Homework 6

GSBA 524

Fall 2022, Term 3

Please submit *two* files for this homework:

1. A report document (with extension `.nb.html`) that is created in Radiant (as we saw in the MLR lab, this is available from Report > Rmd.
2. The state file (with extension `.state.rda`) that is created when you click "Save Radiant state file."

Background

In this problem we will be working with `baseball.csv`, which contains data on Major League Baseball hitters playing in the 1986 and 1987 seasons. Our goal will be to predict the salary of a hitter in 1987 based on information about the player's career leading up to that point. Here is information about each variable in the data set:

```
AtBat
Number of times at bat in 1986

Hits
Number of hits in 1986

HmRun
Number of home runs in 1986

Runs
Number of runs in 1986

RBI
Number of runs batted in in 1986

Walks
Number of walks in 1986

Years
Number of years in the major leagues

CAtBat
Number of times at bat during player's career

CHits
Number of hits during player's career
```

```
CHmRun
Number of home runs during player’s career

CRuns
Number of runs during player’s career

CRBI
Number of runs batted in during player’s career

CWalks
Number of walks during player’s career

League
A factor with levels A and N indicating player’s league at the end of 1986

Division
A factor with levels E and W indicating player’s division at the end of 1986

PutOuts
Number of put outs in 1986

Assists
Number of assists in 1986

Errors
Number of errors in 1986

Salary
1987 annual salary on opening day in thousands of dollars

NewLeague
A factor with levels A and N indicating player’s league at the beginning of 1987
```

Problems

1. Make a plot of salaries versus 1986 home runs. What do you observe?
2. Create a new column in the dataset called `training` that randomly assigns each hitter to either a training or test set[1]. Set the filter to be `training==1` so that all models fit below will be trained specifically on the observations in which `training==1`. Now fit a model where you predict salary based on home runs. Show the output of this model.
3. Is there evidence at the 0.05 level that home runs is a statistically significant feature in this model? Justify your answer.
4. What are the units of β_{HmRun} in this model?
5. Give a 95% confidence interval for the coefficient β_{HmRun}.
6. Is 0 included in the confidence interval from Problem 5? Explain how you could have answered this question without even looking at the confidence interval itself.
7. Now fit a model that includes runs in addition to home runs. Provide the output.

[1] Please keep the defaults how they are, with `Seed` set to 1234 and `Size` set to 0.7.

8. Is there evidence at the 0.05 level that home runs is a statistically significant feature in this model? Justify your answer.

9. Relate your answers in Problems 3 and 8 to each other and explain what may be happening. Make a plot to help in your explanation.

10. Add a column called `pred_hmrun` to the data set that has the predicted salaries using the model in Problem 2. Include in this answer the Radiant output you get after clicking "Store predictions." For the first hitter in the dataset, give the actual salary, predicted salary, and how far off the prediction is from the actual value. Was this hitter in the training set or the test set?

11. Now, perform backward stepwise regression, where you start with all features in the model (but be sure not to include the variables `training` and `pred_hmrun`, which are columns in our data set that are not actually features!). Show the output.

12. What were the first, second, and third features to be dropped?

13. How many features remain in the selected model?

14. Add a column called `pred_step` to the data set that has the predicted salaries using the model selected by stepwise regression. For the first hitter in the dataset, give the actual salary, predicted salary, and how far off the prediction is from the actual value.

15. Make a table showing both the training and test errors for the `pred_hmrun` and `pred_step` models.

16. Which of these two models has a lower test mean absolute error (MAE)?

17. What are the units of the MAE in this problem?

18. Fit a neural network with 3 hidden units[2] that uses all the variables (again, be sure to exclude the `training` and any columns you've created of predictions). Display the output.

19. Include a picture of the network itself.

20. Add a column of predictions called `pred_nn3`. Show the Radiant output you get after clicking "Store predictions."

21. Show the table of training and test errors with the results from this neural net included. How does the test MAE of this neural network compare to the two regression models?

22. Now fit a neural network with 15 hidden units, create a column of predictions called `pred_nn15`, and recompute the table of errors. Show the output of the neural net, show the Radiant output you get after clicking "Store predictions," and show the training/test error table. How does the test MAE of this neural network compare to the one with 3 hidden units?

23. Of all the models we've considered, which appears to be the best at predicting salary?

[2] Keep `Decay` at 0.5 and `Seed` at 1234.