistic regression model and random forest model.

2.3.1 Logistic Regression Model

The logistic regression model is set as follows:

$$\pi(X) = \frac{1}{1 + e^{-(\beta_0 + \beta X)}}$$

- $\pi(X)$ is the probability that a client subscribe to term deposit.
- β_0 is the regression coefficient, β is the regression coefficient vector.
- X is the predictor variables vector including age, job, marital, education, default, housing, loan, contact, month, day of week, duration, campaign, pdays, previous, poutcome, emp.var.rate, cons.price.index, cons.conf.index, euribor3m, nr.employed, we choose these variables based on literature review[1].

Hypothesis testing:
For presence of β_i in the model:

$$H_0 : \beta_0 = 0 \quad v.s. \quad H_a : \beta_i \neq 0$$

Using the z-statistic $z = \frac{\hat{\beta_1}}{s(\hat{\beta_1})}$
For investigating if fitting a logistic linear model is appropriate, we define $\pi_i' = \log \frac{\pi_i}{1-\pi_i}$, then:

$$H_0 : \pi_j' = \beta_0 + \beta_1 X_j \quad v.s. \quad H_a : \pi_j' \textit{ do not line on straight line.}$$

2.4 Model Evaluation

Logistics Model assumption:

- Linearity assumption: linear relationship between continuous predictor variables and the logit of the outcome. This can be done by visually inspecting the scatter plot between each predictor and the logit values.
- Bernoulli distribution: Response variable follows the Bernoulli distribution.

3 Result

3.1 Model Fitting

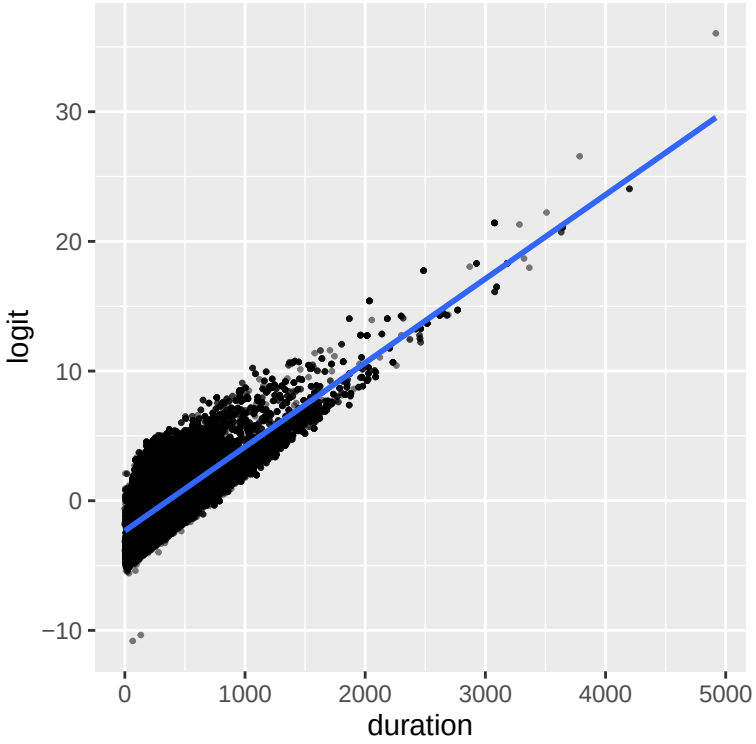
The result of model fitting is shown in the appendix. According to the hypothesis test we introduced before, we may notice that some variables are significant, such as duration, emp.var.rate, cons.price.index. Although some of them are not significant, we do not drop them. Because an automated selection approach is adopted, based on an adapted forward selection method[3], all these variables are necessary to get the best performance.

Compared with the association with the response variable, we think the performance is much more important to a classifier. We prefer a better ROC curve rather than the significance level.

The result of model fitting is shown in the appendix.

3.2 Model Diagnostic

The scatter plot shows that variable is quite linearly associated with the term deposit outcome in logit scale.



The response variable is whether the product (bank term deposit) would be or not subscribed. There are only two values: 'no' and 'yes' and we can rewrite it as 0 and 1. A client subscribed to the product with a certain probability. Hence, the response variable follows a Bernoulli distribution obviously.

3.3 Model Performance

3.3.1 Unbalanced Dataset

Table 1: Table 1: Result on test of original dataset

		Observed	Observed
		no	yes
Predict	no	7168	489
Predict	yes	195	386

For the logistic regression, when we fit the model with the original data, the accuracy of the model is 91% on the test set. It seems to be an ideal result that we only make less than 10% mistakes. However, notice that there are 905 positive samples in total while we only recognize 368 of them. Considering the aim of our

model, It is not well-performed. If we predict all the samples are negative, the accuracy is nearly 90%. This model is meaningless because it doesn't give any information.

