Using Data Mining Techniques to Improve Efficiency for Bank Direct Marketing Compaigns

TIANYU ZHUANG, A53101494
YIHONG ZHANG, A53097346
SIWEN YAN, A53093655
XIYUN LIU, A53099348

ACM Reference format:
Tianyu Zhuang, Yihong Zhang, Siwen Yan, and Xiyun Liu. 2016. Using Data Mining Techniques to Improve Efficiency for Bank Direct Marketing Compaigns. 1, 1, Article 1 (January 2016), 5 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

There are two main approaches for companies to promote their products or services: through mass campaigns, which target the general public population, and directed campaign, which targets only a specific group of people. Directed marketing focuses on targets that assumed to be keener to that specific product and thus enhancing the efficiency of campaigns. Our goal is to improve the rate of success (clients subscribing the deposit), which could further improve the bank's profit during a campaign: do less contacts while keeping an approximately number of successes.

In this report, we apply some data mining approaches to bank direct marketing campaigns. In particular, we used a direct marketing dataset of a Portuguese bank. Three models, linear model using stochastic gradient descent, Support Vector Machine (SVM), Random Forest are used to predict whether a client will subscribe the deposit. The best models, materialized by SVM and Random Forest achieved high predictive performances.

2 DATASET ANALYSIS

The bank marketing dataset from UCI Machine Learning Repository contains 17 direct marketing campaigns of a Portuguese banking institution occurred between May 2008 and November 2010. The marketing campaigns used its own contact-center, through phone calls, getting client information as attribute information. Each contact contains 17 attributes, including three parts. One is the bank client data, which includes personal information of clients such as age, job, marital and education. The second part is the information of the last contact of the current campaign. The rest part includes the number of contacts performed (before) during this campaign for this client, the number of days that passed by after the client was last contacted from a previous campaign and outcome of the previous marketing campaign. The classification goal is to predict whether the client subscribed a term deposit during the campaigns or not.

The raw dataset has 79354 contacts, however, 34143 of them have missing values, which is difficult for a lot of classification models (e.g. SVM) to train, therefore, the contributor discards the examples that contained missing values[1], leading to a dataset with 45211 samples, among which 5289 subscribes a term deposit categorized as an output of 'yes'.

The input variables related to bank client data are:

- age (numeric)
- job : type of job
- marital : marital status (categorical: "married", "divorced", "single")
- education (categorical: "unknown", "secondary", "primary", "tertiary")
- default: has credit in default? (binary: "yes","no")
- balance: average yearly balance, in Euro (numeric)
- housing: has housing loan? (binary: "yes","no")
- loan: has personal loan? (binary: "yes","no")

The input variables related to the last contact of the current campaign are:

- contact: contact communication type (categorical: "unknown","telephone","cellular")
- day: last contact day of the month (numeric)
- month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
- duration: last contact duration, in seconds (numeric)

The other input variables are:

- campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
- previous: number of contacts performed before this campaign and for this client (numeric)
- poutcome: outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")

The output variable (desired target) is:

- y: has the client subscribed a term deposit? (binary: "yes","no")

Digging into the dataset, we could find some interesting observations:

(1) The data in the raw dataset is in time order, from May 2008 to November 2010. Thus, we should shuffle the data before training process.

(2) The number of instances is 45211, however, only 11.698% successfully subscribed the term deposit. This fact gives us some insight that while dealing with unbalanced dataset, we need to focus on balanced error rate rather than MAE or overall accuracy.
(3) Some features (`default, marital, education`) are less related to the result of the campaign than the others, so we could discard them. This is discovered by checking the rate of success for each option. From figure 1, we could see that rate of success for single, married and divorced is all most the same. Similarly, some features do show the preference of a particular client, which is proved in figure 2.

