Harshaanth Thiyagaraja Kumar

Competitive auctions on eBay.com:

The file eBayAuctions.xls contains information on 1972 auctions that transacted on eBay.com during May-June in 2004. The goal is to use these data in order to build a model that will classify competitive auctions from non-competitive ones. A competitive auction is defined as an auction with at least 2 bids placed on the auctioned item. The data include variables that describe the auctioned 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 that the auction closed at. **The goal is to predict whether the auction will be competitive or not.**

Analysis:

1. Based on the goal, will you use all variables in the dataset as predictors? (Hint: are the values of the predictors known at the start of an auction?). What variables should not be included in the predictor set? Explain why?

Goal: Our aim is to find the level of competitiveness of an auction prior to its conclusion.

Excluded Predictor variables:

Competitiveness: This variable is excluded as it serves as our target outcome.
Closed Price: Owing to its availability only after the auction concludes, it cannot be utilized to forecast competitiveness.

Included Predictive Factors:

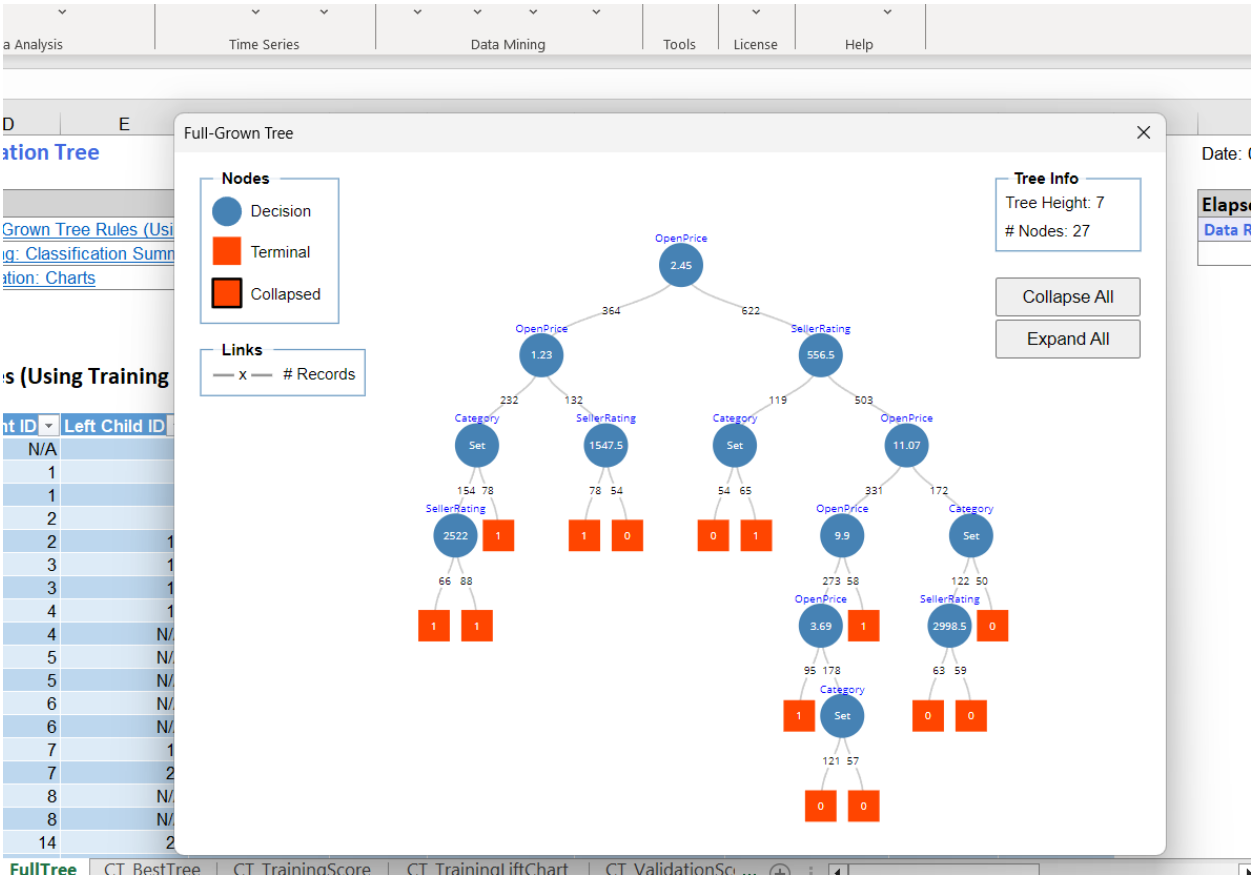
Auction Category: Describing the nature of the item being auctioned, it could potentially influence the competitiveness by attracting differing numbers of bidders across categories.
Seller's Rating: The reputation of the seller may impact buyer trust and consequently affect the competitiveness of the auction.
Duration: The length of the auction could influence bidder participation, with longer durations potentially attracting more bidders and thus affecting competitiveness.
Opening Price: The initial auction price might influence bidder interest, thereby impacting the competitiveness of the auction.
Currency: Although less likely to directly influence competitiveness, it could play a role in specific scenarios, particularly if buyers are from diverse regions with differing exchange rates.
EndDay2: The timing of the auction's end could also influence bidder engagement, with certain days potentially being more conducive to bidding activity.

