Rotman School of Management
UNIVERSITY OF TORONTO

Rotman

Group Number:	7
Assignment Title:	Group Assignment 2
Course Code:	RSM8413
Instructor Name:	Gerhard Trippen

In submitting this **group** work for grading, we confirm:

- That the work is original, and due credit is given to others where appropriate.
- That all members have contributed substantially and proportionally to each group assignment.
- That all members have sufficient familiarity with the entire contents of the group assignment so as to be able to sign off on them as original work.
- Acceptance and acknowledgement that assignments found to be plagiarized in any way will be subject to sanctions under the University's Code of Behaviour on Academic Matters.

Please **check the box and record your student number** below to indicate that you have read and abide by the statements above:

☐	1002183031	☐	1006604701
☐	1002897378	☐	1007045118
☐	1007554745	☐	1005627403

Executive Summary

Background

This report discusses the use of the decision tree classification algorithm offered by Scikit Learn to classify eBay auctions as competitive or not. Auction competitiveness is determined by the number of bids. If an auction has more than one bid, it is considered as a competitive auction. The first step involved exploring the available data by generating a series of count plots. The categorical data was transformed into sets of dummy variables which created a binary variable representing each of the possible categorical data point (for example, the Currency variable had three options, USD, EUR, and GBP, each of which became its own column after the transformation). After the data was explored, preprocessed, and split between training and testing sets using a 60%-40% proportion, it was used to fit two decision tree classification models.

Most Salient Findings

The first tree leveraged the entire preprocessed dataset which contained all dummy and quantitative variables. One of our most salient findings was that closing price, open price and seller rating were some of the significant variables for the model, while all the dummy variables were not influential at all. As this tree uses a variable which is only observed after an auction is completed (ClosePrice), it cannot be leveraged to classify new auctions, so a second tree was created by using all data excluding that variable. After some EDA and identifying sellerRating and OpenPrice as the most impactful, we created a classification tree that used just these two variables. We found that higher sellerRating led to lower chances of being classified as competitive in the tree which is contrary to expectations. We also found that lower OpenPrice led to higher chances of being classified as competitive.

Introduction

eBay is an online auctioning platform where sellers can post their items for sale, and buyers place bids within an auctioning period; with the highest bid at closing winning the auction and purchasing the item. The company, eBay, is interested in understanding the competitive nature of their auctions in order to better optimize their platform and improve the buying and selling experience. A competitive auction is one that involves a minimum of two buyers placing bids on an item within the auctioning time-period. This allows for a more accurate depiction of the item price, since the value is represented by the extent the two, or more, bidders are willing to pay. Data has been provided regarding numerous auctions that have taken place from May - June 2004. The variables included information about the auction itself such as: category the item falls under, the day the auction closed, the duration of the auction, opening price, closing price and

bidding currency. It also includes seller information such as their respective ratings as given by previous buyers and other sellers on the platform. Using this information, auctions will be classified as competitive or not. A supervised classification method will be used to conduct this analysis: Decision Tree. This method entails a collection of decision nodes connected via branches to leaf nodes; the tree is an instance- based learning method, where it uses training (historical) data to formulate an algorithm which leads to prediction on future incoming (test) data.