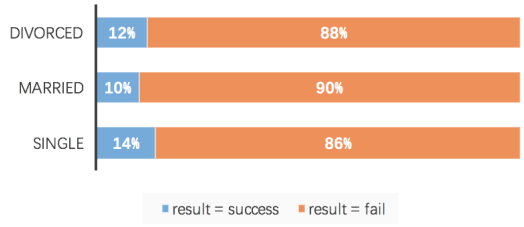

Fig. 1. Rate of success over different marital condition (marital)

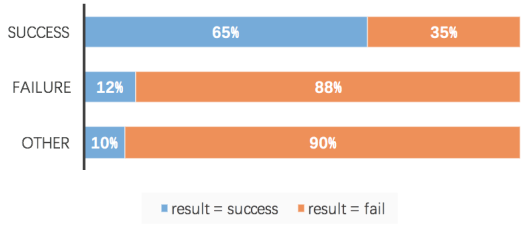

Fig. 2. Rate of success over different precious outcome (poutcome)

3 PREDICTIVE TASK

Our goal is to predict whether the client will subscribe a term deposit, thereby improving the efficiency of direct marketing. We evaluate it by calculating the true positive rate, true negative rate and balanced error rate of the prediction. We also plot the false positive rate vs true positive rate. The one has a higher quality AUC(Area Under Curve) value is identified as a better model for this predictive task. We established a baseline by running linear model using stochastic gradient descent without giving any class weight. By comparing our models with the baseline, we could better evaluate different models.

To train the models, we first shuffled the data and then divide it in a ratio of 6:2:2, that is, 60% of it is the training set, 20% of it is the validation set and the rest 20% is the test set. Validation set is used to adjust unpredicted parameters in order to improve the models performance. We predict if the client will subscribe a term deposit on the test set to validity our model's prediction.

We have tried using three data mining techniques' models: linear model using stochastic gradient descent, Support Vector Machine (SVM) and Random Forest. Linear model is easy to train but it works badly for imbalanced dataset classification. SVM works well for binary classification, but it does not perform well on highly imbalanced data sets, therefore, we need to set class weights. Random Forest is a good model for prediction, which preforms implicit feature selection and provides a good indicator of feature importance and can be trained faster than SVM. One drawback of it is that Random Forest is easy to end up with a huge forest, so we need to set parameters to limit the number of trees, the depth of the trees and so on.

The features we tried in the model is as following (including finally used and finally not used):

- Type of job (`job`). It is transformed to one-hot encoding.
- Has housing loan or not (`housing`). If has, 1, else, 0
- Has personal loan or not (`loan`). If has, 1, else, 0
- Contact communication type (`contact`). If it is 'unknown', 1, else 0.
- Last contact day of the month (`day`)
- Month of contact (`month`). It is also transformed to one-hot encoding.
- Age of client (`age`)
- Last contact duration in seconds (`duration`)
- Number of contacts performed during this campaign (`campaign`)
- Number of days that passed by after the client was last contacted and for this client (`pdays`)
- Number of contacts performed before this campaign and for this client (`previous`)
- Outcome of the previous marketing campaign (`poutcome`). This categorical feature is transformed to one-hot encoding.

4 MODEL

Since it's a classification problem, we select SVM and Random Forest to do prediction as the most appropriate models, and then further use Support Vector Regressor and Random Forest Regressor to associate score with each prediction and plot the ROC curve. The issues we encountered and solutions to them are shown below.

4.1 Unbalanced dataset

We first trained the model using `SVM.SVC` from `sklearn` without giving any weight to different classes. The correction rate on validation set is 88.166%. We changed the parameters and the result came to show little difference and always be 'good'. So we came up with the idea that we should change the evaluation. We then calculate the true positive to be 0 and true negative rate to be 1, thus the balanced error rate is 0.5. Obviously, the trained model predicted all case to be 0. We double checked our training set and found most of training data to be negative, saying 'y' = 'no'. This means giving different classes a weight and choosing a proper metrics to evaluate our model played an important role in this particular dataset.

4.2 Over-fitting

The dataset was split into 60% training set, 20% validation set, and the rest as test set. We trained our model on the training set first and calculated the TPR, TNR and balanced error rate on the validation set to see the different between the two sets. If the BER on training set is far lower than the validation one, which means over-fitting on the training set, we should pick a smaller parameter C, saying a larger regularizer λ, to make our model less complicate and vice versa. We traversed the variable C to find an optimum result at C = 0.1.

4.3 Feature selection

In our data set, there are 16 input features, including 8 bank client features, 4 contact features and 4 other features. As is mentioned in dataset analysis, some features may have less relationship with the result. Because of the large category of our features, it's quite hard to exactly find the useful features. We tried every feature and statistically learned on the feature.

