eBay Auctions Classification

Data Preprocessing

The file `eBayAuctions.csv` contains information on 1972 auctions that transacted on `eBay.com` during May-June 2004. The goal is to use these data to build a model that will classify auctions as **competitive** or **noncompetitive**. 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. The task is to predict whether or not the auction will be competitive.

Step 1: Preparing the data

1. The variable `Duration`, which was originally numerical, was converted into a categorical variable using the `factor()` function in R.
2. The dataset was randomly shuffled for better distribution of data points into training and validation datasets.
3. The dataset was split into two subsets: a training dataset (60% of the data) and a validation dataset (40% of the data) to ensure fair evaluation of the model.

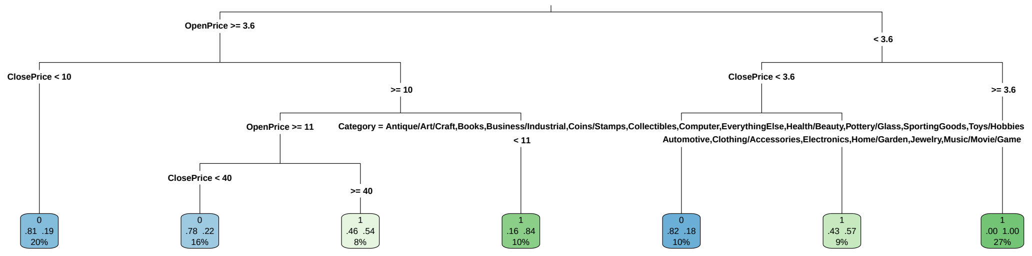
All steps and functions used for preprocessing can be found in the R script.

Step 2: Model Fitting

A classification tree was fitted using the training dataset with the following configurations:

- The dependent variable was `Competitive.`, which indicates whether an auction is competitive or not.
- A minimum bucket size of 50 records was set (`minbucket = 50`) to avoid overfitting.
- The maximum tree depth was limited to 7 levels (`maxdepth = 7`) for better interpretability.
- The tree was pruned based on the complexity parameter (`cp`) to select the optimal model using cross-validation.

The results from the classification tree were visualized using `rpart.plot`.


OpenPrice >= 3.6
< 3.6
ClosePrice < 10
>= 10
ClosePrice < 3.6
>= 3.6
OpenPrice >= 11
< 11
Category = Antique/Art/Craft,Books,Business/Industrial,Coins/Stamps,Collectibles,Computer,EverythingElse,Health/Beauty,Pottery/Glass,SportingGoods,Toys/Hobbies
Automotive,Clothing/Accessories,Electronics,Home/Garden,Jewelry,Music/Movie/Game
ClosePrice < 40
>= 40
0 .81 .19 20%
0 .78 .22 16%
1 .46 .54 8%
1 .16 .84 10%
0 .82 .18 10%
1 .43 .57 9%
1 .00 1.00 27%

The rules were extracted for interpretability. The full implementation details can be found in the R script.

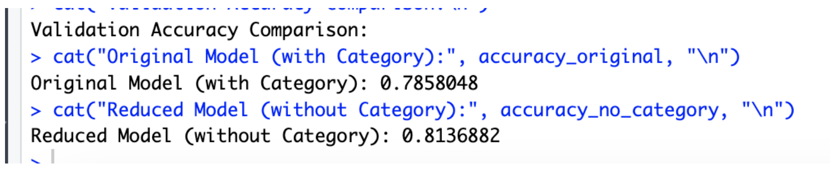
Step 3: Simplification by Reducing Less Effective Variables

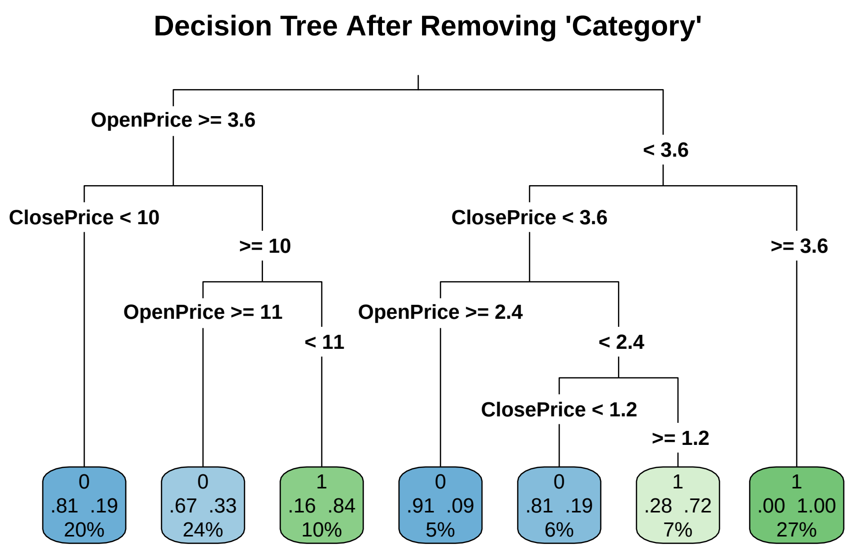
Decision trees select splitting variables based on their ability to reduce impurity (e.g., Gini index or entropy) at each node. A variable will not be chosen for splits if it:

