A Study of Logistic Regression and Random Forest for Bank Marketing Prediction

Ning Hu
MSc Data Science
Thompson Rivers University
Kamloops, Canada
hun21@mytru.ca

Abstract. Marketing campaign is important for company. The accuracy of a model could improve the efficiency of direct marketing. A successful business model also need to be interpretable and robust to. In this project, we applied the logistic regression (LR) and the random forest (RF) model on a bank marketing data set to predict whether a client will subscribe a term deposit. Precision-recall curve was used in evaluating the accuracy of the models, because of the unbalanced classification outcome. The result of the project shows that both method have almost the same performance respect to the precision and recall. The variables that LR selected and RF emphasize on are also almost the same. We interpreted the both models, and explained why the LR model is more interpretable in business decisions.

1 Introduction

Marketing campaign is an important strategic activity that boost the growth of revenue for a company. Companies can reach out their existing customers or new customers through email or telelphone. This activity can be time consuming for both the clients and companies. To maximize the value of the conversation, an optimization model should be used to select the sets of clients that are most likey to accept the offer [1]. With the clients informantion the company collected, data mining method can be used to build this model.

Lots of marketing models have been created over years, but only a small fraction of used in real business decisions. The success of a model critically depends on its simplicity and robustness of [2]. Therefore, more interpretable methods are required in solving marketing compaign problems. In this project, we will discuss the performance and interpretability of the logistic regression (LR) model and random forest (RF) model.

2 Data

2.1 Data explanation

A bank marketing compaign data set that is provided by [3] is used for this study [4]. This is a record of wheather a clients has accepted a term deposit offer after telemaketing phone calls. There are 41188 records in this data set. Each record contains 1 response, which is the contact outcome, and 18 predictors, including client background (e.g., job, education), previous campaign contact

information (e.g., outcome of the previous campaign), and the social and economic context attributes (e.g., consumer price index).

The outcome of the contact is success or not. Thus, this is a binary classification problem. However, the outcome is quite unbalanced. Only 11% of the contanct results are success, which is common in a blind marketing campaign. The unbalanced nature of the data influenced the evaluation of the performace, which we will discuss in section 3.

2.2 Removed variables

The *duration* variable, which is the last contact duration in seconds, is recorded after a phone call is performed. Before deciding which client is the potential buyer, the duration is known. So, this variable is removed from the predictors.

The *default* variable means wheather the client has credit in default. There are only 3 records with the positive answer in over 40,000 records. So, this variable is almost useless in any model, and has been exclude from the current analysis.

2.3 Missing values

Dealing with missing values can be very tricky. For this data set, there are no missing values in the continuous variables. However, in many categorical variables (e.g. job, education, et al.), the missing values are marked as "unkown". The records with missing values account for 8% of the whole data set.

Usually, there are two ways to deal with missing values. The most easiest method is to remove the records with missing values. But it would be wasteful if the record is very few or there are many missing values that ignore them would significantly affect the final model. The other method is to impute the the missing data. Impute method can be replacing missing values by the average values of the existing values or a completion matrix, which is dependent on the nature of the missing values. [5]

With this data set, we can simply remove the missing values as they are only a little fraction of the whole data. However, in a prediction situation, it is inevitable there are still missing values in the same varaibles. Since all the missing values are categorical variables, we will treat them as one different category in each variable. This would not be influence the result too much for the existing value and can make better prediction for the data with unkown categories.

3 Method

3.1 Models

To make the model useful, we must understand the concerns of the bank manager in making business decisions. According to the survey in [4], the questions that the manager most interested in is if the contact result relevant to some certain variables. Furthermore, they would more likely to know which variables have the most influence in the contact result, and how the change of a value would affect the result.

