**Introduction**

Marketing selling campaigns include typical strategies to enhance business. In order to increase a financial asset, companies use various ways to contact customers and ask them to purchase their products. There are two main ways to reach customers. First one is through mass campaign. That is, companies try to contact as many customers as possible without screening. The drawback with this method is that, the success rate is typically very low. The second method is to target customers with specific characteristics. Basically, companies will identify which customer is more likely to buy their products and contact them. The success rate of this strategy is usually higher than mass campaign. However, the drawback of this approach is that customers’ personal information is released, resulting in the loose of trust.

**Dataset**

The current study analyzes the dataset of campaigns where retail bank persuades their customers to buy the deposit term. After the term ends, customers will get their initial investment back plus the interest. There are three campaign in each year. Within each campaign, the human agents execute phone calls to a list of clients to sell the deposit. This study uses real data collected from a Portuguese retail bank from 2008 to 2010. There are 41188 observations and the observations are order by date. I split the dataset into 3 sub-datasets, the year of 2008, 2009 and 2010. For each dataset, the statistics is different from each other. For example, in terms of the success rate (Table 1), campaign 2008 has only success rate of 4.84%. The success rate increases in the following campaign with 19.48% in 2009, reaching up to 52.14% in 2010. I suspect that three campaigns used different strategies to target customers. So I expect that the first two campaigns use mass campaign strategy, where companies contact as many customers as possible and the third campaign uses some screening procedures, where company uses certain characteristics to target potential consumer, resulting in a very high success rate. To back up my speculation, here are some differences in customer characteristic among three campaigns. For example, for the variable of previous sell record (Table 2), campaign 2008 only reached 0.13% customer who successfully subscribed previous deposit term. However, this number increases up to 31.49% in the campaign 2010. Another predictor is marital status (Table 3). We can observe that more single customers were targeted in 2010 campaign (41.3%) than in 2008 campaign (24.6%). I used campaign 2010 as my dataset because the outcome variable is more balanced than campaign 2008 and 2009.

Table 1. The percentage of success outcome and failure outcome for each year

|  |  |  |
| --- | --- | --- |
|  | Failure | Success |
| 2008 | 95.16% | 4.84% |
| 2009 | 80.52% | 19.48% |
| 2010 | 47.86% | 52.14% |

Table 2. The percentage of outcome of the previous campaign for each year

|  |  |  |  |
| --- | --- | --- | --- |
|  | Failure | Nonexistent | Success |
| 2008 | 2.52% | 97.35% | 0.13% |
| 2009 | 26.10% | 67.87% | 6.03% |
| 2010 | 27.55% | 40.96% | 31.49% |

Table 3. The percentage of marital status for each year

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Divorced | Married | Single | Unknown |
| 2008 | 11.68% | 63.54% | 24.61% | 0.17% |
| 2009 | 10.10% | 55.52% | 34.12% | 0.26% |
| 2010 | 10.79% | 47.76% | 41.30% | 0.15% |

One reason of conducting statistical analysis is to understand the story behind datasets. In another word, what can we learn from data? In the current analysis, I want to answer the questions in the Table 4. This would help us to decide which customers are more likely to purchase the deposit.

Table 4. Business questions to answer

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Questions | |
| 1 | | Is age relevant? | |
| 2 | | Is job relevant? | |
| 3 | | Is marital status relevant? | |
| 4 | | Is education relevant? | |
| 5 | | Is default relevant? | |
| 6 | | Is housing relevant? | |
| 7 | | Is loan relevant? | |
| 8 | | Is type of contact relevant? | |
| 9 | | Is the number of contacts performed during this campaign relevant? | |
| 10 | | Is the number of contacts performed before this campaign relevant? | |
| 11 | | Is the outcome of previous marketing campaign relevant? | |

**Result**

There are 2058 observations in the year of 2010. These observations were randomly divided into training set, 80% and test set, 20%. After the models are trained on the training set, their performance will be compared based on the test set.

**Predictor selection**

To answer the business questions, I use logistic regression (LR) to predict the outcome variable by each single predictor in Table 4. The result is in Table 5. We can see that only some of the variables predicted the customer’s decision better than guessing. We can use this result to answer previous questions in Table 4. For example, customers’ past purchasing behavior predicts their current purchasing behavior. That is, customers, who successfully subscribed to previous deposit term, are more likely to subscribe to current deposit term. Another influential predictor is the type of contact, telephone phone or cell phone. It turns out that, customers are more likely to purchase deposit terms if they are contacted by telephone not cell phone (Figure 1). Other predictors do not seem to have good predictive values, such as job, education and loan.

