PROJECT REPORT – ANALYTICS IN BANKING

Niky Chen, Eldon Yeh, Nicholas Kelly, Patrick Howes and Steve Peiriez

Introduction

- The objective was to produce a model that will determine the likelihood that a client will subscribe to a bank term deposit based on various predictors.

- The evaluation metric used was:

$$LogLoss = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)]$$

- The dataset consisted of 14 predictors and 30436 observations.

- 7.7% of people in the sample took a long term deposit.

- Of the 14 predictors in the dataset, only 4 of the predictors were quantitative- the rest were all qualitative values.

Methodology

- Easily interpretable.
- Producing the tree with the minimum cross-validation error resulted in a single-node root tree.
- More parameters led to overfitting.
- Single node tree was likely generated due to the large imbalance in classes.
- The resulting tree when we pruned at CP = 0.0014233.

nsplit	CP	CV error
0	0.0032736	0.0769480
4	0.0019214	0.0770796
6	0.0017079	0.0770796
10	0.0014944	0.0770465
12	0.0014233	0.0769480

Methodology

Random Forest

- Chosen to resolve the imbalance of classes.
- The Recursive Binary Partition tree was essentially discarding the minority class as noise in the model.

Naïve Bayes Classifier

- Naïve Bayes Classifier learns by the distribution of classes available in the training set.
- Predicted well in terms of cross-validation error.
- Poor performance as the test error was large, due to imbalance in the data as well as the assumption of independent variables.

Logistic Regression

- Initially used all predictors.
- Decided to use a subset of predictors that were statistically significant in terms of their p-values at the 5% significance level.
- 10-fold cross validation error was used to gauge effectiveness.

Model	Variables	Number of Variables	CV Error
LR Full	All 14 Predictors	14	0.0672099
LR1	Age, Job, Marital Status, Default Status, Housing Loan, Personal Loan	6	0.0706979
LR2	Age, Job, Marital Status, Default Status, Housing Loan, Personal Loan, Previous Campaign Outcome	7	0.0704332
LR3	Job, Marital Status, Default Status, Education Level, Housing Loan, Previous Campaign Outcome	6	0.0704299
LR4	Job, Marital Status, Default Status, Education Level, Housing Loan, Previous Campaign Outcome, Last Contact Communication Type	7	0.0701542
LR5	Job, Marital Status, Default Status, Education Level, Previous Campaign Outcome, Last Contact Communication Type, Number of Previous Contacts	7	0.0700887
LR6	Job, Marital Status, Default Status, Last Contact Communication Type, Previous Campaign Outcome, Current Campaign Contact Number	7	0.0700220
LR7	Job, Marital Status, Default Status, Education Level, Last Contact Month, Last Contact Communication Type, Previous Campaign Outcome, Current Campaign Contact Number, Number of Previous Contacts	9	0.0673934
LR8	Job, Marital Status, Default Status, Education Level, Last Contact Month, Last Contact Communication Type, Previous Campaign Outcome	7	0.0673652
LR9	Job, Marital Status, Education Level, Last Contact Month, Previous Campaign Outcome, Last Contact Communication Type	6	0.0673817

Results

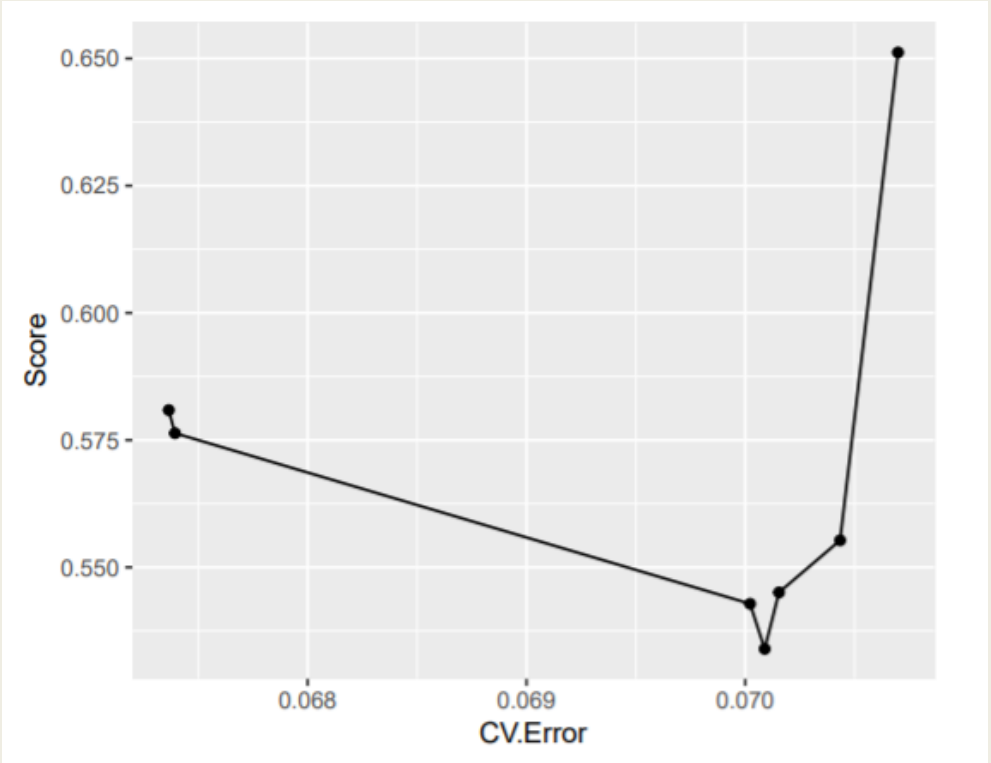
- It was observed that the model with the lowest CV errors did not produce the best competition scores.
- The best models proved to be Logistic Regression models.
- Best performance because the coefficients of the model were chosen to minimise the Log-Loss Function.
- A person who worked in admin, is high-school educated, divorced, whose previous campaign attempt by the institution was a failure, and who had been contacted only once during the current campaign, contacting that client by telephone reduced the probability of the person subscribing to a bank term deposit by 0.029 compared to when they were contacted on their mobile phone.

Results

Table of CV error and Competition Score

Model	CV.Error	Score
LR Full	0.0672099	NA
LR8	0.0673652	0.58090
LR9	0.0673817	NA
LR7	0.0673934	0.57635
LR6	0.0700220	0.54282
LR5	0.0700887	0.53396
LR4	0.0701542	0.54507
LR3	0.0704299	NA
LR2	0.0704332	0.55530
LR1	0.0706979	0.65120

Graph of CV error and Competition Score



Conclusion

- Best performing model, a Logistic Regression model that used the predictors; Job, Marital status, Default status, Education level, Outcome from the Previous campaign, method of last Contact in the current campaign, and number of Previous contacts performed before the current campaign.

	Model	CV.Error	Score
6	LR5	0.0700887	0.53396

- Possible improvements Oversampling.