Predicting Long-Term Deposit in Bank Marketing with Predictive Modeling

1. Introduction

Direct bank telemarketing campaigns are a technique of outreaching customers who are willing to purchase particular products, which is used to improve the financial benefits for stakeholders (Abu-Srhan & Al-Sayyed, 2019). However, it is a very time consuming and financial burden for commercial banks. Therefore, it is important to establish a cost-effective way to find target customers and achieve the goal of maximizing profit.

The purpose of this project is to establish a powerful prediction model to identify the most likely customers who will subscribe to long-term deposits. A Logistic regression model will be used in the project because of its wide application in the field of marketing management and its ability to discriminant analysis and classification (Mentzer et al, 2008). Besides, to validate the accuracy of prediction, we will compare between logistic regression model with a random forest model.

2. Exploratory Data Analysis

This study considers a data set with 41188 phone call records from a Portuguese retail bank, from May 2008 to June 2013. Each record contains the contact outcome (y = "yes" or "no", representing whether the client subscribed a term deposit) and 20 client features, including client basic information (age, marital, job, et al.), last contact information and social and economic context attributes.

Through the summary statistic (*Appendix 1*), the variable pdays (number of days that passed by after the client was last contacted from a previous campaign) has too many missing values (96% missing). The variable default (whether a person has credit in default) shows that most of people have no credit in default since only 3 people states they have. In this case, two features (pdays and default) can be removed from the data. In the original data, the variable duration means the last contact duration in seconds and it is not known before a call is performed. Also, after the end of the call whether the client has long-term deposit is obviously known. Thus, "duration" is not considered for realistic prediction. After removing all the missing data, a data set (named as Data0) containing 38245 observations is explored.

Through the correlation matrix(*Appendix 2*), three variables (emp.var.rate, euribor3m, nr.employed which are all social and economic context attributes) have strong correlations (correlation > 0.7) with other variables, which might cause multicollinearity in logistic regression.

The outcome of "no" (n = 33987) is much more than "yes" (n = 4258), which indicates the data are imbalanced. Since the data is highly imbalanced, which could cause inaccurate prediction, the Synthetic Minority Over-sampling Technique (SMOTE) is conducted to obtain a data set with equal number of two outcomes. In this project, logistic regression model and random forest model are used to make prediction. A new data set (named Data1) which is used for random forest model is obtained through conducting SMOTE on Data0, and another data set (named Data2) which is used for logistic regression is obtained through conducting SMOTE on Data0 after removing the three highly correlated variables. After conducting SMOTE, equal outcomes for "yes" and "no" are obtained without changing the characteristics of other variables. (See *Figure 1*)

Through exploratory data analysis, the bank marketing data set is divided into training data and test data. The training data which is 80 percent of bank marketing data set can be used to fit the models. As for testing

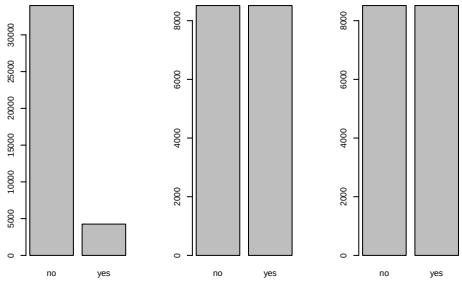

Figure 1: Checking Balanced Situations for Data0(Left),Data1(Middle) and Data2(Right)

data which is 20 percent of bank marketing data set, it is utilized to calculate the accuracy of the model prediction, which is able to show the efficiency and effectiveness of models.

3. Logistic Regression

3.1 Model Setting and Selection

In order to predict the long-term deposit in bank marketing, logistic regression model is suitable to make prediction. We can denote 1 if the client will subscribe a long-term deposit and denote 0 if the client will not. The logistic Regression model is:

$$log(p(x)/(1-p(x))) = \beta_0 + X\beta$$

where $p(x)$ is the probability of one client will be a long-term deposit member, β_0 is the intercept and β is the coefficient vector of relative predictors. In addition, X represents the predictors including bank client data, variables related with the last contact of the current campaign, social or economic context attributes and so on.

Since there are 14 predictors in the model, model selection method based on AIC is applied to reduce model complexity. AIC is an estimator of out-of-sample prediction error, so it can be used as a preliminary criterion of predictive precision and efficiency. In addition, AIC can deal with the problem of overfitting and the model with lower AIC value tends to have good performance of predicting new observations. To fit a better logistic regression model, stepwise model selection method with minimum AIC value is utilized in R.

