COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Prediction of the Success of Bank Telemarketing and Model Comparison

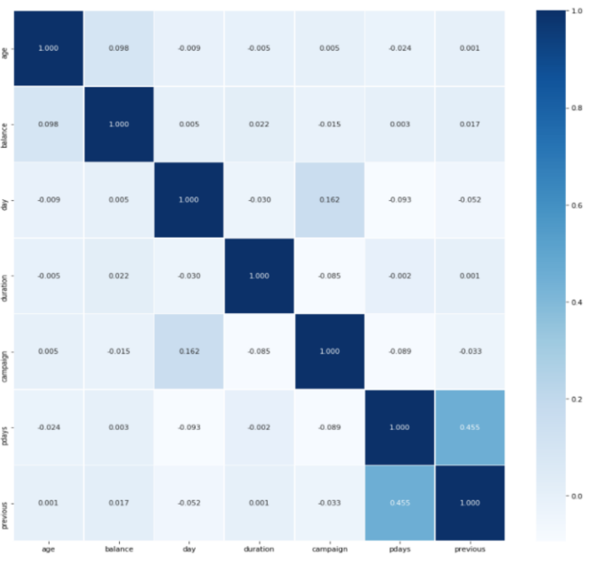
Multivariate
Spring'19

Introduction

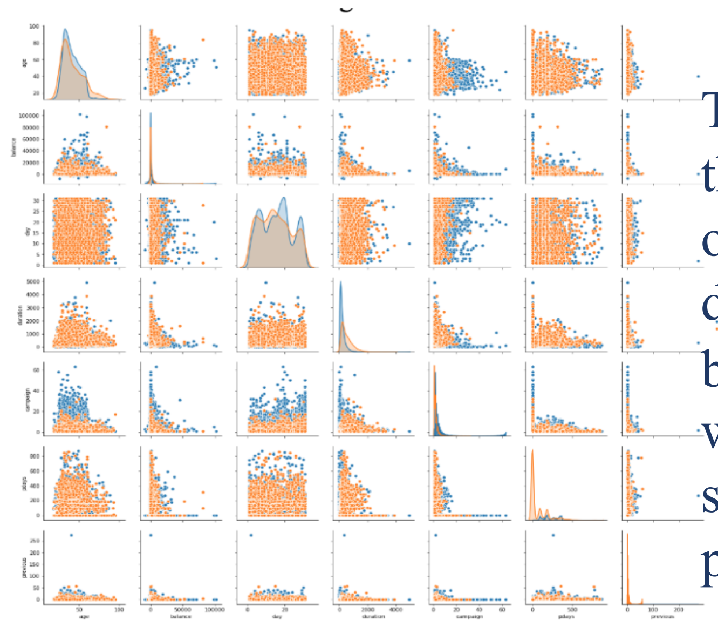
Marketing selling constitute a typical strategy to enhance business. Companies not only use direct marketing when targeting segments of customers by contacting them to meet a specific goal, but also through various indirect channels, telephone (fixed-line or mobile) being one of the most widely used. We compare four Machine Learning models (Logistic Regression, Random Forest, Neural Network and SVM) using the metric of AUC-ROC and classification metric. We find that the best model (Random Forest) could most accurately predict the possibility of the customers buying the products of the bank.

Data

Using data driven methods like the machine learning techniques, it is more efficient and effective for the companies to grasp the information of their customers, thus improving their selling process. We analyze a recent and large dataset (45,211 records) from a Portuguese bank. The data were collected from 2008 to 2013, thus including the effects of the global financial crisis that peaked in 2008。The data include two attributes of the sample. The first is the bank client data including age, married, education, current balance, whether has loan or not, etc. The second is some auxiliary information like the contact duration, number of contacts performed with customers, etc.



The figure on the right is the heat map of the correlation between the variable, which indicates that there does exist a not strong correlation between the variable except the correlation between p-days and previous, which is 0.455.



The figure on the left is the pair wise scatter plot of the variables. The blue dots is the sample of not buying the product , whereas the orange is the sample of buying the products.

Models

Logistic Regression

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). A multiple linear regression function defined as:

$$\operatorname{logit}(\mathrm{p}) = \log\left(\frac{p(y=1)}{1-(p=1)}\right) = \sum_{i=1}^{p} \beta_i x_i$$

Lasso/ Ridge Regression

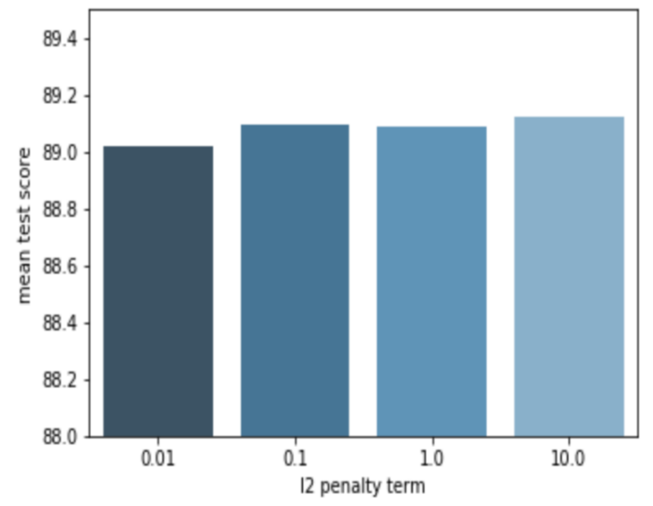
Ridge and Lasso regression are some of the simple techniques to reduce model complexity and prevent overfitting which may result from simple linear regression. The cost function for Lasso (least absolute shrinkage and selection operator) regression can be written as:

$$\sum_{i=1}^{M}(y_i-\hat{y}_i)^2 = \sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p} w_j \times x_{ij}\right)^2 + \lambda\sum_{j=0}^{p}|w_j|$$