After exploring the data, we found that the number of people who subscribe to the long-term deposit is far lower than the number of people who don't subscribe to the long-term deposit. So this is an unbalanced data. When we care about the minority classes, the class balancing techniques are really necessary.

3.3.2 Balanced Dataset

Because of the unbalanced dataset, it is not reasonable to measure the performance of the model with accuracy. In this case, the more positive targets are found out, the better the model is. We choose true positive rate (TPR) as the parameter to measure the performance. The TPR is 40.7%, which is really poor.

Table 2: Table 2: Result on test of balanced dataset

		Observed	Observed
		no	yes
Predict	no	775	115
Predict	yes	153	813

Hence, we use the undersampling method to reconstruct the dataset since we are more interested in the positive samples. We sample randomly from the negative samples to make the amount of positive samples and negative samples are equal. By this method, we wouldn't lose any information on positive samples and at the meantime, balance the dataset.

With the new dataset, the accuracy of the logistic model decreases a little, about 86.5%. But surprisingly, the true positive rate increases extremely, about 89%. It means we can predict most of the positive samples correctly. It is much more meaningful. Actually, we don't care much about the false positive rate because the cost of making a call is much less than the benefit.

3.4 Comparison between Logistic Regression and Random Forest

Random forest is a learning method for classification. It based on generating a large number of decision trees, each decision trees are used to identify a classification consensus by selecting the most common output.

Table 3: Table 3: Result on test of original dataset

		Observed	Observed
		no	yes
Predict	no	7128	235
Predict	yes	396	479

Table 4: Table 4: Result on test of original dataset

		Observed	Observed
		no	yes
Predict	no	767	161
Predict	yes	64	864

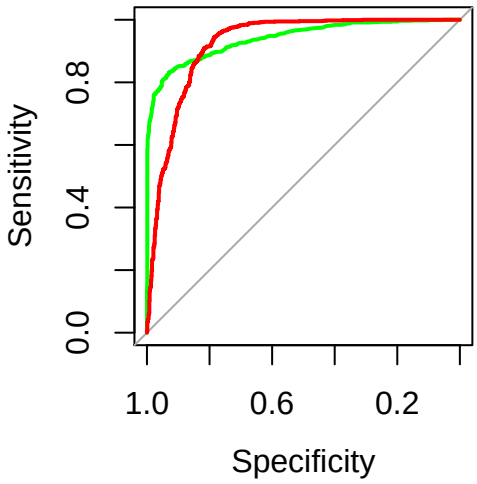
For the random forest model, when we fit the model with the original data, the accuracy of the model is 92% on the test set while the true positive rate is 54.6%. When we consider the balanced data, the accuracy of the model turns to be 88% while the true positive rate is up to 93.1%. It is slightly better than the logistic regression.

Table 5: Table 5: Comparison between Logistic Regression and Random Forest

	Logistic	Logistic	Random Forest	Random Forest
	unbalanced	balanced	unbalanced	balanced
Accuracy	91%	86.5%	92%	88%
TPR	40.7%	89%	54.6%	93.1%

We may get the same conclusion if we consider the ROC curve and the AUC. The AUC of logistic regression is 0.92 and the AUC of random forest is 0.94. As shown in the figure 4.

Figure 4



	Logistic Regression	Random Forest
Area under the curve	0.92	0.94

For the cases of more complex datasets, linear-based algorithms may not be sufficient in segmenting the class labels, leading to poor accuracies. More sophisticated algorithms may then be required like random forest, which can learn a non-linear decision boundary and thus can achieve higher accuracy scores.

Random forests can have a decision boundary with high variability in predictions but low bias. According to the decision boundary, logistic regression poorly segments the two classes while the more flexible decision boundary learned from the random forest model produces a higher classification accuracy.

4 Further Research

In this project, we did not adjust the parameters of the model, especially for the random forest. Moreover, we may use those more advanced sampling methods, such as SMOTE, to balance the dataset. Finally, to get a better performance, some other models worth trying, such as SVM, Adaboost.

5 Reference

[1]. Moro, Sérgio, Paulo Cortez, and Paulo Rita. "A data-driven approach to predict the success of bank telemarketing." Decision Support Systems 62 (2014): 22-31.

[2]. Kirasich, Kaitlin, Trace Smith, and Bivin Sadler. "Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets." SMU Data Science Review 1.3 (2018): 9.

[3].] Isabelle Guyon, André Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning Research 3 (2003) 1157–1182.

