Potential Subscribers Classification

Kenneth Lee, Xialin Sang, Rong Duan, Chen Zhang

1. Introduction

Prior to a bank telemarketing campaign, it is critical and valuable for the campaign managers to collect a list of potential contacts based on rigorous statistical analysis. The first step is to gain some insights into the market from the readily available dataset. The data used in this report is collected from a telemarketing campaign in 2008 started by a Portuguese retail bank. It contains 41188 observations and 21 features, including attributes on client features, past campaign history, the current campaign setting, and social and economic context. The last feature is the campaign outcome stating whether a client subscribed to the term deposit after receiving the marketing phone call.

In this report, we propose two distinct statistical models to predict this outcome. Logistic regression has the advantage of fitting models that tend to be more interpretable, while also providing decent predictions in binary classification tasks. As for Random Forest, better performance is made possible with sacrifice in terms of interpretability. After a more detailed comparison of the two models based on suitable evaluation metrics, we recommend the Random Forest model due to its superior prediction of potential subscribers based on the available information. This would help increase campaign efficiency and allow more strategic investment of resources.

2. Data Analysis

2.1 Dataset Selection To study the retail banking market in Portuguese, we utilize the full bank marketing dataset from the UCI Machine Learning Repository with 41188 observations and 20 inputs. Here are two main reasons for this selection:

- *Drawing insights from other studies*: Since this dataset is very similar to the dataset used by another study relevant to bank telemarketing, we sense that it would be helpful to draw insights from other literature in building our prediction model [1].
- *Rich data tends to provide more ideas*: We believe that more data can often allow us to see a bigger picture of any trends in the market and enrich our findings based on exploratory data analysis.

2.2 Exploratory Data Analysis Our top interest is to investigate which variables are most relevant in making the prediction, especially for the logistic regression model. `Figure 1` shows some representative plots of the effects of categorical on the campaign outcome. To make a fair comparison across the unbalanced sample size, we show the categorical histogram in terms of percentile (instead of count). It is easier to tell whether there is an effect by tracing the proportion of the outcome "yes" in each category. For example, from all the job functions on the left panel, students and the retired group display a higher success rate. The middle panel shows that calls made to cell phones have a higher chance of success. The right panel is a typical representation of no detectable effect from the variable. We are less likely to include this kind of variable into the prediction model.

Notably, we can also tell from `Figure 1` that the dataset is quite imbalanced. In fact, the ratio between success ("yes") and failure ("no") is nearly 1:8. The imbalances need to be considered in the prediction model.

The left panel of **Figure 2** shows the effect of some numerical features on the campaign. It plots the number of contacts performed during this campaign for a certain client against the last contact duration. We can observe a separation of the outcome because **duration** highly affects the campaign outcome. As stated from the data description, the duration is not known before a call is performed. Therefore, it should not be part of a realistic predictive model. The mid panel shows a similar plot of two economic variables. Instead of providing trends regarding campaign success, the jitter plot presents a high correlation between each other. Collinearity might be a concern because it affects the model assumption and variance of the resulting coefficients. We further examined the variables with correlations by generating a correlation matrix plot on the right panel.

