Three data mining models to predict bank telemarketing

Wei Jin\textsuperscript{1,\ast} and Yingying He\textsuperscript{2}
\textsuperscript{1}China Key Lab of Pulse Power Laser Technology, Electronic Countermeasures Institute, National University of Defense Technology, Hefei, China
\textsuperscript{2}Qianshan County Hospital, Anqing, China
\\textsuperscript{\ast}Corresponding author e-mail: kingvee@163.com

Abstract. The bank telemarketing dataset was collected by a Portuguese retail bank when selling a bank long-term deposits product. In order to predict the success of bank telemarketing, a more accurate approach was proposed. Firstly, the dataset was analyzed and processed. The features value of the dataset could be divided into two kinds: continuous variable and categorical variable. The two kinds of variable had been processed with different methods. The continuous variables were transformed to discrete features and then were normalized. The categorical variables were coded with two different methods according to whether they were belong to the ordered categories. The missing value of features had to be processed with two different methods depending on the examples number of the missing value. Secondly, the three data mining models which were support vector machine, neural network and decision trees were adopted to predict the result of the processed bank telemarketing dataset. The classification and regression trees algorithm was adopted in the decision trees model. The BFGS algorithm had been used when the neural network model was programming. The simple sequential minimal optimization (SMO) algorithm was adopted in the support vector machine model. All the three models were executed using python 2.7 and conducted in a Windows7 PC, with an Intel Cerelon G1630 2.80 GHz processor. The results of three models were analyzed and compared. The decision trees presented the best results and cost the least time, and its accuracy, precision, recall value and AUC area of the were all equal to 1.0. It showed that decision trees could predict 100% correct rate. The accuracy, recall value and AUC area of the neural network were better than those obtained by the support vector machine (improvement of 1.5 pp, 2.7 pp and 5 pp respectively), but the neural network cost almost 4.4 times more time than the support vector machine. It showed that not only choosing the data mining models but also processing datasets in advance played a key role in obtaining better prediction results when datasets needed classifying.

1. Introduction
The bank marketing dataset was from the UCI dataset\textsuperscript{[1]}. It was collected by a Portuguese retail bank from 2008 to 2013 when selling bank long-term deposits by telemarketing. All data was in the additional-full file. There were more than 40 thousands examples. In this paper, we used the bank-additional file which included 10\% of the examples (4119) randomly selected from the bank-additional-full file. The dataset included 20 features which related with bank client, product and social-economic attributes\textsuperscript{[2]}. The dataset had one binary classification result which shows whether the customer would subscribe a bank term deposit.

Data mining (DM) could play a key role in personal and intelligent decision support systems, allowing the semi-automatic extraction of explanatory and predictive knowledge from raw data. In particular, classification that learned an unknown underlying function that mapped several input variables with one labeled output target was the most common DM task. There were several familiar
classification models, such as the support vector machine (SVM) [3], neural network (NN) and decision trees (DT). DT model can be easily understood by person and can be fitted to predict good classification results. These are the advantage of DT [4]. NN and SVM are more flexible when compared with classical statistical modeling or even DT, presenting learning capabilities that range from linear to complex nonlinear mappings. However, the two models can be hard understood.

In this paper, the SVM, NN and DT models were used to classify bank telemarketing dataset. The calculation cost, accuracy and receiver operating characteristic (ROC) of the three models were compared. However, to obtain more accurate results, the dataset should be processed in advance before it was used by the three DM models.

2. Processing the dataset
The dataset included two kinds of features: continuous features and categorical features. Some categorical features missed their value in some examples, and their values were displayed as ‘unknown’ in the data file. Therefore, the dataset should be processed to well fit DM models.

2.1. Processing the numerical features
The continuous features were processed by two steps. Firstly, continuous features were transformed to discrete features in order to obtain more stable prediction results of the DM models [5]. The equal distance discrete method was adopted to get $n$ bins that their distance was the same. The continuous value was matched in some bins, and correspondingly coded by discrete value. Secondly, the discrete value was normalized to a zero mean and one standard deviation [6].

2.2. Processing the missing value
The missing value was processed in two routes. In first route, when the number examples of missing value in a feature were less than 100, the corresponding examples would be deleted and not be used in DM models anymore. In second route, when the number examples of missing value in a feature were equal to or more than 100, the missing value should be predicted by using the whole dataset. The feature of which missing value should be predicted was considered as the label, all the other features and the classification result of the dataset were all considered as the features. The random forest model was adopted to give the prediction value of the missing value.

2.3. Processing categorical features
The categorical features should be coded into discrete value. The coding methods might be three according to the value of categorical features.

2.3.1. The binary categorical features coding.
There were 2 categories value in some features, such as ‘default’, ‘housing’ and ‘loan’. The two different categorical value could be code with 0 and 1.

2.3.2. The ordered categorical features coding.
The ‘education’ feature was belong to an ordered categorical feature. The feature had 7 categories which were ‘illiterate’, ‘basic.4y’, ‘basic.6y’, ‘basic.9y’, ‘high.school’, ‘professional.course’, ‘university.degree’ according to their influence. The variables could be coded with 1, 2, 3, ..., 6 and 7 in turn.

