Data Analysis of Bank Market

Quynh Doan

12/14/2020

1. Introduction

Financial market research is a critical component for a company to master the market demand and to seek out potential customers, however, it costs a lot as well. To avoid this, this study decides to analyze the existed public dataset to identify the possible group of customers and to find a powerful prediction model to study the retail banking market in Portuguese. This paper utilizes a typical case from the UCI machine learning repository, the Portuguese retail bank telemarketing campaign data from May 2008 to November 2010, to achieve these two goals. Considering the binary dependent variable, two popular classification prediction models are applied: the logistic regression and random forests with a comparison of their predictive power. Among the four open-sourcing datasets, the bank-additional-full file is selected as the target dataset, because it provides more observations and the most complete potential explanatory variables which could bring more information for analysis.

2. Exploratory Data Analysis

2.1 Data Information

There are 21 variables and 41,188 observations in bank data. In order to predict the characteristics of potential clients, whether the observed client has subscribed a term deposit (y) is chosen to be the response variable, and the rest are the potential explanatory variables. Y is an imbalanced two-levels categorical variable, most of the values (88.7%) taken are "no" and 11.3% of the values are "yes". In data processing, this paper transforms "no" as 0 while "yes" as 1. Among those potential explanatory variables, there are seven variables related to clients' personal information, which are age, job, marital, education, default, housing, loan. Four variables are related to the last contact of the current campaign, which are contact, month, day of week, and duration. Another 4 variables related to other attributes are campaign, pdays, previous, and poutcome. The rest five variables are related to social and economic context attributes, which are Emp.var.rate, Cons.price.idx, Cons.conf.idx, Euribor3m, and Nr.employed.

2.2 Missing Patten

There are 12,718 unknown values in this dataset. Six features (default, education, housing, loan, job, marital) have at least one unknown value (Table 1). Four features (default, education, housing, loan) have more than 330 missing values, which cannot be afforded. Due to countless losing information, these four variables are eliminated. As the information of the rest two (job and marital) are concerned, these two variables are kept with missing values deleted. Moreover, since the goal is to seek the best candidates who will have the best odds to subscribe to a term deposit, the call duration cannot be known before, hence this variable is removed from data.

Variable	unknown
default	8597
education	1731
housing	990
loan	990
job	330
marital	80

Table 1: Unknown variables

2.3 Multicollinearity

For five continuous variables in this dataset, Figure 1 shows that the emp.var.rate has a highly correlation coefficient with the other three variables (coms.price.ind, euribor3m and nr.employed). Therefore, the variable emp.var.rate is removed to avoid the multicollinearity.

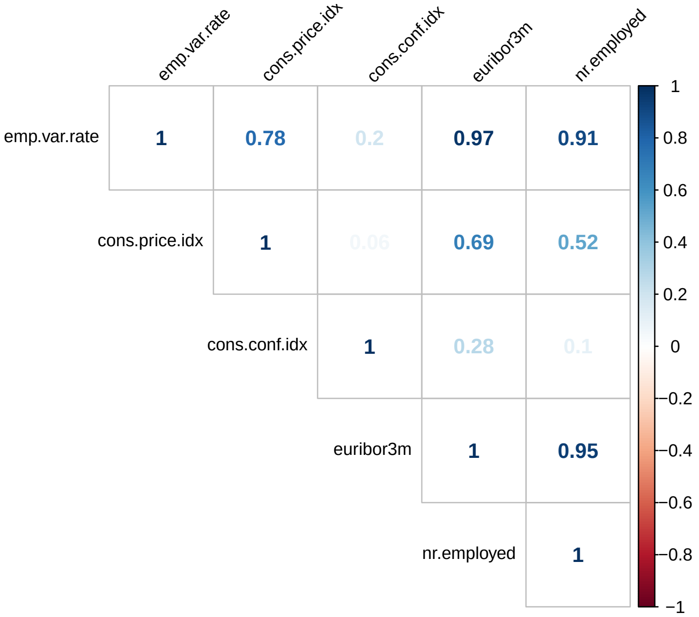

Figure 1: Correlation Matrix

Based on the distribution (Figure 2) and the amount of customers from different ages, this paper divides variable age into four groups: 30-, 30-45, 45-60, 60+.