Support Vector Machine

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. The objective of SVM is to find a hyperplane which maximize the margin separating the two categories of data. We use a kernel method SVM and we choose the Gaussian kernel:

$$k_{\mathrm{RBF}}\left(\mathbf{x},\mathbf{x}'\right) := \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{2\sigma^2}\right)$$

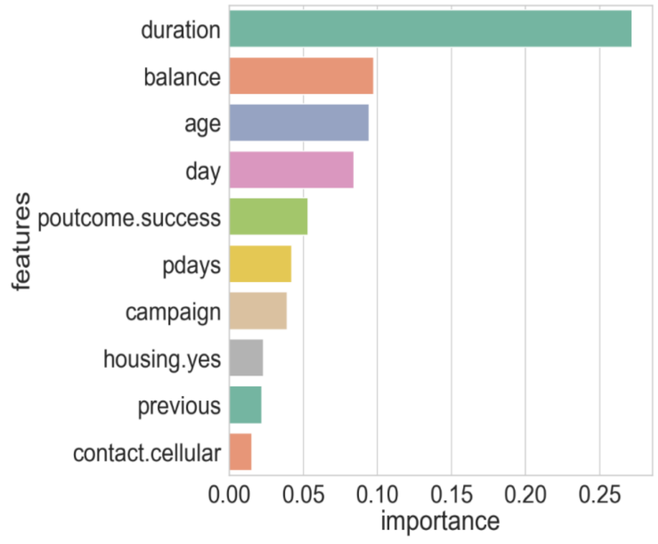


The left figure shows the cross validation on the penalty term of SVM. 0.1 turns out to be the best choice and the accuracy is about 89%.

Random Forest

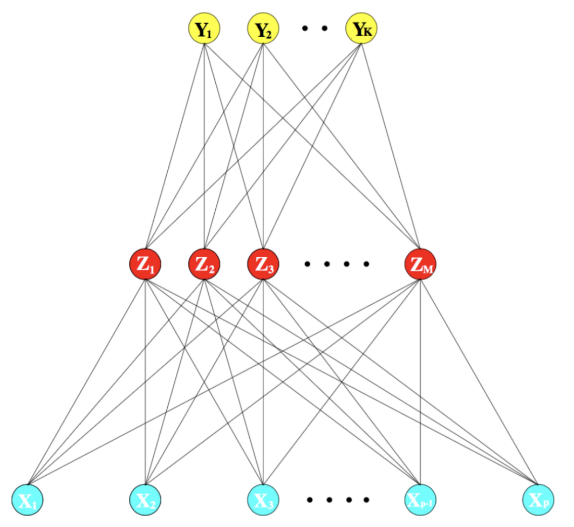
Random forests, or random decision forests are an ensemble learning method for classification and regression. It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

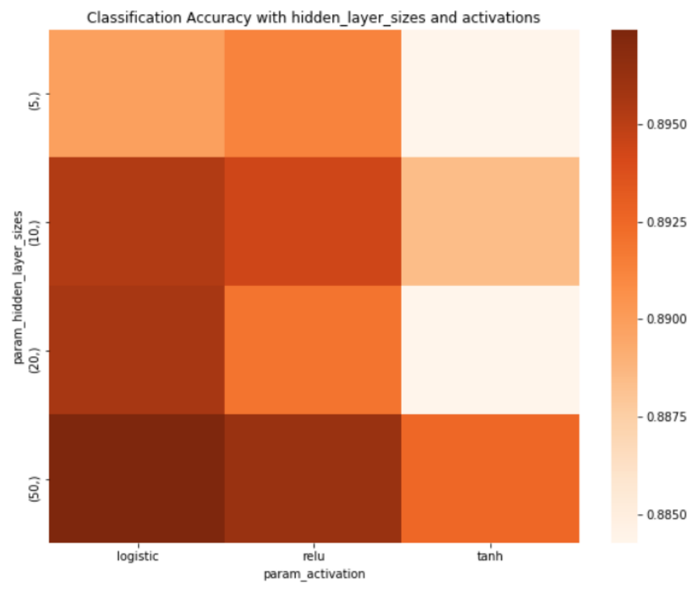
The figure shows the feature importance generated by random forest. It is shown that duration plays an important role, followed by balance, age and day. We also do the cross validation using the grid search method.