3.2 Model Fitting and Evaluation

Based on the result of model selection, the selected model contains 10 predictors with minimum $AIC = 14788$. The estimates of coefficients are listed in the *Appendix 3*, most of the coefficients of variables are statistically significant. The 10 predictors are (1)marital status, (2)housing loan status, (3)personal loan status, (4)communication type, (5)last contact month, (6)number of contact during the campaign, (7)number of contact before the campaign, (8)outcome of the previous marketing campaign, (9)consumer price index and (10)consumer confidence index.

For prediction, the threshold value of classification is designated to be 0.5. In order to evaluate the selected model compared with the full model with all 14 predictors, the confusion matrix is used to analyze the performance of predictive models.

In this study(*Table 1*), successfully subscribing the time posits is regarded as interesting category called "Positive", while the others (fail to subscribe) is considered as "Negative". Compared with the full model, the

Table 1: Confusion matrix for selected model

	Actual NO	Actual YES
Predicted NO	1446(1430)	601(587)
Predicted YES	257(273)	1102(1116)

Note: The values in (*) represent the value of full model.

selected model with minimum AIC value have similar level of sensitivity and specificity(Sensitivity: Selected model = 0.64(Full model = 0.65); Specificity: Selected model = 0.85(Full model = 0.84)). For selected model, the sensitivity is relatively lower than the specificity. It means that this logistic model can detect the member who will not subscribe the long-term deposit more precisely than people who will subscribe. In addition, the prediction accuracy of selected model (= 0.748) is slightly higher than the one of full model (= 0.747). However, the full model and selected model performs similarly. It is better to get a less complicated model with less predictors since it will lower the cost of collecting data. In this case, the selected model with minimum AIC value is suitable to predict the members of long-term deposit.

3.3 Model Diagnostics

In logistic regression, there are some basic assumptions of modeling: (a) Linearity assumption: there is linear relationship between continuous predictor variables and the logit of the outcome; (b) There is no high influential point or extreme value; (c) The explanatory variables are not linear combinations of each other, which means there is no multicollinearity in the model; (d) Logistic regression requires each observation to be independent.

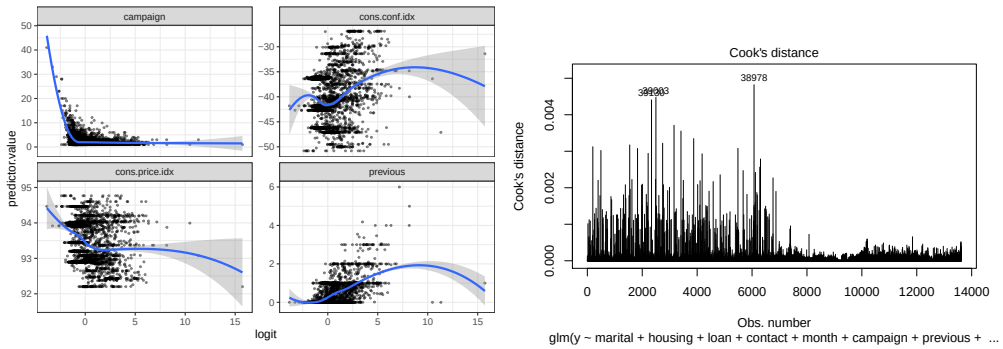

Figure 2: Smoothed scatter Plots(Left) and Cook Distance Plot(Right)

The smoothed scatter plots (*Figure 2 Left*) show that variables consumer price index(cons.price.idx), consumer confidence index(cons.conf.idx) and number of contact(Previous) before the campaign are all slightly linearly associated with the outcome of whether the member will get a long-term deposit in logit scale. However, the variable number of contact during the campaign(Campaign) is not linear and might need some transformations.

Influential values are extreme individual data points that can alter the quality of the logistic regression model. The most extreme values in the data can be examined by visualizing the Cook's distance values. From the *Figure 2 Right*, most of points have low level of Cook distance so there is no high-influential point in the model.

Multicollinearity corresponds to a situation where the data contain highly correlated predictor variables. As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity. In our selected logistic regression model, there is no collinearity since all variables have a value of VIF well below 5 (See *Table 2*).