Table 5. The test error rate for each customer related predictor

|  |  |  |
| --- | --- | --- |
| Variables |  | Test error rate |
| Age |  | **44.9%** |
| Job |  | 49.3% |
| Marital |  | 51.0% |
| Education |  | 51.0% |
| Default |  | 51.2% |
| Housing |  | 51.2% |
| Loan |  | 51.2% |
| Contact |  | **41.3%** |
| Campaign | # contact during campaign | **48.1%** |
| Previous | # contact before campaign | **44.7%** |
| Previous outcome |  | **32.0%** |

In order to check if there is an interaction between two strong predictors: previous outcome and type of contact. I used them to predict the outcome variable together with and without an interaction term. The results are in Table 6. When they predicted by alone, the test error rates are 32% for previous outcome, and 41.3% for contact type. When they predicted together without interaction term, the collective performance is better than individual performance, with test error rate 30.3%. However when an interaction term is added, the test error rate was still 30.3%, indicating that an interaction term did not help with the prediction.

Table 6. The interaction between previous campaign outcome and type of contact. When the interaction term between previous outcome and type of contact was added, the test error rate was the same as when there was no interaction term.

|  |  |  |  |
| --- | --- | --- | --- |
| p.outcome | contact | p.outcome+contact | p.outcome+contact+p.outcome\*contact |
| 32% | 41.3% | 30.3% | 30.3% |

**Visualization**

The two most influential predictors are previous outcome and contact type (Table 5). I plotted the outcome variable on these two variables in order to visualize decision boundary (Figure 1). We can observe that when the customers were contacted by telephone and when they purchased the deposit term before, they were more likely to purchase the current deposit term. However, they were reached by cell phone, previous behavior did not predict current behavior. Conversely, when customers were contacted by telephone and when they did not purchased previous deposit term, they were less likely to purchase current deposit term. Finally, if there was no record on customers’ previous decision, they were unlikely to subscribe to current deposit term.

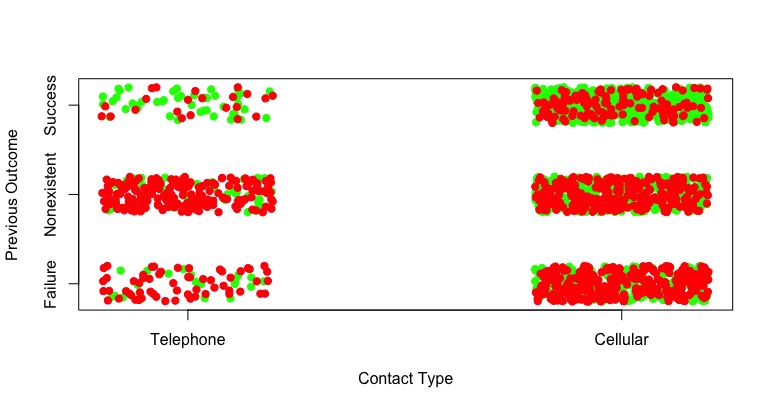
****

Figure 1. The plot of customers’ decision on the type of contact and previous outcome variable. The green dots indicate successful marketing campaign and the red dots indicate unsuccessful marketing campaign.

**Model selection**

In order to pick the model with the best predictive ability, I used all predictors except the duration of phone call because duration is unknown before phone call is made thus it will not help us with targeting potential customers. The performance of different methods will be quantified by the following metrics. Error rate or misclassification rate refers to the mistake what the model makes, compared with the true data. Test error rate is the error rate that model makes in test data. Lower test error rate indicates more accurate prediction. The threshold is set to 0.5. Because test error rate changes according to the threshold, it will also include ROC curve to compare the performance of different models. In statistics, a receiver operating characteristic (ROC), or ROC curve, is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. Area under the ROC curve (AROC) is an important statistics because it indicates how well the model predicts. AROC ranges from 0 to 1. The larger the AROC is, the better the prediction is.

