STATS 202: DATA MINING AND ANALYSIS HOMEWORK #4

INSTRUCTOR: LINH TRAN, HOMEWORK #4, DUE DATE: AUGUST 11, 2023,
STANFORD UNIVERSITY, AND STUDENT: ADAM KAINIKARA

Note: For This HW I Would Like To Use Both of My Late Acceptance Free Passes

Part 1. **Problem 1 (10 points) Chapter 8, Exercise 4 (p. 362).**

See Picture

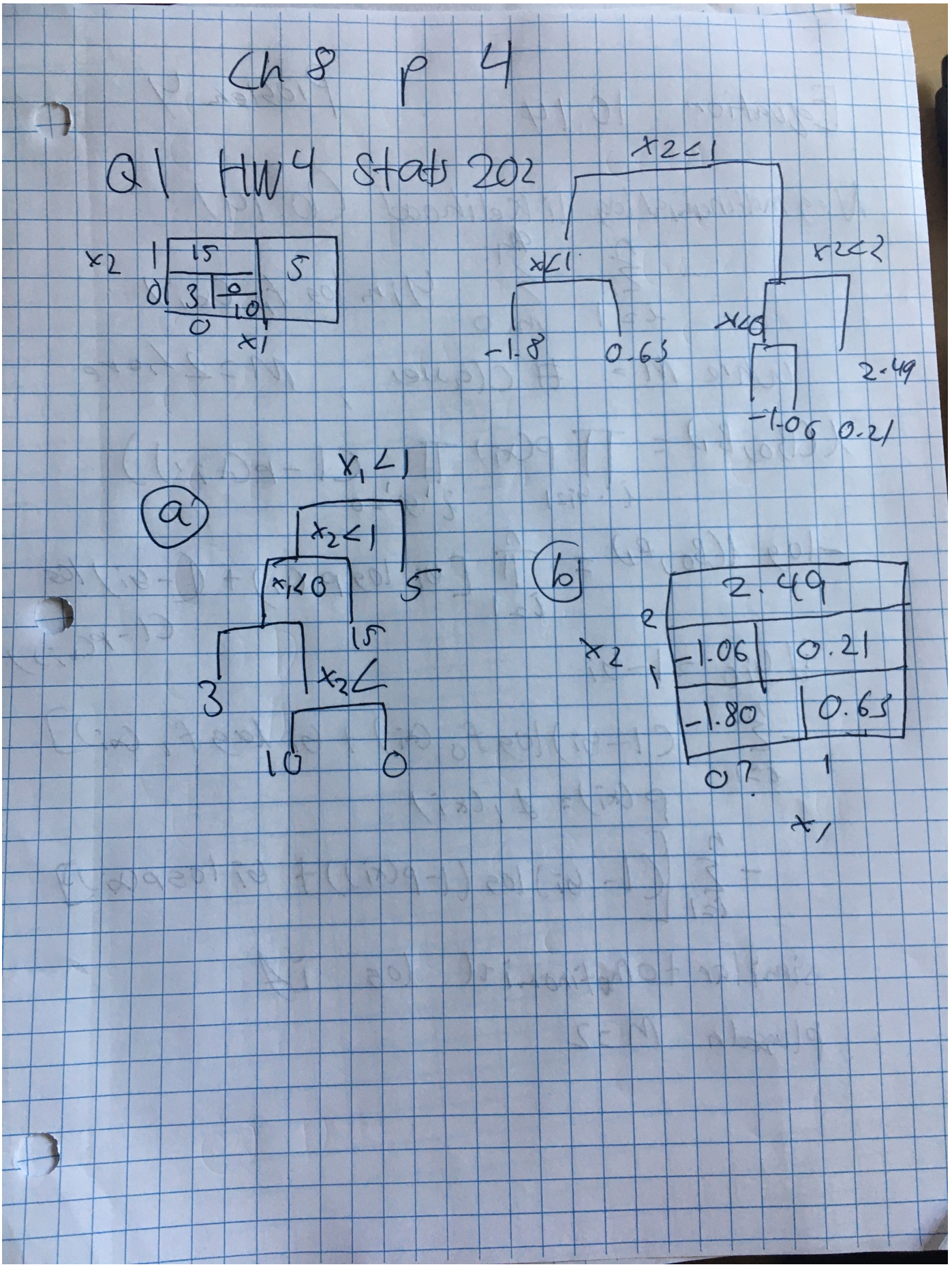


FIGURE 0.1. Partition Predictor Space, Tree Corresponding

Part 2. **Problem 2 (10 points) Chapter 8, Exercise 8 (p. 363).**

a) Loaded the data. Used the first 300 observations as training data and remaining as test data.

b)Below is the decision tree obtained. The text of the decisions is there but there is no way for me to zoom into it and make it visible without compromising viewing the entire tree.

