Predicting the success of bank telemarketing

1 Introduction

In 2007, a global economic crisis emerged in the United States and quickly spread to Europe, leading to new thoughts about financial management (Hodgson 2019). Suspicion on banks results in withdrawing money, frozen investment and credit loss. Therefore, for those banks who were affected by the public debts, a competition for subscribing clients to long-term deposits started to enhance their business. Due to the widespread use of telephones, telemarketing, a marketing method through remote communication channels, became a common and easy way to obtain various aspects of information (Moro, Cortez, and Laureano 2013).

In this project, we propose a logistic regression (LR) model to predict the success of telemarketing calls for selling bank long-term deposits and to identify potential subscribers based on the information from a Portuguese retail bank telemarketing campaign. We also compared the performance of the logistic regression model with a random forest (RF) model. We choose LR and RF since they have the advantage of fitting models that tend to be easily understood by humans, while also providing good predictions in classification tasks (Moro, Cortez, and Rita 2014). Using four performance metrics, true positive rate (TPR), false positive rate (FPR), receiver operating characteristic (ROC) curve, and area under the curve (AUC), the two models are compared using the testing set.

We choose to use the full dataset with 45211 observations and 17 inputs, which is ordered by date from 2008 to 2010. There are a few reasons why we choose the full dataset instead of subsets. First, the full dataset provides more information and guarantees enough observations in both training and testing sets for predicting models. More importantly, all the available datasets are highly imbalanced towards "no" in the decision variable. When balancing the training set, more "yes" results in a larger set. Therefore, we choose the "bank-full" dataset for its highest records of successful contact r table(data$y)[2]‘ (11.7%) among four datasets we potentially choose from (Lusa and others 2015).

2 Data Exploration

There are some variables with unknown values in this dataset. The left histogram in Figure 8 in appendix shows that the result of the previous marketing campaign has the most unknowns (82%). Contact methods, education level and type of job have some unknown values (29%, 4% and 0.6% separately). The right one indicates combination of unknown values between variables, for example, the fourth row from the top means that the number of observations that all these four variables contain missing value. Given the fact that it is diffcult to obtain all known values in real suvey, we will not regard those "unknown"s as missing values, but rather we regard those as a factor level and will be analysed later.

The following part will explore several variables and their effects on the outcome that whether the customer subscribes to a long-term deposit or not. Based on the results in Figure 1, middle-aged people which around 30-40 are more likely to buy the bank product. Management jobs, blue-collar and technicians are top three jobs tend to subscribe since those jobs may earn more money and are more eager to manage their finances. Also for those married people are more willing to buy deposits than others since they want good feedback from a bank (High interest) to support family livelihoods. The same issue happens to those highly educated people due to that they know some basic knowledge about economics.

