Homework 7 Possible Answers

Abra Brisbin

2021-08-03

Intellectual Property:

Important: This document contains R code solutions and example answers for problems posed in the DS740: Data Mining homework. We are intentionally sharing this document with learners *who have already completed the associated homework.* We want you to be able to review and troubleshoot your code, as well as ask questions to prepare you for later assessments. *By using this document, you are accepting responsibility* **not** *to share it with anyone else, including other students in the course or the program who might not have completed the homework yet.* By upholding this agreement, you are helping us use this tool with learners in future terms.

From Problem 1

Question

After loading the OJ data set, make *STORE* and *StoreID* factor variables. Set the random seed equal to 7 and take a random sample of 800 rows of the data. This will be the training data set; the remaining observations will be the validation set.

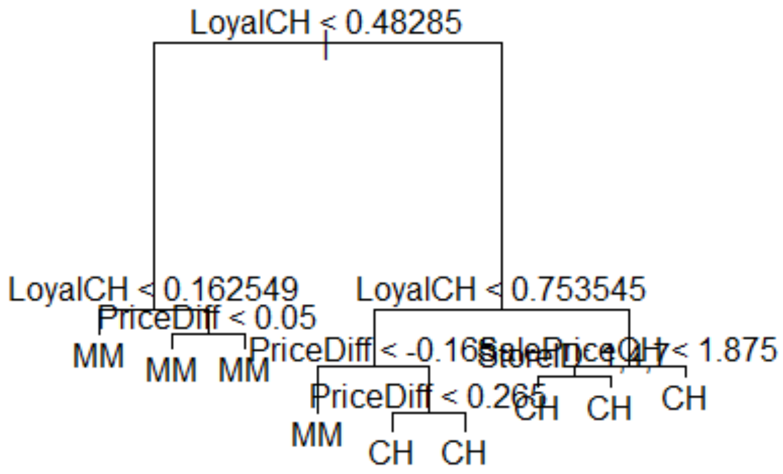
Possible Answer:

```r
OJ <- OJ %>%
  mutate(STORE = factor(STORE),
         StoreID = factor(StoreID))

set.seed(7)
groups = c(rep(1, 800), rep(2, 270)) # 1 represents the training set
random_groups = sample(groups, 1070)

in_train = (random_groups == 1)
```

Question

Plot the decision tree with category labels.

Possible Answer:



Question

For the decision tree you just uploaded, write 3–5 sentences interpreting the fitted model. Your interpretation should include the phrase "This makes sense because…"

- Note that `PriceDiff` represents (Sale Price of MM) - (Sale Price of CH). So, negative values of `PriceDiff` mean that Minute Maid is cheaper.

Possible Answer: The most important factor in which brand of orange juice a customer buys is their loyalty score. If a customer's loyalty to the Citrus Hill brand is less than .483, then we can predict that they will buy Minute Maid, and if their loyalty is more than .754, then we can predict that they will buy Citrus Hill (although the Citrus Hill sale price, price difference, and StoreID are associated with varying levels of confidence in these predictions). This makes sense, because customers with higher values of loyalty to Citrus Hill are predicted to buy Citrus Hill.

Customers with an intermediate loyalty are predicted to buy Citrus Hill if it is not more than 16.5 cents more expensive than Minute Maid. This makes sense, because customers with a mild brand loyalty may not stick with their preferred brand if it is much more expensive than the alternative.

From Problem 2:

Question

After loading the Hitters data set, remove the observations with unknown salary. **Create a new variable** equal to log(*Salary*), and then remove the original Salary variable.

Possible Answer:

```
my_hitters <- Hitters %>%
  filter(!is.na(Salary)) %>%
  mutate(log_salary = log(Salary)) %>%
  select(-Salary)
```

Question

Set the random seed to 7 again. Then perform 10-fold cross-validation to compare boosting (using the same parameters as in the previous question) to multiple linear regression.