Table 2: Variance Inflation Factor

marital	housing	loan	contact	month
1.01	1	2.94	2.07	1.81
campaign	previous	poutcome	cons.price.idx	cons.conf.idx
1.01	1.83	1.03	2.16	1.76

For independence assumption of logistic regression, the data are not generated from any dependent samples design. In this case, each observation cannot be affected by others. In conclusion, the independence assumption can be satisfied.

4. Random Forest Prediction

4.1 The basic of classification tree

Since our predictive variables include both qualitative and quantitative variables, in addition to logistic regression, an effective method to predict classification variables is to use classification trees. For a classification tree, we predict that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs. Classification trees have many advantages over traditional methods, for example, trees are easy to explain to people and trees can easily handle qualitative predictors without the need to create dummy variables. Unfortunately, trees generally do not have the same level of predictive accuracy as other regression approach because of high variance. However, we can improve predictive performance of trees by using methods like random forests.

Table 3: The confusion matrix for classification tree

	Actual NO	Actual YES
Predict NO	1483	590
Predict YES	220	1113

Note: Sensitivity=0.83;Specifity=0.72.

From a tree diagram (*Appendix 4*)obtained without using random forest method. We can see that variables nr.employed is a very strong predictor in this model. *Table 3* is the confusion matrix for prediction. Because it only use one sample to build a model, therefore the variance of the model is large and the prediction accuracy is only 76.22%.

4.2 Random forest

To reduce the high variance of the decision tree and improve predictive accuracy. We can randomly take multiple training sets from the population, build a separate prediction model using each training set, and average the prediction result. In order to decorrelate the trees, when building these classification trees, each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors. Usually m equals the square root of p. In this way, the correlation can be decorrelated because only a subset of the predictive variables is considered at a time, which can effectively overcome the influence of strong predictive variables and reduce the influence of correlation between variables, so this process can be considered as decorrelating. Therefore, we don't need to worry about the correlation between predictors. When we used the random forest method, we only removed three variables (duration, default, pdays) for reasons shown above, and the other variables would be retained in order to get a good predictive model.

Table 4: The confusion matrix for random forest

	Actual NO	Actual YES
Predict NO	1576	362
Predict YES	127	1341

Note: Sensitivity=0.91;Specifity=0.81.

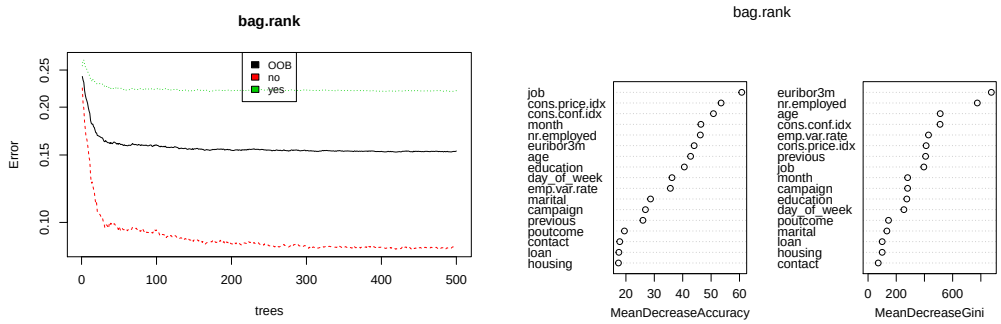

Figure 3: Random forest model outcome(Left) and A variable importance plot(Right)

In the *Figure 3 Left*, the black line is OOB (out of bag) error rate which represents the test error of random forest model. The red line is the test error rate for the client subscribed a term deposit. The green line is the test error rate for client did not subscribe a term deposit. In this case, we can see that as the number of trees increases, the error rate decreases. when the number of trees increase to 100, the error rate remaines almost constant. From the *Figure 3 Right*, we can see the variable importances in the data set (mean decrease in accuracy for each variable in the left and mean decrease in Gini index for each variable in the right). The variables with largest mean decrease in accuracy are job, cons.price.idx and cons.conf.idx. The variables with largest mean decrease in Gini index are euribor3m, nr.employed and age. *Table 4* is a confusing matrix using the training dataset fitting model to predict the results of the test dataset. We can see that the accuracy is about 85.64%.

