

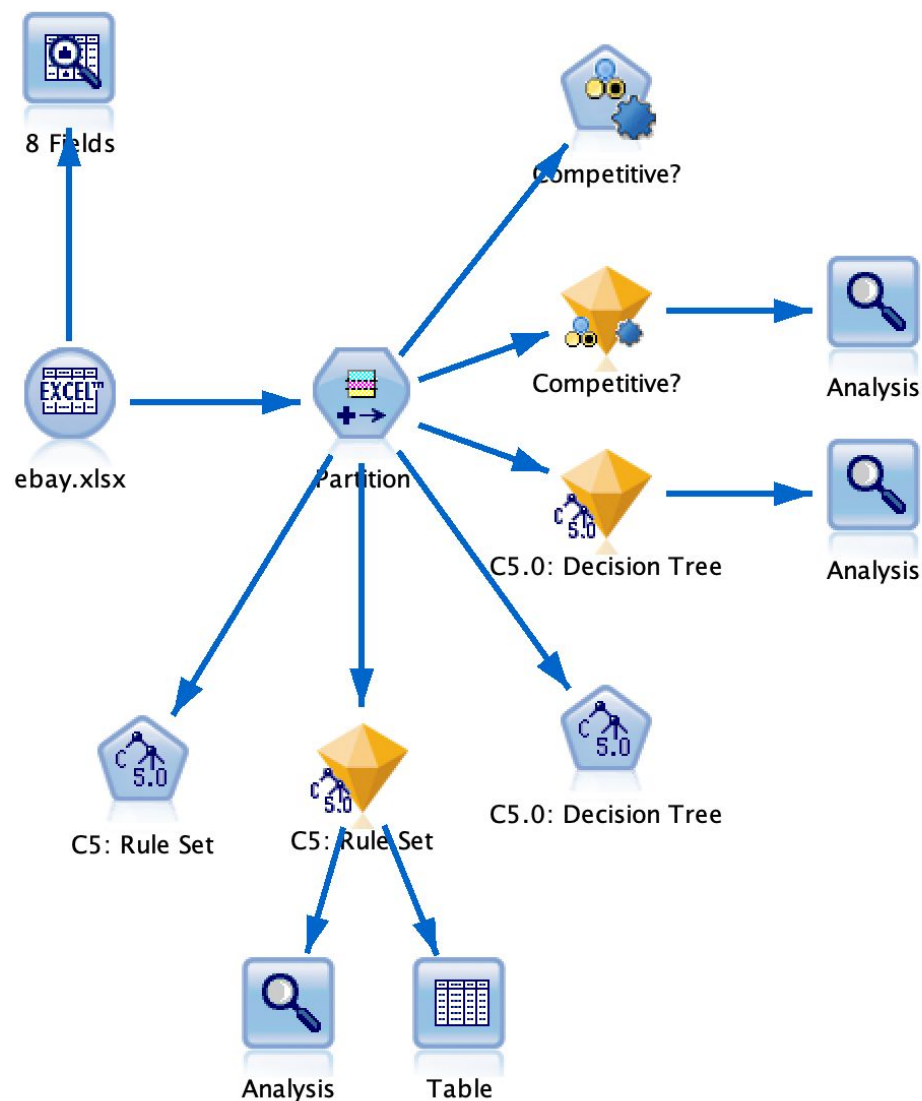
Project #2. Competitive Auctions on eBay.com

(Adapted from Shmueli, Patel, Bruce, "Data Mining & Business Intelligence 2nd Ed")

The file eBay.xlsx contains information on auctions transacted on eBay.com over 2 months. The goal is to use these data to build a model that will classify competitive auctions from noncompetitive ones. A competitive auction is defined as an auction with at least two bids placed on the item auctioned. The data include variables that describe the item (auction category), the seller (his/her eBay rating), and the auction terms that the seller selected (auction duration, opening price, currency, day-of-week of auction close). In addition, we have the price at which the auction closed.

1. Split the data into training and validation datasets using a 70%: 30% ratio.




After performing a data audit to verify there was no missing data and visualize trends in the data set before partitioning the data, we used partitioned it using the partition node. Our stream is below:






2. Fit several classifiers (do not use KNN, as it may be slow, and memory consuming), and create a table comparing the results. In this exercise predictive performance is important: you want to produce the best possible model, checking for class imbalance, choosing your predictive performance metrics and plots wisely, and controlling model overfitting as much as possible.

Using an Auto Classifier node, we were able to fit the data to 12 models, comparing all the results in order to determine the best behaving model. In the end, SPSS determined the C5,0 model to be the most accurate, with CHAID and the Decision Tree to be our next best choices.

