Targeting Customers for Long Term Deposit

Team ID: Team 6

NAME: Connor Rosenberg

NAME: Rongkui Han

NAME: Yuqing Yang

NAME: Nassim Ali-Chaouche

---

1 Introduction

1.1 Background

This study considers real data collected from a Portuguese retail bank, from May 2008 to November 2010, in a total of 41,188 phone contacts. This financial campaign focused on targeting through telemarketing phone calls to sell long-term deposits. The response variable of the dataset is a binary successful or unsuccessful contact. Marketing selling campaigns constitute a typical strategy to enhance business. Technology enables rethinking marketing by focusing on maximizing customer lifetime value through the evaluation of available information and customer metrics, thus allowing organizations to select the best set of clients, i.e., that are more likely to subscribe a product (S. Moro, P. Cortez and P. Rita., 2014). The goal of this study is predict if the client will subscribe a long-term deposit using demographic, campaign and economic information.

Many versions of the dataset was made available on the (UCI Machine learning repository). The one selected for this analysis contained all samples and the highest number of variables. This was done to maximize the information input and return the best predictions possible. The dataset is unbalanced, as only 4.640 (11.27%) records are related with successful outcomes.

It must be stressed that for the purpose of this analysis, a false negative error (Type II error) in prediction incurs a much higher cost than a false positivce error (Type I error). If a customer unlikely to subscribe to the product is falsely identified as a potential customer (false positive), it only costs the bank caller a few minutes to make an unsuccessful phonecall; on the other hand, if a true potential customer is dismissed by the model as an unlikely target, the bank runs the risk of losing thousands of dollars of deposits the customer could put into the bank. Therefore, we make it a priority in our study to build the model that has the highest recall rate for true positive cases.

1.2 Questions of Interest

- Can we predict if a client will subscribe a long-term deposit using demographic, campaign and economic information?

- What model, among (a) logistic regression, (b) random forest classification and (c) naive Bayes classification, perform the best in terms of capturing the highest number of true successful outcomes in its prediction?

2 Analysis Plan

2.1 Population and study design

The population of interest in this study is all adults with the possibility of subscribing a long-term deposit with a specific Portugese retail bank. Using the outcome of the bank's telemarketing campaign, a "yes" or "no" to subscribing the deposit product, as the response variable, and different subsets of 20 demographic, campaign, and economic variables as predictor variables, we hope to build a model to predict whether a new potential customer will subscribe to the long-term deposit.

The "duration" variable, last contact duration in seconds, was dropped from the dataset. This attribute is highly correlated with the output target, yet the duration is not known before a call is performed. Thus, this input was discarded to have a realistic predictive model. Among the variables used in the analysis, demographic predictor variables include: age, job, marital status, education level, presence of credit in default, if they have housing loan, abd if they have personal loan. Predictor variables related to the campaign include: communication type, month, day of the week, number of contacts performed, number of days since customer contacted from previous campaigns, number of contacts before the campaign, and outcome of previous campaigns. Social and economic variables include: quarterly employment variation rate, monthly consumer price index, monthly consumer confidence index, daily euribor 3-month rate, and quarterly number of employees.

After a brief inspection of the data, we find a heavy imbalance between respondents who signed up for the long-term deposit and those who did not (Table 1). Over 88% of called customers did not sign up for our product. Because of this imbalance, models are much more likely to classify a potential customer as "no", solely because of their greater presence in the dataset. To help offset this bias, we will oversample our training data to equalize the proportion of customers who did and did not sign up.

Table 1: Imbalanced response variable.

	Response: Yes	Response: No
Percentage	11.27%	88.73%

Common prediction performance metrics include recall/sensitivity, accuracy, precision, and specificity. The definition and differences are demonstrated in Table 2. For reasons stated in the introduction, we set our model performance metric to be **recall/sensitivity**.

Table 2: Definition of recall/sensitivity, precision, and specificity.

