Example Analysis of LEO Data

Data Description

The data was acquired from the latest excel file. These data were converted to SAS and JMP files. The analysis that follows is solely for illustrative purposes as none of the results are meaningful or necessarily informative about the prediction of coronary events of interest.

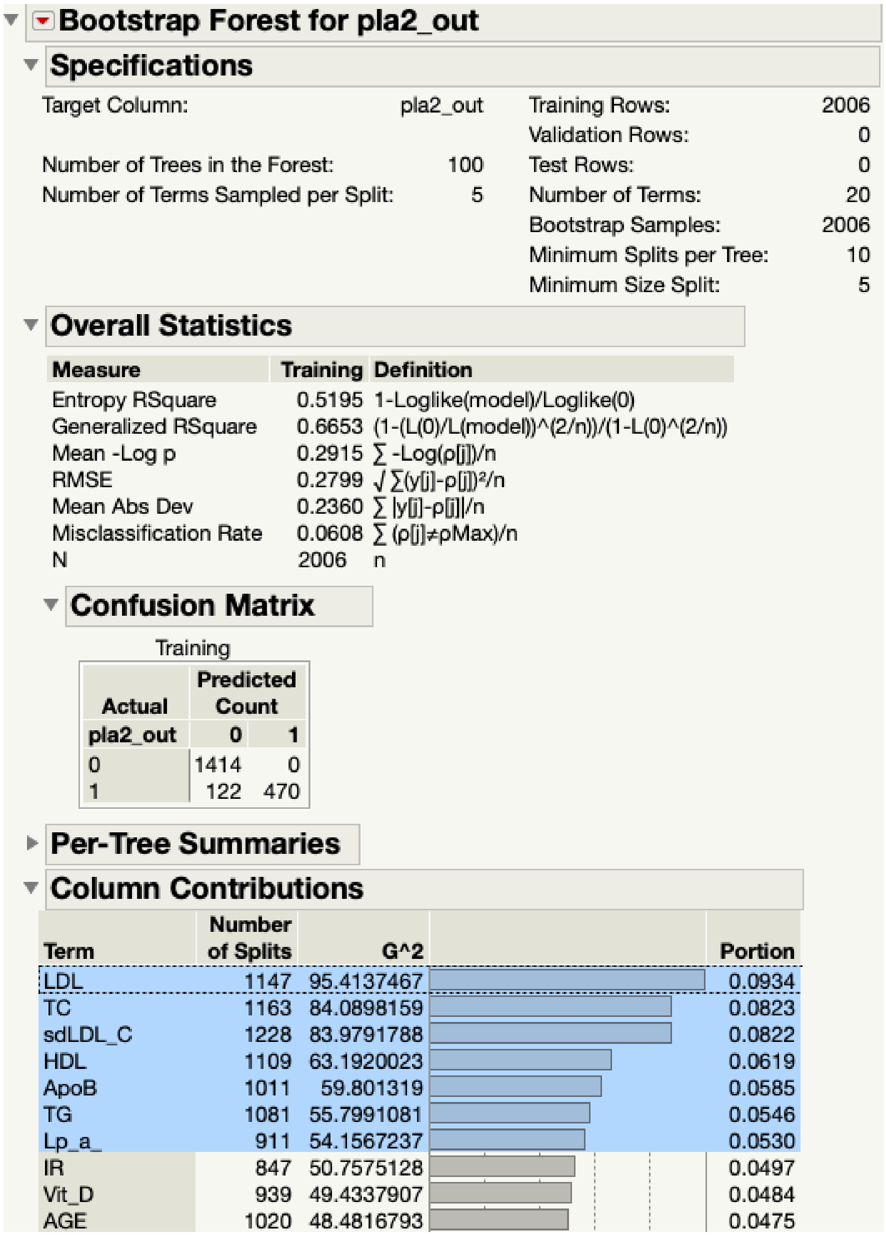
Data Analysis

I used the variable “PLA2 Out of Range” as an indicator event of interest. This variable was binary, and I assumed that PLA2 Out of Range = “TRUE” is the event of interest. My objective was to build a predictive model for this event. My approach was to use a machine learning technique called “random forest” which is an extension of a CART (Classification and Regression Trees) procedure that identifies which variables provide the best predictive information for the binary (in our case) event of interest. Once these variables were identified I used them to build a predictive model for the probability of the event of interest as a function of the selected predictor variables. Models of these types are used in developing potential risk models for the medical condition of interest. I have included some output examples (again all meaningless due to the sparseness of the data set that I used).