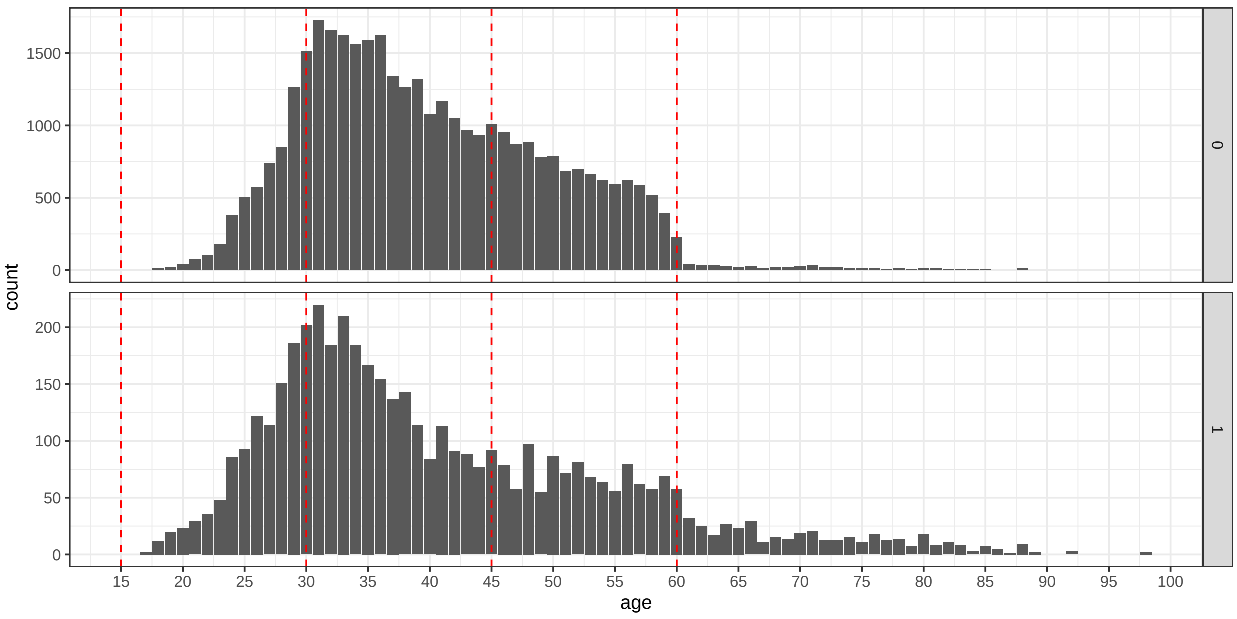

Figure 2: Distribution Plot of Age

For variable campaign (Figure 3), the amount of cases that a person received more than ten times calling is almost zero compared with the amount of cases that a person received less than ten times calling. As a result, this report eliminates the cases that a person received more than ten times calling and reduces the amount of factor levels.

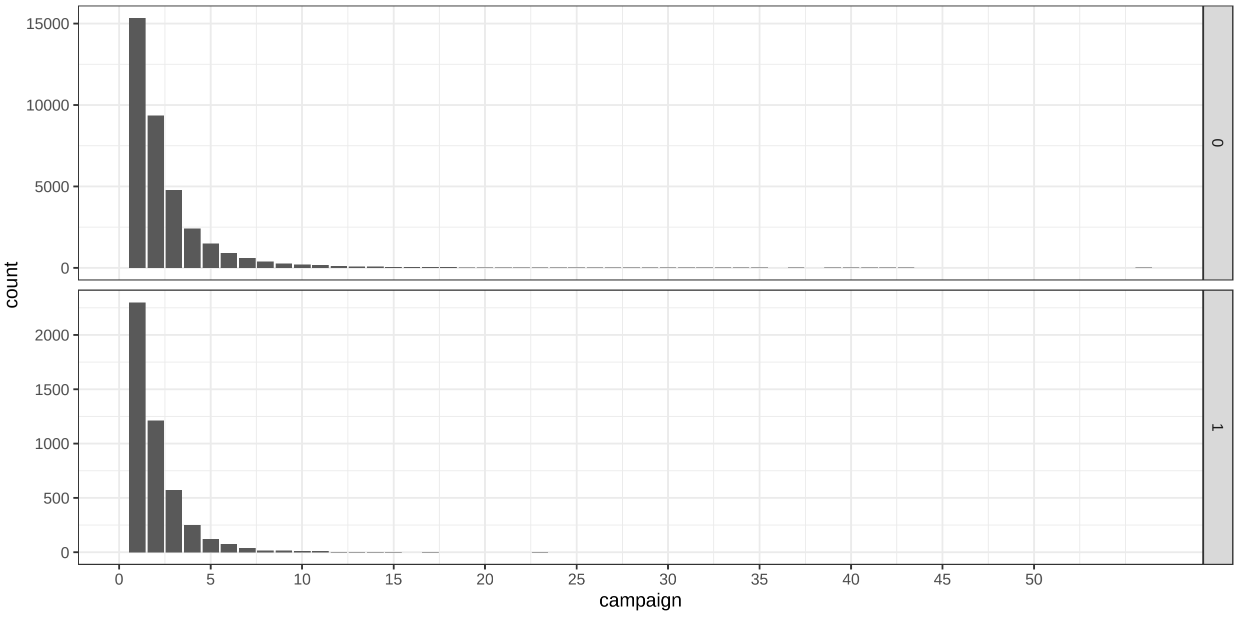

Figure 3: Distribution Plot of Campaign

In brief, after the data exploration analysis, there are overall 14 potential variables kept to do the subsequent model selection, which are age, job, marital, contact, month, day_of_week, campaign, previous, poutcome, cons.price.idx, cons.conf.idx, euribor3m, nr.employed, and pdays_dummy.

3. Predictive Models

In order to measure the prediction abilities of classified models, one can split data into two parts: training subset and test subset. The training data contains 80% observations, and the test data has the rest of the data (validation/test data). Both two have similar predict variables distractions respectively across subset.