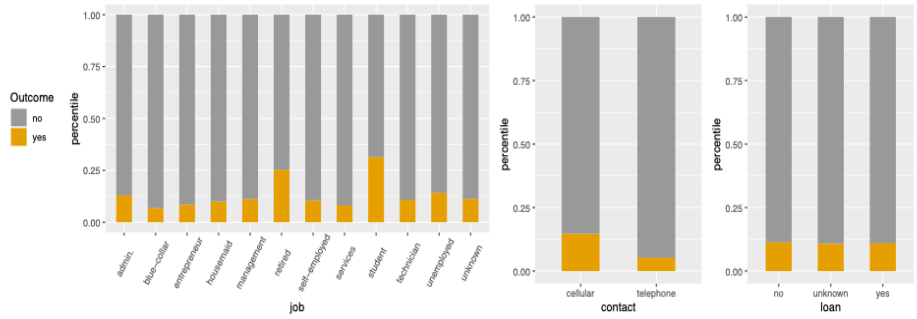

Figure 1: *Investigation of campaign outcome based on categorical variables*

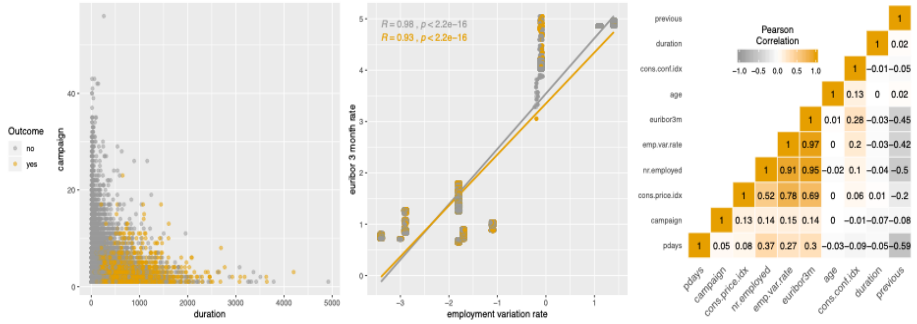

Figure 2: *Left and mid panels: Investigation of campaign outcome based on numerical variables; Right panel: Correlation Matrix*

2.3 Data Preprocessing We will describe any data preprocessing procedures in details before model prediction as follows:

- *Splitting into Training and Test Sets*: We randomly split the dataset into two parts with 80% being the training set and 20% being the test set. The test set is used as a hold-out sample to validate our model prediction performance while the training set is for building the model. Also, both the training and test sets have the same proportion of positive labels, indicating whether the client has subscribed to a long-term deposit.
- *Converting the data type of the response variable*: For building the logistic regression model in R programming language, we specifically convert the data type from factor to numeric.

2.4 Evaluation Metrics As we are faced with a highly imbalanced dataset, we will use precision, recall, and F1-score to be the performance measures instead of classification accuracy. In this study, classification accuracy is not an accurate measure for our interests as we try to see how well the model is able to identify potential subscribers, and there are very few positive labels in the dataset. In a highly imbalanced label setting, simply classifying all the observations to be non-subscribers can result in high accuracy.

We now briefly discuss the concepts of precision, recall, and F1-score. Precision is a measure of how many actual subscribers that we are able to identify out of the total number of predicted subscribers with 1 describing the model correctly identifying all the subscribers based on the total number of predicted subscribers and 0 otherwise.

Recall is a measure of how many subscribers that we are able to identify out of the total number of actual subscribers with 1 describing the model correctly identifying all the subscribers based on the total number of actual subscribers and 0 otherwise.

F1-score is a measure of precision and recall with 1 describing the model being perfect in both precision and recall and 0 describing the model being poor in both precision and recall. Thus, the best model we aim to build should have the F1-score as high as possible.

Another useful metric is the area under the precision and recall curve. We want to select a model that has the largest area under the precision and recall curve, which measures the overall performance of the model in terms of maximizing precision and recall.

2.5 Model Prediction : Logistic Regression To test the probability of a new user to subscribe a long-term deposit, we will use logistic regression model.Our logistic regression model is specified below:

$\pi_i\prime = \beta_0 + \sum_{j=1}^{r} \beta_j X_{ij}$ where $i = 1, ..., n$ and $j = 1, ..., 11$

Explanation of the notation

- $\pi\prime$ denotes the the logit transform of π_i, $log(\frac{p}{1-p})$, where p represented as the probabality of ith client subscribes the long-term deposit.
- β_0 denotes the intercept of the logistic regression model.It is also the log-odds of the event that $Y = 1$, when all X_j are equals to 0.
- β_j denotes the what amount of increase or decrease in the predicted log odds of event $Y = 1$ that would be predicted by a 1 unit increase in jth the predictor, holding all other predictors constant.
- X_{ij} denotes the predictor variable in ith case and jth characterastics. In our model we select 11 variables, which are age, job, marital, education, default, contact, day_of_week, campaign, pdays, previous and cons.conf.idx
- The index i represents the the number of cases we study. In our data, there are 41188 observations.
- The index j represents the variables included in the model. After depth analysis, 11 variables are selected.

2.5.1 Probability Threshold Selection As we use precison and recall as our metric, we aim to find probability threshold value that can give us good performance on both precision and recall. As shown by figure 3, we see that we should pick a probability threshold value around 0.4 in order to balance both precision and recall. The green color on the curve highlights the ideal probability values we should pick in order to have high precision and recall rate.

