Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Bank Marketing Data

Zengyu Yan, Bohan Wang, Qunyou Liu
Electrical Engineering Faculty, Lausanne, Switzerland
CS-401 Applied Data Analysis, Project,Switzerland

***Abstract*—Traditional machine learning methods have achieved promising performance when handling linear models and common events prediction. However, when it comes to rare events, they show poor performance in predicting. Compared with that, random forest, as an ensemble learning method for classification and regression, provide a much higher prediction accuracy compared with other statistical models such as logistic regression [1]. In this project, we compare the performance of random forest with logistic regression in the field of marketing. We would use these two models to predict whether a costumer will subscribe to a term deposit and use ROC curve and F_1 score to assess their prediction accuracy. In order to minimize the negative impact of imbalanced data on the prediction, we apply SMOTE algorithm [3] to oversample the minority class. The ROC curve shows that random forest provides much more accurate predictions than logistic regression. In addition, it also has the ability to explain the casual process of the occurrence of rare events.**

***Index Terms*—Rare Events, Random Forest, Logistical Regression, SMOTE**

I. Introduction

Classic statistical models such as logistic regression are proven to perform poorly in predicting rare events like civil war onset. In the paper "Civil War Onset" [2], the author compared the performance of random forest with three versions of logistic regression of different features and hyperparameters. The results show that random forest outperforms all the logistic regression models in terms of prediction accuracy. Therefore, it is worthwhile to expand the research to other fields to see if the random forest would also provide accurate predictions in out-of-sample data. For this purpose, we proposed an extension project analyzing the performance of random forest on predicting rare events in the marketing field. Specifically, our task is to predict whether a client will subscribe a term deposit or not based on information about the client. To achieve this, we first collect a dataset from the direct marketing campaign of a bank institution in Portugues and preprocess it. Since the dataset is highly imbalanced, it can be viewed as a rare event. We then apply random forest and logistic regression to make predictions and compare the results of these two methods. [4]

The report is organized as follows: In section II, we make a short description of our datasets. Section III introduces the methods used in preprocessing the data. Section IV will implement logistic regression and random forest to the data. Section V compares and analyzes the performance of the models. Finally, we will conclude our project in section VI.

II. Data sets

We collect the Bank Marketing Data Set from the UCI Machine Learning Repository [5]. The dataset is based on a direct marketing campaign of a bank. Basically, the employees from the bank would make phone calls to the clients and convince them to make a term deposit with their bank. After the phone calls, the decision of the client will be noted - whether they accept the subscription or not, together with the information of the client. In our datasets, there are 41188 samples in total, and each with 20 attributes and a final decision: whether they subscribed to the term deposit ("yes") or not ("no"). The attributes can be divided into the following categories:

- Client's personal information: Age, job, marital status, education, loan status, and so on;
- Campaign activities: When and how to contact, the duration of the phone call;
- Social and economic environment data: Employment variation rate, consumer price index, consumer confidence index, euribor 3-month rate, number of employees;

The dataset is highly imbalanced since only a few people would make the bank term deposit. The ratio of accepting the subscription ('yes') and rejecting the proposal ('no') is roughly 1:8 in the datasets. In addition, there are missing values that appeared in some attributes marked with "unknown", since some people are unwilling to disclose their private information to the bank. Therefore, before we apply the statistical models, we have to preprocess the data and select useful features, which is also referred to as feature engineering.

III. Data Cleaning

A. Redundant Data Removal

The first step is to drop redundant data. To do this, we check if there are any duplicated rows and simply discard them.

B. Feature Selection

We now analyze the features and check if there are any irrelevant or inappropriate features to be discarded. Beforehand, just by looking at the data information, we have full reason to discard the attribute "duration" – the duration of each phone call, since it is unpredictable before each call. Besides, the duration is highly correlated to the final decision – the longer the duration, the higher the possibility that a client would subscribe. Therefore, to ensure a realistic predictive model, we would discard this attribute.

The rest of the features are analyzed by visualizing the distribution of the feature as well as its relation to the final decision. We notice that the "default" variable gives an anormal distribution, shown in Fig.1. The number of unknowns take a significant amount that is unlikely to ignore. Besides, people with unknowns default show a high tendency to rejecting the term deposit. Therefore, the unknowns of this attribute are too unpredictable so that it could be an inference to the final prediction. After discussion, we decide to discard this attribute.