2. Fit a classification tree using all predictors(Split the data into training, validation and test datasets using a 50%, 30%, and 20% ratio).To avoid overfitting, set the minimum number of records in a leaf node to 50. Also, set the maximum number of levels to be displayed at 7,

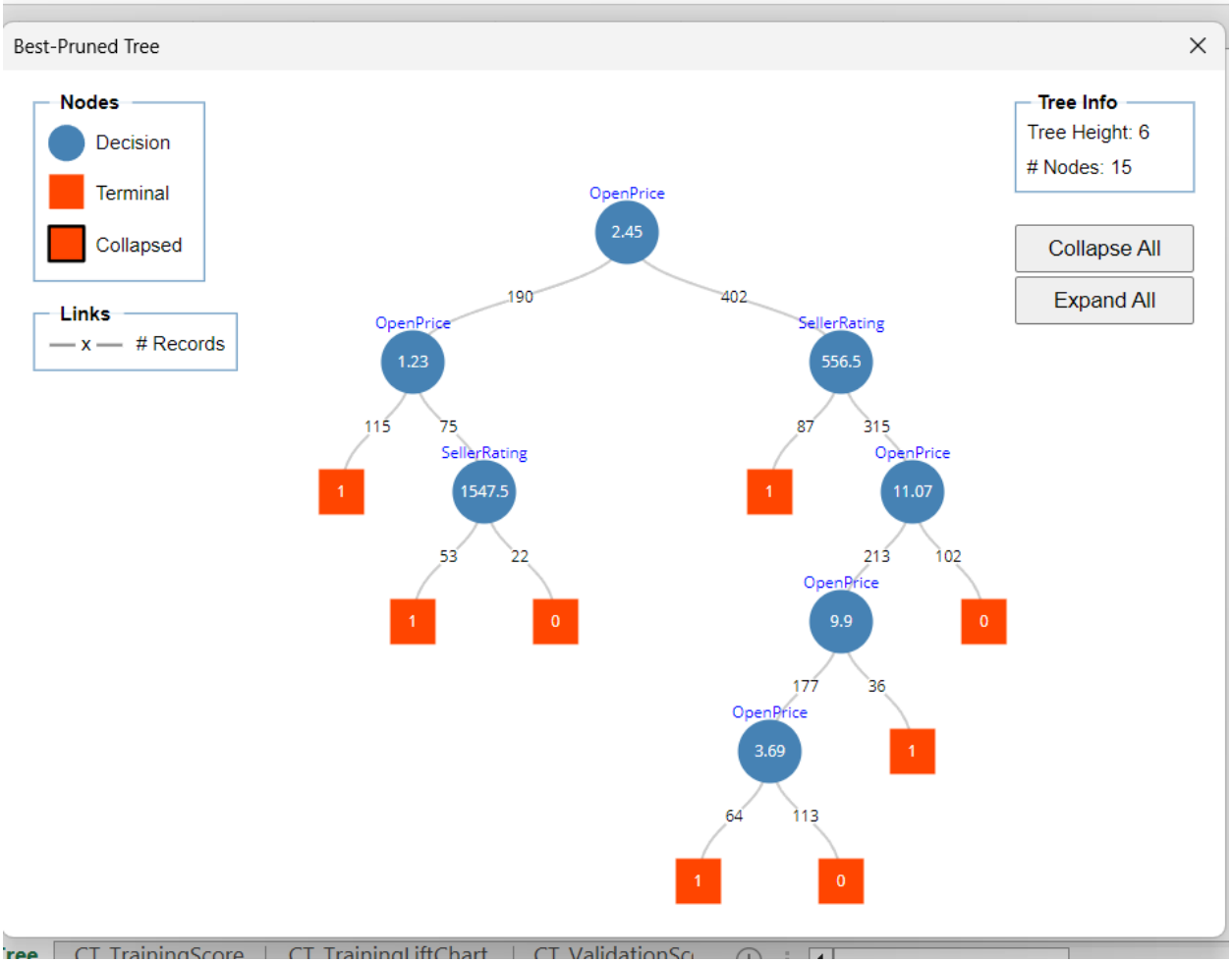
show both the full tree and the best pruned tree.