Data Analysis Results

The data set that I used had 2006 useable records of which I had 592 entries where PLA2 Out of Range = “TRUE” (29.5%). The best result that I could get using the random forest had a correct classification of PLA2 Out of Range = “TRUE” 470 times with a misclassification of 122. None of the cases where PLA2 Out of Range = “FALSE” were misclassified. The results are found in Figure 1. The blue variables (selected) were then used in a Generalized Logistic Regression model from which I can get the predictive probabilities of (in this case) the event PLA2 Out of Range = “TRUE”. Some summary results are seen in Figure 2.

Figure 1



Bootstrap Forest for pla2_out

Specifications

Target Column:	pla2_out	Training Rows:	2006
		Validation Rows:	0
Number of Trees in the Forest:	100	Test Rows:	0
Number of Terms Sampled per Split:	5	Number of Terms:	20
		Bootstrap Samples:	2006
		Minimum Splits per Tree:	10
		Minimum Size Split:	5

Overall Statistics

Measure	Training	Definition
Entropy RSquare	0.5195	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.6653	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.2915	∑ -Log(ρ[j])/n
RMSE	0.2799	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.2360	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.0608	∑ (ρ[j]≠ρMax)/n
N	2006	n

Confusion Matrix

Training

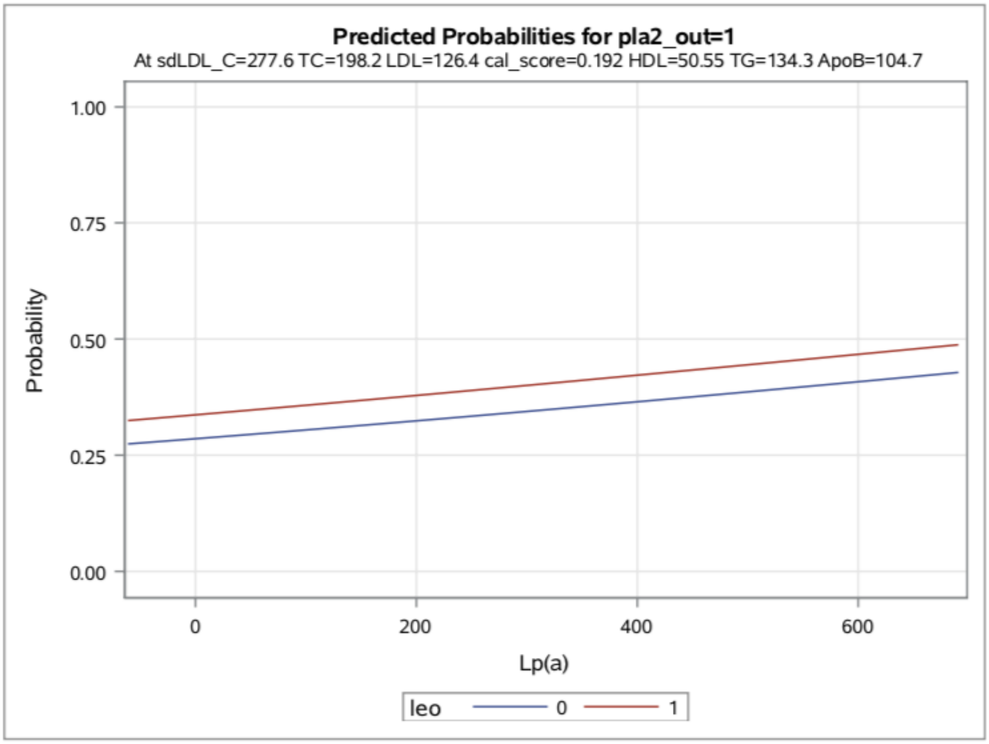
Actual	Predicted Count	
pla2_out	0	1
0	1414	0
1	122	470

Per-Tree Summaries

Column Contributions

Term	Number of Splits	G^2	Portion
LDL	1147	95.4137467	0.0934
TC	1163	84.0898159	0.0823
sdLDL_C	1228	83.9791788	0.0822
HDL	1109	63.1920023	0.0619
ApoB	1011	59.801319	0.0585
TG	1081	55.7991081	0.0546
Lp_a_	911	54.1567237	0.0530
IR	847	50.7575128	0.0497
Vit_D	939	49.4337907	0.0484
AGE	1020	48.4816793	0.0475

The LOGISTIC Procedure


Predicted Probabilities for pla2_out=1
At sdLDL_C=277.6 TC=198.2 LDL=126.4 cal_score=0.192 HDL=50.55 TG=134.3 ApoB=104.7
Probability
Lp(a)
leo