Support Vector Machines

Summary

At the end of our last study on the telephone marketing campaign, it was suggested the best way to move forward was by studying other models that might possibly have a better ROC score, thus being a better classifier. We also wanted to understand how some of the other variables weighed in the decision making for subscribing to a term deposit.

Research Design

Similar to the last study we will be running a logistic regression and comparing it to a support vector machine (SVM). SVM's are very versatile machine learning models; it is great in situations where the data is not well understood or known as you have a lot of tools you can work with to customize the model.

Since we are doing a classification study, we will once again use the area under the ROC curve to verify which of the two is a better model and then determine which variables have more weight in determining subscriptions. The only real difference in the data is the addition of the balance explanatory variable, which unlike the other three variables is continuous and not binary.

Data Analysis

The data structure is the same as it was for the previous study. The binary explanatory variables are converted to 1's and 0's as it is easier to work with numerical variables and it was found that the mean for the balance is 1422 with a minimum value of -3313 and maximum of 71188.

Model Evaluation

A cross validation design was implemented once again with a fold count of 10. The only issue with trying to compare a SVM model to another classifier using the area under the ROC curve is that SVM classifiers

do not out-put probabilities for each class which is needed to calculate the ROC score. A work around was to use the SVC class and to set the Kernel to “linear” and probability to “true”. The boundaries or also known as the support vectors can be altered using the C hyper-parameter, this aids in regularizing the data and preventing over-fitting. In this study we gave C a value of 1.

Figure 1 in Appendix A shows a comparison of the ROC curves with and without the balance variable. You will notice immediately that having a continuous variable causes more rigidity in the line, however that line is still pretty far from the left corner of the box. Taking a look at the ROC scores:

Logistic Regression = 0.6262 SVM = 0.4643

It seems the logistic regression model would still be the favored classifier but one thing to notice is that the addition of a new parameter (balance) the ROC score has slightly increased compared to the previous study (0.60813).

Testing

A hypothetical data set was generated, using 16 different possibilities considering we had 4 different explanatory variables. Studying the variables with the highest probability for “No” this time round and eliminating them, the attribute that stood out the most once again was “Default”. Figure 2 shows a breakdown of the different probabilities.

Recommendation

SVM’s are great for smaller datasets, typically under a 1000 rows, this could be one reason it scored poorly. However it should not be dropped entirely, as we increase the number variables and the models get more complex, SVM’s are great when it comes to customizing the models using different hyper-parameters and finding niche groups.

Appendix A

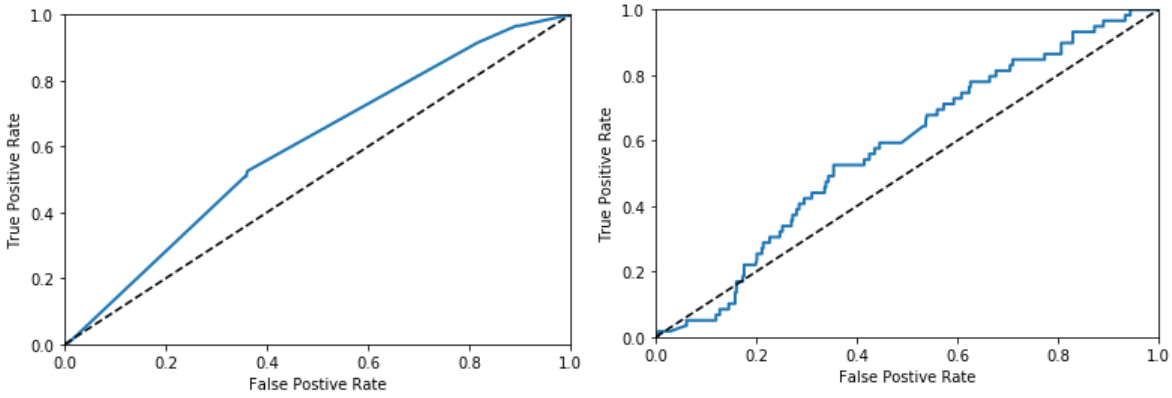

Figure 1 - ROC Curve with and without "Balance"

Training set predictions from Naive Bayes model

Customer	default	housing	loan	balance	response	Prob_NO	Prob_YES
A	0	0	0	0	NO	0.533036	0.466964
B	0	1	1	1000	NO	0.800063	0.199937
C	0	1	0	0	NO	0.681294	0.318706
D	0	1	1	1000	NO	0.800063	0.199937
E	0	1	0	0	NO	0.681294	0.318706
F	0	0	0	1000	YES	0.532987	0.467013
G	0	1	0	1000	YES	0.681252	0.318748
H	0	1	0	0	NO	0.681294	0.318706
I	0	1	1	0	NO	0.800094	0.199906
J	0	1	0	1000	NO	0.681252	0.318748
K	0	1	0	0	YES	0.681294	0.318706
L	0	1	0	1000	NO	0.681252	0.318748
M	0	0	0	0	YES	0.533036	0.466964
N	0	0	0	1000	YES	0.532987	0.467013
O	0	1	1	0	NO	0.800094	0.199906
P	0	0	1	0	NO	0.681252	0.318748

Figure 2 - Prediction/Probabilities