a. Report the full tree and the best pruned decision tree (show the tree diagram).

Full tree:



Best tree:



b. Describe the results in terms of rules of the best pruned decision tree. For example, if variable1<0 AND variable2<2, class=0.

If Opening price < 1.23, class = 1.
If Opening Price >= 1.23 and opening price < 2.45 and Seller Rating < 1547.5, class = 1.
If Opening Price >= 1.23 and opening price < 2.45 and Seller Rating >= 1547.5, class = 0.
If opening price >= 2.45 and Seller Rating < 556.5, class = 1.
If opening price >= 11.07 and Seller Rating >= 556.5, class = 0.
If Opening Price >= 9.9 and opening price < 11.07 and Seller Rating >= 556.5, class = 1.
If Opening Price >= 3.69 and opening price < 9.9 and Seller Rating >= 556.5, class = 0.
If Opening Price >= 2.45 and opening price < 3.69 and Seller Rating >= 556.5, class = 1.

c. Will the auction be competitive? The auction will last for two weeks and end on weekend. The open price is 500. The seller rating is 1000. The product is in the category of Clothing/Toys and Currency is US.

The product is not competitive and belongs to class 0.

d. Report test data scoring-summary report and lift charts (Using best pruned tree). Compare the results to the validation results. Does the model work well?

Harshaanth Thiyagaraja Kumar

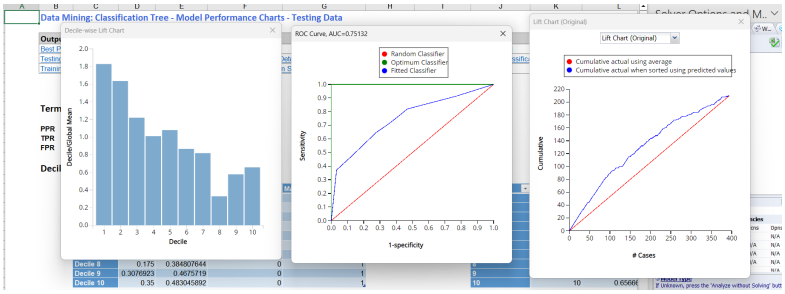
Testing: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	105	79
1	46	164

Error Report			
Class	# Cases	# Errors	% Error
0	184	79	42.93478261
1	210	46	21.9047619
Overall	394	125	31.72588832

Metrics	
Metric	Value
Accuracy (#correct)	269
Accuracy (%correct)	68.27411168
Specificity	0.570652174
Sensitivity (Recall)	0.780952381
Precision	0.674897119
F1 score	0.72406181
Success Class	1
Success Probability	0.5

Test lift chart



Harshaanth Thiyagaraja Kumar

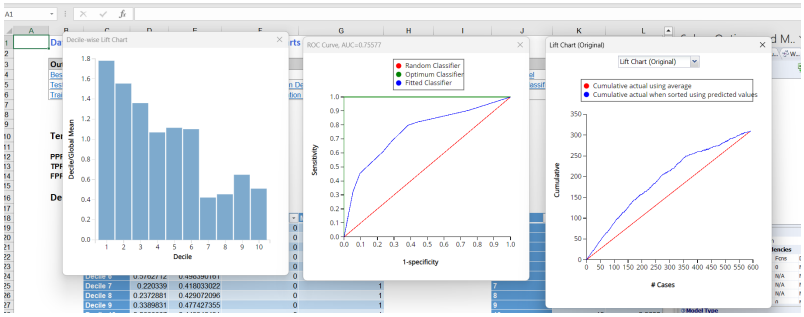
Validation: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	174	108
1	63	247

