Exploring and Visualizing Data

MSPA Survey

Kimberly Kaminsky

Predict 422 Section 59

Introduction

The purpose of this study is to advise a Portuguese bank interested in identifying factors that affect client responses to new term deposit offerings. This will help guide future telephone marketing campaigns. Two modelling techniques, Logistic Regression and Naïve Bayes, will be compared to determine which best segments the clients. Along with a recommendation for a classification technique, recommendations will be provided about classes of clients to pursue in a targeted marketing campaign.

Data Exploration and Visualization

The banking data consists of 4521 observations from past telemarketing campaigns with 17 attributes. One response variable indicates the client's response to a term deposit offer. Three binary predictor variables were selected analysis. The default variable indicates if the client had any credit in default. The housing variable indicates if the client has a housing loan. The loan variable indicates of the client has a personal loan. None of the observations had missing variables.

Figure 1 shows three bar charts that show how each binary variable interacts with the response variable.

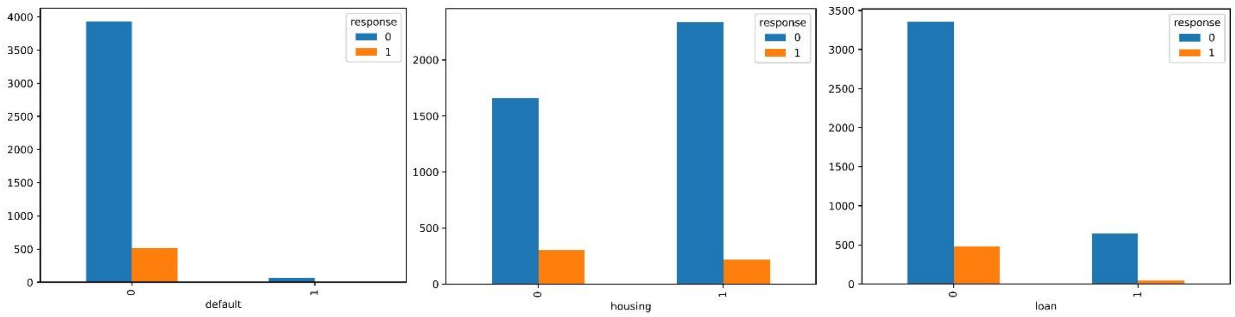

Figure 1: Predictor Variable Interactions with Response Variables

The majority of the clients declined the term deposit offer. Also, very few clients had credit defaults. Positive responses to the offer were about equally split between those clients who had a housing loan and those who didn't. More positive responses were received from those clients who had a personal

loan than those than didn't. A look at a proportion table shows in Table 1 shows that clients that didn't have loans tended to accept the offer at twice the rate as those that did have a loan.

Table 1: Loan Proportion Table

	No_Deposit	Deposit	Row_Total
No_Loan	0.875196	0.124804	1
Loan	0.937771	0.062229	1
Col_Total	0.88476	0.11524	1

Research Design and Modeling Methods

A Logistic Regression and Naïve Bayes model were created and verified utilizing a k-fold, (10 fold), cross validation. The area under the ROC curve, AUC, was used as the measure to compare the two methods. The results were very similar with an AUC value of .611 for Naïve Bayes and .612 for Logistic Regression. Since the Logistic Regression was slightly better and it has more opportunities to add in the remaining continuous predictor variables from the dataset for analysis, this was the model selected for further analysis. After running the logistic regression model against the entire sample dataset each distinct grouping of clients was run against the model to produce the prediction probabilities. Three binary variables create 7 distinct client classes. Table 2 shows the predicted probabilities for each class.

Table 2: Predicted Probabilities for each Class of Clients

	Default	Housing	Loan	Predicted Probability
Class 1	0	0	0	0.16511
Class 2	1	0	0	0.188012
Class 3	1	1	0	0.107651
Class 4	1	1	1	0.054271
Class 5	0	1	0	0.093412
Class 6	0	1	1	0.046723
Class 7	0	0	1	0.085984

Class 1 and 2 had the highest predicted probabilities. In order to utilize the probabilities to make a yes\no prediction a cutoff value must be selected. Table 3 shows the results of 3 cutoff values, (.50, .10, and .09). The .50 cutoff value cannot be selected because it doesn't provide any positive responses.

Table 3: Confusion Matrices for Several Cutoff Values

Confusion matrix for .50 cutoff			
response	0	1	All
0	4000	521	4521
All	4000	521	4521

Confusion matrix for .10 cutoff			
response	0	1	All
0	2575	235	2810
1	1425	286	1711
All	4000	521	4521

Confusion matrix for .09 cutoff			
response	0	1	All
0	636	41	677
1	3364	480	3844
All	4000	521	4521

Both .10 and .09 are good candidates depending on what measure is most important. The .09 cutoff captures more of the true positive cases, but also introduces more false positives. In either case, Figure 2 shows the predicted probabilities against the actual responses for each case providing valuable information about the client classes. Classes 1 and 2 provide the best predictive power. The next best is classes 3 and 5. These classes have in common that the clients do not currently have a personal loan. Particular attention should be paid to clients that don't have both a personal or a housing loan. Both these conditions exist for classes 1 and 2 which are the most promising classes of clients.

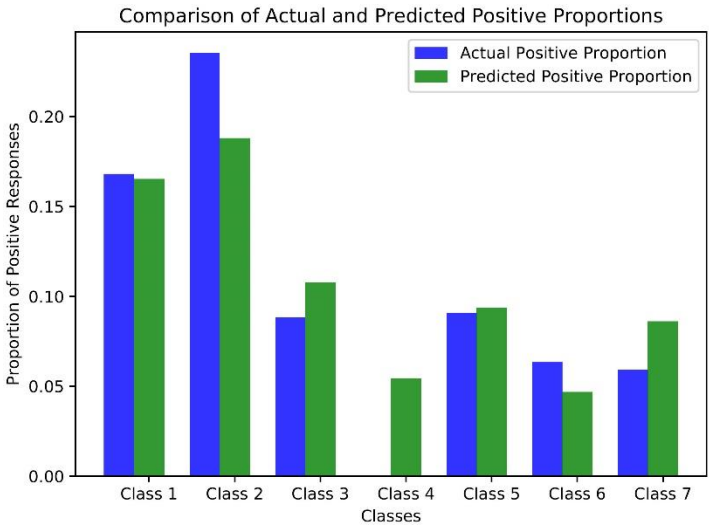

Figure 2: Comparison of Proportions of Positive Responses between Actual and Predicted Cases

Recommendations

From the analysis above, several recommendations may be made. First, neither of the two methods provided a very good AUC score. Therefore, it seems likely the models are underfitting. Since the Logistic Regression model can accommodate non-binary records, it can easily accommodate further investigation of some of the continuous predictor variables. Second, the clients that seem most likely to accept the proposal are those that don't have a personal loan. If they also don't have a housing loan, that is even better. Third, when choosing a cutoff value for the predictive model some cost\benefit analysis should be conducted. A .09 cutoff value provides the most true positives. However, it creates a

large number of false positives, which could be expensive from an advertising perspective. The cost of contacting each client should be evaluated against the profit obtained from each positive response to determine the best equilibrium point.