3.1 Logistic Regression

3.1.1 Model

Logical regression is widely used in marketing applications, such as the prediction of a customer's propensity to purchase a product or halt a subscription (Berry, 1997). In this study, as most of the variables are discrete, a logistic regression is applied instead of linear regression, in order to do the classification. The logistic regression model is proposed as:

$$\ell = log\frac{p}{1-p} = \beta_0 + \beta_i x_i; i = 1, 2, 3...$$

Response variable y (whether a client has subscribed a term deposit) is a binary variable with $p = P(Y = 1)$.

$$p : P(Y = 1)$$

$$\beta_0 : \text{y-intercept}$$

$$\beta_i : \text{regression coefficient of ith variable}$$

$$x_i : \text{the ith potential predictor variable}$$

3.1.2 Model Fitting

- Initial model

According to the result of exploratory data analysis, 14 predictor variables are kept in the initial model.

Many variables do not significantly affect the subscription status in this model. Thus, after applying the step selection method, variables that are not relevant to the model are dropped. Now, the new model contains variable age, job, contact, month, day_of_week, campaign, outcome, cons.conf.idx, nr.employed, and pdays.

- Final model

Figure 4 shows the importance rank of these variables. From this figure, one can also observe that job and campaign do not show much importance while other variables are included in the model, even though these two are significant from the result of the step function. As a consequence, the final model excludes these two variables.

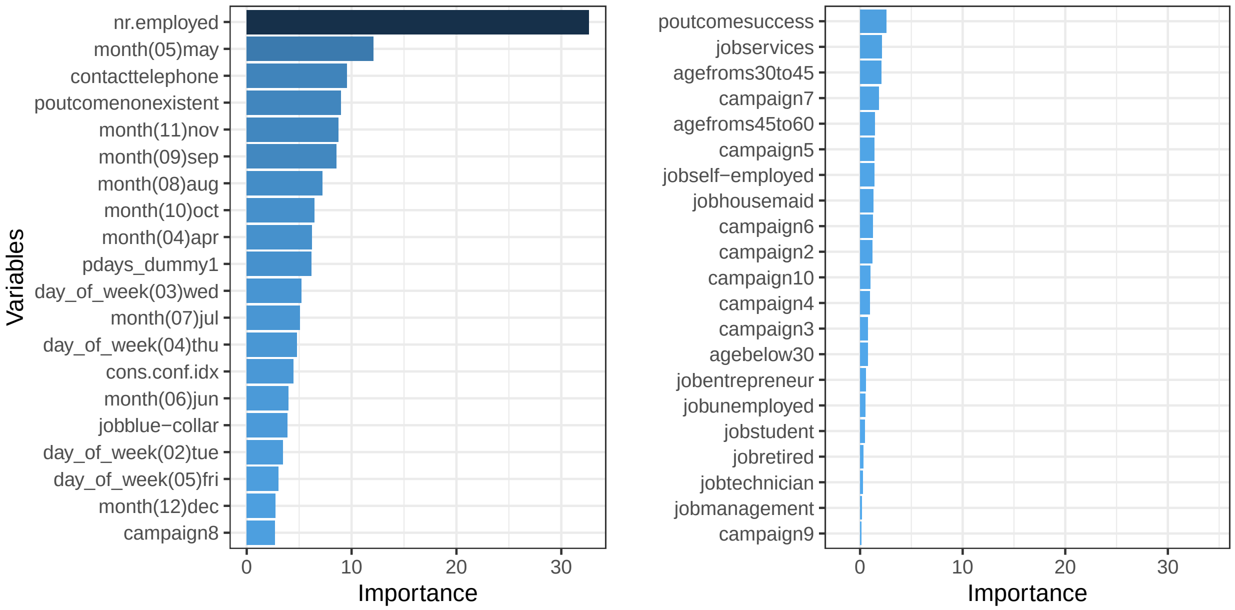

Figure 4: Variable Importance Plot of Logistic Regression Model

After that, predict scores are computed for better evaluating and validating the final model. Since the prediction of a logistic regression model is a probability, a cutoff value (threshold value) has to be chosen as a classifier. The default threshold is 0.5. Nevertheless, in this study, the 0.2 cut seems a better settlement (Figure 5).

