Rotman School of Management
UNIVERSITY OF TORONTO

Rotman

Group Number:	Group 3
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:

☐	1006507512	☐	1005605374
☐	1002140541	☐	1006527741
☐	1004654527	☐	1006604934

Assignments are to be submitted using Student ID Numbers *only*; do not include your name. Assignments that include names or that do not have the box above checked **will not be graded.**

Please pay attention to Course Outline for specific formatting requirements set by instructors.

Rotman School of Management
UNIVERSITY OF TORONTO

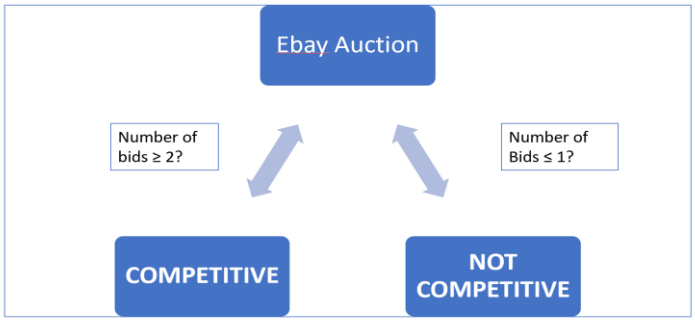
Rotman

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY

The overarching goal of our analysis is to use a decision tree model to predict if an auction on eBay will be competitive. For the purpose of this analysis, a competitive auction is defined as one in which there are at least two bids placed on the item. The dataset contains information on eBay auctions that took place during May-June 2004. The variables included in the data are as follows: auction category, seller eBay rating, the auction terms that the seller selected (duration, open price, currency, and day-of-week auction close), and closing price. In total, there was data on 1972 different auctions.

Exhibit 1: Auction Classification



Given the available data, we developed two decision tree models. The first model used all available predictors. With this model, we are able to classify an auction will be competitive, with 85% accuracy. In the second model, we reduce the number of predictors to simplify our model, so that it better reflects reality. With the second decision tree model, we are able to predict whether an auction will be competitive or not with 71% accuracy. Below we detail our methodology for constructing the decision tree models.

2. DATA PREPROCESSING

The first step of our analysis was to preprocess the data. During this step we completed three main tasks. First we performed data cleaning, then we created dummy variables for categorical variables, and lastly we partitioned the data into training and validation datasets.

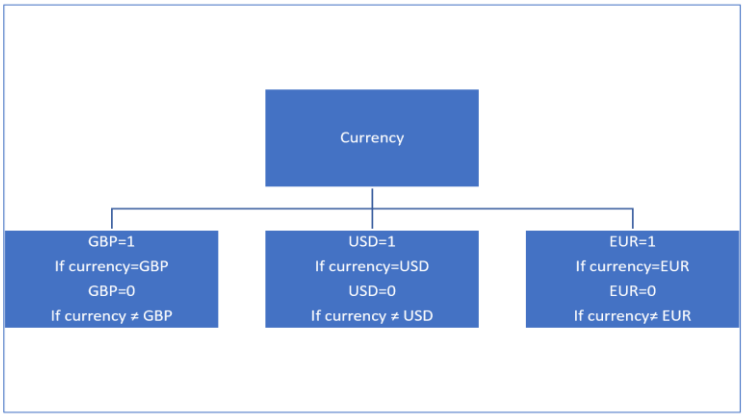
i. Data Cleaning

While cleaning the data, we observed 39 records that were classified as competitive (Competitive = 1), but the opening and closing price were the same. As described in the case that *a competitive auction is one with at least **two** bids placed* on the item being auctioned. Intuitively, we know that for an auction to be competitive, the closing price must be greater than the opening price. Thus, these 39 observations contradict the definition of a competitive auction, and are incorrectly classified as competitive. To correct for the misclassification error, we decided to drop the 39 observations. Thereby making the total number of observations now equal to 1934.

ii. Variable Transformation

We have 4 categorical variables in our dataset. These include Category (18 categories), Currency (USD, GBP, EUR), EndDay (Monday-Sunday) and Duration (1, 3, 5, 7, or 10 days). We converted each categorical variable to dummy variables. Once we get dummy variables for all the four categorical variables, we dropped the original Category, Currency, EndDay and Duration variables. A visual explanation of variable transformation is shown below in the diagram, using the currency categorical variable as an example.

