INDR 450 Homework 4 Report

Umur Berkay Karakaş
69075

Question 1

a) I fitted the regression model using training set with all predictors, but I couldn't add the results to appendix because it doesn't fit to my screen, and I couldn't take a screenshot of it. The results can be seen in the code.

b) At 5 percent significance level, only x1t, x2t, x6t and d1t are significant. R^2 is 0.506 and adjusted R^2 is 0.390. Both are too low, so we can infer that even though we have many predictors, they are not sufficient to explain the Audi sales.

```
RMSE of full model on training set: 489.12
MSE of full model on training set: 239239.06
```

c)

```
RMSE of full model on test set: 1018.20
MSE of full model on test set: 1036730.55
```

d) I used two methods to find an optimal alpha value, manual iteration over selected alpha values and lassocv. Using manual iteration, I was able to eliminate more predictors with a trade-off for slightly higher RMSE. Therefore, I think manual iteration was more reasonable approach.

```
RMSE of reduced model with manual iteration on training set: 584.5474
RMSE of reduced model with manual iteration on test set: 1182.0019
RMSE of reduced model with lassocv on training set: 491.2503
RMSE of reduced model with lassocv on test set: 1035.6985

Predictors after using lassocv:

['t', 'tsq', 'x1t', 'x2t', 'x3t', 'x4t', 'x6t', 'x7t', 'x8t', 'x9t', 'x10t
', 'd1t', 'd2t', 'd3t']

Predictors after using manual iteration:

['t', 'tsq', 'd1t']
```

Question 2

a) I tried [2,3,4,5,8,9,10,11,12,15,20,30] for depth of regression tree. I found the optimal depth value for which the test RMSE is the minimum. If I were to minimize train error, highest depth value would be returned so it would have been pointless. At the end, I found that depth = 8 is the optimal depth value.

```
Train RMSE of the regression tree: 122.89
Train MSE of the regression tree: 15101.44
```

```
Test RMSE of the regression tree: 1512.36
Test MSE of the regression tree: 2287232.21
```

b) I found a method for bagged regression tree, BaggingRegressor in sklearn library. I initialized the regressor with bootstrap_features = True so that features can be drawn with replacement.

```
Train RMSE of the bagged regression tree: 256.42
Train MSE of the bagged regression tree: 65751.14
Test RMSE of the bagged regression tree: 1321.82
Test MSE of the bagged regression tree: 1747204.34
```

c) For random forest, initially I tried 1 to all features and I was minimizing over test RMSE. The algorithm returned that test RMSE is minimized with using only 1 feature. Then, I tried 3 to all features and the optimal number of maximum features was returned as 5.

For max_features = 5:

```
Train RMSE of the random forest: 216.05
Train MSE of the random forest: 46677.32
Test RMSE of the random forest: 1123.59
Test MSE of the random forest: 1262464.75
```

d) From Figure 1 and the code, it is observed that top 3 important predictors in random forest are **'t', 'd1t', 'tsq'.**

e) Among learning rate values of ['0.1', '0.2', '0.3', '0.4', '0.5', '0.6', '0.7', '0.8', '0.9', '1.0'] and depth values of [2,3,5,10,15,20], I minimized over test RMSE and found that lr = 0.1 and depth = 2 are the optimal parameters.

```
Train RMSE of the boosted tree: 230.59
Train MSE of the boosted tree: 53173.17
Test RMSE of the boosted tree: 1134.49
Test MSE of the boosted tree: 1287065.39
```

f) From Figure 2 and the code, it is observed that top 3 important predictors in boosted tree are **'tsq', 'd1t', 't'.**

Question 3

a)

Method	Train RMSE	Test RMSE	Predictors	Spec.
Full Regression	489.12	1018.20	'x1t', 'x2t', 'x6t', 'd1t'	-
Lasso	584.55	1182.00	't', 'tsq', 'd1t'	alpha = 20
Regression Tree	122.89	1512.36	-	max_depth = 8
Bagged Tree	256.42	1321.82	-	-
Random Forest	216.05	1123.59	't', 'tsq', 'd1t'	max_features = 5
Boosted Tree	230.59	1134.49	't', 'tsq', 'd1t'	lr = 0.1, depth = 0.2

b) It is clear that trend ('t'), square of trend ('tsq') and 1-month lag ('d1t') are the most important predictors since lasso, random forest and boosted tree returned them as the most important ones.

All tree methods can fit the training data really well, but they do not predict the test set as good as they fit the training set, so there is clearly an overfitting or data problem, even though I selected the parameters to minimize test RMSE. It is strange to have the least test RMSE in full regression since the adjusted R-squared was only 0.390. Since full regression has not an overfitting issue and has the least test RMSE, I would still use full regression until better predictors are presented to the problem.

Appendix

Figure 1: Variable importance for random forest

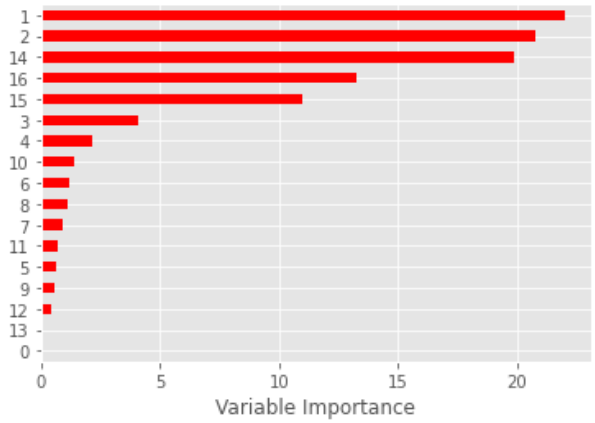

Figure 2: Variable importance for boosted tree