5. Model Comparison

To compare the two models, we chose the ROC curve. The ROC curve is a popular graphic because it can simultaneously displaying two types of errors(both false positive rate and true positive rate) for all possible thresholds. The overall performance of a classifier is given by the area under the (ROC) curve (AUC). An ideal ROC curve will hug the top left corner, so the larger the AUC the better the classifier.

As can be seen from the *Figure 4*, the AUC of the random forest model is 0.912, which is very close to the maximum, while the AUC of the logistic regression model is 0.799, which is smaller than that of the former.Therefore, we can conclude that the method of random forest has better classification performance and is more suitable for prediction model.

6. Conclusion and Discussion

By the comparison of the results of AUC and accuracy rate, the random forest model has a better predictive ability than the logistic regression model in this project. In general, people use a logistic regression model as a standard method for binary variables when dealing with low-dimensional data (Ranganathan et al, 2017). However, random forest models can handle high-dimensional data because it can reduce variance compared

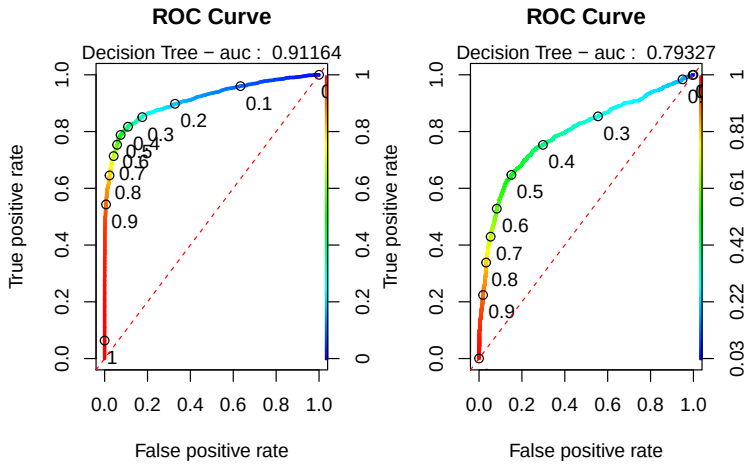

Figure 4: Model Comparison(Left: Random Forest; Right: Logistic Regression)

to the single decision trees through the aggregation of a large number of decision trees. Our data set is a high dimensional data set, therefore it's likely to have very strong collinearity and/or have nonlinear relationship between predictor and response variables. In this case the tree-based approach is a little bit better than the traditional way with logistic regression. This may be the reason why the random forest outperforms the logistic regression.

However, there are several limitations to this project. First, the data is obtained only from a Portuguese retail bank. Therefore, the results or conclusions can be only generalized to this type of bank. Second, the logistic regression model is effective to a small space of variables that indicates there may exist ill-fitting with multiple variables in this model (Ranganathan et al, 2017). Third, we only compare the logistic regression model with the random forest model due to the time limit. It is better to compare more classification models to improve fitness, accuracy, and prediction.

Reference

Abu-Srhan, A., & Al-Sayyed, R. (2019). Visualization and Analysis in Bank Direct Marketing Prediction. Int. J. Adv. Comput. Sci. Appl, 10(7), 651-657.

Asare-Frempong, J., & Jayabalan, M. (2017, September). Predicting customer response to bank direct telemarketing campaign. In 2017 International Conference on Engineering Technology and Technopreneurship (ICE2T) (pp. 1-4). IEEE.

Mentzer, J. T., Stank, T. P., & Esper, T. L. (2008). Supply chain management and its relationship to logistics, marketing, production, and operations management. Journal of business logistics, 29(1), 31-46.

Ranganathan, P., Pramesh, C. S., & Aggarwal, R. (2017). Common pitfalls in statistical analysis: Logistic regression. Perspectives in clinical research, 8(3), 148–151.

Appendix 1: Summary statistics for original data

X

	age	job	marital	education	default
	Min. :17.00	admin. :10422	divorced: 4612	university.degree :12168	no :32588
	1st Qu.:32.00	blue-collar: 9254	married :24928	high.school : 9515	unknown: 8597
	Median :38.00	technician : 6743	single :11568	basic.9y : 6045	yes : 3
	Mean :40.02	services : 3969	unknown : 80	professional.course: 5243	
	3rd Qu.:47.00	management : 2924		basic.4y : 4176	
	Max. :98.00	retired : 1720		basic.6y : 2292	
		(Other) : 6156		(Other) : 1749	