Exhibit 2: Variable Transformation



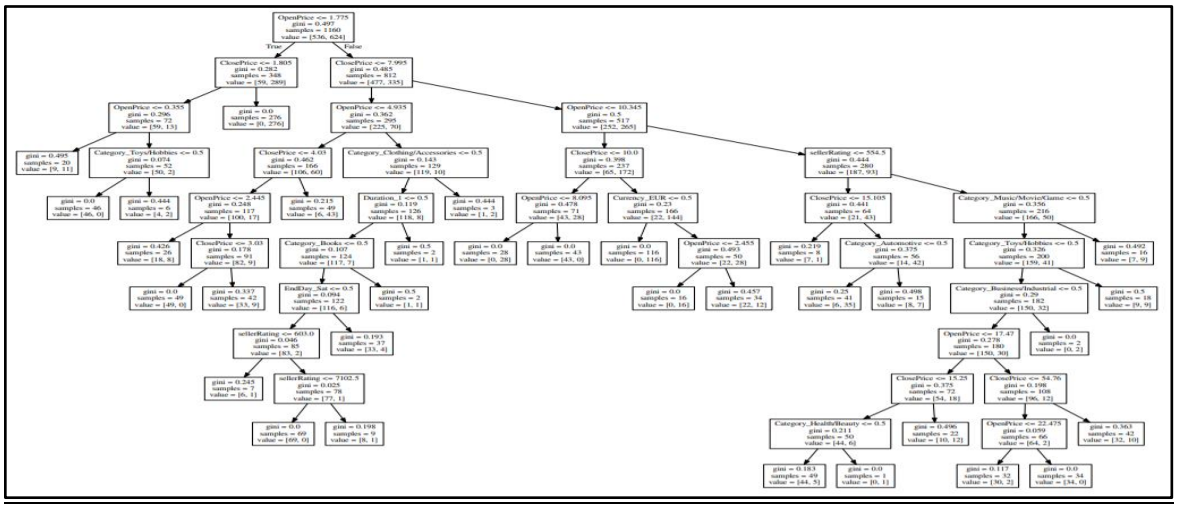
iii. Partitioning the Data

The final step in data preprocessing is to split the data into training and validation datasets. In alignment with the project requirements, the data was split using a 60% : 40% ratio.

3. MODEL 1 RESULTS

i. Classification Tree & Results in Terms of Rules:

Exhibit 3: Decision Tree Classifier (Model 1)

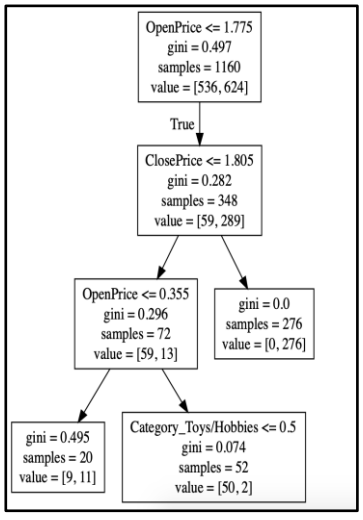


The above graph is the entire decision tree with all predictors. One of the results we get from the decision tree is that given the OpenPrice <= "1.775" -> ClosePrice <= "1.805" -> OpenPrice <= "0.355", then for a new observation, we predict that it is a competitor given the aforementioned conditions.

The following graph is one of the results from the entire decision tree. It describes how the tree is constructed, where each step acts as a condition for the following step. The top node is the "root node", the middle nodes act as "decision nodes", and finally the last node is the "leaf node". The leaf node is the final decision (it is a competitor) given the decision nodes are true.

One of the results based on these rules are given as follows:

Exhibit 4



The root node is (OpenPrice <= "1.775"), if the new observation satisfies the root node, it moves on to the decision node (ClosePrice <= 1.805), the transition is labeled as "True" (left arrow) since the condition is met. Next, it would move on to the next decision node (OpenPrice <= 0.355) if the node 1 condition is met, otherwise it would move to the leaf node (right arrow) with the conclusion that it is a competitor (276/276 = 100% accuracy) . Next, if the (OpenPrice <= 0.355) condition is met, we move to the leaf node (11/20 = 55% accuracy).

