Question1.

a)

1. The data should be partitioned into training and validation sets because we need two sets of data: one to build the model that depicts the relationship between the predictor variables and the predicted variable and another to validate the model's predictive accuracy.
2. The training data set is used to build the model. The algorithm 'discovers' the model using this data set.
3. The validation data is used to 'validate' the model. In this process, the model (built using the training data set) is used to make predictions with the validation data - data that were not used to fit the model. In this way, we get an unbiased estimate of how well the model performs. We compute measures of 'error,' which reflect the prediction accuracy.

b)

The equation to predict median house price:

MEDV = -28.81068 + (-0.26072 * CRIM) + (3.76304 * CHAS) + (8.27818 * RM)

c)

CRIM = 0.1, CHAS = 0 and RM = 6

Hence, Median house price (Y) = -28.81068 + (-0.26072 * CRIM) + (3.76304 * CHAS) + (8.27818 * RM)= -28.81068 + (-0.26072 * 0.1) + (3.76304 * 0) + (8.27818 * 6) =20.832

The prediction error is the difference between the predicted value and the actual value (if any) in the dataset.

d) Use stepwise regression with the three options (backward, forward, both) to reduce the remaining predictors as follows: Run stepwise on the training set. Choose the top model from each stepwise run. Then, use each of these models separately to predict the validation set. Compare RMSE, MAPE, and mean error, as well as lift charts. Finally, describe the best model.

You should get a final selected model in each of the three cases (backward, forward, and stepwise). Then, you use RMSE to compare them and select the best one.

Question2.

a)

i. logit = -14.7208 + (89.8326 * TotExp_Assets) + (8.3713 * TotLnsLses_Assets)

ii. Odds = e^(logit) = e^(-14.7208 + 89.8326 * TotExp_Assets + 8.3713 * TotLnsLses_Assets

iii. p = odds / (1+odds) = 1 / (1 + e^(-logit)) = 1 / (1 + Exp[-(-14.7208 + (89.8326 * TotExp/Assets) + (8.3713 * TotLns&Lses/Assets)])

b)

TotLns&Lses/Assets = 0.6; TotExp/Assets =0.11

The Logit:

=-14.118+79.964*TotExp/Assets+9.173*TotLns&Lses/Assets =-14.118+79.964*0.11+9.173*0.6 =0.112

The Odd:

= exp(-14.118+79.964*0.11+9.173*0.6)=0.894

The Probability:

= e^(-14.118+79.964*0.11+9.173*0.6)/[1+ e^(-14.118+79.964*0.11+9.173*0.6)]=0.528

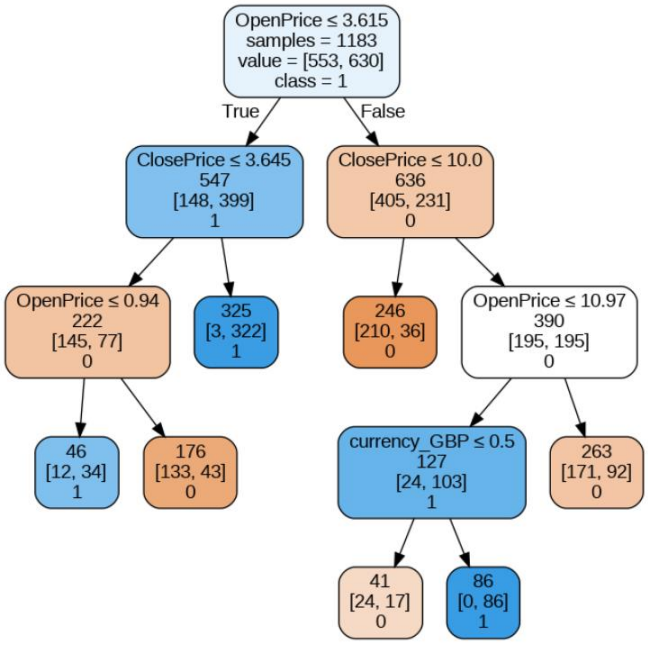
c) The cutoff value of p=0.5.
Odds = (p) / (1-p) = (0.5) / (1-0.5) = 1
If odds > 1, then classify financial status as "weak" (otherwise classify as "strong").
Logit = ln (odds) = ln (1) = 0
If Logit > 0, then classify financial status as "weak" (otherwise, classify it as "strong")
Therefore, a cutoff of 0.5 on the probability of being weak is equivalent to a threshold of 1 on the odds of being weak, and to a threshold of 0 on the logit.

d) β1 is positive. So an increase in (TotLns/Assets) is associated with an increase in the odds.

e) Cutoff should be decreased to minimize the number of outcomes that are predicted as 0 (Financially strong) for more accurate predictions.