	Reference Positive	Reference Negative	
Predicted Positive	a: True positive	b: False positive, Type I error	$Precision = \frac{a}{a+b}$
Predicted Negative	c: False negative, Type II error	d: True Negative	
	$Recall = Sensitivity = \frac{a}{a+c}$	$Specificity = \frac{d}{b+d}$	$Accuracy = \frac{a+d}{a+b+c+d}$

We will compare three different types of models: logistic regreession, random forest, and naive Bayes in their prediction performance. To build the models, we will:

1. Randomly select 70% of the datapoints as the training dataset, leaving the rest 30% as the testing dataset.
2. In order to accommodate the unbalanced outcome variable, upsample the training dataset so the outcome variable had equal counts of "yes" and "no".
3. train the model on the upsampled training data.

4. evaluate the model recall/sensitivity on the unaltered testing dataset.

2.2 Descriptive Analysis

To better understand the relationships between each predictor and the proportion of called customers who signed up, we will focus our exploratory analysis on the change in the proportion of these customers who sign up for the long-term deposit. This will provide a starting point for variable selection when constructing our predictive models.

2.3 Model building

2.3.1 Logistic Regression

Logistic regression is suitable for analyzing datasets with binary/categorical response varaibles. The logistic regression model used for prediction will be:

$log(\frac{\vec{\pi}}{1-\vec{\pi}}) = X^T\vec{\beta}$

where π is a column vector with $\pi_i = P(Y_i = 1)$, probability of positive outcome for subject i; X is a $n \times p$ matrix of predictor variables; and $\vec{\beta}$ is the effect of each predictor variable.

We will use three different subsets of predictor variables for the logistic regression model: (a) a subset of predictor identified by a LASSO regression process, (b) a subset identified by a ridge regression process, and (c) all predictor variables available in the dataset. When deployed for prediction, cases with predicted $\pi^* \geq 0.5$ are identified as "positive", and the ones with predicted $\pi^* < 0.5$ are identified as "negative".

2.3.2 Random Forest Classification

We also use the random forest to build the prediction model. Random forest is a tree-based ensemble learning method. It uses a modification of the bagging technique, which could improve models in terms of stability and classification accuracy, reduces variance and helps to avoid overfitting. There are two sources of randomness in the random forest method. On the one hand, each tree is grown based on a bootstrap resampled data set. On the other hand, each time a split is to be performed, the search for the split variable is limited to a random subset of variables.

Since the outcome of the data set is very imbalanced, weights for the sampling of training observations were assigned proportionally to the inverse of its frequency. Observations with larger weights will be selected with a higher probability in the bootstrap samples for the trees.

2.3.3 Naive Bayes Classification

Another model to consider is the naive Bayes classifier. This model, based on Bayes' theorem, provides a way to calculate the posterior probability of each observation. From these posterior probabilities, we can predict whether a potential customer will sign up for a long-term deposit.

The naive Bayes classifier is a very versatile model as it can both handle categorical and continuous variates. The model assumes *class conditional independence* between our predictors on the response (Rish, 2001). That is, there are no interactions between predictor variables and the effect of each predictor on the response class is independent of all other predictors. Even though this assumption is rarely certified in practice, the Naive Bayes Classifier has shown to be an effective classifier regardless if the conditional independence assumption truly holds in the data (Rish, 2001). Since we are only concerned with the predictive capability of our model, we can move forward with the Naive Bayes Classifier even though the model assumption is not strictly certified.

We will present two Naive Bayes Classification models. The first model, **the full model**, will be fit on the full list of predictors from the data set. The second model, **the optimal model**, will include only the variates which returned the highest proportion of true-positive outcomes when tested on the validation dataset.

2.4 Model Diagnostics

A scatterplot between each continuous predictor variable and the log odds of the response variable will be used to examine the assumption of the logistic regression model that the predictors are linearly related to the log odds of the response variable. VIF values will be used to examine the assumption that there is low multicollinearity among the predictors. Influential observations will also be discussed.

2.5 Model Evaluation

For the logistic regression model, we will rank the importance of predictors in the logistic regression based on the absolute value of the t-statistics corresponding to each predictor. A Pseudo R^2 measure will be used to test the goodness-of-fit of the model.

Across the different models, the recall of the models will be assessed, and the areas-under-curve of the ROC and Precision-Recall curves will be used to compare the predictive abilities of the models.