Note: *The rest of the results follow the same pattern.*

ii. Discussion of Interesting and Uninteresting Results

The Interesting information we observed is:

(1) The decision tree only use 12 out of 36 predictors, which are ClosePrice, OpenPrice, Caregory_Books, Category_Business/Industrial, Category_Clothing/Accessories, Category_Health/Beauty, Category_Music/Movie/Game, Category_Toys/Hobbies, Currency_EUR, Duration_1, sellerRating and EndDay_Sat, but the other predictors are not included in the trees, so the decision trees does not use them to make decisions

(2) Only Currency_EUR is shown in the trees nodes, but the other currencies: Currency_GBP and Currency_US are not included in the tree

(3) Only EndDay_Sat is shown in the tree nodes, but the EndDay from Monday to Friday, and Sunday are not included in the tree

(4) Only Duration_1 is shown in the tree nodes, but the Duration 3, 5, 7, 10 are not included in the tree

(5) SellerRating is shown in the bottom nodes. We found it is interesting because we thought the auction's competitiveness rely more on the reputation of sellers, but the tree shows that the Price and Category have more impact in the decision trees

(6) Only 6 categories are included in the tree nodes, but the other 12 categories are not, which tells us only these 6 categories will affect the auction as competitive or non-competitive

The Uninteresting information we observed is:

(1) OpenPrice is used as the root node, and ClosePrice is also used in the top nodes, which means they have the most information gains. This is not surprising because OpenPrice and ClosePrice affect most on whether an auction is competitive or not compared to the other predictors. Also, we use OpenPrice not equal to ClosePrice to determine the competitiveness of an auction, so they should be listed as the top nodes in the tree.

iii. Model Tuning: Reduction of Predictors

The initial model has 37 predictor variables, in which a large number of predictor variables are dummy variables (Category, Duration, Currency, and EndDay). This abundance of predictor variables makes our decision tree model overfit the existing data. Thus, we decide to tune our decision model by:

(1) Removing the unnecessary predictor that we do not use in our decision tree

(2) Pruning our decision tree by cutting some of the predictor variables from the tree to improve the accuracy of the model.

Please see the details below for each of those steps:

(1) Firstly, we remove all the 24 predictor variables that we do not use in the decision tree.

Exhibit 5

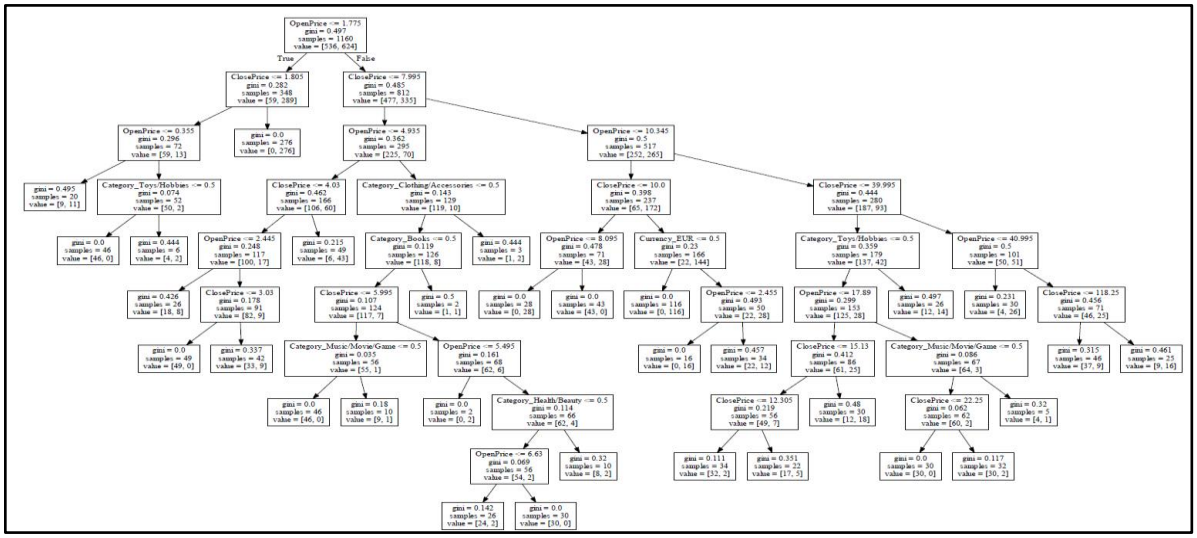
Dropped Variables			
Category_Antique/Art/Craft	Category_EverythingElse	Currency_GBP	EndDay_Tue
Category_Automotive	Category_Home/Garden	Currency_US	EndDay_Wed
Category_Coins/Stamps	Category_Jewelry	EndDay_Fri	Duration_3
Category_Collectibles	Category_Photography	EndDay_Mon	Duration_5
Category_Computer	Category_Pottery/Glass	EndDay_Sun	Duration_7
Category_Electronics	Category_SportingGoods	EndDay_Thu	Duration_10

After we drop all these unnecessary variables, we arrive at the same decision tree and accuracy score as in our initial model (84.75%). Explanation for this is because, as long as we do not remove the variables used in the decision tree, our decision tree remains intact and our accuracy is the same. The next step is to prune the tree in order to improve the accuracy.