Performance of Training Set:

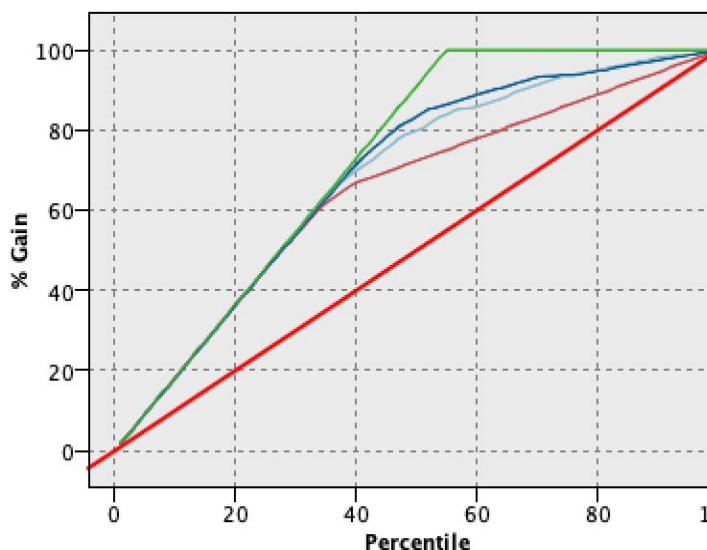
Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift{Top 30%}	Overall Accuracy (%)	No. Fields Used	Area Under Curve
	C5 1	< 1	2,945.0	45	1.857	89.779	5	0.938
	CHAID 1	< 1	2,598.974	48	1.858	84.559	6	0.919
	Decision List 1	< 1	2,173.239	37	1.816	78.309	3	0.810

Performance of Testing Set:

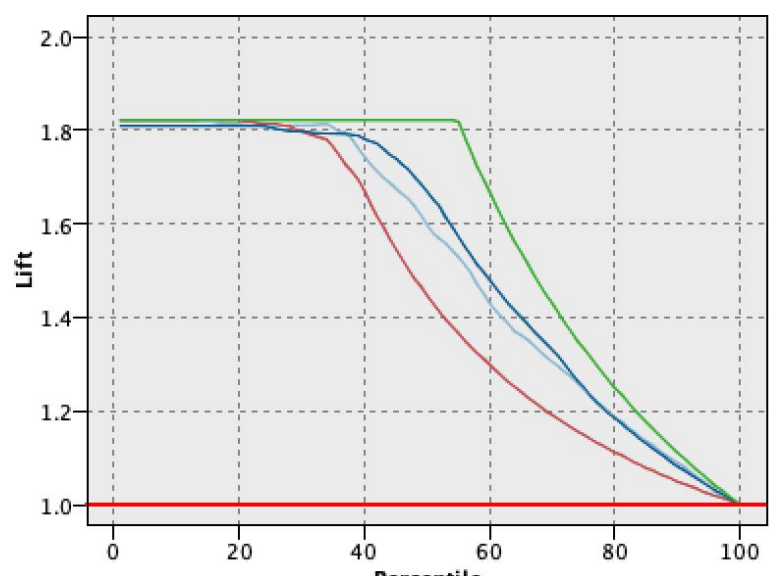
Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift{Top 30%}	Overall Accuracy (%)	No. Fields Used	Area Under Curve
	C5 1	< 1	1275.000	47	1.797	86.438	5	0.917
	CHAID 1	< 1	1178.158	47	1.812	83.66	6	0.901
	Decision List 1	< 1	1,028.0	38	1.797	78.758	3	0.818

— CHAID 1
— Decision List 1
— C5 1
— \$BEST

Evaluation of Top 3 Models: Gain



Evaluation of Top 3 Models: Lift



3. For the informative models (DTs, logistic regression) explain:

Rule Set generated from C5.0 Model (0 = Competitive // 1 = Not Competitive)