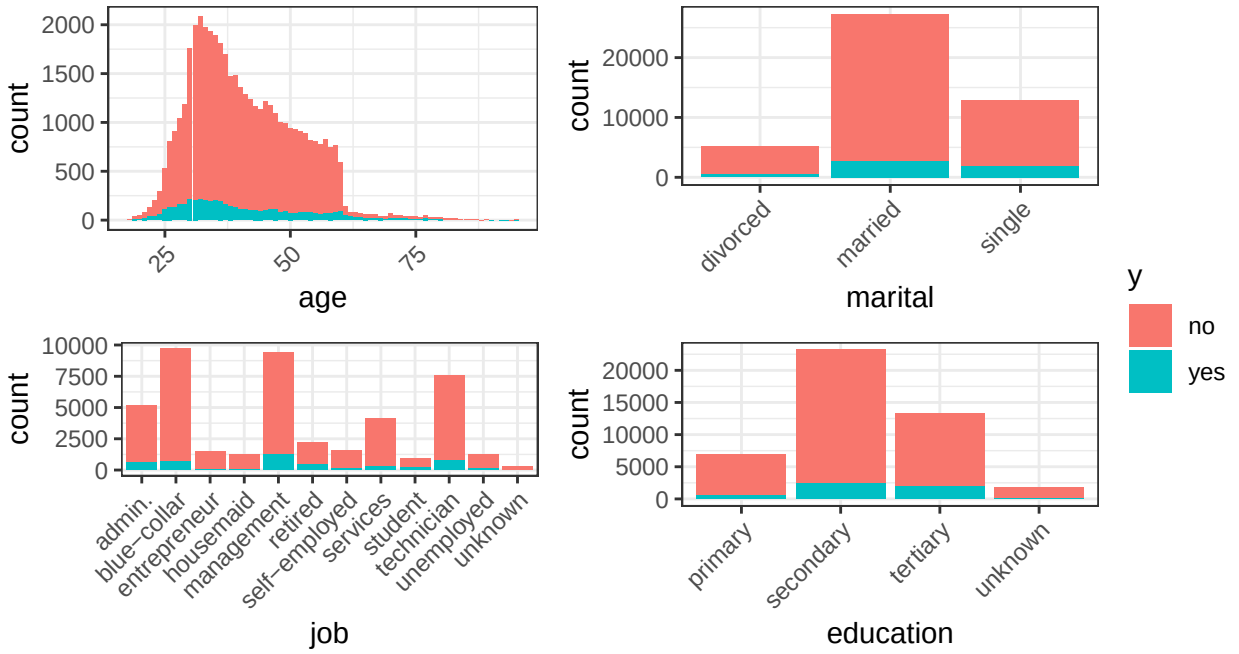

Figure 1: Deposit Subscription based on key variables

The full dataset is randomly partitioned into 70% being utilized to train the model while the remaining 30% (13564 observations) are left for testing the model. Since the dataset is highly imbalanced, and the standard classifier algorithms, such as logistic regression and random forest we use in the project, will have a bias which means that it tends to only predict the majority class data and regards minority class as noise and are usually ignored. As a result, the minority class may have a high probability of misclassification compared with the majority. To avoid such a problem, we balance the training data by maintaining "yes" (3721) and randomly selecting the same number of "no" from the initial training set to form a new training set with 7442 observations.

3 Methodology

3.1 Logistic Regression

Logistic regression is a common statistical model, which is used to model the probability of a binary class or event. It uses a logistic function to measure the relationship between the categorical dependent variable and independent variables. The detailed logistic regression model is shown below:

$$ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

where k is the number of predictors used in the model, p=P(Y=1|X) is the probability of response variable Y. In our case, Y is whether the client subscribed a term deposit and 1 means the client did subscribe a term deposit; β_0 is the intercept; β_1 to β_k are coefficients for predictors, which are decided through Akaike information criterion (AIC). This is because AIC measures predictive accuracy while BIC measures goodness of fit (Sober 2002; Shmueli and others 2010). Since our main purpose is to predict whether the client subscribed to a term deposit, AIC would be preferred for feature selection. Based on AIC results, 11 independent variables are used in the logistic regression model, which are job, marital status, education, average yearly balance, whether having housing loan, whether having personal loan, contact information, last contact month, number of contacts during this campaign, number of contacts before this campaign, and outcome of the previous marketing campaign. The coefficients are estimated based on the Maximum

Likelihood method, which maximizes the likelihood (conditional probability of the data given parameter estimates) of the sample data.

The major assumptions of logistic regression include (Kassambara 2018): 1) the outcome is a binary or dichotomous variable like yes vs no; 2) there is a linear relationship between the logit of the outcome and each predictor variables; 3) there is no influential value (extreme values or outliers) in the continuous predictors; 4) there is no high intercorrelation (i.e. multicollinearity) among the predictors.

3.2 Random Forest

The random forest (RF) is an "ensemble learning" technique consisting of the aggregation of a large number of decision trees, resulting in a reduction of variance compared to the single decision trees (Couronné, Probst, and Boulesteix 2018). We use the 'randomForest' and 'caret' packages in R to train and tune our random forest model. The most important parameters for RF are ntree and mtry. The parameter ntree denotes the number of trees in the forest with a default value of 500. The parameter mtry denotes the number of features randomly selected as candidate features at each split. The default value is $\sqrt{p}$ for classification with p number of features in the dataset.

First, we select features as input by using variable importance from the random forest algorithm with the default values of ntree and mtry parameters. The variable importance for each feature is shown in figure 2. To match the features number of logistic model, we select 11 variables with highest meanDecreaseGini value as input features to the RF model, which are last contact day, last contact month, average yearly balance, age, job, outcome of the previous campaign, contact type, number of contacts during this campaign, number of days that passed by after the client, education, whether having housing loan. Gini in RF algorithm means the importance of a particular variable in partitioning the data into the defined classes; therefore, variables with higher MeanDecreaseGini play a more important role in classification and prediction.