- Does not create significant reductions in impurity, or
- Provides splits with lower predictive power compared to other variables.

In this case, the variables `endDay`, `Duration`, `currency`, and `sellerRating` were not chosen for any splits in the tree. This indicates that these variables do not meaningfully reduce impurity or have predictive power compared to stronger variables like `OpenPrice` and `ClosePrice`, which have much stronger relationships with the target variable (`Competitive.`).

The results indicate that removing the `Category` variable improved the validation accuracy of the model from **78.58%** (with `Category`) to **81.37%** (without `Category`). This suggests that `Category`, despite appearing in the original tree, was not a strong predictor and may have introduced noise or unnecessary complexity. If the number of predictors needs to be reduced due to software limitations or for clarity of presentation, `Category` would be a good candidate for removal. Simplifying the model by removing `Category` resulted in better generalization and improved predictive performance.


```
Validation Accuracy Comparison:
> cat("Original Model (with Category):", accuracy_original, "\n")
Original Model (with Category): 0.7858048
> cat("Reduced Model (without Category):", accuracy_no_category, "\n")
Reduced Model (without Category): 0.8136882
>
```



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

The model, after removing `Category`, demonstrates improved accuracy (**81.37%**) and simplicity, making it practical for analyzing completed auctions. It primarily relies on `OpenPrice` and `ClosePrice`, with `OpenPrice` being known before the auction starts.

However, the inclusion of `ClosePrice`, which is only available after the auction ends, limits the model's practicality for real-time prediction of new auctions. To make the model fully practical for predicting new auctions, `ClosePrice` would need to be excluded or replaced with features available before the auction begins.

In its current form, the model is best suited for analyzing past auctions or predictions based on estimated `ClosePrice`.

c. Describe the interesting and uninteresting information that these rules provide.

Interesting Information

- **Role of `OpenPrice` and `ClosePrice`**: The rules show that `OpenPrice` and `ClosePrice` are the most important variables for predicting auction competitiveness. For example, the first split at `OpenPrice >= 3.6` shows that auctions with lower starting prices tend to be less competitive.
- **Useful Thresholds**: Key thresholds, such as `ClosePrice < 10` and `OpenPrice >= 11`, provide clear patterns:
  - Auctions with `OpenPrice >= 11` and `ClosePrice >= 10` are very likely to be competitive.
  - Auctions with `OpenPrice < 3.6` and `ClosePrice < 1.2` are less likely to be competitive.
- **Practical Insights for Sellers**: Sellers can use these rules to set opening prices strategically. Higher starting prices (`OpenPrice`) often lead to higher competitiveness.