Possible Answer:

```
set.seed(7)

n = dim(my_hitters)[1]
k = 10 # Using 10-fold CV
groups = rep(1:k, length = n)

cvgroups = sample(groups, n)
boost_predict = numeric(length = n)
linear_predict = numeric(length = n)
data_used = my_hitters

for(ii in 1:k){
  groupi = (cvgroups == ii)
  # Perform boosting on everything not in groupi
  # Then predict values for groupi
  boost = gbm(log_salary ~ ., data = data_used[!groupi, ],
              distribution = "gaussian", n.trees = 5000,
              shrinkage = .001, interaction.depth = 4)


  boost_predict[groupi] = predict(boost, newdata = data_used[groupi, ],
n.trees = 5000, type = "response")

  # Perform linear regression on everything not in groupi
  # Then predict values for groupi
  linear = lm(log_salary ~ ., data = data_used[!groupi, ])
  linear_predict[groupi] = predict(linear, newdata = data_used[groupi, ])
}
```

From Problem 3:

Question

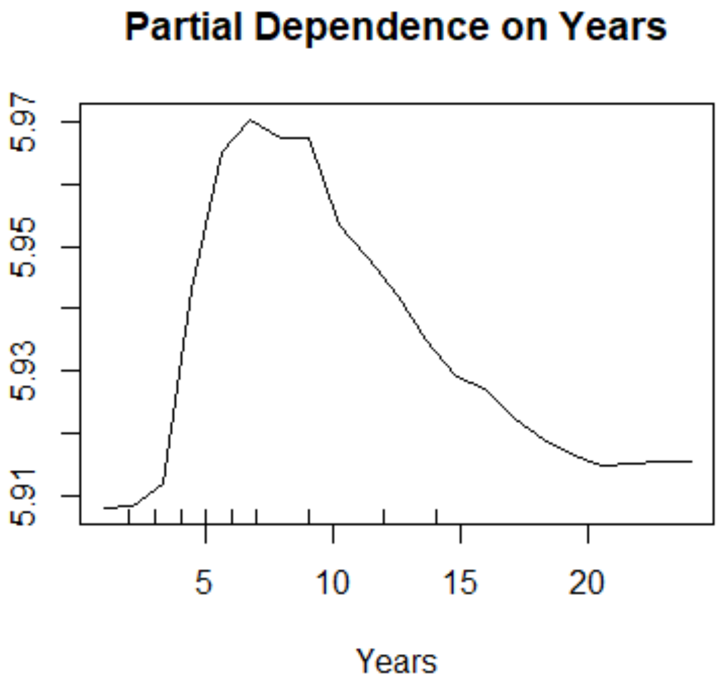
Use ?Hitters to view what each variable in the data set represents. Examine some preliminary graphs and summaries of the data. If we want to use decision trees to analyze this data set, why are random forests a good idea? Explain in 2-4 sentences.

Possible Answer: The data set contains several variables that people who are familiar with baseball would expect to be correlated (such as at-bats, hits, and runs); high correlations are confirmed using pairs() and corrplot(). Using random forests will enable us to reduce the correlations in the predictions that stem from having correlated variables, by choosing from among a subset of variables at each node. This is likely to reduce the variance in our estimates.

Question

Make a partial dependence plot of the variable you chose in the previous problem.

Possible Answer:



Question

Write 2-4 sentences interpreting the relationship between predicted log(Salary) and the variable you selected in the previous two problems. Include a possible explanation for the shape of the relationship.

Possible Answer: The relationship between Years and predicted log(Salary) is non-monotonic. Players with between 5 and 10 years of experience are predicted to have the highest salaries, while players with more and less experience are predicted to have lower salaries. One possible explanation is that after players reach a certain age, it may be harder–or perceived to be harder–to maintain a top level of physical fitness, so their teams pay them less.

However, this is based on data from only two seasons, so this “trend over time” may not hold. Another possibility is that each player’s salary is roughly constant during their career, but the more experienced players started their careers before inflation caused a big jump in starting salaries.