LR, decision trees (DT), neural networks and support vector machines are tested in [4]. Among those classification approaches, DT and LR are most interpretable. DT can be implemented by IF-

ELSE statement in programming, and it can be understood by non-expert. But the DT model tends to be overfiting the training data and can be very non-robust [5]. In our test, we will use the LR and RF approaches. The RF is an enhanced method of DT. Compared to DT, it loses some interpretability, but can still satisty the concerns of the manager and increase the robustness at the same time.

LR assumes the response follows a bernoulli distribution, and uses sigmoid function to transform the linear combination of predictors into a probability between 0 and 1. Suppose the probability of $y_i = 1$ is p_i, the LR model can be written as

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})}}$$

where p is the number of predictors. In prediction, a probability threshold is required to separate the two classes. The interpretation of LR is that, the change of the predictors, will result in a change in the log odds. Although the probability of being one class is not changed evenly coresponding to the change of linear part, they are still comparative between different variables.

To simplify the LR model, step-wise feature selection method is performed. The result of the feature selection also highlight the most revelant variables.

RF is a based on a bootstrap aggregation procedure. Based on bootstrap resampling, many decision trees can be built. The final decision is determined by voting result of each decision tree. This procedure can reduce the variance of the model significantly. Apart from this, when building each level of the trees, only one part of the predictors is used, the purpose of which is to reduce the similarity of each tree. The result of a prediction of RF is also a probability of how likely it belongs to each class. The interpretation of RF is not as clear as DT, but we can get important factors to explain which factor has more influence in making decisions.

3.2 Evaluation

Confusion matrix is very helpful to evaluate the performance in classification problems. It is a cross-tabulation of actual classes and predicted classes with associated statistics. For two class problems, the confusion table is defined as Table (1).

		Actual	
Predicted		Event	Not Event
	Event	A	B
	Not Event	C	D

Table (1) Two classes confusion table

Accuracy that associated with confusion matrix is the most useful assessment of the performance. It is caculated by (A+D)/(A+B+C+D), and reflects the probability that one model could correctly classify one observation. However, it is not useful for imbalanced classification distributions [6]. In our data set, the failed contact result is not interested by the compaign manager, but it account for most of the part in accuaracy.

Precision-recall curve is a ideal tool to assess the imbalanced classification problem. It shows the relationship between precision and recall associated with confusion matrix. Using the notions

above, precision (also positive predicted value) is A/(A+B), and recall (also sensitivity) is A/(A+C). In the bank marketing problem, we treat clients accpeting offer as positive. Then the precision means the proportion of clients will actually accept the offer in our whole candidates predicted by the model. The higher the value, the higher the efficiency is in the maketing calling activity. Recall means the proportion of the candidates that will accept the offer to the whole population that will accept the offer. High recall values mean more potential clients are reached.

The code can be found at https://github.com/Spacy/DASC5420.

4 Result

4.1 Model

Before modeling, the data set was separated to training set and test set. We use the same training set and test set to build both LR and RF model. A 10-fold cross validation is used in building logistic regression model to perform the backward step-wise feature selection. No cross validation is used for RF model, because the prediction of RF is based on voting, which could reduce the variance that is caused by random choice of the training sample.

The features that are selected in the final LR model and the important factors that outcomes from RF are shown as Table (2) and Figure (1) respectively. There are no importance information in LR model, so we will only compare the 13 variable selected in LR and the first 13 most important variables in RF. In RF's important factors figure, both MeanDecreaseAccuracy and MeanDcreaseGini means how much a factor will affect the RF prediction. The higher the values, the higher the importance of variable in the model [8].

The 13 most important factors defined by MeanDecreaseAccuracy are almost the same, except *age* is considered in RF, while it is not selected by LR. The diffference between LR and MeanDecreaseGini are much bigger. Only the *loan* variable is not important in either model.