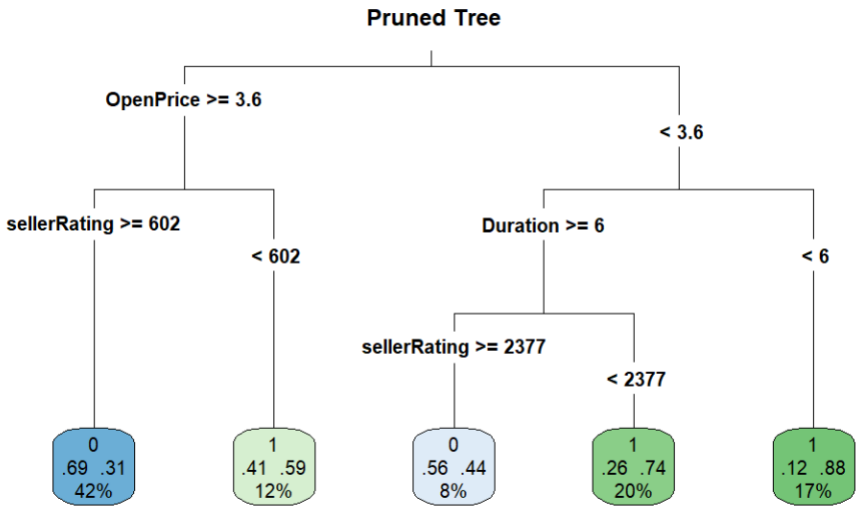
Uninteresting Information

- **Limited Variety of Predictors**: The tree only uses `OpenPrice` and `ClosePrice`, which limits its ability to provide insights about other factors, such as seller reputation or auction timing.
- **Deep Splits with Little Value**: Deeper splits, such as `OpenPrice < 2.4` and `ClosePrice < 1.2`, add complexity but do not give much new or useful information.
- **No Influence of Item Types**: Removing the `Category` variable makes the tree simpler, but it also means the model cannot explain how different item types affect auction competitiveness.

d. Fit another classification tree, this time only with predictors that can be used for predicting the outcome of a new auction.

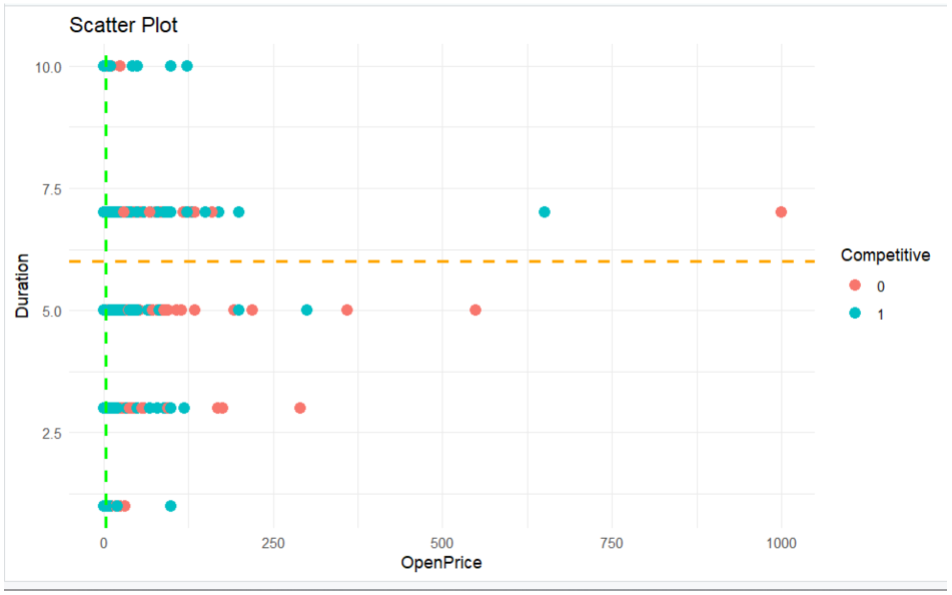
To avoid overfitting on the training data the decision tree model is pruned. The complexity parameter, cp, helps in finding the optimum size of the tree by regulating the degree of branching of the tree. If a tree is too large, it risks fitting the noise rather than the patterns in the training data and will generalize poorly on unseen data. The cp value is chosen by crossvalidation xerror so that the model is not unnecessarily complex while the error on unseen data is smallest.

The pruned tree generated is the best since it is simple with few splits. Also has a good predictive accuracy by selecting a cp that minimizes the cross-validation error. The model uses predictors that are available at the start of an auction which include OpenPrice, Duration, and sellerRating. These features are actionable and relevant because they can be used by auction participants to predict the competitiveness of a listing. These are the features the pruned tree focuses on in order to make decisions. The pruned tree is less complex and more interpretable which allows for good for decision-making.



e. Plot the resulting tree on a scatter plot.

The scatter plot from the decision tree shows the relationship between OpenPrice (x-axis) and Duration (y-axis). The points are separated based on whether the auction is competitive (1) or non-competitive (0). This splitting seem reasonable with respect to the meaning of the two predictors. An auction is less likely to be competitive as the opening price of an auction increases. The time duration of the auction is also a reasonable feature to predict competitiveness. A longer duration could lead to lower competitiveness since bidders might wait for better deals or move to other auctions. This analysis shows that the decision tree effectively captures the relationship between OpenPrice, Duration, and the competitiveness of auctions. Additionally, the splits seem meaningful with the predictors.


Scatter Plot
Duration
OpenPrice
Competitive