(2) Secondly, we remove one predictor variable per trial and test if the accuracy of the model improves. If the accuracy of the model remains the same or drops in value, we ***will not drop that variable*** otherwise, we drop the variable to arrive at a new model with improved accuracy. We start with the variable at the end of the decision tree and move upwards. We identify 2 variables which are"SellerRating" and "Duration_1" whose removal causes the accuracy of the model to improve. We decide to drop these 2 variables together and arrive at a new model with only 10 predictor variables and an improved accuracy score of 85.79%. We make no further drops.

The new tuned decision tree is as follows:

Exhibit 6



iv. Model Practicality

At this stage, the classification tree is not practical for predicting the outcome of a new auction. This is due to the predictors used in model 1 and the overall complexity of the model. The dataset that was used to construct the decision tree classifier included information on opening and closing price of ebay auctions. Evident in Exhibit 3, the decision tree classifier, the most information gain comes from the close price of an auction. This result is obvious, if the close price is known, then you can easily compare the open and close prices to determine if an auction is competitive. *However, when predicting the outcome of a new auction, the closing price is not known until the auction closes. Thus, when predicting the outcome of a new auction, the model renders useless, as the majority of the information gain comes from the close price.*

A second issue with model 1 is the overall complexity. As shown in Exhibit 3, the decision tree classifier has numerous nodes, and requires considerable computing power to classify a new observation. Furthermore, the model complexity also hurts the clarity of presentation. We correct both these issues, related to close price and complexity, in model 2. The second model is more practical and better reflects reality.

4. UPDATED CLASSIFICATION TREE

i. Model Tuning Process:

Since the "ClosePrice" variable should not be included in our predicting model, we need to conduct the classification accuracy by running the model again without ClosePrice variable. The accuracy we get is 70.9%. In order to make sure overfitting or underfitting is avoided, we need a better model, for which we decided to conduct Cross-Validation. By performing Cross-Validation, we identified the optimal depth to be depth -> 6 and then we continued to construct the decision tree with the maximum depth 6, which the model achieved the classification accuracy to be approximately 71.45%. By comparison, our new model provides a better accuracy than the old model, which makes sense considering the old model includes all the variables which might be overfitting.

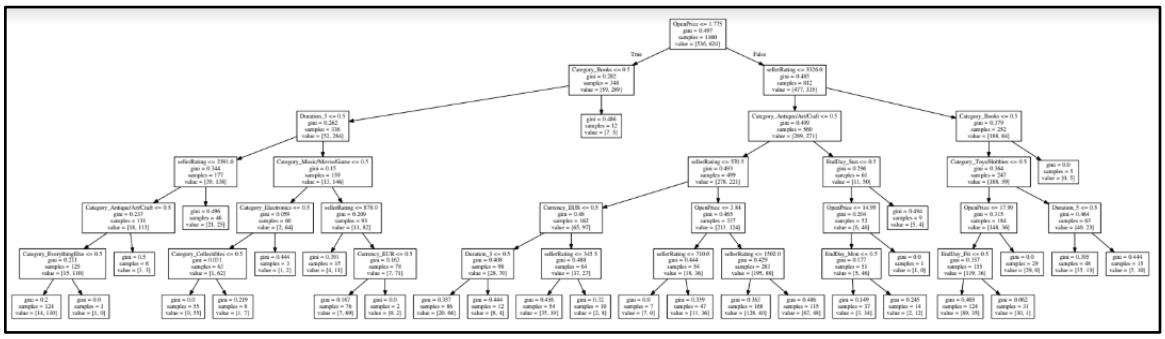
Also, the decision tree picked the following fifteen variables for consideration into the decision making process:

Exhibit 7

Picked Variables		
OpenPrice	Category_Antique/Art/Craft	Duration_3
Category_Books	Category_Electronics	EndDay_Sun
Duration_5	Category_EverythingElse	EndDay_Mon
sellerRating	Category_Collectibles	EndDay_Fri
Category_Music/Movie/Game	Currency_EUR	Category_Toys/Hobbies