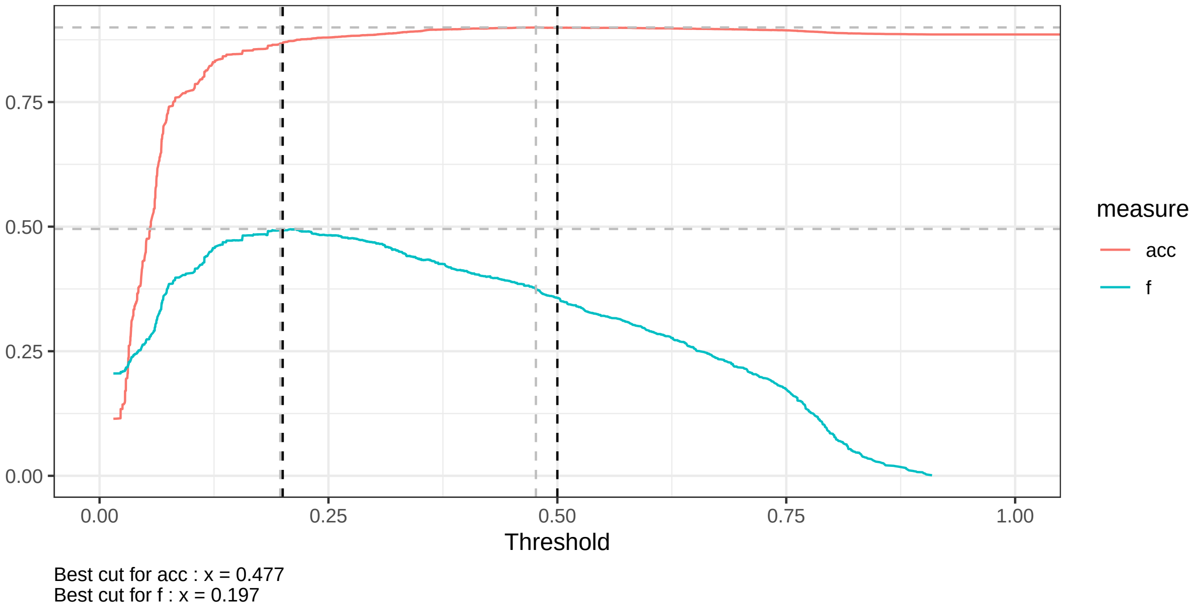

Figure 5: Threshold Plot of Logistic Regression Model

acc: accuracy rates (for accuracy)

f: F1 rates (for precision)

A good F1 score is preferred without dropping too much on accuracy (trade-off), and the 0.2 cut seems a good settlement.

To summarize the prediction results on a classification problem, a confusion matrix is required, which shows the sensitivity and accuracy of the model.

Table 2 shows that on the training set, the accuracy of the final logistic regression model reaches 86.76% and the sensitivity rate is close to 56.72%, which means that the model manages to correctly label 86.76% of the times and 56.72% of the willing customers are correctly detected.

Accuracy	Sensitivity
0.8676	0.56717

Table 2: Accuracy and Sensitivity of Logistic Regression Model

To evaluate this model, this paper analyzes the hold out (validation/test) data. The same procedures are executed, according to Figure 6, 0.2 is also chosen as the cutoff value. The comparison between Table 2 and Table 3 do not show an obvious overfitting in the final model as the performance values are close to the training values.

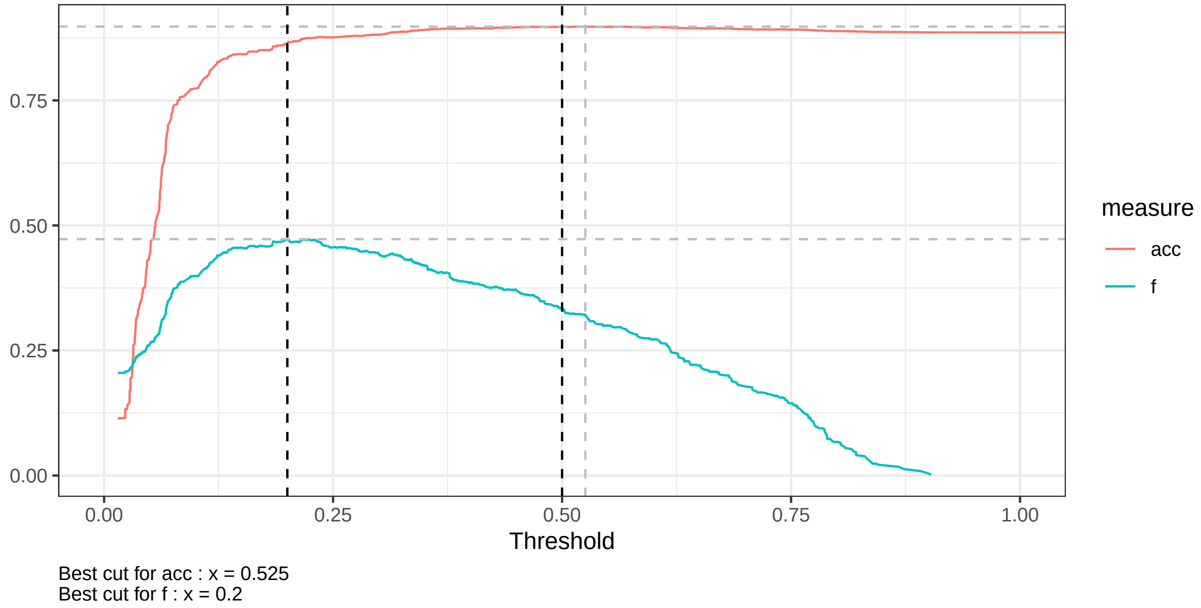

Figure 6: Threshold Plot of Initial Model

Accuracy	Sensitivity
0.8639	0.53231

Table 3: Accuracy and Sensitivity of Initial Model

3.1.3 Model Diagnostic

In order to diagnose the model, the goodness-of-fit, influential outliers, and multicollinearity are checked. The Hosmer-Lemeshow test (the HL-test) is used to check the goodness-of-fit with a null hypothesis that the observed and expected proportions are the same across all doses. The HL-test result shows the p-value is 0.001203, which rejects the null hypothesis at a significant level of 0.01, while cannot reject the null hypothesis at a significant level of 0.001.

The Cook's distance (Figure 7) is used to check whether there is an influential outlier. Since not all outliers are influential, one can also inspect the standardized residual error. Data points with an absolute standardized residuals above 3 represent possible influential outliers. The result shows that there is no influential outlier.