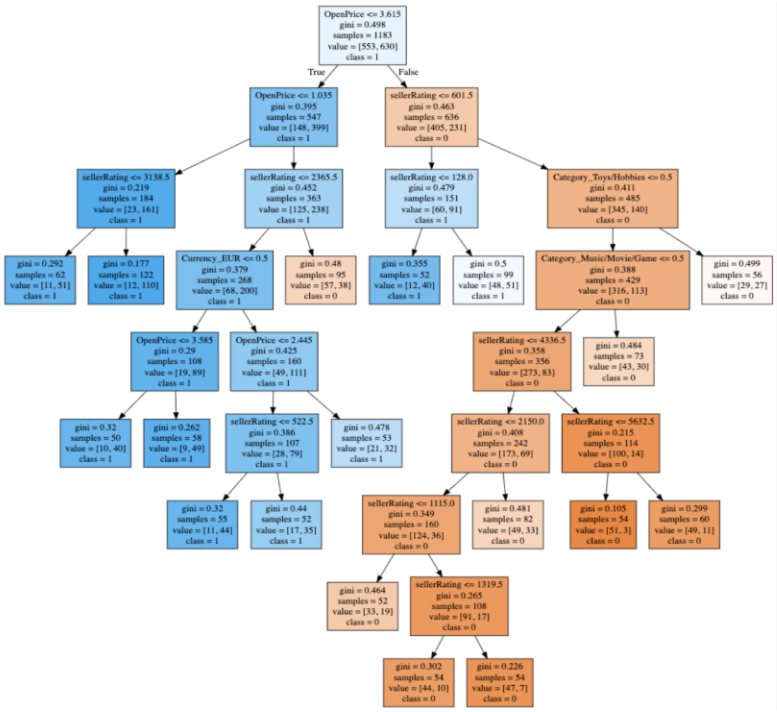
Data Preprocessing

The dataset contains four categorical predictors: *Category*, *Duration*, *endDay*, and *Currency*. There are 18 auction categories, 3 available currencies (USD, EUR, and GBP), 5 auction duration options, and each day of the week as an auction close day. In order to evaluate the importance of each unique value in the categorical predictors on auction competitiveness, the categorical variables are translated into sets of dummy variables. The dataset creates binary variables for each possible categorical value and drops the original categorical variables. The dataset uses prefixes when naming dummy variables to keep the original category information.

After converting the original data into a suitable dataset for the model, the data are split into training and test datasets. In our model, we decided to choose 60% of the full dataset as a training dataset and choose the remaining 40% as a validation dataset. The observations for each dataset are selected randomly in order to reduce bias. There were no missing values found in the dataset.

Classification Tree with All Feasible Variables

We created a decision tree classifier using the training dataset split previously. To make sure we did not overfit, we set a minimum of 50 samples for each terminal node. This resulted in our decision tree model to not have any pure leaf nodes.
Below, the graph shows the decision tree using all the predictors excluding ClosePrice. ClosePrice is an attribute that is unknown until the auction has been concluded; hence it is not feasible to use this predictor to evaluate the competitiveness of an auction and as such it has been removed from our first decision tree.
In the graph below, the shade or intensity of the color demonstrates the level of confidence in each leaf node.



Write down the results in terms of rules.

Summary of the decision rules have been demonstrated in the below table:

Antecedent	Consequent
If OpenPrice ≤ 3.615 **&** OpenPrice ≤1.035 **&** SellerRating ≤ 3138.5	Auction is competitive
If OpenPrice ≤ 3.615 **&** OpenPrice ≤1.035 **&** SellerRating ≥ 3138.5	Auction is competitive
If OpenPrice ≤ 3.615 **&** OpenPrice ≥1.035 **&** SellerRating ≤ 2365.5 **&** Currency≠ Euro **&** OpenPrice ≤ 3.585	Auction is competitive
If OpenPrice ≤ 3.615 **&** OpenPrice ≥1.035 **&** SellerRating ≤ 2365.5 **&** Currency≠ Euro **&** OpenPrice ≥ 3.585	Auction is competitive
If OpenPrice ≤ 3.615 **&** OpenPrice ≥1.035 **&** SellerRating ≥ 2365.5	Auction is not competitive
If OpenPrice ≤ 3.615 **&** OpenPrice ≥1.035 **&** SellerRating ≤ 2365.5 **&** Currency=Euro **&** OpenPrice ≤2.445 **&** SellerRating ≤ 522.5	Auction is competitive
If OpenPrice ≤ 3.615 **&** OpenPrice ≥1.035 **&** SellerRating ≤ 2365.5 **&** Currency=Euro **&** OpenPrice ≤2.445 **&** SellerRating ≥ 522.5	Auction is competitive
If OpenPrice ≤ 3.615 **&** OpenPrice ≥1.035 **&** SellerRating ≤ 2365.5 **&** Currency=Euro **&** OpenPrice ≥2.445	Auction is competitive
If OpenPrice ≥ 3.615 **&** SellerRating ≤ 601.5 **&** SellerRating ≤ 128.0	Auction is competitive
If OpenPrice ≥ 3.615 **&** SellerRating ≤ 601.5 **&** SellerRating ≥ 128.0	Auction is competitive
If OpenPrice ≥ 3.615 **&** SellerRating ≥ 601.5 **&** Category≠Toy **&** Category ≠Music **&** SellerRating ≤ 4336.5 **&** SellerRating ≤ 2150.0 **&** SellerRating ≤ 1115.0	Auction is not competitive
If OpenPrice ≥ 3.615 **&** SellerRating ≥ 601.5 **&** Category≠Toy **&** Category ≠Music **&** SellerRating ≤ 4336.5 **&** SellerRating ≤ 2150.0 **&** SellerRating ≥ 1115.0 **&** SellerRating ≤ 1319.5	Auction is not competitive
If OpenPrice ≥ 3.615 **&** SellerRating ≥ 601.5 **&** Category≠Toy **&** Category ≠Music **&** SellerRating ≤ 4336.5 **&** SellerRating ≤ 2150.0 **&** SellerRating ≥ 1115.0 **&** SellerRating ≥ 1319.5	Auction is not competitive
If OpenPrice ≥ 3.615 **&** SellerRating ≥ 601.5 **&** Category≠Toy **&** Category ≠Music **&** SellerRating ≤ 4336.5 **&** SellerRating ≥ 2150.0	Auction is not competitive
If OpenPrice ≥ 3.615 **&** SellerRating ≥ 601.5 **&** Category≠Toy **&** Category ≠Music **&** SellerRating ≥ 4336.5 7 SellerRating ≤ 5632.5	Auction is not competitive
If OpenPrice ≥ 3.615 **&** SellerRating ≥ 601.5 **&** Category≠Toy **&** Category ≠Music **&** SellerRating ≥ 4336.5 7 SellerRating ≥ 5632.5	Auction is not competitive
If OpenPrice ≥ $3.615 **&** SellerRating ≥ 601.5 **&** Category≠Toy **&** Category =Music	Auction is not competitive
If OpenPrice ≥ $3.615 **&** SellerRating ≥ 601.5 **&** Category=Toy	Auction is not competitive

Describe the interesting/unexpected and uninteresting (= rather obvious) information that these rules provide.

Based on above decision rules, it shows that lower OpenPrice would result in higher competitiveness of the auction. It is rather obvious as lower OpenPrice indicates lower barrier of entry and could attract more buyers.

However, there is an interesting observation in the above decision tree where a higher SellerRating decreases the competitiveness of an auction. Compared with the range of sellerRating, the split numbers for sellerRating indicated in the above rules are relatively on the lower end. One end node created on the 3rd split shows a bad Gini score (0.499), with almost 50-50 in competitive and non-competitive. This led to a white labeled end node due to the uncertainty.

Is this model practical for predicting the outcome of a new auction?

The above tree has too many predictors, which makes it impractical to interpret and use the outcome to predict the competitiveness of an auction. And all the dummy variables are imbalanced. For a more practical interpretation, it is better to identify the most important predictors and just use those predictors to form the decision tree.