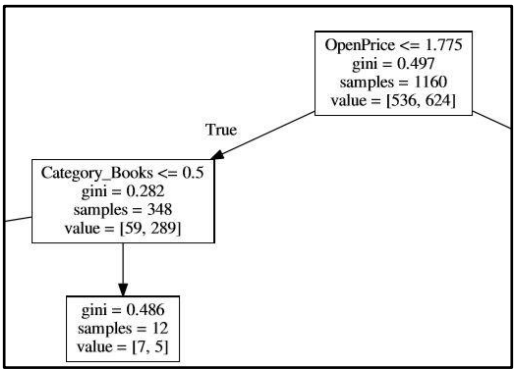
ii. Model 2

Exhibit 8



From the above graph, we can see that the new decision tree is much shorter with the depth 6 after conducting the cross validation process.

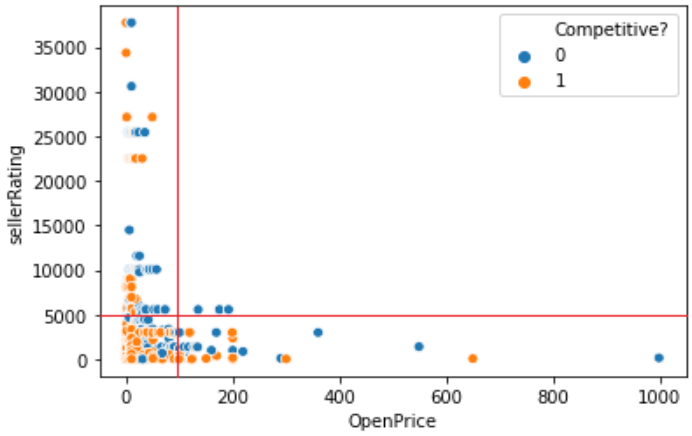
One of the results we get from the decision tree shown on the right is that given the OpenPrice <= “1.775” -> Category_Books > 0.5 (meaning it is in this category), the leaf node contains 7 “non-competitive” and 5 ”competitive” as indicated by value [7, 5], then we predict that it is not competitive with 58.33% confidence because 7/12 records are classified as non-competitive.
Note: *The rest of the results follows the same pattern.*



iii. Scatter Plot & Explanation:

We decided to use two best quantitative predictors to draw our scatter plot, which are “sellerRating” and “OpenPrice”. Then we draw these 2 lines to split the “Competitive?” equal to 1 versus 0. The first line is whether “sellerRating” is higher than 5,000 or not. The second line is whether “OpenPrice” is higher than 100 or not. Please see the graph below for the scatter plot and the splitting lines.

Exhibit 9



This splitting seems reasonable due to the following reasons:

1) There is a significant proportion of "Competitive?" equal to 1 with "sellerRating" less than 5,000. If the "sellerRating" is higher (over 5,000), it means that the seller has a better reputation as he/she has more "good" transactions compared to other sellers. Thus, the buyers are more likely to believe in the "good" seller and that the price the seller demands is reasonable. The auction will then be shorter and both the seller and buyer will come up with the final price quickly. On the other hand, if the seller is considered "bad", he has to go through multiple rounds of bids with the buyer to come up with the final price.

2) If the "sellerRating" is less than 5,000, there is a significant proportion of "Competitive?" equal to 1 with "OpenPrice" less than 100. Possible explanation for this is that when the seller has a lower rating, he/she has a "lower" reputation and tends to open the auction with a low opening price and then push the auction price higher via the following bids that lead to the auction being a competitive one with several rounds of bids.

iv. Analysis of Predictive Performance

The classification/confusion matrix on test dataset to analyze predictive performance can be seen in the table below:

	Predicted Category			
Actual Category		**0**	**1**	**Total**
	0	281 (TN)	89 (FP)	370
	1	132 (FN)	272 (TP)	404
	Total	413	361	774

Fields corresponding to '1' in the table above denote competitive bids. The columns represent the predicted classifications and the rows represent the actual classifications of total 774 observations in the test data. Also, 370 bids are not competitive and 404 bids are competitive. Of the 361 records which were predicted competitive bids by the model, 272 of

them were truly competitive. However, the model incorrectly classified 89 of these records as unsuccessful conversions. The model gives an accuracy of 71.44% with a sensitivity of 75.3% and a specificity of 68.03%. Therefore, if we are interested in obtaining higher sensitivity or higher accuracy of the model to determine actual conversions, then our future task would be to consider different data mining models.

v. Discussion of Competitive Auctions

By analyzing the decision tree, we see that Currency, Duration, OpeningPrice, and EndDay are influential in predicting if an auction will be competitive. This is an impactful result for sellers, as it indicates that the auction settings set by the seller impact if an auction is competitive or not. Based on this, we provide a few recommendations to the sellers, that should be used to maximize the likelihood of an auction being competitive.