	housing	loan	contact	month	day_of_week
	no :18622	no :33950	cellular :26144	may :13769	fri:7827
	unknown: 990	unknown: 990	telephone:15044	jul : 7174	mon:8514
	yes :21576	yes : 6248		aug : 6178	thu:8623
				jun : 5318	tue:8090
				nov : 4101	wed:8134
				apr : 2632	
				(Other): 2016	

	duration	campaign	pdays	previous	poutcome
	Min. : 0.0	Min. : 1.000	Min. : 0.0	Min. :0.000	failure : 4252
	1st Qu.: 102.0	1st Qu.: 1.000	1st Qu.:999.0	1st Qu.:0.000	nonexistent:35563
	Median : 180.0	Median : 2.000	Median :999.0	Median :0.000	success : 1373
	Mean : 258.3	Mean : 2.568	Mean :962.5	Mean :0.173	
	3rd Qu.: 319.0	3rd Qu.: 3.000	3rd Qu.:999.0	3rd Qu.:0.000	
	Max. :4918.0	Max. :56.000	Max. :999.0	Max. :7.000	

	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
	Min. :-3.40000	Min. :92.20	Min. :-50.8	Min. :0.634	Min. :4964
	1st Qu.:-1.80000	1st Qu.:93.08	1st Qu.:-42.7	1st Qu.:1.344	1st Qu.:5099
	Median : 1.10000	Median :93.75	Median :-41.8	Median :4.857	Median :5191
	Mean : 0.08189	Mean :93.58	Mean :-40.5	Mean :3.621	Mean :5167
	3rd Qu.: 1.40000	3rd Qu.:93.99	3rd Qu.:-36.4	3rd Qu.:4.961	3rd Qu.:5228
	Max. : 1.40000	Max. :94.77	Max. :-26.9	Max. :5.045	Max. :5228

X

Appendix 2: Correlation matrix of numeric variables

	age	campaign	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
age	1.00	0.01	0.03	0.00	0.00	0.13	0.01	-0.02
campaign	0.01	1.00	-0.08	0.15	0.13	-0.01	0.13	0.14
previous	0.03	-0.08	1.00	-0.42	-0.21	-0.06	-0.45	-0.49
emp.var.rate	0.00	0.15	-0.42	1.00	0.78	0.21	0.97	0.91
cons.price.idx	0.00	0.13	-0.21	0.78	1.00	0.07	0.69	0.52
cons.conf.idx	0.13	-0.01	-0.06	0.21	0.07	1.00	0.29	0.12
euribor3m	0.01	0.13	-0.45	0.97	0.69	0.29	1.00	0.95
nr.employed	-0.02	0.14	-0.49	0.91	0.52	0.12	0.95	1.00

Appendix 3: Fitting Results of Coefficients of Logistic Regression

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	46.9967018	4.2637243	11.022453	0.0000000
maritalmarried	-0.2125072	0.0633317	-3.355461	0.0007923
maritalsingle	0.1680207	0.0671820	2.500978	0.0123851
housingyes	-0.1009724	0.0402284	-2.509975	0.0120740
loanyes	0.8178540	0.0481906	16.971219	0.0000000
contacttelephone	-0.1673440	0.0572107	-2.925045	0.0034441
monthaug	-1.3299479	0.1029394	-12.919721	0.0000000
monthdec	0.5434712	0.2973692	1.827597	0.0676100
monthjul	-0.6511487	0.0918145	-7.092000	0.0000000
monthjun	-0.2568534	0.0935764	-2.744853	0.0060538
monthmar	1.2051582	0.1708694	7.053094	0.0000000
monthmay	-1.1392551	0.0791894	-14.386458	0.0000000
monthnov	-1.1830905	0.0953281	-12.410724	0.0000000
monthoct	0.9117790	0.1668767	5.463790	0.0000000
monthsep	0.2616095	0.1718559	1.522261	0.1279437
campaign	-0.0695855	0.0099341	-7.004711	0.0000000
previous	1.1000381	0.0666803	16.497196	0.0000000
poutcomenonexistent	0.4311689	0.0838933	5.139494	0.0000003
poutcomesuccess	2.3380492	0.1240869	18.842026	0.0000000
cons.price.idx	-0.4797032	0.0459695	-10.435258	0.0000000
cons.conf.idx	0.0519643	0.0056474	9.201532	0.0000000

Appendix 4: Tree diagram