Evaluating Classification Models

**Summary**

With the ability to capture vast amounts of data on specific individuals, targeted marketing has become popular in the industry. The traditional “spray and pray” model treated all customers as one group. However identifying the power of personalized messaging/marketing, and the ability to capture big data sets in a cost effective way, many companies have opted to store as much data as possible on their customers. During the telephone marketing campaigns conducted between 2008 and 2011, a lot of data on the bank’s customers were recorded with some possible determining factors, that can aid us in our marketing approach to get clients to subscribe to new term deposits. The goal of this study is to identify some common variables within customers that have already subscribed to new term deposits, and use this information to target similar customers and aid in other outreach methods.

**Research Design**

Considering we are trying to understand or predict a variable given certain attributes, what we have is a study of conditional probability. For this specific case the probability of subscribing to new term deposits given three explanatory variables, default, housing and loan.  And since we are predicting classes and not values we will be using two common classification methods, logistic regression and naïve bayes.

We will determine which of the two models is a better fit given the data obtained and leverage this model in determining which variables have more weight in determining subscriptions.

**Data Analysis**

The data obtained was in a single column, therefore it was first necessary to separate the different variables. When using a classification method it is a lot easier to work with binary (numerical) variables over categorical variables, for this reason we had to convert the explanatory variables to 1’s and 0’s. After dropping the null values we were ready to build a model and evaluate.

**Model Evaluation**

To get a good and precise understanding of the two different models, we used cross-validation. This is basically where we take the training set and split it into a smaller training and validation set. This can occur multiple times, and each iteration is known as a fold. In the following study we used a 10 fold cross-validation design. The measure used to differentiate the two models was the area under the ROC (Receiver Operating Characteristic) curve. The closer the value is to 1, the more precise the model is.

The ROC scores were: Naïve Bayes = 0.60813 , Logistic Regression = 0.60791

Considering the scores, both models seem to have relatively low scores however, it seems that the Naïve Bayes model is slightly more preferable. Figure 1 in Appendix A gives us a visual as to what the ROC curve looks like in this case.

**Testing**

A hypothetical data set was generated using 8 different possibilities and the model was run on this data.

Studying the variables with the highest probability for yes the attribute that stood out the most in determining subscriptions was “Default”. Figure 2 shows a breakdown of the different probabilities.

**Recommendation**

Ideally we would want to investigate other models that might have a better ROC score. Depending on the scope and the budget of the marketing campaign, different models are used to meet different needs. Once the models are established the next step would be to take action on the insights gathered. What teams and resources would have to be leveraged to get the message out to the targeted individuals and can this process be automated as well for effectiveness.

**Appendix A**

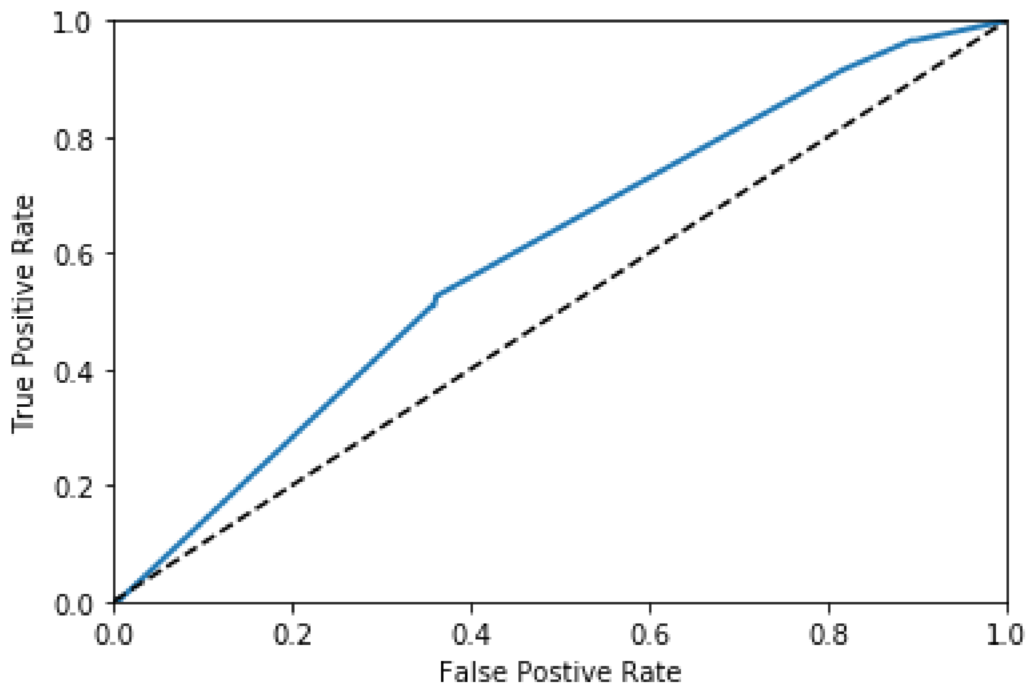


Figure 1 - ROC Curve

Training set predictions from Naive Bayes model

| Customer | default | housing | loan | response | Prob_NO | Prob_YES | Prediction |
|---|---|---|---|---|---|---|---|
| A | 0 | 0 | 0 | NO | 0.598237 | 0.401763 | NO |
| B | 0 | 0 | 1 | NO | 0.672654 | 0.327346 | NO |
| C | 0 | 1 | 1 | NO | 0.766311 | 0.233689 | NO |
| D | 0 | 1 | 1 | NO | 0.766311 | 0.233689 | NO |
| E | 0 | 0 | 1 | NO | 0.672654 | 0.327346 | NO |
| F | 1 | 0 | 1 | YES | 0.399991 | 0.600009 | YES |
| G | 1 | 0 | 1 | YES | 0.399991 | 0.600009 | YES |
| H | 1 | 0 | 0 | YES | 0.325724 | 0.674276 | YES |

Figure 2 - Naive Bayes Prediction/Probabilities