Got training MSE of: training MSE [6.92555, 6.01548, 4.51463, 3.63198592, 2.780706106, 1.9490206, 1.3444905, 0.9582072, 0.6184801, 0.318548768, 0.1670528, 0.0808429, 0.034383, 0.008528, 0.0008045002, 1.5000000000000248e-06]. I used a max range of 1 to 100. After 1.5e-0.6 it becomes 0. As max range increases, training MSE decreased.

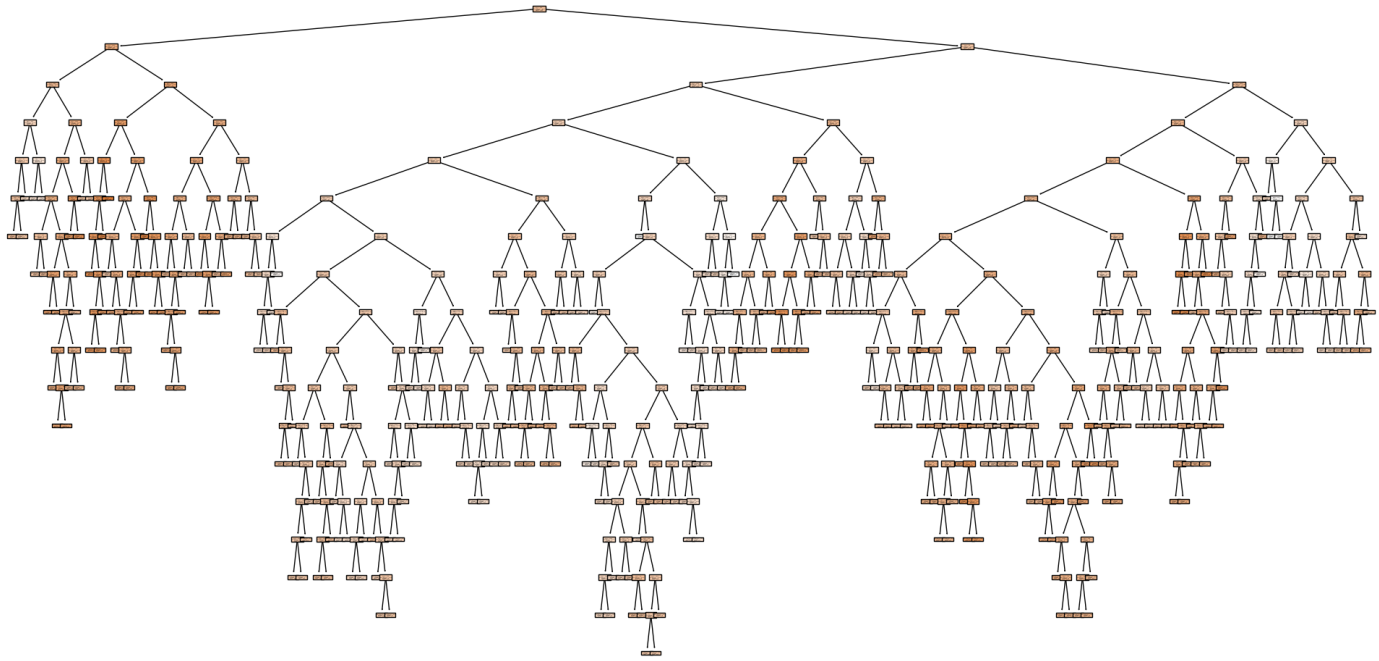


FIGURE 0.2. Decision Tree

c) Used 50 fold cross validation. Test MSE: [6.5926, 6.3392, 6.7435, 6.4871, 6.8642, 7.2016, 6.9819, 6.9257, 6.9609, 7.3169, 6.8377, 6.8694, 7.2913, 7.1097, 6.8807, 7.0151, 7.7205, 6.6586, 7.3512, 7.7440, 7.1296, 7.4179, 7.8666, 7.4043, 7.4034, 7.2810, 7.5770, 7.5597, 7.2096, 7.4301, 6.8801, 7.2097, 7.4453, 7.3694, 7.6687, 7.7158, 7.7056, 7.4671, 7.2502, 7.4823, 7.8073, 7.5877, 7.3345, 7.6565, 7.0715, 7.3730, 7.6791, 7.0353, 8.1583, 7.4158, 7.5074, 7.8412, 7.4064, 7.2881, 7.0706, 7.6808, 7.7743, 7.4709, 7.6717, 7.2558, 7.7232, 7.2517, 7.4788, 7.1370, 7.7501, 7.5733, 7.1611, 7.3242, 7.5635, 7.1397, 7.0876, 7.7972, 7.1674, 7.4782, 7.6583, 7.3527, 7.6408, 7.4755, 7.3450, 7.3133, 7.2930, 7.2004, 7.6946, 7.4337, 7.7125, 7.4135, 7.6839, 7.0104, 6.8647, 7.2573, 7.3605, 7.3134, 7.3692, 7.6356, 7.4716, 7.2165, 7.5625, 7.2234, 7.3998, 7.1178]

Cross-validation MSE: [7.7341, 7.4336, 6.6419, 7.4464, 7.6584, 8.2210, 8.3577, 8.5913, 8.4691, 8.4290, 8.9589, 9.0193, 8.9341, 9.3736, 9.0285, 9.3532, 8.9794, 8.8623, 9.0660, 8.8862, 9.4002, 9.3255, 9.2470, 9.3459, 9.7372, 9.0191, 9.0686, 9.1765, 8.8849, 9.2638, 9.1233, 9.0015, 9.3625, 9.1690, 8.8440, 9.2550, 9.4318, 8.9734, 9.2675, 9.1688, 9.6290, 9.7384, 8.9791, 9.4063, 9.1986, 8.8951, 9.1064, 9.0307, 8.6922, 8.8019, 9.2963, 9.4685, 9.4004, 9.3235, 9.1433, 8.9465,

8.8973, 9.3494, 8.9791, 8.9821, 9.0734, 8.9869, 8.6723, 9.2343, 9.6795, 9.0792, 9.0142, 9.1403, 9.2006, 9.1159, 9.1023, 8.9102, 9.2750, 9.2088, 8.8536, 8.8738, 8.8296, 9.1930, 9.5657, 9.2780, 9.6003, 9.5085, 9.0811, 9.1254, 9.7180, 9.0156, 9.1386, 9.0511, 9.2512, 9.0739, 9.4404, 9.3263, 9.1220, 9.2052, 9.1429, 9.3893, 9.1047, 9.1564, 9.1626, 9.2333]. In the long run Test MSE gradually increases. Here is a graphical representation of it.