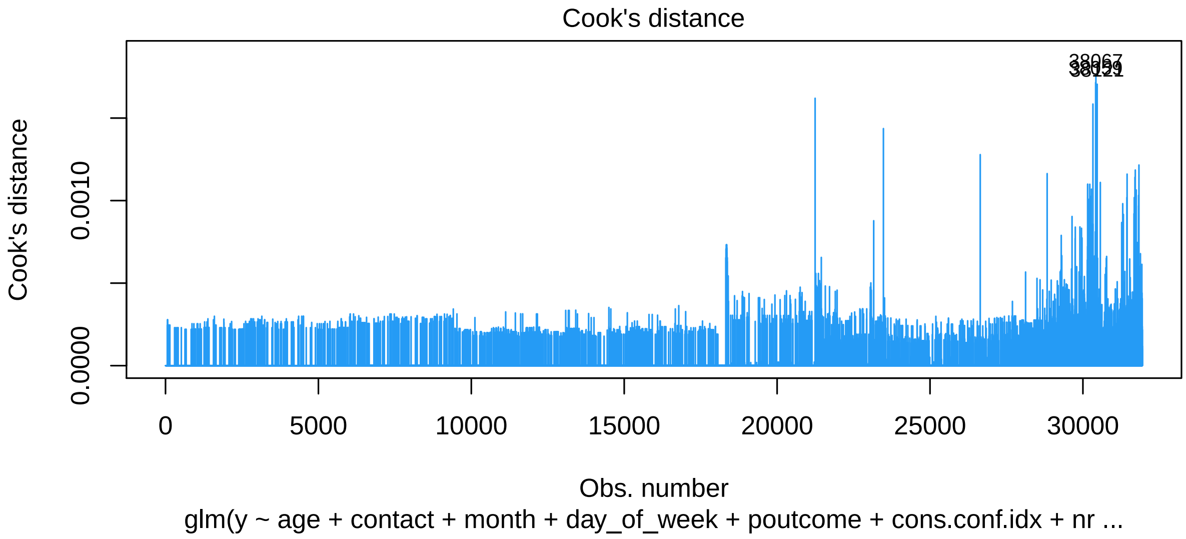

Figure 7: Plot of Cook's Distance for Each Observation

The output of Table 4 shows three columns: GVIF, Df, and GVIF^(1/(2xDf)) where it is calculated to check multicollinearity. Set GVIF ≥ 0 as the threshold of strong multicollinearity. From Table 4, one can observe that the GVIF of the outcome is greater than 10. However, consider that the outcome is a dummy variable, the interpretation of GVIF is different from VIF. In this case, this study refers to $GVIF^{\frac{1}{2*Df}}$.

Since $GVIF^{\frac{1}{2*Df}} = 2$ is usually considered to be equivalent to VIF = 4, one can say that the selected variables do not have any strong multicollinearity.

	GVIF	Df	GVIF^(1/(2*Df))
age	1.156307	3	1.024500
contact	1.436606	1	1.198585
month	3.716152	9	1.075652
day_of_week	1.037345	4	1.004594
poutcome	10.977510	2	1.820229
cons.conf.idx	1.975128	1	1.405393
nr.employed	2.084442	1	1.443760
pdays_dummy	9.775361	1	3.126557

Table 4: Generalized Variance Inflation Factors Calculated for the Explanatory Variables

We suggested using $GVIF^{\frac{1}{2*Df}}$, where Df is the number of coefficients in the subset of the variables. In effect, this reduces the GVIF to a linear measure, and for the VIF, where Df = 1, is proportional to the inflation due to collinearity in the confidence interval for the coefficient.

3.2 Random Forest

Decision tree is a popular method of creating and visualizing predictive models. In this paper, to compare the prediction ability of the logistic regression model and decision tree, data is set with discrete-valued target variables. Thus, a classification tree is proposed. However, when the classification tree depends on many of the irrelevant features of the training data, its predictive power will reduce due to overfitting (Bramer, 2013). To solve the overfitting problem, random forests are build instead of a single classification tree. Because of the Law of Large Numbers, random forests do not overfit and the proof is shown by Breiman in 2001.

In this paper, we use the ranger random forest algorithm, which utilizes the memory in an efficient manner with less runtime comparing to other random forest algorithms (Rao, et al., 2020). In the ranger model, the tuning parameters are the number of variables randomly sampled as candidates at each split (mtry), splitting rule, and minimal node size. As mtry value is usually set as a third of the number of variables by default, the maximum value of mtry is set to 70% of variables in this paper, and all potential mtry values are saved in a sequence to be compared. When choosing the splitting rules, considering random forests are built based on classification trees, gini index, extra trees, and Hellinger distance are considered. In addition, Hellinger distance can be a better measurement among these three rules due to the imbalanced data set (Chaabane, et al., 2019). In order to reduce the runtime, the minimal node size is set to a constant one. Figure 8 shows the cross-validation scores comparing the different mtry values and splitting rules. One can also notice that to attain the best prediction result, mtry value is set to be 3 with gini index as the splitting rule.