Classification Tree with Important Variables

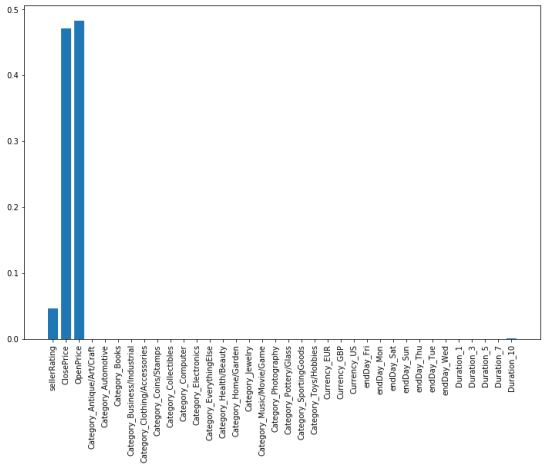
Variables Selection

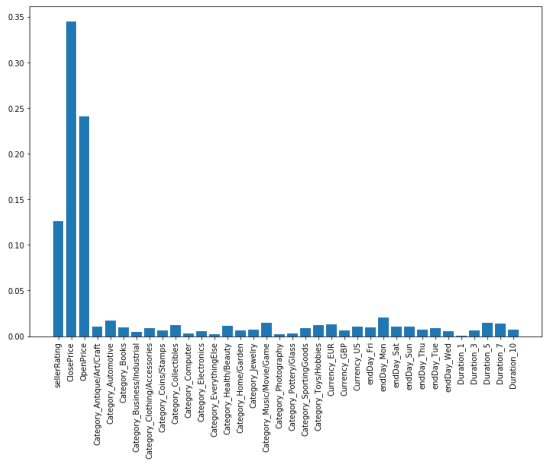
As stated previously, ClosePrice is not included in the model because it is not available until after the auction has occurred. From our analysis of the feature importance score from the DecisionClassifier algorithm, we found that the only features relevant were sellerRating and OpenPrice. Feature importance score is a metric that tracks the relative importance of input features when making a prediction. From the graph below, none of the categorical dummy variables were rated highly in terms of feature importance. Even when using different random states, we found only a few categorical variables showing a very small percentage of importance in terms of feature score. As the stochastic nature of the algorithm or evaluation procedure influences the feature importance, it is more advisable to only use sellerRating and OpenPrice as they are consistently the most important. Additionally, to make sure the algorithm was not at fault, a random forest classifier was used, and the resulting feature score again showed only OpenPrice and sellerRating as significant.

Assumptions

Due to us dropping the currency dummy variables, we assume that the currency of the OpenPrice does not matter, just the absolute value. In this case, a low OpenPrice value regardless of currency will imply a higher likelihood of a competitive auction. As the currencies are US, GBP, and EUR, the relative values of all these currencies to each other are small enough for this statement to hold true.

Decision Tree Classifier Feature Importance Score





Random Forest Classifier Feature Importance Score

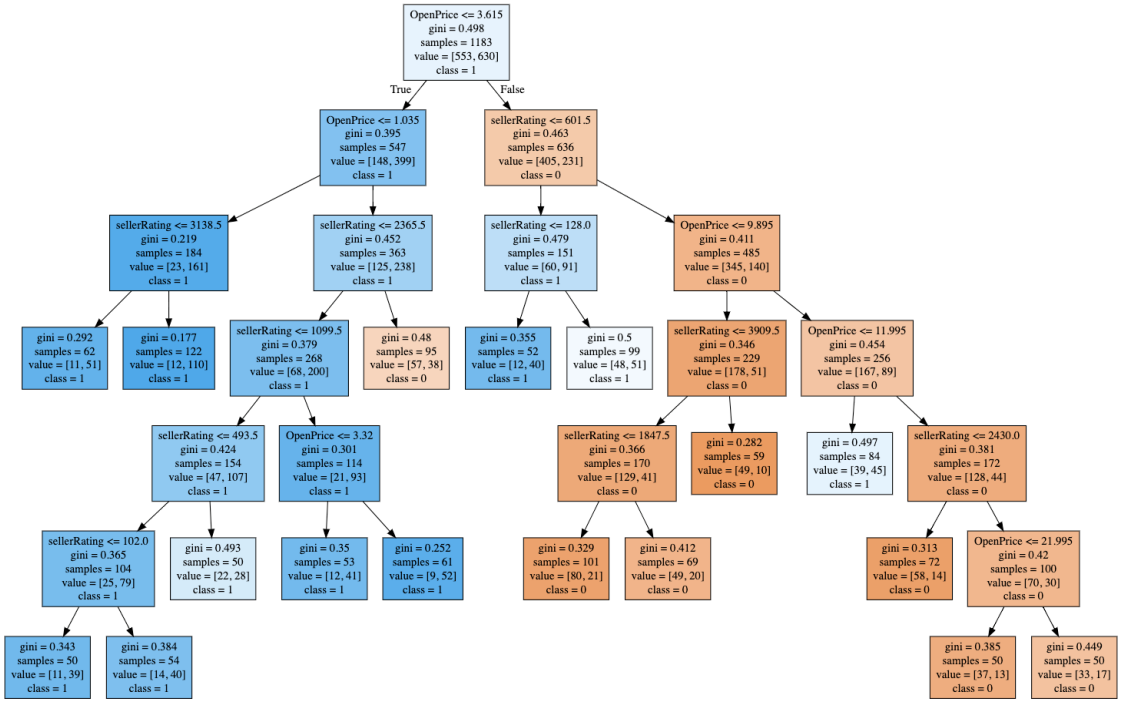
Is this model practical for predicting the outcome of a new auction?

