Predictive Modeling on Bank Marketing for Term Deposit

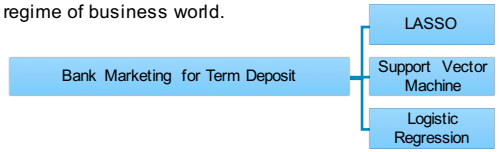
Menglan Jiang, Xinyuan Zhang, Mengrun Li, & Yaqian Cheng

Abstract

Bank Marketing has been a important topic in recent banking strategy sets. In order to explore factors which will have significant impacts on the decision to make term deposit, we employ a bunch of statistical techniques to identify variables of interest. More importantly, we aim to predict the behavior of whether to make term deposit based on personal information of potential customers. We use three techniques, including Logistic Regression, LASSO and SVM. The candidate factors varies from demographic information, financial status to credit history. In specific, we have 16 variables of interest, including age, maritale, education, default, balance, housing, loan, contact, day, month, duration, campaign, pdays, previous, poutcome. Based on the results, we conclude that the SVM is the best method for predicting whether the clients pool will make term deposit in the bank within the research. In the meantime, the other two methods also make significant estimation of the term deposit probability. In LASSO, we find 27 significant covariates out of 42.

Motivation

In the paper, we aim to explore the marketing strategy for attracting term deposit from potential clients. The major task is to identify the targeting customers who have the desire to invest in term deposit. Believing in the power of the different statistical techniques, we want to explore the power of the classification and regression methods in the regime of business world.



LASSO

The idea behind the lasso procedure is to minimize

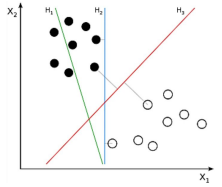
$$\hat{\beta} := \arg\min_{\beta \in R^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \|\beta\|_1$$

where $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$.

The regression coefficient β for the lasso with regularization parameter λ is a p-dimensional vector with many of its values set to zero for larger values of λ. The idea behind the regularization path is to help select how many variables to keep in the model. In the ridge model it is hard to interpret a regularization parameter as coefficients are not sent to zero and the changes are slow. This is somewhat mitigated in the lasso model.

Support Vector Machine

Support vector machines are highly used in classification problems. A support vector machine constructs a hyperplane or set of hyperplanes in a high-dimensional space. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin).



In this figure, H_1 does not separate the classes. H_2 does, but only with a small margin. H_3 is the optimal hyperplane since it separates them with the maximum margin.

Logistic Regression

We have a binary output variable Y, and we want to model the conditional probability Pr(Y|X) as a function of x; any unknown parameters in the function are to be estimated by maximum likelihood.

The main idea is to model the log likelihood ratio by the function f(x)

$$f(x) = \log\left(\frac{P(y=1|x)}{P(y=-1|x)}\right)$$

Which implies $P(y=\pm 1|x) = \frac{1}{1+\exp(yf(x))}$

Given data D and a class of functions f the MLE is

$$f^*_{MLE} := \arg\max_{f \in H} \left[\prod_{i=1}^{n} \frac{1}{1+\exp(y_i f(x_i))}\right]$$

Results

Prediction results vs. Real results

	0	1
no	38940	982
yes	3456	1833

Table1. results of logistic regression without penalty

	0	1
no	39085	837
yes	3612	1677

Table2. results of logistic regression with lasso

	0	1
no	39304	618
yes	3130	2159

Table3. results of SVM

	Logistic regression without penalty	Logistic regression with Lasso	SVM
Prediction Rate	90.18%	90.16%	91.71%

Table4. Prediction rate of three models

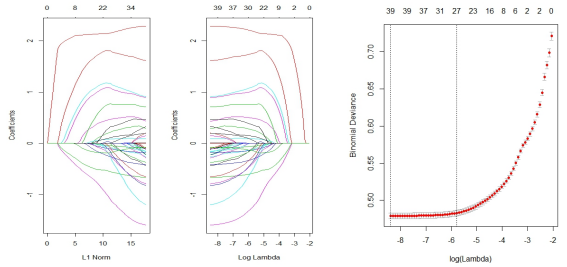
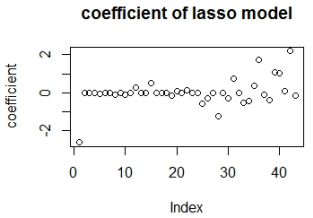


Figure1. trace plot of lasso

Figure 2. lasso select lambda



This session presents the result for the three methods researched in the paper. The tables include the results for predicting whether the customers will make term deposit based on personal information collected.

Conclusion

- The SVM presents the best estimation for whether making the term deposit based on the personal information.
- The LASSO and Logistic also obtain great estimations with relatively less precise results.
- In LASSO, we find out that among 42 candidate variables, approximately 27 of them make significant contribution to the decision.

Works Cited

- [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014
- https://en.wikipedia.org/wiki/Support_vector_machine#/media/File:Svm_separating_hyperplanes_(SVG).svg