Question3.

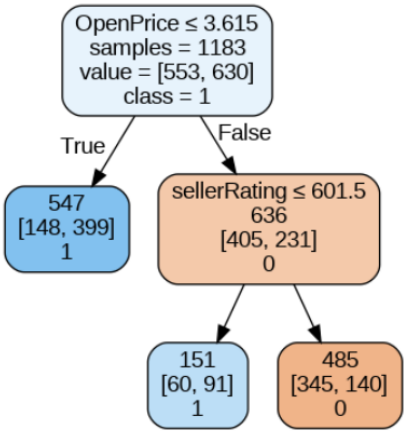
a)



b) No, because the close price is not known at the start of the auction.

c) The main effect of a competitive auction is to raise the item's price. So, by including both the opening and closing price as predictors, our tree is naturally going to focus repeatedly on those two variables since they, by themselves, are the most powerful predictors of a competitive auction. In order to make the tree pay attention to more useful variables, we will need to remove the closing price as a predictor.

d)



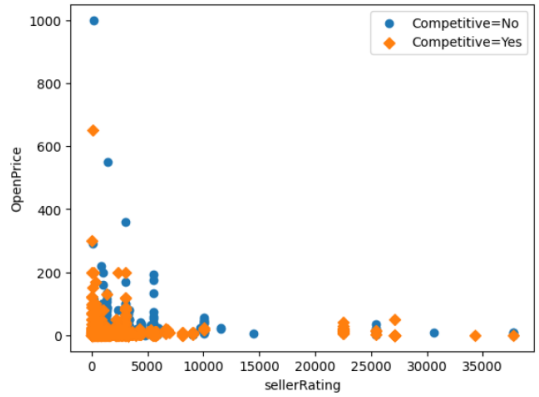
Set of rules

If (OpenPrice <= 3.615) then class = 1

If (OpenPrice > 3.615) and (sellerRating <= 601.5) then class = 1

If (OpenPrice > 3.615) and (sellerRating > 601.5) then class = 0

e) The splitting points are located way down in the lower left corner so the scatterplot does not reveal much to answer the question. We could do a log transform, or restrict the scatterplot to the smaller values.

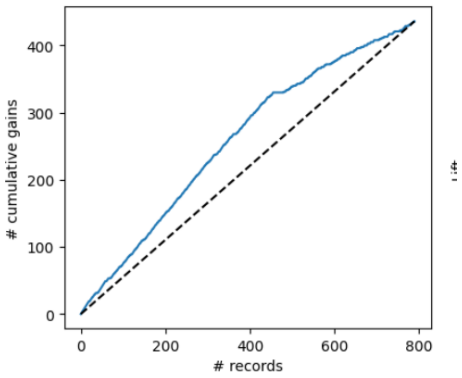


f)

```
Validation Set: Confusion matrix

Confusion Matrix (Accuracy 0.7072)

       Prediction
Actual   0   1
     0 228 125
     1 106 330
```



From the lift chart we see that the model's predictive performance (i.e. correctly capturing the auctions that are most likely to be competitive) is better than the baseline model, since its lift curve is higher than that of the baseline model. The decile lift chart has a limited set of lift values because the tree is a simple one, producing only three potential predicted probabilities.

g)

To get a competitive auction, the most important factor controlled by the seller is the opening price, with lower opening prices attracting more bidders. From the tree we see that if the opening price < $3.615, then it will lead to a competitive auction. In particular, it appears that setting the opening price to the minimum of $0.01 (which is eBay's default) is most likely to lead to a competitive auction.