Feature	**Meaning**	**Selected**
job	type of job	yes
marital	marital status	yes
education	education	yes
contact	contact communication type (cellular or telephone)	yes
month	last contact month of year	yes
day_of_week	last contact day of the week	yes
campaign	number of contacts performed during this campaign	yes
pdays	number of days that passed by after the client was last contacted from a previous campaign	yes
poutcome	outcome of the previous marketing campaign	yes
emp.var.rate	employment variation rate	yes
cons.price.idx	consumer price index	yes
cons.conf.idx	consumer confidence index	yes
nr.employed	number of employees	yes
age	age	no
housing	whether the client has housing load	no
loan	whether the client has personal loan	no

previous	number of contacts performed before this campaign	no
euribor3m	euribor 3 month rate	no

Table (2) Feature selected in step-wise logistic model regression

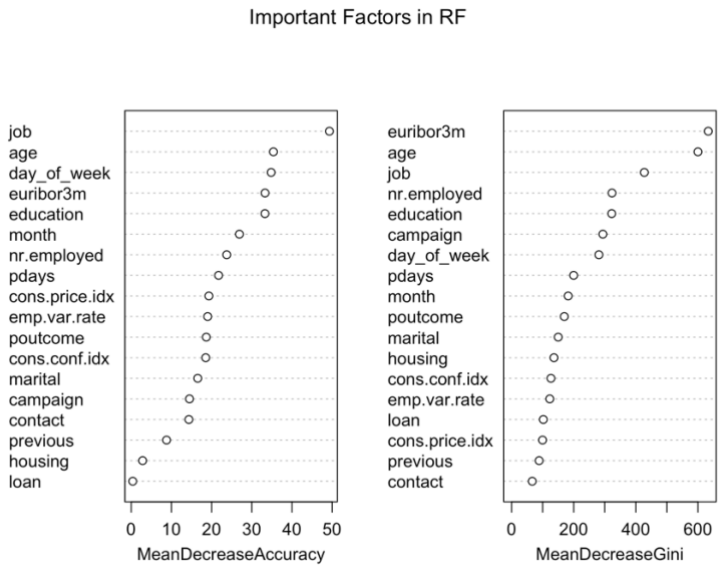

Figure (1) Important factors assessed by accuracy and gini index

By comparing the coefficients of the LR model, we can tell the value of each variable affect the prediction. For example, Table (3) shows a part of the coefficients. Without concidering the confidence interval, client with job is student has higher probability to accept the offer than the other jobs. Market calls excuted in March have higher chance to success than the other months.