Overall, decision tree (DT) with boosting technique yielded the best results. In order to obtain the parameters, which yield the lowest test error rate, I tried different interaction.depth numbers and different number of trees. When interaction.depth equals to 1 and tree number equals to 5000, the test error rate was the lowest. Then I looped through different shrinkage values from 0 to 0.006 and found that when the shrinkage equals to 0.001097347, the test error rate was the lowest 26.0% (Figure 1). The DT with boosting resulted the lowest test error rate, 26% (Table 7). The performance of LR and Ridge was the same, which was 28.4%. The test error rate for SVM is 28.2% with kernel sigmoid, cost of 10, gamma of 0.01. Finally, KNN had the highest test error rate, 30.8%. The ROC curve (Figure 3) and the area under ROC curve (Table 8) are telling us the same story. DT performed only slightly better than LR and Ridge regression. KNN performed the worst. From table 8, we can see that AROC was the largest for DT with 0.79. AROC for LR and Ridge was the same with 0.78. KNN had the smallest AROC with 0.57%. In Figure 2, the ROC curves of different methods were plotted together.

Figure 2. Turning parameters of decision tree with boosting. The parameters which yielded the lowest test set error rate are interaction.depth=1, shrinkage= 0.001097347, and ntree=5000, with test error rate 26.0%.

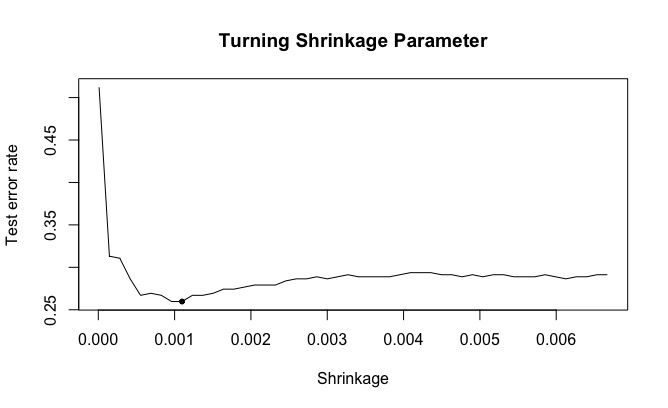


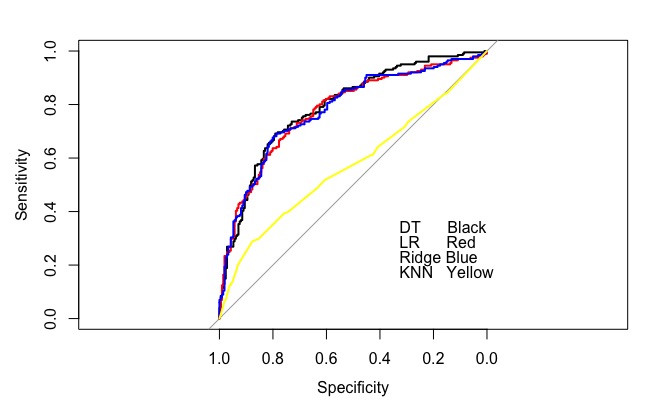
Table 7. The test error rate of four models. Overall, decision tree with boosting yielded the lowest test error rate with 26.0%. LR and Ridge regression preformed the same. KNN had the highest error rate. The threshold is 0.5.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | DT | LR | Ridge | SVM | KNN |
| Test error rate | 26.0% | 28.4% | 28.4% | 27.2% | 30.8% |

Table 8. The area under ROC curve of different model. The performance of all model is similar, except for the KNN.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | DT | LR | Ridge | KNN |
| AROC | 0.79 | 0.78 | 0.78 | 0.57 |

Figure 3. ROC curve for different methods. The ROC curve for DT, LR, and Ridge regression overlapped with each other, indicating similar performance.



The confusion matrix of the best DT with boosting is in the Table 9. I also plotted the importance of each predictor from DT model in Table 10 and Figure 4. As shown in the plot, although pdays has the highest relative influence, previous outcome should still be considered as the most influential predictor because of its practice value.

Table 9. The confusion matrix of decision tree, where interaction.depth =1 and tree number=5000, shrinkage=0.001097347

|  |  |  |  |
| --- | --- | --- | --- |
| Decision Tree with Boosting | | Actual Class | |
| No | Yes |
| Predict Class | No | 167 | 63 |
| Yes | 44 | 138 |

Table 10. The relative influence of the predictors in DT with boosting.

|  |  |
| --- | --- |
| Variables | Relative influence |
| pdays | 30.36121327 |
| poutcome | 25.80713473 |
| contact | 18.71855479 |
| day\_of\_week | 8.37529451 |
| month | 4.46669319 |
| age | 3.48945321 |
| campaign | 3.1396474 |
| marital | 1.93885881 |
| job | 1.85354242 |
| euribor3m | 1.44079394 |