The model is useful as it only takes into consideration variables that can be observed before an auction occurs. The low number of predictors also makes the model more generalizable and not prone to overfitting. There is low multicollinearity as well as the predictors not being highly correlated. From the confusion matrix, the model has an accuracy of around 71%, which is adequate, but not great.

In terms of practicality, knowing sellerRating beforehand can create credibility for an auction as a higher rating is indicative of multiple auctions being completed in the past. Additionally, low open prices can lead to more competitive auctions as the barrier to entry is also lower. The variables make logical sense as tools for predicting competitiveness and they are significant as well.

Explaining Decision Tree Visualization

After creating the new decision tree classification model, we created a visualization of the decision tree below.

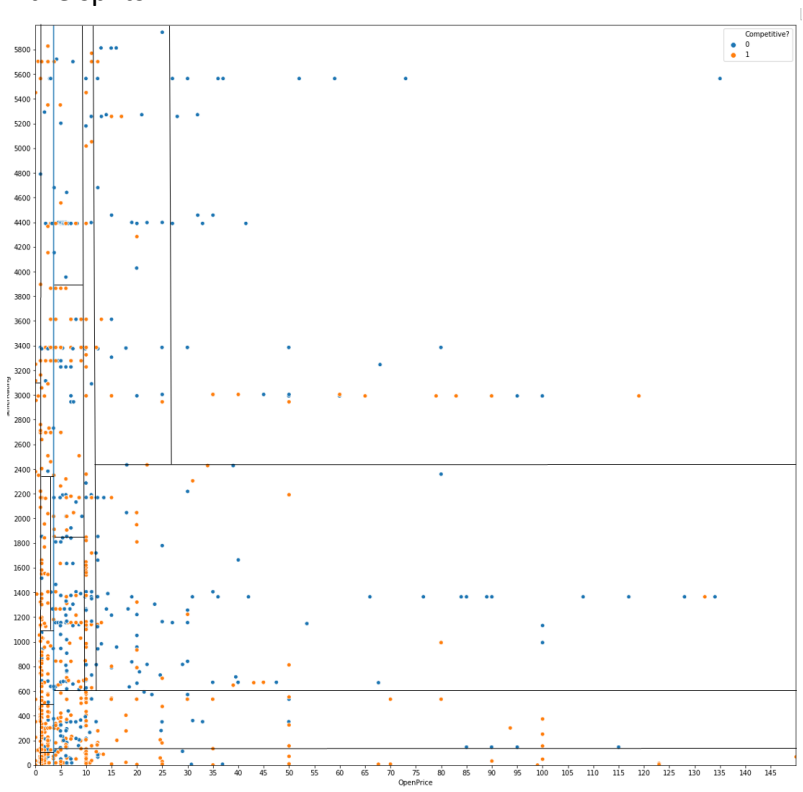


Antecedent	Consequent
If OpenPrice ≤ 3.615 **&** OpenPrice ≤1.035 **&** SellerRating ≤ 3138.5	Auction is competitive
If OpenPrice ≤ 3.615 **&** OpenPrice ≤1.035 **&** SellerRating ≥ 3138.5	Auction is competitive
If OpenPrice ≤ 3.615 **&** OpenPrice ≥1.035 **&** SellerRating ≤ 2365.5 **&** SellerRating ≥ 1099.5	Auction is not competitive
If OpenPrice ≤ 3.615 **&** OpenPrice ≥1.035 **&** SellerRating ≤ 2365.5 **&** SellerRating ≤ 1099.5 **&** SellerRating ≤ 493.5 & SellerRating ≤ 102.0	Auction is competitive
If OpenPrice ≤ 3.615 **&** OpenPrice ≥1.035 **&** SellerRating ≥ 2365.5 **&** SellerRating ≤ 1099.5 **&** SellerRating ≤ 493.5 & SellerRating ≥102.0	Auction is competitive
If OpenPrice ≤ 3.615 **&** OpenPrice ≥1.035 **&** SellerRating ≤ 2365.5 **&** OpenPrice ≤3.32	Auction is competitive
If OpenPrice ≤ 3.615 **&** OpenPrice ≥1.035 **&** SellerRating ≤ 2365.5 **&** OpenPrice ≥3.32	Auction is competitive
If OpenPrice ≤ 3.615 **&** OpenPrice ≥1.035 **&** SellerRating ≤ 2365.5 **&** SellerRating ≤ 1099.5 **&** SellerRating ≥ 493.5	Auction is competitive
If OpenPrice ≥ $3.615 **&** SellerRating ≤ 601.5 **&** SellerRating ≤ 128.0	Auction is competitive
If OpenPrice ≥ 3.615 **&** SellerRating ≤ 601.5 **&** SellerRating ≥ 128.0	Auction is competitive