Error Report			
Class	# Cases	# Errors	% Error
0	282	108	38.29787234
1	310	63	20.32258065
Overall	592	171	28.88513514

Metrics	
Metric	Value
Accuracy (#correct)	421
Accuracy (%correct)	71.11486486
Specificity	0.617021277
Sensitivity (Recall)	0.796774194
Precision	0.695774648
F1 score	0.742857143
Success Class	1
Success Probability	0.5

Validation lift chart



The model performs well on test and validation data, the ROC curve of test data 0.75132 similar to that of validation data 0.75577.

e. What is the overall accuracy for the test dataset? Show how precision, recall and specificity are calculated based on the number in the confusion matrix.

Confusion Matrix		
Actual\Predicted	0	1
0	105	79
1	46	164

Accuracy = (TP + TN) / (TP + TN + FP + FN)
= (164 + 105) / (164 + 105 + 79 + 46)
= 269 / 394
≈ 0.6827 or 68.27%

Precision:
Precision = TP / (TP + FP)
= 164 / (164 + 79)

≈ 0.6749 or 67.49%

Recall:
Recall = TP / (TP + FN)
= 164 / (164 + 46)
≈ 0.7800 or 78.00%

Specificity:
Specificity = TN / (TN + FP)
= 105 / (105 + 79)
≈ 0.5702 or 57.02%

TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

3. Fit a logistic regression (using training, validation and test data). Use stepwise for variable selection and choose the last model among models recommended by variable selection. (You may choose your own baseline)
    a. Report the **selected model's** regression model (the equation and the coefficient table)

Best Subsets Details				
Subset ID	#Coefficients	RSS	Mallows's Cp	Probability
Subset 1	1	1100.824883	85.7032397	9.09914E-14
Subset 2	2	1080.075701	67.54066229	1.05134E-10
Subset 3	3	1065.700048	55.5714257	1.08471E-08
Subset 4	4	1053.483199	45.69996175	4.98469E-07
Subset 5	5	1043.048984	37.56073525	1.17306E-05
Subset 6	6	1033.200741	29.99091381	0.000220396
Subset 7	7	1023.488675	22.55341903	0.003801901
Subset 8	8	1018.737801	19.93685811	0.011137585
Subset 9	9	1013.206914	16.56233597	0.043102274
Subset 10	10	1007.850322	13.35718182	0.155774589

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Odds	Standard Error	Chi2-Statistic	P-Value
Intercept	1.460477568	0.92038753	2.00056761	4.308016	0.275561208	28.09012184	1.2E-07
SellerRating	-3.88675E-05	-6.334E-05	-1.439E-05	0.999961	1.24865E-05	9.68927771	0.00185

Duration	-0.090556534	-0.157148	-0.0239651	0.913423	0.033975865	7.103927899	0.00769
Category_Coins/Stamps	-1.653275339	-2.6922959	-0.6142548	0.191422	0.530122278	9.726089555	0.00182
Category_Computer/Electronics	0.981767269	0.3587245	1.60481004	2.669169	0.317884804	9.538444638	0.00201
Category_EverythingElse	-0.910631379	-1.3483895	-0.4728732	0.40227	0.223350107	16.62313346	4.6E-05
Category_Health/Beauty	-1.840568578	-2.751084	-0.9300532	0.158727	0.464557208	15.69732327	7.4E-05
Category_Jewelry	-0.973850963	-1.6585652	-0.2891367	0.377626	0.349350406	7.770742042	0.00531
Category_SportingGoods	0.714388286	0.10777957	1.320997	2.042937	0.309499 93	5.327795421	0.02099
EndDay2_Weekend	-0.762310698	-1.0381105	-0.4865109	0.466587	0.140716751	29.34758925	6E-08

Having Currency – Non US, Category_Art/Collectibles, EndDay2_Weekday as Baseline.