We first tried these three kinds of features respectively, and found the bank client features to be least useful compared with the other two, indicating that whether a campaign is successful or not has little relevance with the background of candidate. On SVM model, bank client features with 'balanced' class weight and C set to 0.1 gave out a BER of 0.4174, while contact and other features with the same parameter gave a better one as 0.2706.

Therefor, all the last 8 features, saying contact features and other features, were used in our model. However, some features in bank client features such as 'age', 'housing' and 'loan' also did a job on our model. Saying 'age', we saw a high percentage of success in candidates under 25 and over 60. Obviously, 'age' can somehow improve our model.

4.4 Model selection

Since what we tried to solve is a classification problem, and the classification error matters, the first idea came to us is Support Vector Machines. SVM can be a useful tool in case of non-regularity in the data, saying our dataset is not regularly distributed as we had far more negative cases. It optimizes the classification error rather than the likelihood. We trained SVM model and adjusted the parameters to get an optimum result as balanced error rate to be 0.2532.

However, the result is still not satisfying. SVM is time-consuming and computationally expensive. We had a large dataset and feature category so we tried to find a more efficient model, that's Random Forest. Random Forest performs more efficient and satisfying on large dataset. What's more, we don't bother to pick out the useful features and decide the features manually. Random forest can give out the importance of each feature and pick the feature through natural selection.

5 LITERATURE

The dataset is from UCI Machine Learning Repository. It was preprocessed by discarding the instances with missing values. We randomly choose data as the training, validation and testing set, focus on useful features and analyze them with different data mining techniques' models. Since direct campaign focuses only on a small group of people who are predicted to be interested in the product, this dataset has been a classical dataset for the research of direct marketing and more researches on similar datasets have been carried out as well. One example of similar datasets that has been studied is 2008 USA Obama vs McCain Voting Pollfis dataset. The dataset has 1000 counts, and has similar fields such as political party, ideology, race, gender, religion, family income, education, age, region and Bush approval. The classification goal is to predict whom the voters vote for.

The methods employed to study this type of data are generally the same, that is, drawing the ROC curve and selecting the model with the maximum AUC as the ideal model, filtering out useless features and selecting the appropriate inputs of the model. If the data is imbalanced, the evaluation is selected as precision and recall, balanced error rate, F1 score etc. Some researches carried out an ensemble technique by combining different classifiers to improve the performance.

6 RESULTS

6.1 Features

Based on several experiments, although almost every feature could give information, for Support Vector Machine, there are six features that are most relevant.

- Last contact duration in seconds (`duration`).
- Number of contacts performed during this campaign (`campaign`).
- Number of days that passed by after the client was last contacted and for this client (`pdays`)
- Number of contacts performed before this campaign and for this client (`previous`)
- Outcome of the previous marketing campaign (`poutcome`). This categorical feature is transformed to one-hot encoding.
- Month of contact (`month`). It is also transformed to one-hot encoding.

For Random Forest, in addition to the features above, there are some features that could give additional information, which could help improve the performance of random forest tree, but not useful in SVM.

- Age of client (`age`)
- Contact communication type (`Contact`)
- housing
- loan

The top six relevant features and their corresponding importance in our Random Forest model are shown in figure 3.

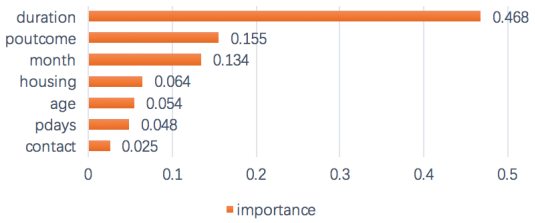

Fig. 3. Six top relevant feature of the Random Forest model

Additionally, based on our experiments, the features `job, days, default, marital, education` are not useful for both SVM and Random Forest. This proves our statement in dataset analysis that those feature are not useful since the rate of success for different categories under each features is almost the same.

6.2 Classification

To predict whether the client subscribes a term deposit during the campaigns or not, some classification models are implemented first. The baseline is the linear model using stochastic gradient descent without giving any class weight. The true positive rate (TPR), true negative rate (TNR) and balanced error rate (BER) using baseline, Support vector classifier and random forest classifier are compared in table 1.