If OpenPrice ≥ 3.615 **&** SellerRating ≥ 601.5 **&** OpenPrice ≤ 9.895 **&** SellerRating ≤ 3909.5 **&** SellerRating ≤ 1847.5	Auction is not competitive
If OpenPrice ≥ 3.615 **&** SellerRating ≥ 601.5 **&** OpenPrice ≤ 9.895 **&** SellerRating ≤ 3909.5 **&** SellerRating ≥ 1847.5	Auction is not competitive
If OpenPrice ≥ 3.615 **&** SellerRating ≥ 601.5 **&** OpenPrice ≤ 9.895 **&** SellerRating ≥ 3909.5	Auction is not competitive
If OpenPrice ≥ 3.615 **&** SellerRating ≥ 601.5 **&** OpenPrice ≥ 9.895 **&** OpenPrice ≤ 11.995	Auction is competitive
If OpenPrice ≥ 3.615 **&** SellerRating ≥ 601.5 **&** OpenPrice ≥ 9.895 **&** OpenPrice ≥ 11.995 **&** SellerRating ≤ 2430.0	Auction is not competitive
If OpenPrice ≥ 3.615 **&** SellerRating ≥ 601.5 **&** OpenPrice ≥ 9.895 **&** OpenPrice ≥ 11.995 **&** SellerRating ≥ 2430.0 **&** OpenPrice ≤ 21.995	Auction is not competitive
If OpenPrice ≥ 3.615 **&** SellerRating ≥ 601.5 **&** OpenPrice ≥ 9.895 **&** OpenPrice ≥ 11.995 **&** SellerRating ≥ 2430.0 **&** OpenPrice ≥ 21.995	Auction is not competitive

Rules

As we are required to set the minimum number of records in a terminal node to 50, none of the samples (shown in the tree diagram) at the end notes is bigger than 50. This is to prevent overfitting problems that tend to happen when data are less representative of the population. The Gini score reduces as the tree creates splits and develops from root to the end notes. The optimal split is determined by the measure of "goodness". The resulting trees have a maximum of six splits in this case until the restriction is reached. Decision rules table is shown above.

We plot the resulting tree on a scatter plot, using the two best predictors are OpenPrice and sellerRating. Each auction is a point, with coordinates corresponding to its values on the two predictors. Blue points with the value of zero show the auctions are non-competitive. Orange points with the value of one show the auctions are competitive. The lines show the splits.



Accuracy and Predictive Performance

Examine the classification table for the tree. What can you say about the predictive performance of this model?