Rule	Result	Confidence
if ClosePrice > 2.040 and OpenPrice <= 2.450	then 1.000	(353; 0.997)
if ClosePrice > 1.025 and OpenPrice <= 1.228	then 1.000	(186; 0.995)
if ClosePrice > 3.800 and ClosePrice <= 10 and OpenPrice <= 3.745	then 1.000	(147; 0.993)
if ClosePrice > 4.505 and ClosePrice <= 10 and OpenPrice <= 4.919	then 1.000	(136; 0.986)
if ClosePrice > 10	then 1.000	(673; 0.649)
if Category = Automotive and sellerRating > 63 and sellerRating <= 562 and Duration <= 6	then 0.000	(14; 0.938)
if ClosePrice <= 3.800 and OpenPrice > 2.450	then 0.000	(82; 0.929)
if ClosePrice <= 10 and OpenPrice > 4.919	then 0.000	(218; 0.895)
if ClosePrice <= 15.125 and OpenPrice > 11.069	then 0.000	(61; 0.889)
if ClosePrice <= 2.040 and OpenPrice > 1.228	then 0.000	(65; 0.881)
if ClosePrice <= 4.505 and OpenPrice > 3.745	then 0.000	(38; 0.85)
if currency = US and OpenPrice > 10.495 and OpenPrice <= 11.069	then 0.000	(4; 0.833)
if currency = EUR and OpenPrice > 2.450 and OpenPrice <= 10.423	then 0.000	(108; 0.782)
if sellerRating > 562 and OpenPrice > 11.069	then 0.000	(225; 0.762)
if ClosePrice <= 1.025	then 0.000	(41; 0.744)
if Duration > 6 and OpenPrice > 11.069 and OpenPrice <= 12.230	then 0.000	(27; 0.724)

- The regression coefficients and odd ratios

Coincidence Matrix for C5.0 Decision Tree Model

Results for output field Competitive?

Comparing \$C-Competitive? with Competitive?

'Partition'	1_Training		2_Testing	
Correct	1,221	89.78%	529	86.44%
Wrong	139	10.22%	83	13.56%
Total	1,360		612	

Coincidence Matrix for \$C-Competitive? (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	611	19
1.000000	120	610
'Partition' = 2_Testing	0.000000	1.000000
0.000000	261	15
1.000000	68	268

Partition	1_Training	2_Testing
Accuracy	89.78%	86.44%
Recall	96.98%	94.57%
Specificity	83.56%	79.76%
Precision	83.58%	79.33%
F-Statistic	16.44%	20.24%

Variables in the Equation:

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a						
sellerRating	.000	.000	2.783	1	.095	1.000
Duration	-.039	.032	1.484	1	.223	.962
ClosePrice	.134	.012	115.940	1	.000	1.143
OpenPrice	-.152	.013	131.661	1	.000	.859
Constant	-.083	.238	.122	1	.727	.920

The values found under the column B are our regression coefficients in this logistics regression. We also see S.E or Standard error for our regression coefficients.

The next column is the Waks test which is used for predictive significance. The regression here shows our p - values of Wald significance test, so we

understand that $p \sim 0.00$ are SellerRating, ClosePrice, and OpenPrice, meaning these values are not important and can remove these variables in revisions.

Odd Ratios are displayed in the last column under Exp(B). So this tells us that a unit increase in ClosePrice increases the odds of a competitive auctions by 1.143 times.

From here we can find our logit function:

$$\text{Logit} = 0.0001 * \text{SellerRating} + -0.039 * \text{Duration} + 0.134 * \text{ClosePrice} + -0.152 * \text{OpenPrice} - 0.083$$

Omnibus Tests of Model Coefficients and Model Summary

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	442.992	4	.000
	Block	442.992	4	.000
	Model	442.992	4	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1435.008 ^a	.278	.371

The tables is showing the Model Chi-square Test. This test is similar to the F-test in linear regression and is a likelihood ratio test. From this we can also see the p - value again is ~ 0.00 meaning that the model is significant.

Classification Table:

Classification Table

			Predicted		
			Competitive?		Percentage Correct
			0.0	1.0	
Step 1	Competitive?	0.0	586	44	93.0
		1.0	250	480	65.8
Overall Percentage					78.4

This last table is reporting the classification accuracy over the training data set as well as the confusion matrix.

Results for output field Competitive?

Individual Models

Comparing \$L-Competitive? with Competitive?

'Partition'	1_Training		2_Testing	
Correct	1,066	78.38%	496	81.05%
Wrong	294	21.62%	116	18.95%
Total	1,360		612	

Coincidence Matrix for \$L-Competitive? (rows show actuals)

'Partition' = 1_Training		0.000000	1.000000
0.000000		586	44
1.000000		250	480
'Partition' = 2_Testing		0.000000	1.000000
0.000000		258	18
1.000000		98	238

Performance Evaluation

'Partition' = 1_Training	
0.000000	0.414
1.000000	0.534
'Partition' = 2_Testing	
0.000000	0.474
1.000000	0.527

Evaluation Metrics

'Partition'	1_Training		2_Testing	
Model	AUC	Gini	AUC	Gini
\$L-Competitive?	0.894	0.787	0.907	0.815

From this we can see that the accuracy is high.