2.3.3. The unordered categorical features coding.
When the influence of categories were not different each other, the corresponding features would be belonged to unordered categorical features, such as ‘job’, ‘marital’, ‘contact’, ‘month’ and ‘day_of_week’. If an unordered categorical features had $n$ categories, $n - 1$ dummy variables were required. For example, the ‘marital’ feature had 3 categories including ‘divorced’, ‘married and single’, so 2 dummy variables which were $v1$ and $v2$ were needed. Table 1 exhibited the coding method.
Table 1. The unordered categorical features coding.

| marital     | V1 | V2 |
|-------------|----|----|
| divorced    | 0  | 0  |
| married     | 1  | 0  |
| single      | 0  | 1  |

3. Experiments and results

3.1. Modeling

We used python 2.7 language to program the models code. All the code was run in a Windows7 PC, with an Intel Celeron G1630 2.80 GHz processor. The dataset was further divided randomly. 3/4 dataset was used as training dataset, and 3/4 dataset was used as validation sets. Each DM model was executed 20 runs. The results of each model were the mean of the 20 runs results.

When the DT model was used, the classification and regression trees (CART) algorithm was adopted. One type of tree, the model tree was built when programming. The variable $tolS$ was a tolerance on the error reduction, and $tolN$ was the minimum data instances to include in a split. $tolS$ and $tolN$ were set to 1 and 20 respectively.

The BFGS algorithm would be used when the NN model was programming. Generally, the counts of NN layers are more, the results are more accurate. To consider the cost time, we set layers as 3. The hidden nodes variable $H$ was set to half of the terminal features number. The iterative step and iterative times were set to 0.01 and 200000 respectively.

The simple sequential minimal optimization (SMO) algorithm was adopted when the SVM model was programming. The constant variable $C$, the tolerance variable $toller$ and the maximum number of iterations before quitting variable $maxIter$ were set to 200, 0.01 and 30000 respectively.

3.2. Results

In the two-class problem, if we correctly classified something as positive, it was called a True Positive (TP), and it was called a True Negative (TN) when we properly classified the negative class. Of course, there were the other two possible cases False Negative (FN) and False Positive (FP). The accuracy $a$, precision $p$ and recall $rec$ were calculated with formula (1), (2) and (3). In formula (1), $num$ was represented as the total examples number.

\[
a = \frac{(TP + TN)}{num} \quad (1)
\]

\[
p = \frac{TP}{(TP + FP)} \quad (2)
\]

\[
rec = \frac{TP}{(TP + FN)} \quad (3)
\]

The results of the three DM models were shown in table 2. Although the DT model cost the least time, its accuracy, precision and recall were the best of all the three models. Its accuracy, precision and recall value were all equal to 1.0. It showed that the DT models could predict 100% accurate results of the bank telemarketing. The NN model cost almost 4.4 times more than the SVM model, but its accuracy and recall outperformed the SVM model (improvement of 1.5 pp and 5 pp respectively).

Table 2. The results of three DM models.

| DM models | Cost time $t$ | Accuracy $a$ | Precision $p$ | recall $rec$ |
|-----------|---------------|--------------|---------------|--------------|
| DT        | 4s            | 1.0          | 1.0           | 1.0          |
| NN        | 390s          | 0.993        | 1.0           | 0.939        |
| SVM       | 89s           | 0.978        | 1.0           | 0.889        |

The AUC curves of the three model was shown in figure1. Apparently, the AUC curve plotted using the DT model prediction results was a straight line, shown as figure 1 (a). The AUC area was equal to 1.0. The AUC curves of the NN and SVM model were shown figure 1 (b) and (c) respectively. The AUC area of the NN model was 0.975 better than the area of which the SVM model was 0.948.

Compared to the results of [2], the three DM models predicted more accurately. The main reason was that the bank telemarketing dataset had been processed before it was used by the three DM models.
4. Conclusion
In this study, we proposed how to predict the success of bank telemarketing more accurately. Firstly, the telemarketing dataset was processed to fit the DM models well. Secondly, three DM models were adopted to classify the processed telemarketing dataset. The accuracy, precision and recall of the three models were higher than those of previous work obtained by other researchers.

References
[1] Information on http://archive.ics.uci.edu/ml/datasets/Bank+Marketing.
[2] S. Moro, P. Cortez and P. Rita, A data-driven approach to predict the success of bank telemarketing, Decision Support Systems, 2014 62(1246) pp. 22-31.
[3] P. H. Chen, C. J. Lin, and B. Schölkopf, A tutorial on $\nu$-support vector machines, Appl. Stoch. Models. Bus. Ind., 2005 21(2) pp. 111-136.
[4] David L Olson, Dursun Delen, and Yanyan Meng, Comparative analysis of data mining methods for bankruptcy prediction. Decis, Support Syst., 2012 52(2) pp. 464-473.
[5] Classification problem of machine learning in action(based on UCI bank marketing dataset), Information on http://www.cnblogs.com/llhthinker/p/7101572.html.
[6] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed., Springer-Verlag, New York, 2008.