|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Method*** | ***Train RMSE*** | ***Test RMSE*** | ***Predictors*** | ***Spec.*** |
| ***Full Regression*** | 489.120 | 1862.522 | All of them are used. | From trend terms, t and tsq values together are statistically insignificant. Monthly dummies at month 3, month 4, month 5, month 9, month 10, month 11 are statistically insignificant. Lastly, difference term at lag 3 is statistically insignificant |
| ***Best Lasso\**** | 516.115 | 1010.747 | All of them instead of predictors 5, 11 and 13. | Predictors 5, 11 and 13 are eliminated |
| ***Best Regression Tree\**** | 439.765 | 1654.158 | All of the predictors are used | Maximum depth is 4. |
| ***Bagged Tree*** | 234.228 | 1185.621 | All of them are used | All of the features are used |
| ***Best Random Forest\**** | 207.682 | 954.076 | Predictors’ significance in decreasing order:  1-0-13-15-14-2-3-5-9-6-7-11-4-10-8-12 | Used 2 features |
| ***Best Boosted Tree\**** | 225.985 | 909.433 | Predictors’ significance in decreasing order: 1-0-13-15-14-2-3-9-5-7-6-10-8-4-11 | Maximum depth is 3, and learning rate is 0.001 |

***SUMMARY TABLE & Homework Report***

***Doruk ÖZER***

***70192***

R^2 is 0.5 which means the %50 of the training data is able to be predicted precisely . According to P>|t| values t and tsq values together are statistically insignificant. It may indicate that there might be no trend in the dataset so there is no trend term is necessary for this prediction or it may also indicate that only one of the trend term is enough to predict the trend if it exists so one of them can be discarded. In other words, t and tsq can be creating an autocorrelation so that one of them could be eliminated by Lasso or Rigid or there might be no trend terms in the data. Monthly dummies at month 3, month 4, month 5, month 9, month 10, month 11 are statistically insignificant. Lastly, difference term at lag 3 is statistically insignificant. The regression may be giving poor performance also because of lack of amount of observations in training set. Some approaches that reduces variance should be implemented such as tree based approaches.

The worst results at the test data and the best result on training data are given by full regression. The full regression contained too many statistically insignificant predictors and R^2 value were 0.5. The insignificant variables and complexity of variables gives the sign of overfitting and it reflected to the test RMSE. By sacrificing the training RMSE, some of the predictors such as predictor 5, 11 and 13 are discarded by lasso. Lasso discarded some of the insignificant predictors and increased the test RMSE significantly (test RMSE for full regression were 1862 and it improved to 1010). Some of the overfitting issue is solved by lasso. Increase in the test set performance is accomplished by sacrificing the performance at the training set.

Then I tried the tree based approaches. First trial was with the regression tree which uses all of the predictors. I experimented with different kind of depths and it is found out that best performance at the test set is given by the full regression tree that has maximum depth of 4. Increase in the depth after depth 4 increases the training set performance but it decreases the test set performance. Which means, overfitting starts after depth 4. The second approach was using bagged full-regression tree that uses all of the predictors. The results of bagged tree on test set is much better than the best regression tree. Because bagged tree has a feature that reduces the variance and creates many subsets. And each set is used to train the decision trees, so it reduces the overfitting issue. The random forest approach also decreases the variance and by limiting the number of features that are used in trees. It also reduces more the overfitting issue. From the experiments best performance at the test set is given by using only 2 predictors. It improved the test set performance significantly compared to bagged tree approach. It is found out that predictors’ significance at random forest, in decreasing order is 1-0-13-15-14-2-3-5-9-6-7-11-4-10-8-12. Final approach was using boosted tree. Boosted tree gives better results compared to random forest that uses 2 features at test set. The best performance of boosted tree is given by restricting the depth to 3 and determining the learning rate to 0.001. It gives the best test set performance among all of the approaches that are used in this homework. It is found out that predictors’ significance at boosted tree, in decreasing order is 1-0-13-15-14-2-3-5-9-6-7-11-4-10-8-12.

I would be using boosted tree that has maximum depth 3 and learning rate 0.001. It gives the best performance at the test set and solves the overfitting issue. The tree based approaches performed better in this homework because we had really lack of observations at training set. It is best to use tree based approaches. Methods such as, bagging, random forest and boosting are giving better performance at test set because those approaches creates different kind of subsets from the current data, limits the number of predictors, solves overfitting issues and decreases the variance with lack of amount of observations. To sum up, boosted tree that has maximum depth 3 and learning rate 0.001 gave the best result among the all approaches.