3 Results

3.1 Descriptive Analysis

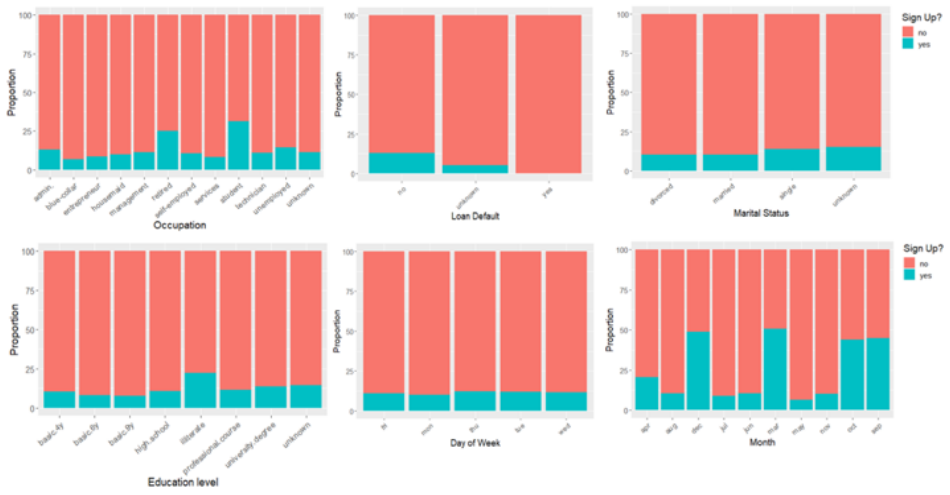

Figure 1: proportion of responses in each category of six predictor variables

To best leverage a predictive model to target potential customers, we want to ensure that the proportion who sign up for our long-term deposit product varies across the different levels. If this proportion remains constant across all levels of each variate, it would be useless to deploy a predictive model as each customer would have an equal chance of signing up.

In our case, a predictive model will help our employees better target potential customers since the proportion of those called who signed up change between levels of variates (Figure 1). We can see that students and retirees have a much higher probability of signing up compared to blue-collar and service workers. Furthermore, we can see that customers called in december, march, October, and September have a much higher proportion of sign-ups compared to the remaining months.

Conversely, we can see that variates, like the day of the week, do not have a dramatic change in proportion between the different days. It also appears the distribution of age is not related to the probability they sign up (Figure 1).

While our exploratory analysis provided a bit of insight into which characteristics make consumers more likely to sign up for our new product, these insights may not be included in our final model. With the main goal of prediction, we will select the model which performs the best on our testing dataset.

3.2 Model Fitting and Prediction Performance

The size and response variable balance is displayed in Table 3.

Table 3: Sizes and distributions of the response variable in the original training dataset, upsampled training dataset, and testing dataset

	Total	Response: Yes	Response: No
Original training	28832	3262	25570
Upsampled training	51140	25570	25570
Testing	12356	1378	10978

3.2.1 Logistic Regression

LASSO and ridge regularizations were used to eliminate predictor variables unlikely to contribute to the model. LASSO regression with optimum $\lambda = 0.000262$ (λ is a weight parameter for penalizing additional variables) eliminated economic variable "nr.employed", the quarterly updated count of number of employers, from the set of predictor varaibles. Ridge regression with optimum $\lambda = 0.0233$ did not eliminate any variable from the list.

Logistic regression fitted without the "nr.employed" variable correctly predicted 839 out of 1378 true positive cases in the testing dataset, resulted in $Recall = 0.6081$. With the "nr.employed" variable, logistic regression correctly recalled 840 out of the 1378 true positive cases, $Recall = 0.6089$, slightly higher than the LASSO-regularized model. Based on the absolute value of the t-statistic corresponding to each predictor, the most important variables in the regression are the quarterly employment variation rate, the monthly consumer price index, and contact method.

3.2.1.1 Model Diagnostics of Logistic Regression Scatterplots between the log odds of the response variable and the continuous predictor variables suggested non-linear pairwaise relationships. Under time constraints we did not have the oppurtunity to try transformations of the variables.