6 Appendix

```
##
## Call:
## glm(formula = factor(y) ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.7076  -0.3046  -0.1878  -0.1351   3.3134
##
## Coefficients: (1 not defined because of singularities)
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -2.108e+02  4.216e+01  -4.999 5.75e-07 ***
## age                           1.248e-03  2.686e-03   0.465 0.642267
## jobblue-collar               -2.184e-01  8.871e-02  -2.462 0.013828 *
## jobentrepreneur              -1.040e-01  1.368e-01  -0.760 0.447200
## jobhousemaid                 -5.869e-02  1.662e-01  -0.353 0.723938
## jobmanagement                 1.702e-02  9.395e-02   0.181 0.856231
## jobretired                    3.205e-01  1.182e-01   2.710 0.006727 **
## jobself-employed             -1.549e-01  1.320e-01  -1.173 0.240738
## jobservices                  -1.285e-01  9.547e-02  -1.346 0.178257
## jobstudent                    2.855e-01  1.235e-01   2.312 0.020779 *
## jobtechnician                 2.926e-02  7.947e-02   0.368 0.712717
## jobunemployed                 9.696e-02  1.397e-01   0.694 0.487505
## jobunknown                   -1.121e-02  2.558e-01  -0.044 0.965055
## maritalmarried               -7.462e-02  7.473e-02  -0.999 0.318000
## maritalsingle                 4.904e-03  8.550e-02   0.057 0.954262
## maritalunknown                1.184e-01  4.504e-01   0.263 0.792553
## educationbasic.6y             8.135e-02  1.350e-01   0.603 0.546657
## educationbasic.9y             1.200e-03  1.062e-01   0.011 0.990986
## educationhigh.school          4.865e-02  1.022e-01   0.476 0.633968
## educationilliterate           1.006e+00  8.648e-01   1.163 0.244856
## educationprofessional.course  5.650e-02  1.128e-01   0.501 0.616306
## educationuniversity.degree    1.924e-01  1.023e-01   1.881 0.060013 .
## educationunknown              1.723e-01  1.319e-01   1.307 0.191375
```

```
## defaultunknown               -3.545e-01  7.486e-02  -4.735 2.19e-06 ***
## defaultyes                   -7.250e+00  1.134e+02  -0.064 0.949043
## housingunknown                1.378e-02  1.534e-01   0.090 0.928433
## housingyes                    1.789e-02  4.587e-02   0.390 0.696430
## loanunknown                          NA         NA      NA       NA
## loanyes                      -2.432e-02  6.327e-02  -0.384 0.700634
## contacttelephone             -5.581e-01  8.425e-02  -6.624 3.49e-11 ***
## monthaug                      7.342e-01  1.334e-01   5.505 3.68e-08 ***
## monthdec                      2.039e-01  2.344e-01   0.870 0.384410
## monthjul                      1.968e-01  1.060e-01   1.857 0.063382 .
## monthjun                     -4.355e-01  1.388e-01  -3.138 0.001703 **
## monthmar                      1.948e+00  1.598e-01  12.189  < 2e-16 ***
## monthmay                     -5.173e-01  9.094e-02  -5.688 1.28e-08 ***
## monthnov                     -4.586e-01  1.342e-01  -3.417 0.000632 ***
## monthoct                      7.170e-02  1.710e-01   0.419 0.674918
## monthsep                      2.784e-01  1.989e-01   1.399 0.161708
## day_of_weekmon               -1.275e-01  7.311e-02  -1.744 0.081224 .
## day_of_weekthu                4.612e-02  7.086e-02   0.651 0.515149
## day_of_weektue                6.512e-02  7.315e-02   0.890 0.373305
## day_of_weekwed                1.701e-01  7.256e-02   2.345 0.019036 *
## duration                      4.689e-03  8.292e-05  56.544  < 2e-16 ***
## campaign                     -4.553e-02  1.297e-02  -3.511 0.000447 ***
## pdays                        -8.781e-04  2.452e-04  -3.581 0.000342 ***
## previous                      1.702e-02  6.610e-02   0.257 0.796819
## poutcomenonexistent           5.373e-01  1.049e-01   5.122 3.02e-07 ***
## poutcomesuccess               1.022e+00  2.382e-01   4.292 1.77e-05 ***
## emp.var.rate                 -1.597e+00  1.569e-01 -10.180  < 2e-16 ***
## cons.price.idx                1.957e+00  2.776e-01   7.049 1.80e-12 ***
## cons.conf.idx                 1.876e-02  8.609e-03   2.179 0.029350 *
## euribor3m                     2.648e-01  1.438e-01   1.842 0.065487 .
## nr.employed                   4.643e-03  3.430e-03   1.354 0.175861
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 23417  on 32949  degrees of freedom
## Residual deviance: 13872  on 32897  degrees of freedom
## AIC: 13978
##
## Number of Fisher Scoring iterations: 10

##          observed
## predicted   no  yes
##       no  7168  489
##       yes  195  386

##
## Call:
## glm(formula = factor(y) ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -7.2801  -0.3897  -0.0491   0.4836   3.0259
```