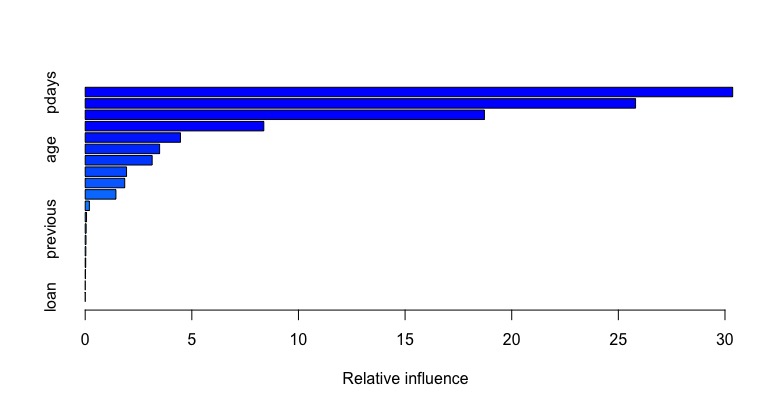


Figure 4. The importance of the predictors in DT with boosting

Because best parameter for boosting is interaction.depth = 1, interactions between predictors may not be too important. I used LR combined with generalized additive models (GAMs) to automatically search over nonlinear transformations of each individual predictor. The variable, previous outcome, use step function to predict outcome variable (Figure 4). The success of current campaign is related to the success of previous campaign.

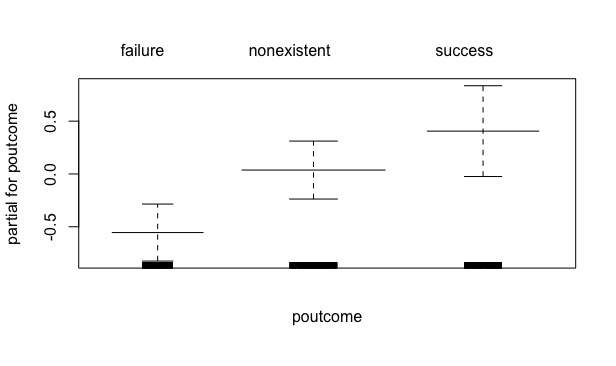


Figure 4. The step function of outcome variable, fitting to variable previous outcome.

Since DT with boosting preformed better than LR, what if we use it to answer our business questions. It turns out that, the test error rate of DT with boosting for each customer related predictor is similar to LR model, except for predictor age LR 44.9%, DT 48.3 and predictor campaign LR 48.1, DT 50%.

In 2009, because the dataset is imbalanced, that is, the success rate is only 19.48%, it is difficult to find a model to make good prediction. To deal with imbalancedness, I used the bootstrap to generate a balanced dataset by resampling from success group with replacement. Then I trained LR model on the balanced dataset and test the model on imbalanced dataset. Table 11 shows the results of LR model, trained on imbalanced and balanced dataset in terms of test error rate and AROC. For test error rate, I set the threshold as 0.5. The test error rate is higher for the model trained on balanced dataset. But this number is just one point on the ROC curve. The test error rate will change if I set the threshold differently. Next, on AROC, I can observe a slight better performance on the model trained on balanced dataset, with balanced 0.78 and imbalanced 0.76. The model trained on balanced dataset is more likely to predict “yes” group than the model trained on the imbalanced dataset.

Table 11. The performance of logistic regression trained on imbalanced dataset and balanced dataset.

|  |  |  |
| --- | --- | --- |
|  | Imbalance | Balance |
| Test error rate (threshold=0.5) | 17.2% | 26.7% |
| AROC | 0.76 | 0.78 |

In conclusion, the telemarketing dataset contains three campaigns from 2008 to 2010. Different campaign has different customer makeup. The analysis focuses primarily on the campaign 2010. For the performance of individual predictor, customers’ previous outcome is the most influential variable. That is, if one successfully purchased the deposit term before, he/she is more likely to purchase the deposit in the current campaign. The second most influential predictor is the type of contact: customers are more likely to deposit their money when they are contacted by telephone. There is not much interaction between different predictors. For model selection, decision tree with boosting technique has the slight advantage over other methods. For the issue with imbalanced data in 2009, the model trained on balanced dataset by bootstrapping performed slight better than the model trained on imbalanced dataset, resulting larger area under ROC curve.