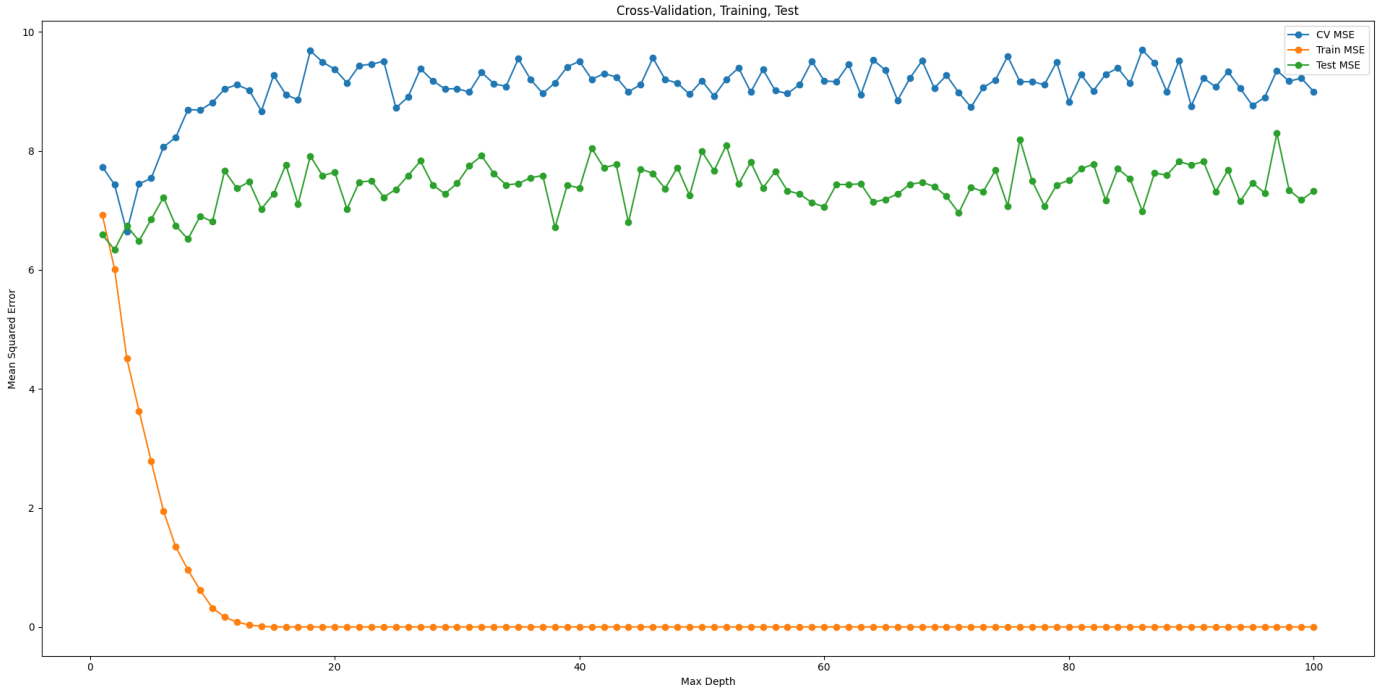


FIGURE 0.3. Cross Validation, Training, Test

d) Bagging reduced the error. CompPrice, Income, and Advertising where the 3 most important features. Got a bagging MSE of 4.8781.

e) Used a Random forest and varied max features.

Max Features: 0.2 Feature Importance: [0.13856 0.13733 0.12714 0.13383 0.23324 0.15025 0.07965]

Max Features: 0.4 Feature Importance: [0.14698 0.11921 0.12795 0.11191 0.28469 0.14311 0.06611]

Max Features: 0.6 Feature Importance: [0.17436 0.09798 0.13299 0.08696 0.32181 0.13276 0.05311]

Max Features: 0.8 Feature Importance: [0.18431 0.08943 0.13647 0.08047 0.32886 0.12985 0.05058]

In order this is CompPrice, Income, Advertising, Population, Price, ShelveLoc, Age, Education, Urban, US

Here is a graphical representation of how changing max features affected MSE.