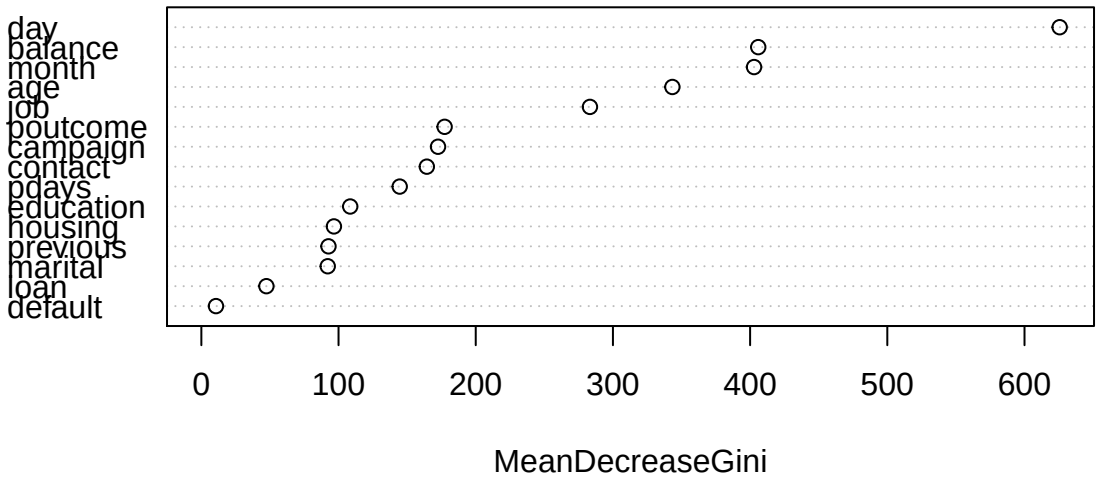

Figure 2: RF model feature importance

3.3 Metrics for Model Comparison (Kirasich, Smith, and Sadler 2018)

Accuracy, true positive rate (TPR), false positive rate (FPR), the receiver operating characteristic (ROC) curve, area under the curve (AUC) are often considered as the core metrics when comparing overall model performance. Table 2 in appendix shows the evaluation metrics for comparison of model performance. Accuracy, which is the percentage of correct classification, is a nice overall average of how well a model can predict and simple to compute. However, if there is a class imbalance, for example, 88% of failure in our testing set, it may not be useful.

In cases where there is a high class imbalance we need to use metrics such as true positive rate (TPR) and false positive rate (FPR). TPR is calculated as the portion of positives that are correctly identified, and FPR is the portion that was incorrectly identified as positive but is negative. They can be graphically represented using the receiver operating characteristic (ROC) curve, which is a graph with the x-axis of the FPR and the y-axis of the TPR at various threshold settings. A perfect prediction would have a false positive rate of 0 and a true positive rate of 1. When graphed over a series of thresholds, the area under the curve (AUC) can provide a single value for providing insight into how well the model classification is: the higher the AUC, the better the model performs. The AUC is more descriptive than accuracy because it is a balance of accuracy and false positive rate.

4 Analysis and Results

4.1 Logistic Regression

For logistic regression, the confusion matrix from test data is shown in figure 3. From the table, we can calculate the prediction accuracy is 78.83%, TPR is 61%, FPR is 18.8%. The ROC curve is shown in figure 3 below, with AUC equals 0.7749.

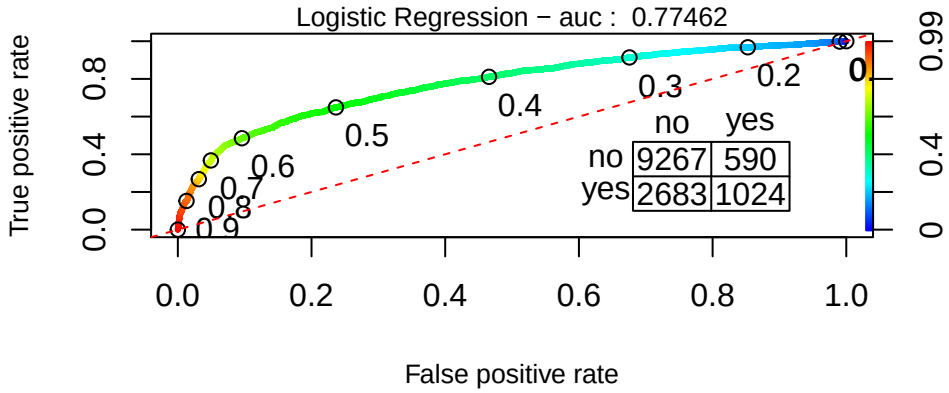

Figure 3: ROC curve with confusion matrix (columns are predicted, rows are target) for LR

