Text, letter

Description automatically generated

9.1

a)

Diagram

Description automatically generated

As we can see in this tree, the set of rules are as follows:

Text

Description automatically generated

First it shows us the distribution of the total unclustered data, of the 1183 data, 524 are competitive, that is, 44.3% of the data are competitive.

The first cluster that is created is based on the openprice variable, if this is greater than or equal to 3.7 the probability that they are competitive is 62.7% while if it is less than 3.7 the probability that they are competitive is 24.1%.

Within the cluster of those with openprice greater than or equal to 3.7, two other groups are created, this time based on the closeprice variable. If this variable is less than 20.2, there is a 75.8% probability that they will be competitive, while if it is greater than or equal to 20.2, there is a 41% probability that they will be competitive.

Within the group of those with an openprice greater than or equal to 3.7 and a closeprice greater than or equal to 20.2, two other groups are formed, those with an openprice greater than or equal to 20.5, which have a 63.1% probability of being competitive, and those with an openprice less than 20.5, which have a 20% probability of being competitive.

On the other hand, from the group with an openprice less than 3.7, two groups are created, those with a closeprice value greater than or equal to 3.7, which have a 60% probability of being competitive, and those with a closeprice less than or equal to 3.7, which have a 1% probability of being competitive.

Within the first group, those with an openprice less than 3.7 and a closeprice greater than or equal to 3.7, the last two groups are created, those with an openprice greater than or equal to 2.4, which have a probability of being competitive of 89.6% and those with an openprice less than 2.4, which have a probability of being competitive of 44%.

b)

Yes, because knowing the openprice of the auction it is possible to know the probability that it will be competitive, for example, in the first group, which distinguishes between those with an openprice greater than or equal to 3.7 have a 63% probability of being competitive, while those with an openprice less than 3.7 have a 24% probability of being competitive.

c)

What we can conclude after looking at the rules of this model is that the most interesting variables are openprice and closeprice, while the rest of the variables do not affect the construction of the model at all.

d)

The variables that we have eliminated for this are all those that cannot be known until after the auction is over, which are endday and closeprice.

The resulting model after eliminating these variables is as follows:

Diagram

Description automatically generated

A black screen with white text

Description automatically generated with low confidence

In the rules of this model, as in the previous one, a first classification is made into two clusters, those with an openprice greater than or equal to 3.7, which have a 62.8% probability of being competitive, and those with an openprice less than 3.7, which have a 24.1% probability of being competitive.

Also within the first cluster, those with an openprice greater than or equal to 3.7, we can make a new classification into two new clusters, those with a sellerrating greater than or equal to 562, which have a 69.2% probability of being competitive and those with a sellerrating less than 562, which have a 38.6% probability of being competitive.

e)

We ran a plot using seller rating on the x-axis and open price on the y-axis, nevertheless we can see that all the observations are grouped around the origin, which makes it hard to read. Therefore, we converted the scale to a log scale, which makes reading the graph much clearer.

We can see that on the competitive observations (red) usually come from a seller with a lower rating, when compared with the non-competitive ones (blue) which tend to come from a higher rated seller.

f)

Text

Description automatically generated

As we can see in the confusion matrix the model has an accuracy of 0.69, i.e. it successfully predicts 69% of the data.

In addition, we can see that the model has more sensitivity than specificity, which means that it has more errors when predicting the auctions that are not competitive, since there have been 144 observations that the model has predicted to be competitive and actually were not, compared to 97 observations that the model has predicted to be non-competitive and actually were.

g)

A black screen with white text

Description automatically generated with low confidence

As we have seen previously in the rules there are two fundamental aspects when evaluating the probability of an auction being competitive or not, the openprice and the sellerrating, so that a strategy that raises the probability of an auction being competitive would be to reduce the price of the auction to be less than 3.7, since there is a 62.8% probability that more than two bids will be made, while otherwise there would only be a 24.1% probability.

As for the sellerrating, we can also see that auctions in which the seller has 562 or more sellerrating have a 69.1% probability of being competitive, while if the sellerrating is less than 562 this probability decreases to 38.6%.

For this reason, another strategy that could have a significant impact would be to take into account and work on certain aspects that increase the seller rating, such as: item description, communication with the buyer, shipping time, shipping and handling chargers.

Text, letter

Description automatically generatedText, letter

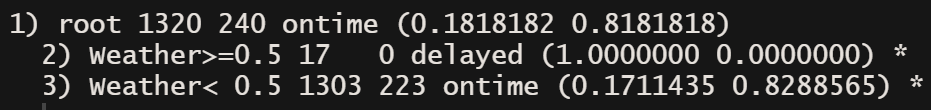
Description automatically generated

9.2

a)

Diagram

Description automatically generated



According to the set of rules of this model, the probability of a flight being delayed or not depends only on the weather variable.

b)

We cannot use this tree because we must know the Weather. The redundant information is the day of week (Monday) and arrival airport (EWR). The tree requires knowing whether the weather was inclement or not. We may not know the weather in advance

c) i)

In the pruned tree we get a single terminal node labeled "ontime". Therefore, any new flight will be classified as being "on time". This is called naïve rule.

c) ii)

This is equivalent to the naive rule, which is the majority rule. In this dataset most of the flights arrived on time, and therefore the naive rule is to classify a new flight as arriving on time.

c) iii)

The top three predictors according to this tree are “CARRIER = CO, DH, MQ, RU”, “DAY\_WEEK = 1, 2, 5, 7”, and “DEST = JFK”. Therefore, they are the most influential factors in the forecast of the punctuality of a flight.

c) iv)

The model's complexity is mostly to blame for the tree's single node outcome. Not a single branch falls below the complexity threshold. This is due to the fact that weather before wasn't as complicated and thus functioned. We no longer have a mechanism to create separations in our tree once we deleted it, rendering the model naïve.

c) v)

Since there is no complexity cap on the unpruned tree, the model may have incredibly complicated top levels without us ever realizing it because we would treat them as excellent predictors.

The tree is then pruned so that we can determine how impactful and difficult a specific split would be because we do have a complexity restriction.

c) vi)

Decision trees can make mistakes in classification situations with several classes and few training samples. Decision trees are less appropriate for estimation scenarios like logistic regression when the goal is to predict the value of a continuous feature.