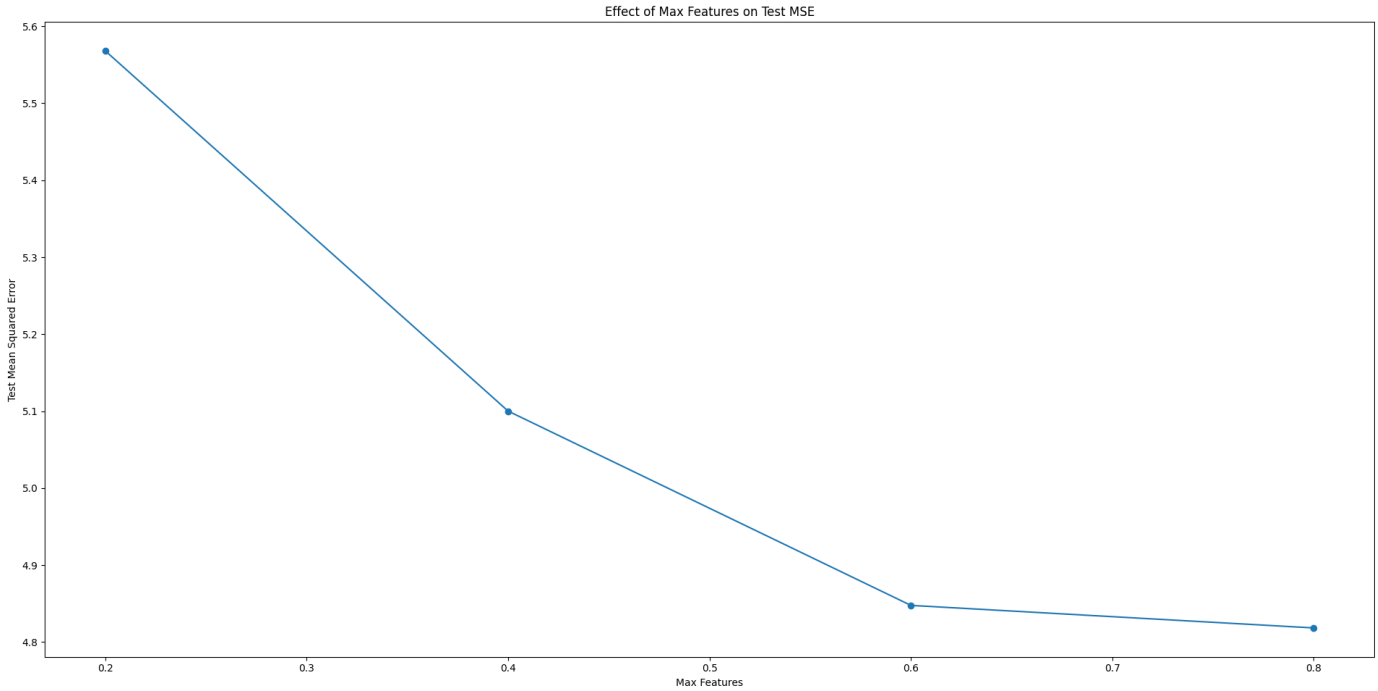


FIGURE 0.4. Random Forest

f) Not sure if BART completely worked. Went through a lot of loop holes for it to start. Got a MSE of 0.0479

Part 3. **Problem 3 (10 points) Chapter 8, Exercise 10 (p. 364).**

a) Imported the data. Made a loop that went over the salary column. If 'NA' was in the column, that individual was removed. In total about 63 people were removed. The remaining people had their salaries log transformed.

b) Split the data where training was first 200 observations, test was remaining. X was all the quantitative values. Y was log transformed salary.

c) Performed gradient boosting with 1000 trees. The range of values for λ where [0.001, 0.005,0.01, 0.05, 0.1, 0.15, 0.5, 1]. Below is a plot of different shrinkage values on the x-axis and the corresponding training set MSE on the y-axis.