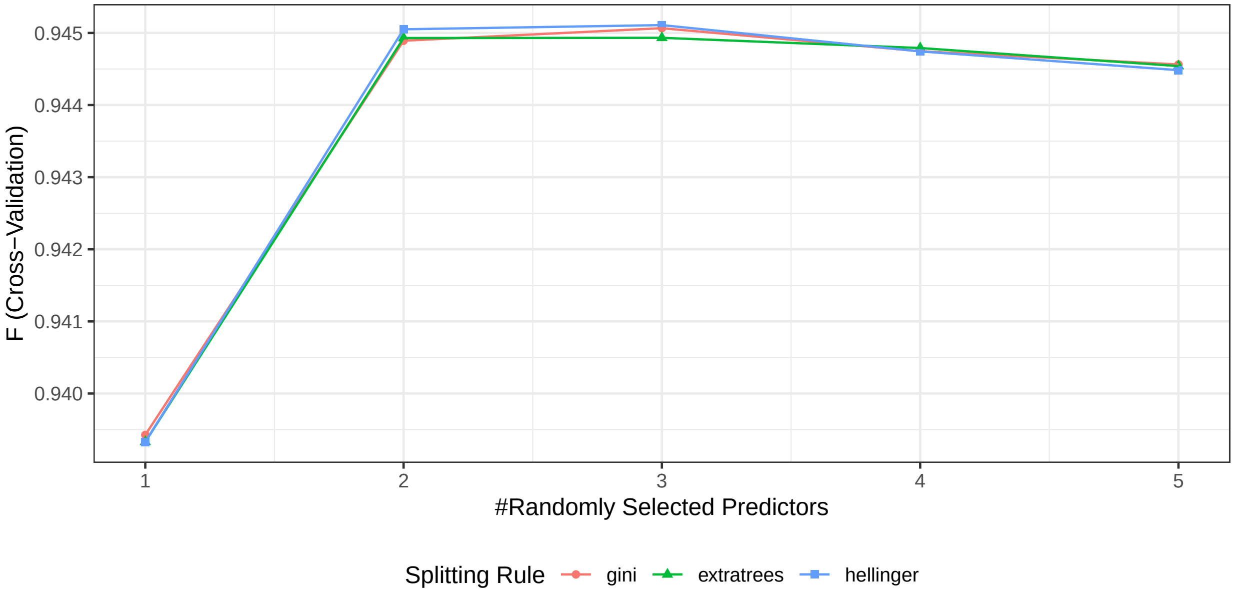

Figure 8: Plot of Gini Index, Extra Trees, and Hellinger Distance

Figure 9 provides the visualization of the importance score among all variables, and the result shows a slight difference comparing to the significant variables selected by logistic regression.

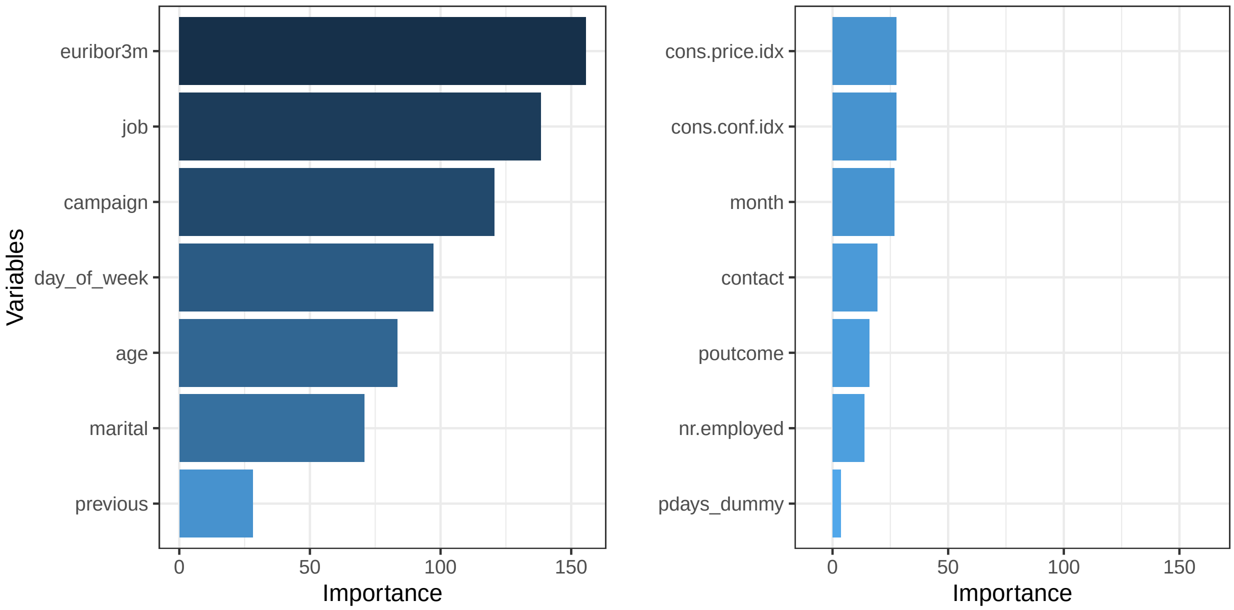

Figure 9: Variable Importance Plot of Random Forest Model

Similar to testing the sensitivity and specificity of the logistic regression model, the cut-off point is chosen as 0.3. Comparing to the logistic regression model, the accuracy, sensitivity, and specificity are all improved by random forests (Table 6). In the next section, the prediction ability of logistic regression and random forests will be compared in detail.