The McFadden Score, which is a measure of goodness-of-fit/reduction in deviance, is 0.23. According to McFadden (1977), values of 0.2 to 0.4 represent an excellent fit. Cook's distance was calculated for all entries and no outliers were identified. Most variables have VIF values of less than 10. A few variables ("emp.var.rate", "euribor3m", and "nr.employed") have VIF values very slightly above 10. Overall, multicollinearity does not pose a big issue in the analysis.

3.2.2 Random Forest Classification

All the variables except 'duration' were included when building the model. The scale permutation importance by standard error for each variable were shown in Figure @ref(fig: ipt). It can be seen that variables campaign,

marital, default, pdays, and contact have negative permutation importance. So these variables were dropped to fit a new random forest model.

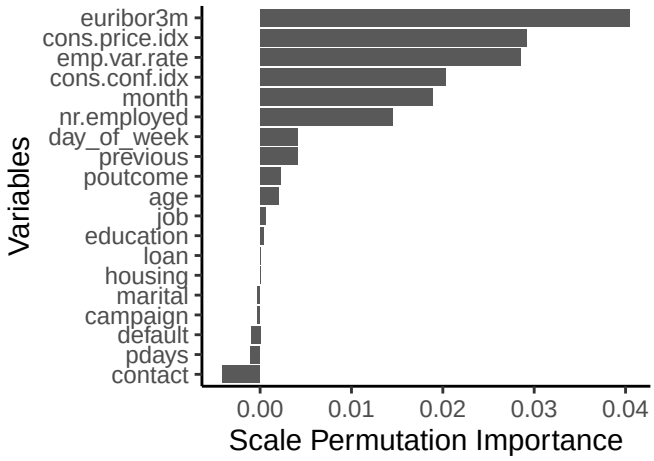

Figure 2: Permutaion Feature importances in the Random Forest Model

Since the new model has a higher sensitivity, we choose the new model as the final version for the random forest method. The final model correctly predicted 781 out of 1378 true positive cases ($Recall = 0.5668$).

3.2.3 Naive Bayes Classification

From the full Bayes model, which used all 20 predictors, our optimal model outperformed the full model while reducing its size to 11 variates. The model used consumer information regarding age, marital status, job, education level, default, housing, and loan; in addition to the economic variables employment variation rate, consumer price index, Euribor 3 month rate, and the number of employees to make the conditional predictions on a customers response.

The optimal Bayes model achieved a recall of 0.7025, which improved performance by a full 5% over the full model (Table @ref:(tab:compare)).

4 Discussion

The ROC curve is largly uninformative because all four models follow approximatly the same path and have the same area under the curve. From the Precision-Recall curve, we can observe that each of the four models has a slightly different performacne on the testing dataset (@ref:(fig:curve)). If we cared equally about the precision and recall, calcualting the area under this curve would be the most appropriate performacne metric. This metric would show that logistic regression and random forest models outperformed both Naive Bayes models. Because we want to avoid false negatives as all cost, recall, on its own, is much more powerful measure. With that measure, the optimal Naive Bayes perfomred the highest.

The Naive Bayes is a classification algorithm based on Bayes rule and a set of conditional independence assumptions. The algorithm makes the assumption that each variate is conditionally independent of all other predictors given the response variable. Logistic Regression assumes a parametric form for the distribution, $P(Y|X)$, to directly estimates the parameters. The parametric form used by Logistic Regression is the same form implied by the Guassian Naive Bayes approach, and thus logistic regression is a close alternative to it. When the assumptions of the Guassian Naive Bayes model do not hold, Logistic Regression and Guassian Naive Bayes "typically learn different classifier functions" (Mitchell, 2015). Thus, the gap between performances of the Logistic Regression ($Recall = 0.6089$) and Naive Bayes ($Recall = 0.6089$) models can be attributed to the data not following all the assumptions of Logistic Regression. According to Ng and Jordan (2002), both Logistic Regression and Guassian Naive Bayes models approach their asymptotic accuracies