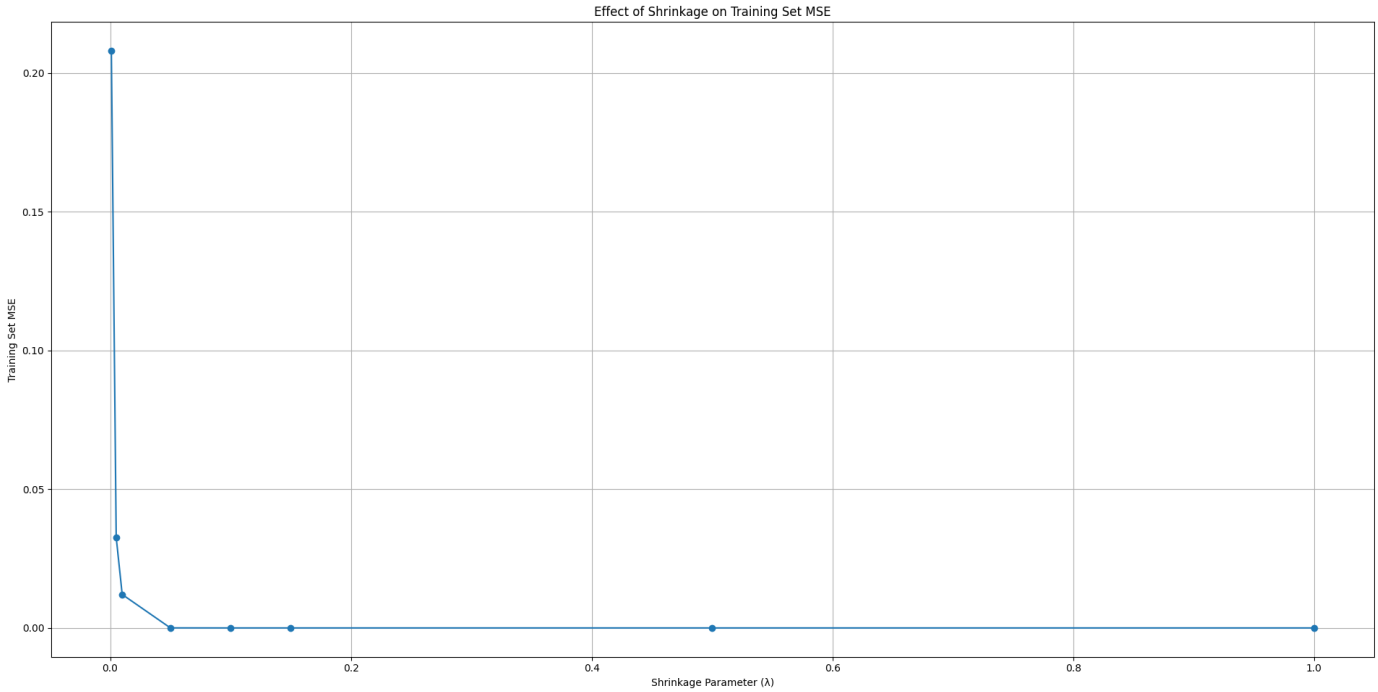


FIGURE 0.5. Shrinkage on Training

d) Same as above but now with test MSE.

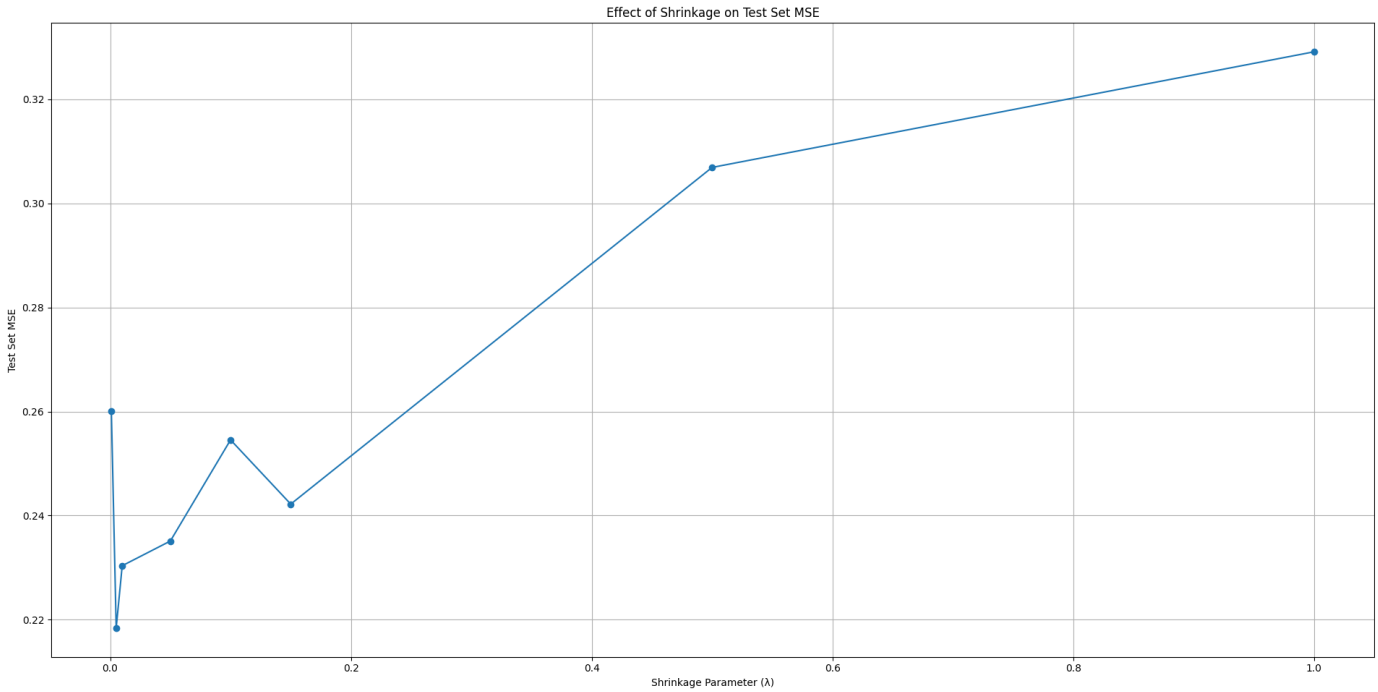


FIGURE 0.6. Shrinkage on Test

e) Train MSE Boosting [0.2079, 0.0326, 0.0120, 1.9360e-05, 1.5571e-08, 6.4497e-12, 2.2169e-16, 2.1421e-16]

Test MSE Boosting [0.2600, 0.2184, 0.2304, 0.2351, 0.2546, 0.2422, 0.3069, 0.3291]

The two regression approaches I used where Linear regression and Ridge regression.