4.2 Logistic Regression Diagnostics

Influential values are extreme individual data points that can alter the quality of the logistic regression model. We inspect the residuals to check whether the data contains potential influential observations. Since in logistic regression the data are discrete and so are the residuals, plots of raw residuals from logistic regression are generally not useful. Instead, the binned residuals plot, after dividing the data into categories (bins) based on their fitted values, shows the average residual versus the average fitted value for each bin (Gelman and Hill 2006). As shown in figure 9 in appendix, the strong pattern in the traditional residual plot arises from the discreteness of the data, and there is no obvious pattern shown in the binned residual plot. Therefore, there are no influential observations in the data. The standardized residuals are plotted in figure 4, suggesting that there are no influential points or outliers.

Multicollinearity is an important issue in regression analysis and should be fixed by removing the concerned variables. We calculate the Variance Inflation Factor (VIF) of each variable. All VIF values shown in table 1 are below 5, suggesting that there are no multicollinearity problems in the model.

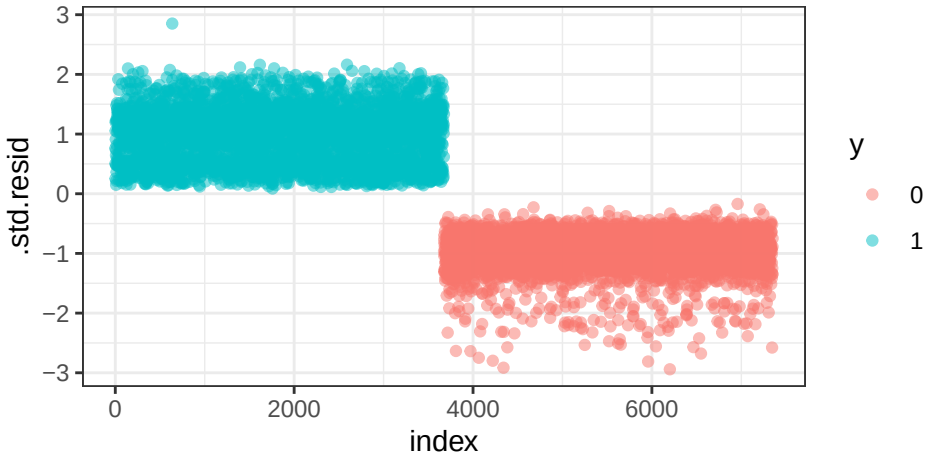

Figure 4: Standardized residual plots

Table 1: VIF for covariates

job	marital	education	balance	housing	loan	contact	month	campaign	previous	poutcome
2.9	1.15	2.3	1.04	1.36	1.05	1.94	2.71	1.08	1.81	2.02

Linear Relationship is to make sure the independent variables are linearly related to the logit of the dependent variable. we visually inspect the scatter plot between each predictor and the logit value, which is shown in the figure below. From figure 7 in the appendix, we notice that all variables except campaign are quite linearly associated with the deposit outcome in the logit scale; however, campaign shows some non-linear relationship.

4.3 Random Forest

To tune the RF model, we use a random search algorithm from 'caret' package in r, which uses 5-fold cross-validation based on training data to optimize the mtry values in the RF model. The optimal mtry value is 7, which provides the highest prediction accuracy and Kappa value. Using optimal mtry value, we try different ntree values varying from 5 to 5000, and find that ntree = 500 provides the highest prediction accuracy and Kappa value.

Similar with logistic regression, the confusion matrix is calculated based on the test dataset, which is shown in figure 5. From the table, we can calculate the prediction accuracy is 75.7%, TPR is 67%, FPR is 23.1%. The ROC curve is shown in figure 6 below, with AUC equals 0.7835.