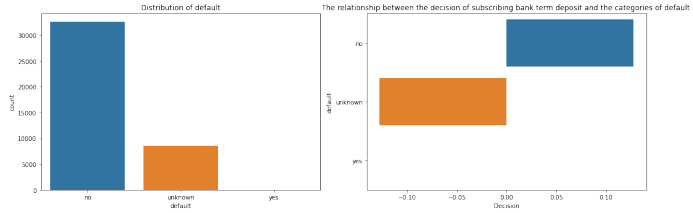

Fig. 1. Percentage of unknown data in "$default$"

C. Unknown Filling

The most challenging part of this project is to deal with the unknowns. There are many ways to handle it. Firstly, and most easily, we can directly delete all the samples that contain the unknowns. However, those unknowns take a noteworthy part in the whole dataset, and simply discarding them would definitely have a negative impact on the prediction accuracy. The second method we try is to use a decision tree to predict the unknowns. But the results also do not seem ideal (since the cross-validation score for predicting unknowns is only 0.3-0.5). Finally, we decide to use the function "SimpleImputer" in sklearn to replace the unknowns with the most frequent value along each column. As we can see later, the performance of our statistical models is satisfactory on our processed dataset.

D. Data Standardization

After filling the unknown data, we plotted the distribution of features with numerical values:

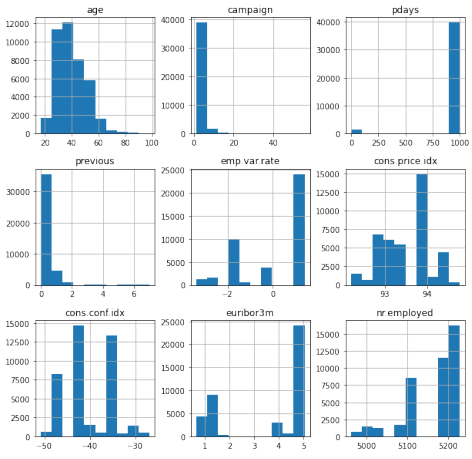

Fig. 2. distribution of the dataset before standardization

We can easily find that the distribution of the dataset is not uniform. For example, the client's age is ranging from 15 to 90, whereas the values of "campaign" stay mostly within the interval 0 to 20. Therefore, the standardization of the data is needed. In addition, the attribute "pdays" indicates the number of days that passed by after the client was last contacted from a previous campaign, and 999 means that the client was not previously contacted. For a better interpretation of the data, we convert the numerical values into category by replacing all the value 999 with "no", and the rest with "yes".

IV. Model Implementation

A. Data Oversample

Before applying the models to the data, We need to handle the problem of imbalance. As Machine Learning algorithms tend to increase accuracy by reducing the error, they do not consider the class distribution. In imbalanced datasets, the predictor would tend to generate outcome that has the larger portion in the dataset (in this case, the model will tend to predict "no", and the bank would be at risk of losing their potential clients). To address this problem, we use SMOTE algorithm to oversample the minority class. Basically, it takes each minority class sample as reference, and generates a random synthetic minority class example along with the line segments between the minority sample and a random neighbor of its k minority class nearest neighbors [4]. The generation of synthetic samples is based on the equation below:

$$x' = x + \text{rand}(0,1) * |x - x_k| \tag{1}$$

where:

x' is the random synthetic minority sample

x is the minority sample

x_k is a random one of the k-nearest neighbors of x

Normally, we choose 5 nearest neighbors of each minority sample to construct synthetic samples, which will be used in later prediction. After oversampling, the ratio between majority class (not subscribing) and minority class (subscribing) will become 4:1, which is much more balanced than the original ratio 8:1. With the help of SMOTE algorithm, the problem of poor prediction performance caused by imbalance datasets is overcome, and the potential of logistic regression and random forest can therefore be better exploited.

B. Logistic Regression

After oversampling, we apply logistic regression to the data. Concretely, we construct three logistic regression models with different settings (no penalty, L_1 norm penalty, L_2 norm penalty), and apply them to the data. We use grid search to find the best parameter for each model and use 10-fold cross-validation to avoid overfitting. After that, the FPR and TPR will be computed respectively.

C. Random Forest

Next, we implement random forest algorithm to the data. As before, we use grid search to find the best hyperparameters (in this case, the number of trees and the max depth of each

tree), and use 10-fold cross-validation to make predictions. Similarly, the FPR and TPR score will be calculated.

V. Performance Comparison and Feature Analysis

A. Performance Comparison

We plot ROC curve and F_1 score to compare the performance of random forest and logistic regression. The result is shown in Figure 3.

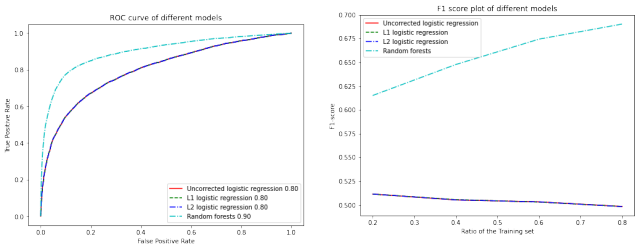

Fig. 3. ROC curve and F1 score

As we can see from the left plot, the random forest has higher AUC score (90%) than the logistic methods (80%), which is similar to the results in the paper "Civil War Onset". In the F_1 score plot, we adjust the ratio of the training set such that the percentage of the training data ranges from 0.2 to 0.8. As we see in the right figure, the F_1 score of random forest is much higher than logistic regressions and will continue to increase as we enlarge the ratio of the training set.