Logit(Competitive=1) = 1.460477568- 0*SellerRating - 0.0906*Duration – 1.6533*Category_Coins/Stamps + 0.9818 * Category_Computer/Electronics – 0.9106 * Category_EverythingElse – 1.8406 * Category_Health/Beauty – 0.9739 * Category_Jewelry + 0.7144 * Category_SportingGoods - 0.7623 * EndDay2_Weekend

Odds(Competitive=1) =

$$e^{1.460477568 - 0*\text{SellerRating} - 0.0906*\text{Duration} - 1.6533*\text{Category\_Coins/Stamps} + 0.9818*\text{Category\_Computer/Electronics} - 0.9106*\text{Category\_Everyth}}$$

Probability (Competitive=1)

$$\frac{1}{1+e^{-(1.460477568 - 0*\text{SellerRating} - 0.0906*\text{Duration} - 1.6533*\text{Category\_Coins/Stamps} + 0.9818*\text{Category\_Computer/Electronics} - 0.9106*\text{Category\_EverythingElse} - 1.8406*\text{Category\_Health/}}}$$

b. Will the auction be competitive? The auction will last for two weeks and end on weekend. The open price is 500. The seller rating is 1000. The product is in the category of Clothing/Toys and Currency is US.

Probability (Competitive=1)

$$\frac{}{1 + e^{-(1.460477568 - 0*\text{SellerRating} - 0.0906*\text{Duration} - 1.6533*\text{Category\_Coins/Stamps} + 0.9818*\text{Category\_Computer/Electronics} - 0.9106}}$$

Probability (Competitive=1)

$$\frac{1}{1+e^{-(1.460477568-0*500-0.0906*14-1.6533*0+0.9818*0-0.9106*0-1.8406*0-0.9739*0+0.7144*0-0.7623*1)}}$$

Probability (Competitive=1)

$$\frac{1}{1+e^{-(-0.57)}}$$

Probability (Competitive=1)

$\frac{1}{1+1.77} = 0.36$

Default Cutoff probability = 0.5

Probability (Competitive=1) < *Default Cutoff probability*

0.36 < 0.5, Therefore it belongs to class 0. Hence not competitive.

c. What is the overall accuracy for the test dataset? Compare to the accuracy in 2.e. And compare the lift charts of using logistic regression and classification tree. Which model would you use, decision trees or logistic regression?

Test lift chart and test data summary using Logistic regression

Testing: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	79	105
1	41	169

Error Report			
Class	# Cases	# Errors	% Error
0	184	105	57.06521739
1	210	41	19.52380952
Overall	394	146	37.05583756

Metrics	
Metric	Value
Accuracy (#correct)	248
Accuracy (%correc	62.94416244
Specificity	0.429347826
Sensitivity (Recall)	0.804761905
Precision	0.616788321
F1 score	0.698347107
Success Class	1
Success Probability	0.5

Harshaanth Thiyagaraja Kumar

Test lift chart and test data summary using Classification decision tree

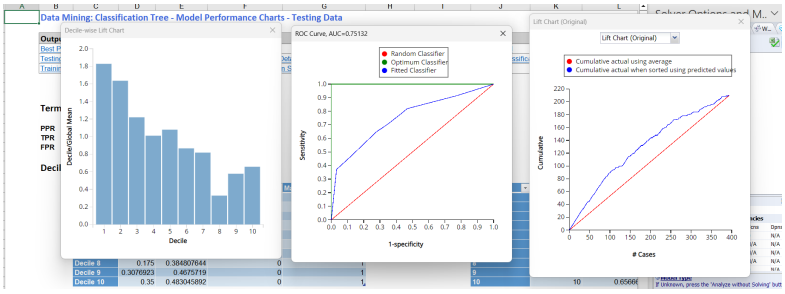
Testing: Classification

Summary

Confusion Matrix		
Actual\Predicted	0	1
0	105	79
1	46	164

Error Report		
Class	# Cases	# Errors
0	184	79
1	210	46
Overall	394	125

Metrics	
Metric	Value
Accuracy (#correct)	269
Accuracy (%correct)	68.27411168
Specificity	0.570652174
Sensitivity (Recall)	0.780952381
Precision	0.674897119
F1 score	0.72406181
Success Class	1
Success Probability	0.5



Overall accuracy for the test dataset using logistic regression = 62.94%
The accuracy in 2.e. using Classification Tree for the test dataset = 68.27%

Comparing the Accuracy, ROC curve, Lift charts of decision trees, and logistic regression, I would choose classification decision trees since it has a better Overall Accuracy of 68.27%, better ROC curve of 0.75132 and a Lift chart curve indicating that it is an overall better predictor for Competitive items in the Auction.