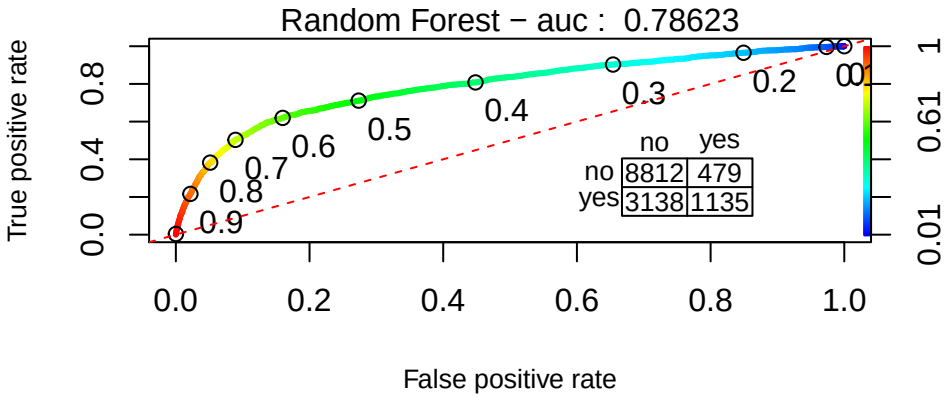

Figure 5: ROC curve with confusion matrix (columns are predicted, rows are target) for RF

4.4 Performance comparision

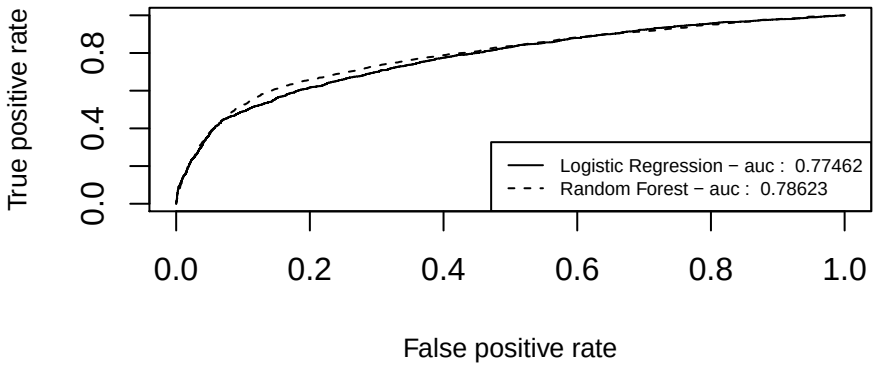

Figure 6: ROC curves comparison

Figure 6 compares the ROC curves for LR and RF models tested. The random forest ROC curve is related to a higher AUC of 0.7852 and outperforms the logistic regression model within most of the FPR range. For the range FPR within [0.08,0.65], the RF gets a higher TPR value ranging from 0.45 to 0.90. The TPR for random forest (67%) is higher than logistic regression (61%) and yields a higher false positive rate (23.1% vs 18.8%). In the case of bank telemarketing, it is better to produce more successful sells even if this involves losing some effort in contacting non-buyers. Therefore, the RF model performs better in predicting the success of bank telemarketing.

5 Discussion

We find that RF performs better than LR according to the Area Under the Curve (AUC) with a difference of 0.0133 (1.33% higher). The TPR for the random forest with 500 trees is 9.8% higher than logistic regression, while FPR also higher in random forest. Since we emphasize more on successful contact, we still suggest a better performance of RF. The results are consistent with previous studies (Couronné, Probst, and Boulesteix 2018; Kirasich, Smith, and Sadler 2018). The better performances of the RF model might come from its ability to capture non-linear relations, while the LG model could only capture the linear relations of explanatory variables.

Reference

Couronné, Raphael, Philipp Probst, and Anne-Laure Boulesteix. 2018. "Random Forest Versus Logistic Regression: A Large-Scale Benchmark Experiment." *BMC Bioinformatics* 19 (1). Springer: 270.

Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge university press.

Hodgson, Geoffrey M. 2019. "The Great Crash of 2008 and the Reform of Economics." In *The Handbook of Globalisation, Third Edition*. Edward Elgar Publishing.

Kassambara, Alboukadel. 2018. "Logistic Regression Assumptions and Diagnostics in R." *STHDA*. http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/#logistic-regression-diagnostics.

Kirasich, Kaitlin, Trace Smith, and Bivin Sadler. 2018. "Random Forest Vs Logistic Regression: Binary Classification for Heterogeneous Datasets." *SMU Data Science Review* 1 (3): 9.

Lusa, Lara, and others. 2015. "Joint Use of over-and Under-Sampling Techniques and Cross-Validation for the Development and Assessment of Prediction Models." *BMC Bioinformatics* 16 (1). BioMed Central: 363.