2.5.2 Feature Selection It is suggested that discarding irrelevant features tend to give better predictive performances [5]. We also conduct feature selection based on the following approaches:

- *Literature Reviews*: It has been shown that some factors are particular useful in building predictive models on a dataset that is similar to ours. These factors include interest rate, gender, client-bank relationship, phone call context, date and time, bank profiling indicators, social and economic indicators [1]. Thus, we built a model based on these factors by excluding the variables ("campaign", "pdays","previous","poutcome") that are specific to campaigns.
- *Hypothesis Testing*: We have also considered dropping some variables that are not statistically significant at the significance level 0.05 based on student t-test. These variables are the marital status ("marital"") and the number of contacts performed before the marketing campaign ("previous").

2.5.3 Prediction Result

- As shown by the second row in `table 1`, we see that by dropping the marital status ("marital"") and the number of contacts performed before the marketing campaign ("previous") can slightly improve F1-score in comparison with the full model. We also see that dropping campaigns-specific variables results in a poor classification performance.

2.6 Logistic Model Diagnostics Next, we discuss various approaches to verify whether the model assumptions hold.

- **Appropriate Outcome Structure :** Based on the attribute information on our desired target, the outcome is the answer to whether the client subscribed to a term deposit, which is yes or no. Thus, we know the output variable is a binary variable.
- **Independence of the observations :** In the `Bank Marketing` dataset, every client was only observed once. Besides, there is only one measurement of the same client. Therefore, the observations in the dataset are independent of each other.
- **Linearity:** We will check the linearity by visually inspecting the scatter plot between each predictor and the logit values. In `Figure 4`, the scatter plots show that variables `Age`, `Campaign`, and `Employment Variable Rate` are quite linearly associated with the outcome in the logit scale. However, the variable `Consumer Confidence Index` and `Previous` seem not to be linear with the logit values. In that case, we may need other methods like random forest below to build the model, or transform the non-linearity variable into linearity with 2 or 3 power terms, fractional polynomials and spline function to improve the goodness of the model.
- **Multicollinearity:** One way of diagnosing multicollinearity is through the calculation of variance inflation factors (VIFs). As a rule of thumb, multicollinearity exists if a VIF value exceeds 5. We have found that for each predictor is less than VIF score of 5, indicating that there is no perfect multicollinearity.
- **Influential Values:** We will first visualize the Cook's distance values to examine the most extreme values in the data. Since not all outliers are influential observations, we will then inspect the standardized residual error to check whether the data contains potential influential observations. If none of the absolute standardized residual above 3 is present, we can conclude that there is no influential observation in our data. In `Figure 5`, it shows that the values of standardized residuals range from -2 to 3, which means there is no absolute standardized residual above 3. Therefore, we will conclude the occurrence of the influential variables is rare in our data so that the assumption is satisfied
- **Large Sample Size:** Comparing to the ten predictors in our model, we have 32951 observations in our training dataset. Following a general guideline, we conduct our model with enough large sample size.

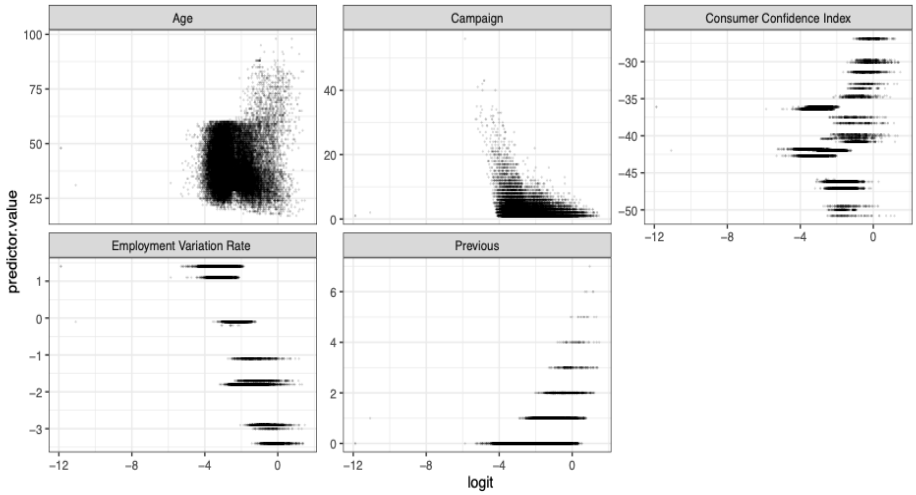

Figure 4: Scatter Plot between predictors and Logit Values

2.7 Model Comparison: Logistic Regression vs. Random Forest In addition, we attempt to find a better predictive model by making use of random forest algorithm [8]. Without any hyperparameters tuning, random forest outperforms all the logistic regression models in this study with a F1 score of 0.58 by using all the features in the data. We provide a few potential reasons why random forest could be more robust for this classification task:

- *Random Forest is more robust in dealing with imbalanced datasets*: It has been shown that random forest generally performs better than logistic regression with imbalanced datasets [4].
- *Random Forest is a non-linear algorithm*: While logistic regression is a linear model, random forest is a non-linear algorithm, which is potentially equipped with greater learning capabilities that range from linear to complex nonlinear mappings [1].