Confusion Matrix (Accuracy 0.7219)

Actual \ Prediction	0	1
0	402	151
1	178	452

Full Model-Train Set

Confusion Matrix (Accuracy 0.7250)

Actual \ Prediction	0	1
0	261	92
1	125	311

Full Model-Test Set

Confusion Matrix (Accuracy 0.7270)

Actual \ Prediction	0	1
0	363	190
1	133	497

Final Model-Train Set

Confusion Matrix (Accuracy 0.7148)

Actual \ Prediction	0	1
0	222	131
1	94	342

Final Model-Test Set

The tree has a test accuracy rate of 0.7148 which is a considerably high amount and it slightly underperforms the first classification tree which had an accuracy rate of 0.7250. With a 71.5% accuracy, the tree performs well at predicting which auctions will have over two bids and is useful as it provides more information than randomly assigning into competitive and non-competitive. The training set of the final model has an accuracy rate of 0.7270 which is not too far off from that of the test set so there is no overfitting. However, it has a certain number of errors which consists of less False Negatives (FN rate= 0.2155) than False Positives (FP rate= 0.3711) but the tree generally performs well overall.

Finally, we found that including the closing price in the tree considerably improves accuracy. The accuracy rate of the first full model without Closing Price is 0.7250 and with Closing Price it is 0.8242. This is to be expected as we see that Closing Price is positively correlated to Being competitive in the correlation matrix. Also, when included, its feature importance is considerably high so it would be a great predictor of competitiveness. However, it is unavailable before the auction ends so there is no way it could be utilized when predicting.

Conclusions and Recommendations

Does this splitting seem reasonable with respect to the meaning of the two predictors? Does it seem to do a good job of separating the two classes?

The splitting seems reasonable with respect to the meanings of the two predictors (*OpenPrice* and *sellerRating*). The splitting uses the two most important and available predictors. For an auction, if the open price is really high, it may scare some buyers away. So, there will be fewer bids and become less competitive for higher open prices. If the seller's rating is high, it could attract more buyers since it indicates better auction

quality. So, they will have more bids and become more competitive for higher seller's ratings. However, it shows the opposite relationship in our model.

However, it does not seem to do a good job of separating the two classes. Since the model limits the number of samples in a terminal node, the leaf nodes are not pure in the optimal model. From the scatter plot above, it is a little bit messy. Also, the accuracy level is not high enough.

Based on this last tree, what can you conclude from these data about the chances of an auction obtaining at least two bids and its relationship to the auction settings set by the seller (duration, opening price, ending day, currency)?

According to the last tree, the chances of an auction getting at least two bids is not affected by auction settings the seller sets except opening price. The rules used by the tree are based on Seller Rating and Opening Price, but they do not use duration, ending day or the currency variables to determine competitiveness. This implies no relationship between obtaining at least two bids and duration, ending day and currency.

The tree does imply a relationship between opening price and obtaining two or more bids and from the tree and correlation matrix (where OpenPrice and Competitive? Have a negative correlation of -0.097), it does seem that a lower opening price increases the chances of an auction being competitive. This is sensible as lower prices tend to attract more customers who then try to outbid each other.

What would you recommend for a seller as the strategy that will most likely lead to a competitive auction?

From our analysis, we found that opening price and seller rating are the best attributes to target when trying to increase chances of having a competitive auction so the strategy we would recommend to the seller is to avoid setting higher opening prices to attract more bids from customers.

As there is a positive correlation between SellerRating and competitiveness, it would also be a great strategy to increase seller rating but unfortunately this is not something the seller can set by themselves. Higher seller ratings reflect the better quality of the goods so there would be more interest in auctions held by such sellers. However, we would also recommend the seller to act in ways that would improve their ratings in the long run at least. For example, they could hold a lot of smaller auctions and carry out efficient transactions for these auctions to increase trust and then hold their main auctions afterwards when they have better ratings. Also, they could reply quickly when interacting with buyers and carry out other activities that reflect a quality seller.

References

Larose, D., & Larose, C. (2015). *Data mining and predictive analytics* (2nd ed.). John Wiley & Sons, Inc.

Final Page

Grade: _____