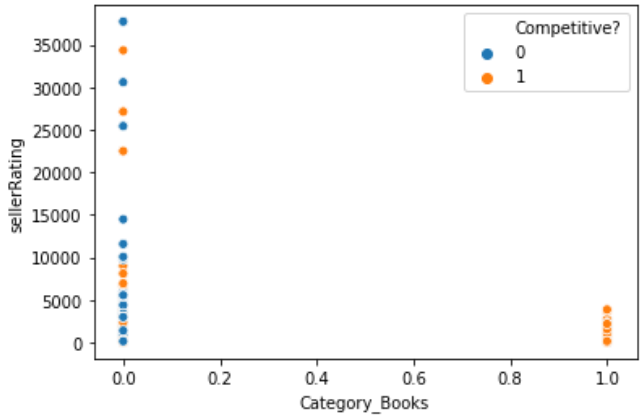
5. SELLER STRATEGY & CONCLUSION

i. Seller Strategy

Given the results of our model, we have determined four strategies that sellers can implement that will most likely lead to a competitive auction. The strategies are detailed below:

a) Evident in the scatterplot below, competitive auctions are likely to occur when the category is books.

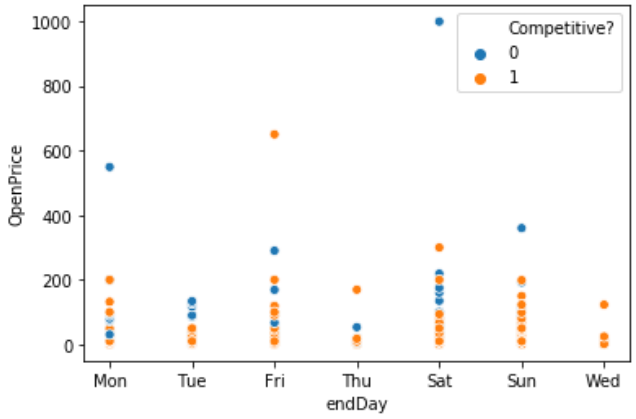
Exhibit 10



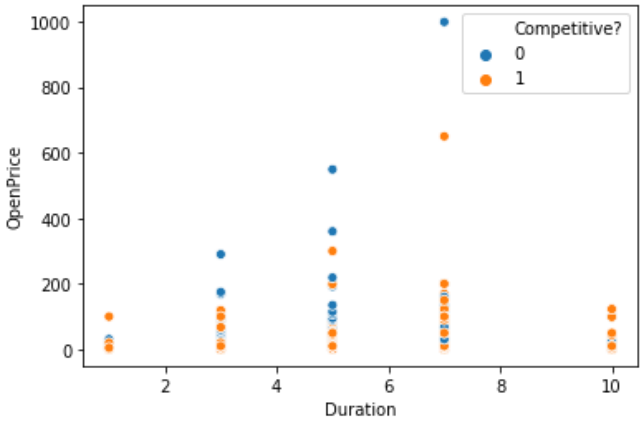
b) If a seller has a low seller rating, then we recommend that the seller sets the open price low. The low opening price will attract bids and drive up the price, leading to a competitive auction. This is evident in *Exhibit 9.*

c) We also recommended that the sellers set the End Day of the auction to be either Wednesday or Sunday to maximize the chance of the auction being a competitive one. This is shown in the scatter plot below.

Exhibit 11



d) The final recommendation is to extend the duration of the auction. Evident in the scatterplot, auctions that have a duration of 10 days are most likely to be competitive.



ii. Conclusion

Multiple iterations of the decision tree model gave us more accurate and robust results. The initial model was run with 36 predictors, which resulted in an accuracy of 84.75%, but at the cost of model complexity. Further, the first model violated one of the fundamental rules of data science, by including closing price in the model. This is because the closing price of an auction is not known until after the auction closes. Thus, it can't be used to classify a new auction. We updated and improved this model by reducing the number of predictors to just 15. Although the accuracy came down to 71.44% the model was much more robust and practical. We utilized the results of this model to generate well rounded strategies that sellers can implement to increase the likelihood of competitive auctions.