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -3.110e+02  3.234e+01  -9.617  < 2e-16 ***
`jobblue-collar`        -1.387e-01  6.624e-02  -2.094 0.036227 *
jobmanagement           -1.512e-01  8.884e-02  -1.702 0.088703 .
jobretired               2.263e-01  8.937e-02   2.533 0.011325 *
jobstudent               2.387e-01  1.127e-01   2.118 0.034183 *
maritalsingle            8.790e-02  4.936e-02   1.781 0.074969 .
educationbasic.9y       -1.600e-01  7.457e-02  -2.145 0.031920 *
educationuniversity.degree  1.155e-01  5.072e-02   2.276 0.022825 *
contacttelephone        -6.945e-01  7.767e-02  -8.942  < 2e-16 ***
monthaug                 6.793e-01  1.136e-01   5.981 2.21e-09 ***
monthdec                 7.301e-01  2.227e-01   3.279 0.001042 **
monthjun                -8.244e-01  1.171e-01  -7.040 1.92e-12 ***
monthmar                 1.809e+00  1.435e-01  12.602  < 2e-16 ***
monthmay                -3.212e-01  7.515e-02  -4.274 1.92e-05 ***
monthnov                -2.450e-01  8.560e-02  -2.862 0.004205 **
```

Table (3) Part of coefficients of logistic regression model

4.2 Precision and recall

Figure (2) shows the precision-recall curve for LR and RF on the same test data set. Generally, precision is a tradeoff of recall. With the same model, increasing precision ususally resulted in the decrease of recall. The area under curve can be used to compare two classification performance. In this test, the performance of precision-recall of LR and RF are almost the same.

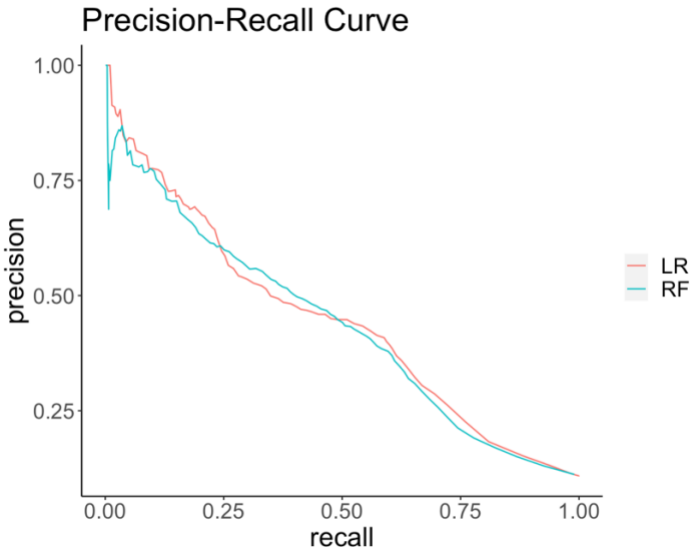

Figure (2) Precision-recall curve for logistic regression and random forest

5 Discussion

In this project, the performance of LR and RF is almost the same. When we applied step-wise method to perform feature selection, AIC is used to assess the model accuracy. However, as we discussed the section 3, our performance is evaluated based on precision-recall curve. Therefore, a metrics that combines the precision and recall is more accuracy for model selection. F1 score, which is the harmonic mean of precision and recall [9], would be a better choice for us. This metrics replacement might also be necessary for important factors of RF.

6 Conclusion

By fitting both LR and RF on the bank marketing data set, we compared the interpretability and performance of the two models. Both model get almost the same important factors in terms of accuracy and the same performance in terms of precision and recall. But the LR model is more interpretable. The change of value is more perdictable while using LR model.

References

[1] Talla Nobibon F, Leus R, Spieksma FC. Optimization models for targeted offers in direct marketing: Exact and heuristic algorithms. European Journal of Operational Research. 2011 May;210(3):670-83.

[2] Hanssens DM, Leeflang PSH, Wittink DR. Market response models and marketing practice. Appl Stochastic Models Bus Ind. 2005 Jul;21(4-5):423-34.

[3] UCI Machine Learning Repository: Bank Marketing Data Set. [cited 2023Apr13]. Available from: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

[4] Moro S, Cortez P, Rita P. A data-driven approach to predict the success of bank telemarketing. Decision Support Systems. 2014 Jun;62:22-31.

[5] James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: With applications in R. Boston: Springer; 2022.

[6] Brownlee J. Failure of Classification Accuracy for Imbalanced Class Distributions [Internet]. Machinelearningmastery.com. 2020 [cited 2023 Apr 15]. Available from: https://machinelearningmastery.com/failure-of-accuracy-for-imbalanced-class-distributions/

[7] Precision-recall curves – what are they and how are they used? [Internet]. Acutecaretesting.org. [cited 2023 Apr 15]. Available from: https://acutecaretesting.org/en/articles/precision-recall-curves-what-are-they-and-how-are-they-used

[8] Martinez-Taboada F, Redondo JI. Variable importance plot (mean decrease accuracy and mean decrease Gini). [Internet]. PLOS ONE; 2020 [cited 2023Apr15]. Available from: https://plos.figshare.com/articles/figure/Variable_importance_plot_mean_decrease_accuracy_and_mean_decrease_Gini_/12060105/1

[9] Korstanje J. The F1 score [Internet]. Towards Data Science. 2021 [cited 2023 Apr 15]. Available from: https://towardsdatascience.com/the-f1-score-bec2bbc38aa6