Test MSE Linear 0.51394

Test MSE Ridge 0.5139, 0.5139, 0.5139, 0.5139, 0.5139, 0.5139, 0.5139, 0.5138, 0.5136

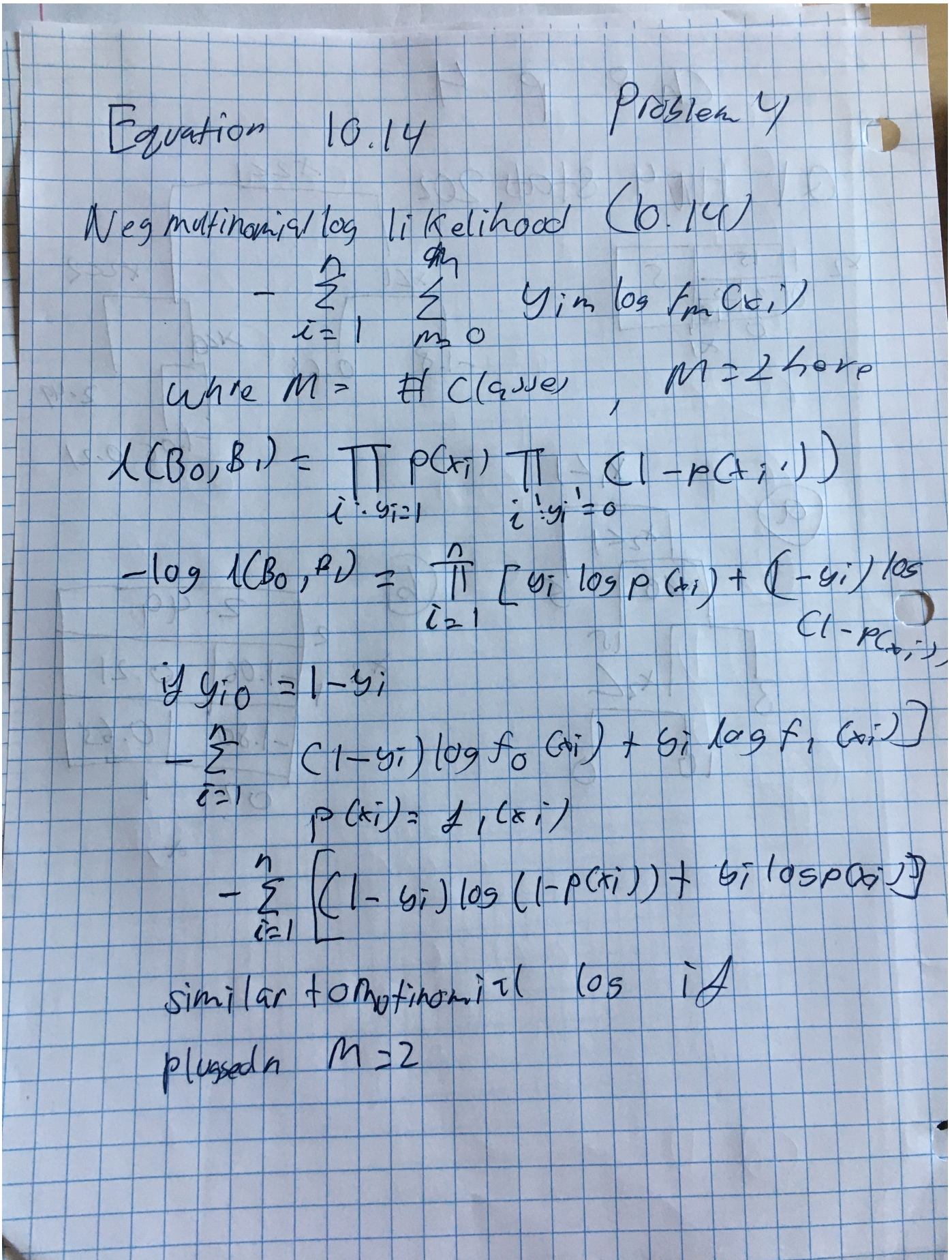
The test MSE of linear and ridge regressions where greater than that of boosting

f) Did feature importance. These where the values. The 3 most important variables where Catbat, Chits, Cruns.