Finally, we should also note that there is a trade-off between model prediction and interpretability. When the random forest is more prone to identifying subscribers, it is also more difficult to understand what factors that may explain the result compared to logistic regression.

Table 1: Model Performance Comparison

Method	Precision	Recall	F1-Score	Area Under PR curve
Logistic Regression (excluded campaign features)	0.50	0.60	0.54	0.57
Logistic Regression (selected features)	0.52	0.62	0.57	0.57
Logistic Regression (all features)	0.52	0.62	0.56	0.57
Random Forest (all features)	**0.53**	**0.65**	**0.58**	**0.64**

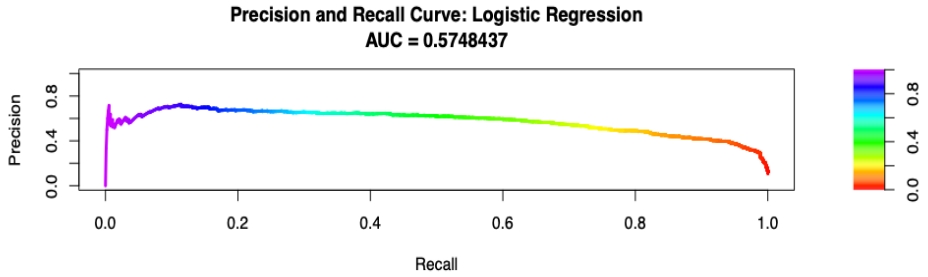

Figure 3. Left *Precision and Recall Curve for Logistic Regression* We select the probability threshold values roughly around 0.4 as it gives us a good balance between precision and recall.

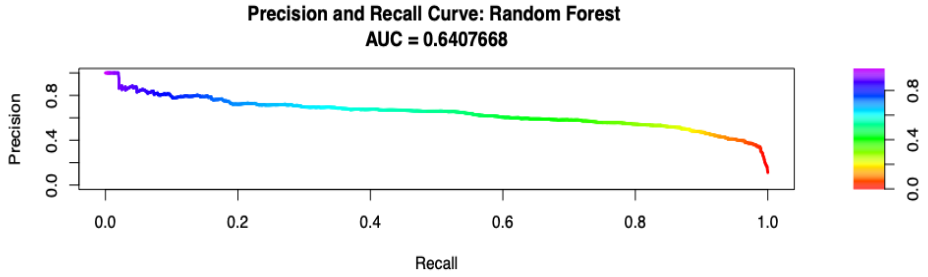

Figure 4. *Precision and Recall Curve for random forest.* We select the probability threshold values roughly around 0.4 as it gives us a good balance between precision and recall for random forest.

3. Conclusion In conclusion, we have found that random forest is better than logistic regression in terms of classification performance with precision, recall, and F1-score as the metrics. Without any hyparameter tuning, random forest is able to achieve 0.58 F1-score by using all the features. Also, we have found that the marital status may be not be useful for logistic regression model to identify potential subscribers. In the future, we can also consider using other algorithms such as aritficial neural network, support vector machine, etc. to compare with random forest to see if we can have classification performance gain.

Appendix I. Reference

[1] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

[2] Davis, J. and Goadrich, M. The relationship between precision-recall and roc curves. In Proceedings of the 23rd international conference on Machine learning, pp. 233–240. ACM, 2006.

[3] Saito, T. and Rehmsmeier, M. The precision-recall plotis more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. PloS one, 10 (3):e0118432, 2015.

[4] Muchlinski, David, et al. "Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data." Political Analysis 24.1 (2016): 87-103.

[5] Isabelle Guyon, André Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning Research 3 (2003) 1157–1182.

[6] Stoltzfus, Jill C. "Logistic regression: a brief primer." Academic Emergency Medicine 18.10 (2011): 1099-1104.

[7] Hosmer DW, Lemeshow SL. Applied Logistic Regression. 2nd ed. Hoboken, NJ: Wiley-Interscience,2000.

[8] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

Github Repository https://github.com/kenneth-lee-ch/STA-207/tree/master/project4