	TPR	TNR	BER
Baseline	0	0.999	0.500
Support vector machine	0.770	0.723	0.253
Random forest tree	0.857	0.842	0.150

Table 1. Comparison between different models on True Positive Rate(TPR), True Negative Rate(TNR) and Balanced Error Rate(BER)

Using SVM model with rbf kernel and degree of 3, we got the final balanced error rate of 0.253 on the test set. In SVM model, there's a parameter, class weight, which is used to put additional focus on minor class and cost on misclassification. In our case, we had far more negative class than positive one. So we put more weight on '1' class. We first used the default class weight 'balanced', which is setting the class weight inversely proportional to class frequencies in the input data, and got a quite satisfying result with balanced error rate of 0.2706 compared with 0.3535 without class weight. Then we adjusted the weight and the final class weight was set to 0:1, 1:5 to achieve a best result. It can bias the model towards paying attention to minority `class 0`, which is y = 1 in our case. The regularization term C is set to 0.1. Smaller C means larger regularization, which help to avoid over-fitting problem.

Using random forest tree with `n_estimators= 40, max_depth= 8, max_features='log2', bootstrap=True, class_weight= 'balanced'`, we got the final balance error rate of 0.169 on the test set. The number of trees in the forest. `n_estimators` means the number of trees in the forest, `max_features` means the number of features to consider when looking for the best split. `max_depth` means the maximum depth of the tree. `bootstrap` means whether bootstrap samples are used when building trees. `class_weight` is similar as what it is in SVC.

In conclusion, from our experiment, Random Forest helps the banks better predict which group people might more likely to subscribe a term deposit. Theoretically, when the dataset is relatively large and there are some outliers making the dataset not quite clean, Random Forest works better than SVM. Random Forest also performs implicit feature selection, which could avoid noisy feature influence the model. SVM works well when the dataset could be small (less than a few hundreds) and outliers free. Since our dataset has around 50,000 samples, many of them may be outliers, and it has 16 attributes, some of which may be noise, Random Forest should work better and it does perform a better result than SVM.

6.3 Regression

To show the predictive results more clearly and identify which model is better for this prediction, we could assign scores to each prediction and plot the True Positive Rate vs False Positive Rate curve, which is shown in figure 4.

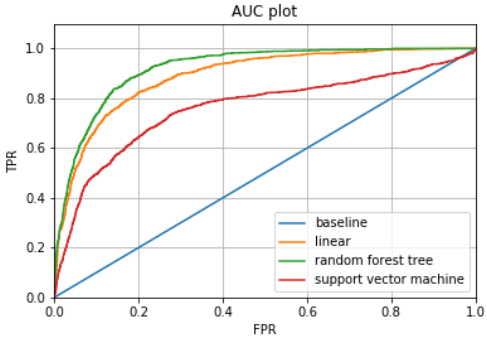

Fig. 4. ROC curves for different predicting models

The higher the line is, the larger the AUC value is and the better the model is. From figure 4, we could see that the AUC value of Random Forest is 0.934, the AUC value of linear regression is 0.872, and the AUC value of SVM is 0.694. Therefore, Random Forest works better than Linear Regression, Support Vector Machine and baseline.

7 CONCLUSION

Bank direct marketing is significant in seeking the potential customer groups. To predict if the client subscribed a term deposit during the campaigns, we use three different models, including linear model using stochastic gradient descent, Support Vector Machine and Random Forest campaigns. We evaluate it by comparing the balance error rate of different models and AUC value of the False Positive Rate vs True Positive Rate plot. We finally achieved 15.0% balanced error rate by using Random Forest Classification, which is much lower than 25.1% using SVM, and the AUC value is highest

using Random Forest. In conclusion, the experimental results have shown that SVM and Random Forest Models outperformed other models.

8 REFERENCES

(1) Moro S, Laureano R, Cortez P. Using data mining for bank direct marketing: An application of the crisp-dm methodology[C]//Proceedings of European Simulation and Modelling Conference-ESM'2011. Eurosis, 2011: 117-121.
(2) Elsalamony H A. Bank direct marketing analysis of data mining techniques[J]. International Journal of Computer Applications, 2014, 85(7).
(3) Gupta T, Xia T, Lee D. Understanding the effectiveness of bank direct marketing.