('catbat', 0.5562) ('chits', 0.0875) ('cruns', 0.0525) ('walks', 0.0481) ('cwalks', 0.0419) ('chmrun', 0.0403) ('atbat', 0.0385) ('crbi', 0.0336) ('hits', 0.0279) ('years', 0.0241) ('putouts', 0.0160) ('rbi', 0.0132) ('assits', 0.0072) ('errors', 0.0070) ('hmrun', 0.0060)

g) Test MSE Bagging 0.22657

Part 4. **Problem 4 (10 points) Chapter 10, Exercise 3 (p. 459).**



Part 5. Problem 5 (Bonus 10 points) Let xi : i = 1, ..., p be the input predictor values and a(2s) k : k = 1, ..., K be the K-dimensional output from a 2-layer and M-hidden unit neural network with sigmoid activation sv(a) = {1 + e—a}—1 such that a(1s) j = w(1s) j0 + p∑ i=1 w(1s) ji xi : j = 1, ..., M a(2s) k = w(2s) k0 + M∑ j=1 w(2s) kj sv (a(1s) j) Show that there exists an equivalent network that computes exactly the same output values, but with hidden unit activation functions given by tanh(a) = ea—e—a ea +e—a , i.e. a(1t) j = w(1t) j0 + p∑ i=1 w(1t) ji xi : j = 1, ..., M a(2t) k = w(2t) k0 + M∑ j=1 w(2t) kj tanh (a(1t) j) Hint: first derive the relation between sv(a) and tanh(a). Then show that the parameters of the two networks differ by linear transformations.

Given:
$\sigma(a) = \frac{1}{1+e^{-a}}$
$tanh(a) = \frac{e^a - e^{-a}}{e^a + e^{-a}}$
Want to find the relationship between $\sigma(a)$ and $tanh(a)$
Factor out e^a as a common factor
$tanh(a) = \frac{e^a(1-e^{-2a})}{e^a(1+e^{-2a})} = \frac{1-e^{-2a}}{1+e^{-2a}} = \frac{1}{1+e^{-2a}} - \frac{e^{-2a}}{1+e^{-2a}}$
$\sigma(2a) = \frac{1}{1+e^{-2a}}$
Therefore
$tanh(a) = \sigma(2a) - \frac{e^{-2a}}{1+e^{-2a}}$
Now dealing with $\sigma(a) = \frac{1}{1+e^{-a}}$
$1 - \sigma(a) = 1 - \frac{1}{1+e^{-a}}$
$1 - \sigma(a) = \frac{1+e^{-a}}{1+e^{-a}} - \frac{1}{1+e^{-a}}$
$1 - \sigma(a) = \frac{e^{-a}}{1+e^{-a}}$
$1 - \sigma(2a) = \frac{e^{-2a}}{1+e^{-2a}}$
Therefore
$tanh(a) = \sigma(2a) - (1 - \sigma(2a))$
$tanh(a) = 2\sigma(2a) - 1$
$\sigma(2a) = \frac{tanh(a)-1}{2}$
Show that the parameters of the two networks differ by linear transformations
$a_k^{(2s)} = w_{k0}^{(2s)} + \Sigma_{j=1}^{M} w_{kj}^{(2s)} \sigma(a_j^{(1s)})$
$a_k^{(2t)} = w_{k0}^{(2t)} + \Sigma_{j=1}^{M} w_{kj}^{(2t)} tanh(a_j^{(1t)})$
Replace $tanh$ and σ with relationships found earlier
$a_k^{(2s)} = w_{k0}^{(2s)} + \Sigma_{j=1}^{M} w_{kj}^{(2s)} 2\frac{tanh(a)-1}{2}(a_j^{(1s)})$
Multiply by 2 cause it is $2a$
$a_k^{(2t)} = w_{k0}^{(2t)} + \Sigma_{j=1}^{M} w_{kj}^{(2t)} 2\sigma(2a-1)(a_j^{(1t)})$
$a_k^{(2s)} = w_{k0}^{(2s)} + \Sigma_{j=1}^{M} w_{kj}^{(2s)} tanh(a) - 1(a_j^{(1s)})$
If this was multiplied out and simplified it would become
$w_{kj}^{(2t)} = 2w_{kj}^{2s}$
$w_{k0}^{(2t)} = w_{k0}^{(2s)} - \Sigma w_{kj}^{2s}$

A linear transform can be seen between the two with the 2 and the constant