at different rates. As "the number of training examples m is increased, one would expect generative naive Bayes to initially do better, but for discriminant logistic regression to eventually catch up to, and quite likely overtake, the performance of Naive Bayes." Thus, it is possible that if we had a larger sample size for the training set, the performance with Logistic Regression would be better than that of the Naive Bayes approach. In terms of Random Forest, it is an ensemble-based learning algorithm which is comprised of n collections of de-correlated decision trees, and uses multiple trees to average or compute majority votes in the terminal leaf nodes when making a prediction (Kirasich et al., 2018). Thus, it is a completely different approach compared to Logistic Regression and Naive Bayes. Couronne et al. 2018 state that "the superiority of Random Forest tends to be more pronounced for increasing" the number of predictor variables and the ratio of predictors to its sample size. Since the number of predictors in our data set is relatively small, less than 20, this is most likely the reason why the Random Forest ($Recall = 0.5668$) did not perform as well as Logistic Regression and Naive Bayes.

The optimal Naive Bayes model outperformed all other models regarding recall of positive cases (Table 4). Positive cases were the point of focus of this study because of the high cost of false negative errors and the low impact of false positive errors. Moving forward, implementing this model in our business stucture will allow our marketing employees to make educated decisions about the clients they target. After its deployment, we expect to see an increase in enrollment. Furthermore, by reducing the number of calls to unlikely customers, we will increase our overall customer satisfaction and maintain a more positive relationship with our potemntial customers for future products.

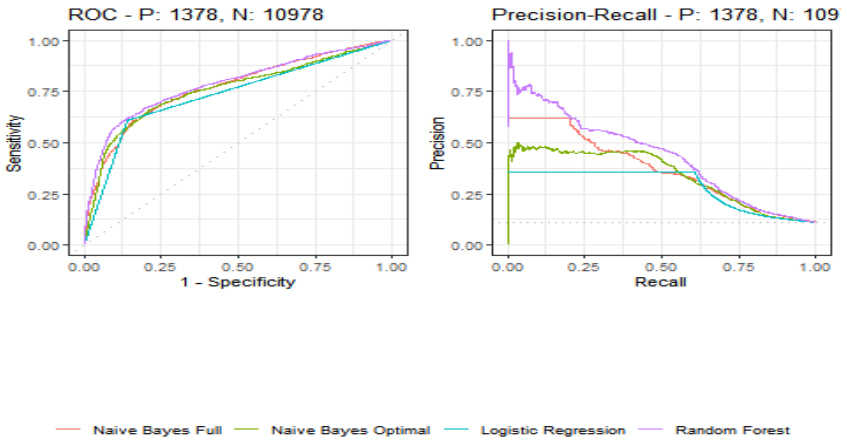

Figure 3: ROC and Precision-Sensitivity curves for all four models: Logistic Regression, Random Forest, Naive Bayes Full, and Naive Bayes Optimal

Table 4: Prediction performance of four models

Model	class	precision	recall
Logistic	no	0.95	0.86
	yes	0.35	0.61
Random Forest	no	0.94	0.90
	yes	0.43	0.57
Naive Bayes Full	no	0.95	0.79
	yes	0.28	0.65
Naive Bayes Optimal	no	0.95	0.72
	yes	0.24	0.70

5 Reference

Couronne, R., Probst, P., & Boulesteix, A.-L. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics, 270(19). Retrieved from https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2264-5

Kirasich, K., Smith, T., & Sadler, B. (2018). Random Forest vs Logistic Regression: Binary Classification for Heterogenous Datasets. Retrieved from https://scholar.smu.edu/cgi/viewcontent.cgi?article=1041&context=datasciencereview

McFadden, D. (1977, November 22). Quantitative Methods for Analyzing Travel Behaviour of Individuals: Some Recent Developments. Retrieved from http://cowles.yale.edu/sites/default/files/files/pub/d04/d0474.pdf

Mitchell, T. M. (2017, September 23). Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression. Retrieved from https://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf

Ng, A. Y., & Jordan, M. I. (2002). On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes. Retrieved from http://ai.stanford.edu/~ang/papers/nips01-discriminativegenerative.pdf

Rish, I. (2001). An Empirical Study of the Naïve Bayes Classifier. IJCAI, 41–46.

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