Type	Probability estimation
Number of trees	1000
Sample size	31945
Number of independent variables	14
Mtry	3
Target node size	1
Variable importance mode	impurity
Splitrule	1
OOB prediction error(Brier s.)	0.07

Table 5: Random Forest Model Object

3.3 Support Vector Machine - Extra Credit

Support Vector Machine, or SVM, is a nonparametric supervised learning model that constructs hyperplanes which can be used for classification, regression or other tasks. The purpose of SVM is to identify a line that separates the two classes. When there are more than one possible hyperplane, we search for the Maximum Margin Hyperplane that creates the greatest separation between the two classes. A good separation is achieved by the hyperplane that has the largest distance to the nearest training data points (Guenther, 2016). The SVM model is very effective in high dimensional spaces, which is particularly applicable for our dataset since it is large.

3.3.1 Initial model

		Reference	
Prediction		no	yes
	no	644	56
	yes	94	6

Table 6: Confusion Matrix

Accuracy	Sensitivity
0.8125	0.87263

Table 7: Accuracy and Sensitivity of Initial Model

Here we notice that the accuracy is less than the random forest, but the sensitivity is a lot higher. It means that the proportion of observed positives that were predicted to be positive is 0.87263.

For SVM, we can adjust Cost and Gamma to improve the model. For the above model, we are trying to find a hyperplane, which perfectly divide the data points into two parts, which is called "hard margin SVM". But it is very risky since it could be overfitting. As a result, we can allow some cost, which is the wrong prediction made by this model, exists. With C, we can add penalty to those data we classify wrong. Then we can control the influence of support vectors. The smaller the C, the margin is narrower, which is closer to the hard-margin SVM. The greater the C, the margin is wider, which means to allow more error in order to create a better model. For the gamma, there are three types: linear, polynomial, and RBF. Basically, we are mapping those overlapping data into a higher dimensions in order to classify them better. "Linear" and "polynomial" are like creating a linear/ polynomial function to map data points into higher dimension. But RBF is like creating a function to measure the relativity between each data points by their distance. The formula is $e^{(-(\text{gamma})(\text{a-b})}2)$. If the gamma is small, it would be a Gaussian function with large variance. Thus, even two data points are far apart, they could be consider similar. If the gamma is large, it would be a Gaussian function with small variance. Thus, even two data points are very close, they could be consider different.

3.3.2 Tune the model

By using tune(), we are able to find the best model with optimized cost and gamma. As shown in Figure 10, the performance of SVM has a turning point at 10, therefore the cost is equal to 10. It also suggest that the best model will be radial.

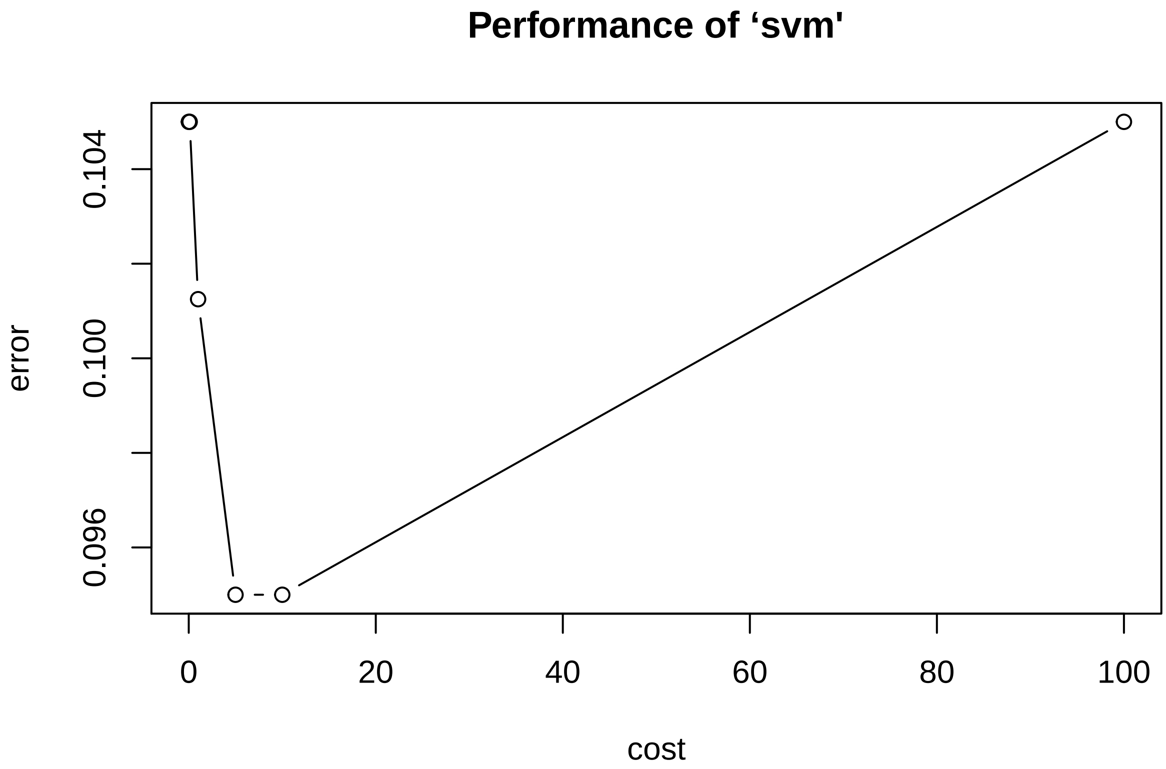

Figure 10: Plot of SVM

3.3.3 Test model

		Reference	
Prediction		no	yes
	no	168	7
	yes	12	13

Table 8: Confusion Matrix

Accuracy	Sensitivity
0.905	0.9333

Table 9: Accuracy and Sensitivity of Test Model

Table 9 shows accuracy of 90.5% and sensitivity of 93.33% which means the model manages to correctly label 90.5% of the times and 93.33% of the willing customers are correctly detected.