Neural Network

A neural network is a supervised learning method which can be applied to both regression and classification problems. The idea of neural network is to extract linear combinations of the inputs as derived features, and then model the target as a nonlinear function of these features. The figure in the left bottom shows the architecture of Neural Network. There are three layers: the input layer, hidden layer and the output layer. We use the cross validation(right bottom figure) to check the units of the hidden layer, as well as the activation function. Sigmoid function turns out to be the best choice.



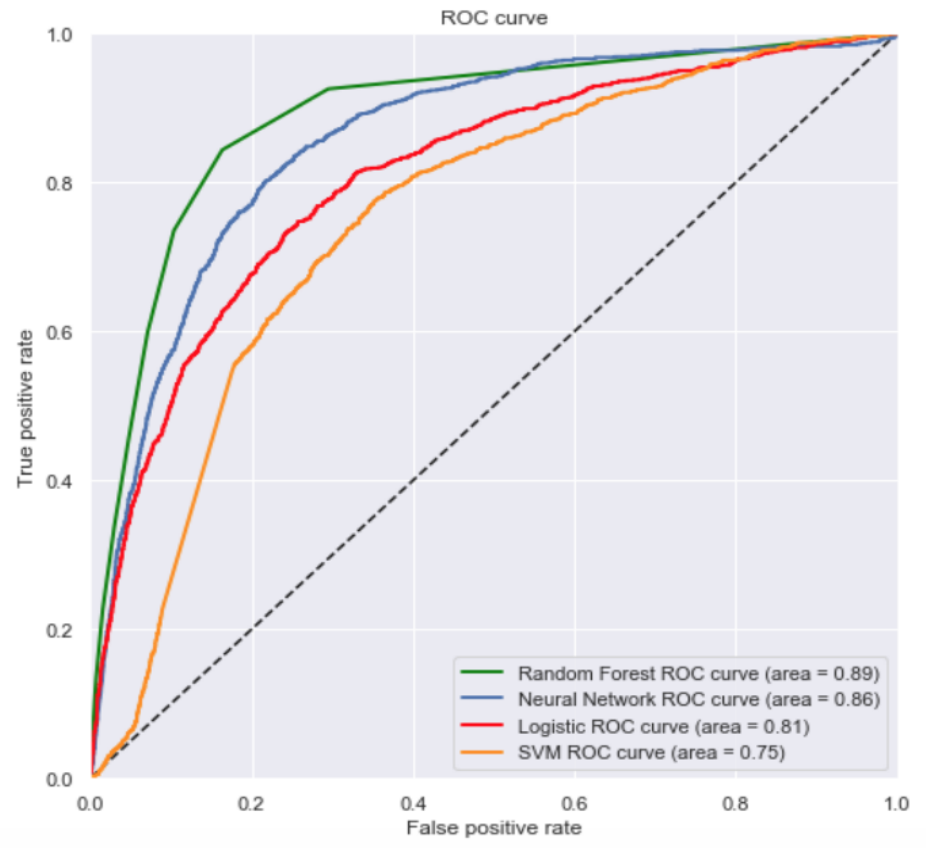


Result

These models were compared using two metrics, area of the receiver operating characteristic curve (AUC) and the prediction accuracy. The best results were obtained by the RFs, which resulted in an ROC-AUC of 0.89, and Classification Accuracy of 0.92. Such AUC corresponds to a very good discrimination.

Comparison of models.(bold denotes the best value).

Metric	Logistic Regression	Random Forest	SVM	Neural Network
ROC-AUC	0.81	**0.89**	0.75	0.86
Classification Accuracy	0.85	**0.92**	0.88	0.90



Conclusion

In the logistic regression model, we can conclude that the most important features are married status, no housing loan, no personal loan, cellular and telephone contact status, last contact day, last contact month, contact duration, number of contacts and the success outcome status of the last campaign. For example, married status has a negative effect to the model.

As an interesting outcome from Random Forest model, the duration of the call, which highly affects the probability of a success contact or the amount that is deposited in the bank, is outstandingly significant. Also the average yearly balance, and age variables.

Also, four Machine Learning models were compared: logistic regression (LR), random forest (RFs), neural networks (NNs) and support vector machines (SVMs). These models were compared using two metrics, area of the receiver operating characteristic curve (AUC) and the prediction accuracy. The best results were obtained by the RFs, which resulted in an ROC-AUC of 0.89, and Classification Accuracy of 0.92. Such AUC corresponds to a very good discrimination.

Future work

In future work, the dataset may provide history of telemarketing behavior for cases when clients have previously been contacted. Such information could be used to enrich the dataset (e.g., computing frequency and monetary features) and possibly provide new valuable knowledge to improve model accuracy. Also it would be interesting to consider the possibility of splitting the sample according to two sub-periods of time within the range of 2008–2012, which would allow us to analyze impact of hard-hit recession versus low recovery.

Reference

Saptashwa. (Sep 26, 2018). Ridge and Lasso Regression: A Complete Guide with Python Scikit-Learn.

Ian H.Witten, Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd edition Morgan Kaufmann, 2005.

Producer

Yuting He yh3054
Zhiying Li zl2697
Xiaoming Chen xc2479