```
##
## Coefficients: (1 not defined because of singularities)
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -2.862e+02  6.667e+01  -4.293 1.76e-05 ***
## age                           -8.693e-04  4.351e-03  -0.200  0.84165
## jobblue-collar                -5.953e-02  1.442e-01  -0.413  0.67977
## jobentrepreneur               -1.028e-01  2.116e-01  -0.486  0.62708
## jobhousemaid                  -7.771e-02  2.556e-01  -0.304  0.76106
## jobmanagement                 -1.611e-01  1.527e-01  -1.055  0.29148
## jobretired                     6.468e-01  2.044e-01   3.164  0.00155 **
## jobself-employed              -2.238e-01  2.042e-01  -1.096  0.27308
## jobservices                   -1.093e-01  1.515e-01  -0.721  0.47080
## jobstudent                     3.713e-01  2.237e-01   1.660  0.09693 .
## jobtechnician                 -8.001e-02  1.284e-01  -0.623  0.53305
## jobunemployed                  9.621e-02  2.184e-01   0.441  0.65956
## jobunknown                     2.708e-01  4.066e-01   0.666  0.50538
## maritalmarried                -1.148e-01  1.246e-01  -0.921  0.35701
## maritalsingle                 -2.094e-02  1.424e-01  -0.147  0.88314
## maritalunknown                -8.691e-01  6.105e-01  -1.424  0.15456
## educationbasic.6y             -1.837e-01  2.160e-01  -0.851  0.39500
## educationbasic.9y              5.597e-02  1.674e-01   0.334  0.73805
## educationhigh.school           1.248e-01  1.666e-01   0.749  0.45396
## educationilliterate            2.347e+00  1.721e+00   1.364  0.17260
## educationprofessional.course   2.364e-01  1.840e-01   1.285  0.19882
## educationuniversity.degree     4.785e-01  1.685e-01   2.840  0.00452 **
## educationunknown               3.966e-02  2.203e-01   0.180  0.85713
## defaultunknown                -4.689e-01  1.161e-01  -4.038 5.39e-05 ***
## housingunknown                 3.144e-01  2.570e-01   1.223  0.22136
## housingyes                     1.543e-01  7.453e-02   2.071  0.03839 *
## loanunknown                           NA         NA      NA       NA
## loanyes                       -1.807e-01  1.031e-01  -1.754  0.07948 .
## contacttelephone              -3.800e-01  1.388e-01  -2.739  0.00617 **
## monthaug                       1.314e+00  2.454e-01   5.355 8.58e-08 ***
## monthdec                      -1.070e-03  4.107e-01  -0.003  0.99792
## monthjul                       2.028e-01  1.747e-01   1.161  0.24556
## monthjun                      -9.002e-01  2.254e-01  -3.993 6.52e-05 ***
## monthmar                       2.458e+00  2.924e-01   8.404  < 2e-16 ***
## monthmay                      -7.491e-01  1.463e-01  -5.119 3.08e-07 ***
## monthnov                      -4.713e-01  2.181e-01  -2.161  0.03069 *
## monthoct                       1.738e-01  2.764e-01   0.629  0.52949
## monthsep                       5.919e-01  3.322e-01   1.782  0.07481 .
## day_of_weekmon                 1.245e-01  1.184e-01   1.052  0.29294
## day_of_weekthu                 2.416e-02  1.162e-01   0.208  0.83531
## day_of_weektue                 6.444e-02  1.193e-01   0.540  0.58923
## day_of_weekwed                 2.332e-01  1.185e-01   1.969  0.04896 *
## duration                       6.829e-03  1.845e-04  37.021  < 2e-16 ***
## campaign                      -2.727e-02  1.943e-02  -1.403  0.16057
## pdays                         -8.298e-04  4.306e-04  -1.927  0.05397 .
## previous                      -1.998e-01  1.333e-01  -1.499  0.13383
## poutcomenonexistent            3.723e-01  1.887e-01   1.973  0.04844 *
## poutcomesuccess                1.384e+00  4.254e-01   3.253  0.00114 **
## emp.var.rate                  -2.338e+00  2.531e-01  -9.237  < 2e-16 ***
## cons.price.idx                 2.661e+00  4.449e-01   5.980 2.23e-09 ***
## cons.conf.idx                  1.241e-03  1.552e-02   0.080  0.93628
```

```
## euribor3m                    5.553e-01  2.379e-01   2.334  0.01960 *
## nr.employed                  6.360e-03  5.442e-03   1.169  0.24254
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10291.8  on 7423  degrees of freedom
## Residual deviance:  4902.3  on 7372  degrees of freedom
## AIC: 5006.3
##
## Number of Fisher Scoring iterations: 6

##          observed
## predicted  no yes
##       no  775 115
##       yes 153 813
```