Moro, Sérgio, Paulo Cortez, and Raul Laureano. 2013. "A Data Mining Approach for Bank Telemarketing Using the Rminer Package and R Tool." Instituto Universitário de Lisboa (ISCTE-IUL).

Moro, Sérgio, Paulo Cortez, and Paulo Rita. 2014. "A Data-Driven Approach to Predict the Success of Bank Telemarketing." *Decision Support Systems* 62. Elsevier: 22–31.

Shmueli, Galit, and others. 2010. "To Explain or to Predict?" *Statistical Science* 25 (3). Institute of Mathematical Statistics: 289–310.

Sober, Elliott. 2002. "Instrumentalism, Parsimony, and the Akaike Framework." *Philosophy of Science* 69 (S3). The University of Chicago Press: S112–S123.

Appendix

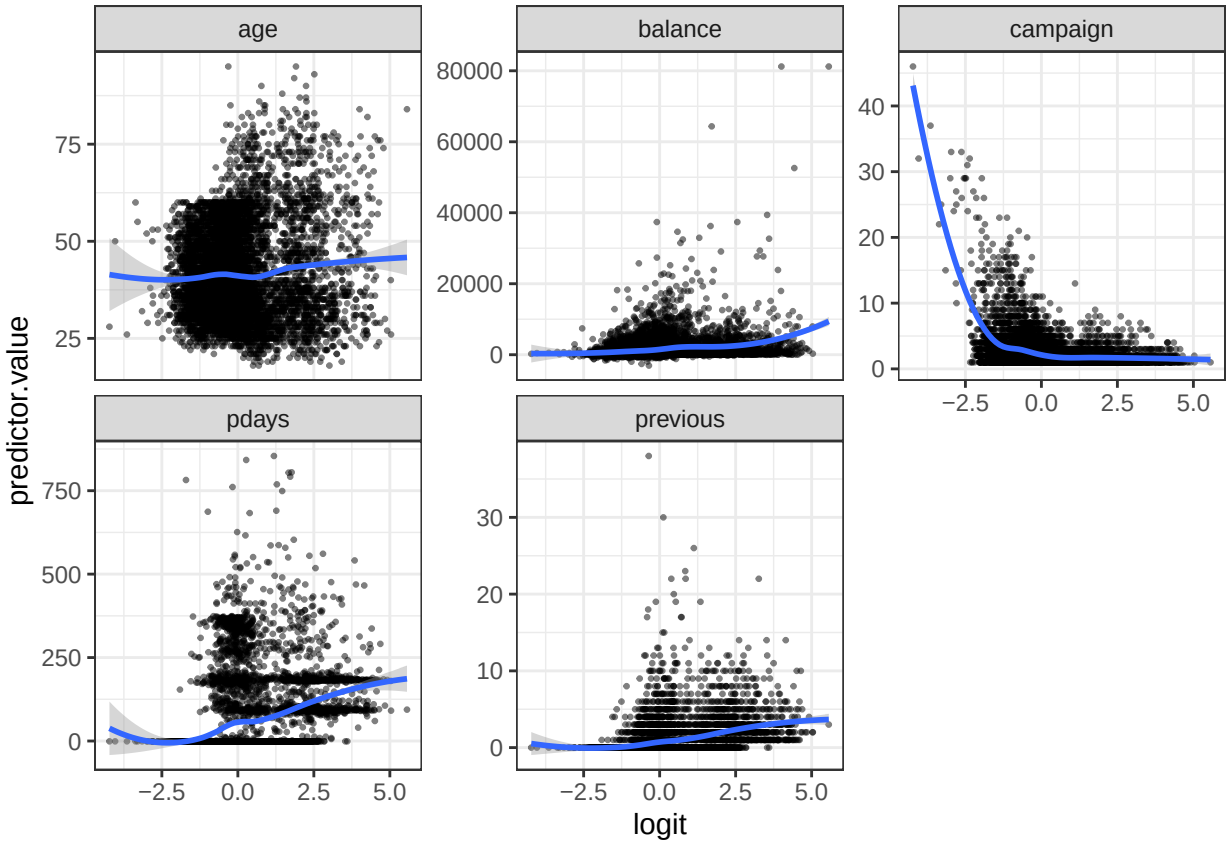

Figure 7: Linear relationship between variables

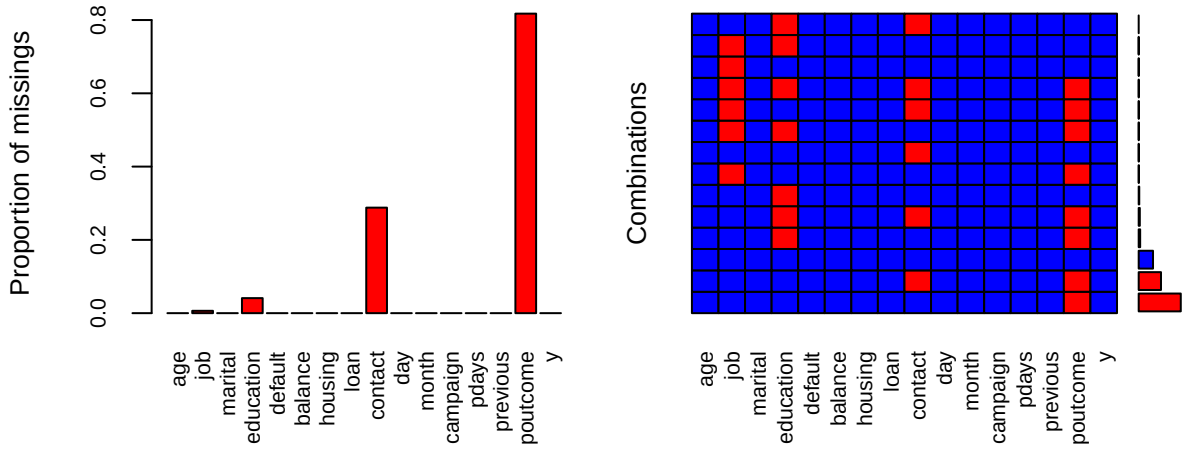

Figure 8: Distribution of Missing data

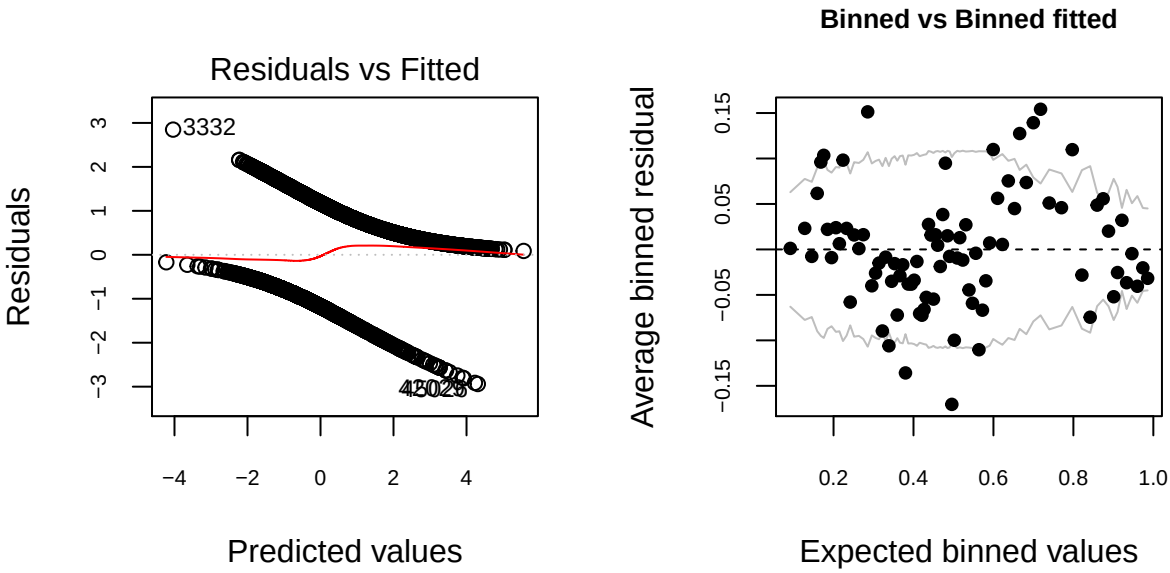

Figure 9: residual plot

Table 2: Evaluation metrics for comparison of model performance, TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative

Metric	Formula
Accuracy	(TP + TN)/(TP + TN + FP + FN)
True Positive Rate (TPR)	TP/(TP + FN)
False Positive Rate (FPR)	FP/(FP + TN)
Area Under the Curve (AUC)	Integral area of plotting TPR vs FPR