B. Feature Analysis

To further analyze the impact of the data features, we compute the importance of features based on its Gini score. The mean decrease in the Gini Score is the predictive accuracy lost by omitting a given predictor from the trees. It can be used to describe the importance of features, shown in Figure 4.

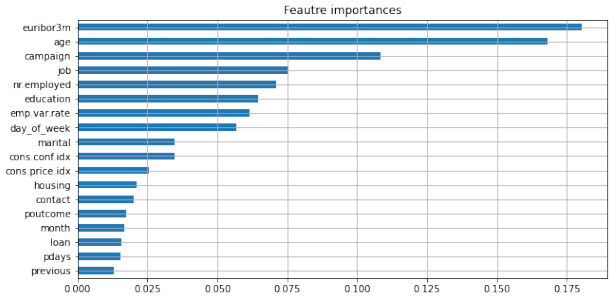

Fig. 4. variable importance

We see that the features 'euribor3m', 'age' give the most importance to the decision of subscription, followed by the attributes 'campaign', 'job', 'nr. employed', 'education', etc. Interestingly, the number of contacts performed before this campaign for the client affects the decision the least.

We next plot the partial dependence of each individual feature to see how individual variables can affect the result. Figure 5 visualized the first 9 important features.

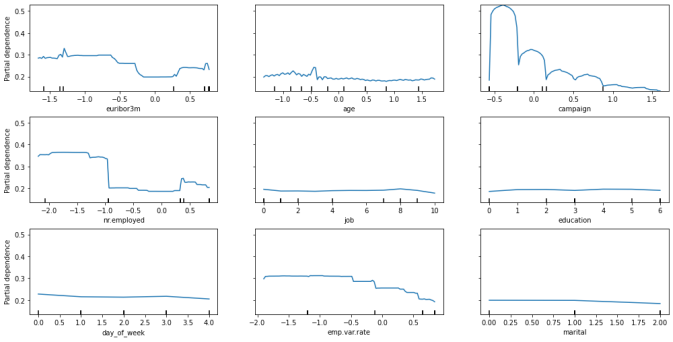

Fig. 5. Partial Dependence of Individual Features

From the partial dependence plot we have the following conclusions. With lower euribor 3-month rate, people would prefer to make term deposit. Young people are also more likely to subscript the term deposit. The 'campaign' feature, which is the number of contacts performed during this campaign and for this client, shows that the less phone call being performed, the higher the possibility for client to subscribe. The number of employees also highly affect the result, with lower number of employees giving more subscriptions.

Since the partial dependence plot cannot capture the influence of category features, we analyze them by visualize the distribution. For example, figure 6 tells that people with higher education degree would be more likely to make a term deposit.

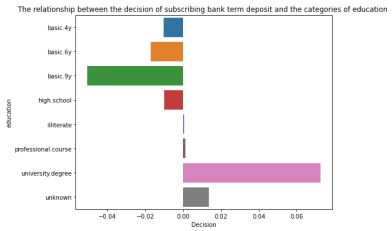

Fig. 6. Education Value Distribution

Those casual processes being visualized explained when using random forest algorithm provide more insight to the data, in which random forest surpass logistic regression. The dependency of categorical variables to the decision can not be captured by logistic regression models, in which random forest again shows its superiority.

VI. Conclusion

We compare the performance of random forest with logistic regression by apply them to the marketing field. We proposed a research question of whether a client will subscribe a term deposit or not. The prediction accuracy is assessed by ROC curve and F_1 score. The results show that random forest provides more accurate predictions than logistic regression. In addition, the ability of interpreting casual effect of important variables, such as euribor 3-month rate and age, is also another advantage of random forest implementation.

REFERENCES

[1] Muchlinski, D., Siroky, D., He, J., & Kocher, M. (2016). Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data. Political Analysis, 24(1), 87-103. doi:10.1093/pan/mpv024

[2] Chawla, Nitesh & Bowyer, Kevin Hall, Lawrence & Kegelmeyer, W.. (2002). SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. (JAIR). 16. 321-357. 10.1613/jair.953.

[3] Khoshgoftaar, Taghi & Golawala, Moiz & Van Hulse, Jason. (2007). An Empirical Study of Learning from Imbalanced Data Using Random Forest. Tools with Artificial Intelligence, 2007. 2. 310-317. 10.1109/ICTAI.2007.46.

[4] Chen, C., A. Liaw, and L. Breiman. 2004. Using random forest to learn imbalanced data. Berkeley: University of California.

[5] https://archive.ics.uci.edu/ml/datasets/Bank+Marketing