3.3.4 Larger trainning set

It seems like SVM is a good idea, but do not forget that we just throw 1000 observations into the model in order to run faster. We learn that SVM doesn't perform as well in a imbalanced dataset as it does in a balanced dataset. This is a very imbalance dataset, so we need to check its performance with more observations. This time we throw 5000 observations into it.

		Reference	
Prediction		no	yes
	no	842	36
	yes	63	59

Table 10: Confusion Matrix

Accuracy	Sensitivity
0.901	0.9304

Table 11: Accuracy and Sensitivity of Larger Trainningset Model

This time we throw 7000 observations into it and let check the confusion matrix and accuracy and sensitivity.

		Reference	
Prediction		no	yes
	no	1215	33
	yes	95	57

Table 12: Confusion Matrix

Accuracy	Sensitivity
0.9086	0.9275

Table 13: Accuracy and Sensitivity of Larger Trainningset Model

Table 12 and Table 13 shows that on the larger training set, 5000 and 7000 observations respectively, the accuracy of the SVM model reaches up to 90% and the sensitivity rate reaches up to 93%. That means the model manages to correctly label 90% of the times and 93% of the willing customers are correctly detected.

4. Conclusion

Model	AUROC	AUPR	Cut	Accuracy	F1
Random forest (ranger)	0.79	0.46	0.3	0.88	0.48
Logistic regression (final)	0.78	0.43	0.2	0.86	0.47

Table 14: Comparison Table of Random Forest vs. Logistic Regression Model

Model	accuracy	sensitivity
Logistic	0.8639	0.53231
RF	0.8125	0.87263
SVM	0.905	0.9333

Table 15: Accuracy and Sensitivity Table of Logistic, RF, and SVM Models

The ultimate goal of growth for a bank is to reduce the cost while increase revenues in an effort to boost profitability. In order to reduce the costs and increase profits, it is essential to find target clients for telemarketing, and so does an appropriate statistical model that can support bank campaign managers for client selection. In this report, a dataset about the direct phone call marketing campaigns which aim to promote term deposits among existing customers, by a Portuguese banking institution is used.

Throughout the reports, we analyze a total of 39,930 records with 15 explanatory variables. There are two approaches which are Logistical Regression (LR) model and Random Forests (RF) that proposed to predict long-term deposit subscription based on these 15 features. F1 scores and Receiver Operating Characteristic Curve (ROC) curve are used as the indicator to determine the prediction ability. For both indicators, RF shows a better result with ROC score 79% (vs. 78%) and F1 score 48% (vs. 47%). F1 score links to two measurements, which are precision and recall. Precision indicates that among all the clients who subscribe, the model manages to correctly label near half of them. Recall indicates near half of those who have been labeled have a tendency to subscribe indeed.

There are three potential reasons to explain why the RF model has a more power prediction ability in this case. First of all, this is an imbalanced dataset, with 88.7% of values taken “no” (or 0), and only 11.3% of the values “yes” (or 1). In other words, there is a small ratio of positive cases (1) to negative cases(0). Second, logistic regression performs worse under high dimensionality conditions. Third, there is no linear separation of classes along with the given variables. Therefore, those three characteristics of the given dataset should be able to explain the gap in the performances of the two models. In future work, one can run more models with different algorithms to test which one is better for particular scenarios. Also, it would be interesting to include time series as a variable to analyze the impact of a hard-hit recession in contrast to slow recovery.

Results for SVM

For SVM, it has a fairly high accuracy and sensitivity comparing to logistic regression model and random forest model. The main reason is that this dataset has high dimension, and SVM performs good on high dimension data. Since it is finding a hyperplane, and it is not sensitive to the dimension. But as we throw more observations into the model, its accuracy and sensitivity also decrease, but very slow. Therefore, SVM works fine on this unbalance dataset. Part of the reason is that this dataset is very sparse. Since the hyperplane is related to the support vectors, and data far away from the plane can’t influence the hyperplane.

References

Berry, M. J. A (1997). Data mining techniques for marketing, sales, and customer support, Wiley, 10.

Bramer, M. (2013). Avoiding overfitting of decision trees, Principles of Data Mining, Springer, London, 121-136.

Breiman, L. (2001). Random forest, Machine Learning (45), 5-32.

Chaabane, I, Guermazi, R., Hammami, M. (2019) Enhancing techniques for learning decision trees from imbalanced data, Springer.

Guenther N., Schonlau M. (2016). Support vector machines, 917-937.

Rao G.M., Ramesh D., Kumar A. (2020) RRF-BD: Ranger random forest algorithm for big data classification. Computational Intelligence in Data Mining. Advances in Intelligent Systems and Computing, vol 990. Springer, Singapore, 15-25.

Theodoros Evgeniou, Massimiliano Pontil. (2001) Support Vector Machines: Theory and Applications