**BANA 7047 – Prof. Yan Yu**

**Individual Case I**

Last Name \_\_\_Newkold\_\_\_\_\_\_\_\_\_\_\_

First Name \_\_\_Tess\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

M#\_\_\_\_11434445\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

A picture containing invertebrate

Description automatically generatedPlease use your **M#** to set the seed to draw a random sample.

Signature \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Please use this as your cover page and follow **exactly** the required format below. Points will be deducted otherwise.

Please submit a WORD copy via blackboard with file name 7047-00X case#-last 4 digits of your M#.docx.

**Case reports**: Please have a cover page including above; **one-page executive summary** on Page 1, clearly stating:

* Goal and Background -- What is the problem?
* Approach -- What have you done?
* Major findings -- What do you find and what is your conclusion?

Organize, report, and interpret your **major** outputs with labeled figures and tables in your detailed report. Please do NOT include R code, data, and raw outputs.

**Boston Housing data**.

Random sample a training data set that contains 75% of original data points. Fit a linear regression, and various tree models including regression tree, bagging, random forests and boosting tree. Write a brief report up to eight pages including all labeled figures and tables, focusing on tree models. Compare different model performance based on fits of the training data (in-sample).

Test the out-of-sample performance. Using final model built from the 75% of original data, test with the remaining 25% testing data. Compare the model fits.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Case One  Data Mining II  Tess Newkold  03-12-2019 | A tall building in a city  Description automatically generated | |  | A close up of a sign  Description automatically generated | |
| Boston Housing Data  Advanced Tree Comparison with Regression Tree, Bagging, Random Forest, and Boosting |

# Executive Summary

Goal:

Our goal is to evaluate all the advanced tree models we have learned in class by comparing the out of sample prediction performance in each.

Approach and Major Findings:

A best linear model was chosen through a stepwise variable selection method based on the lowest AIC. That model is:

medv ~ rm + lstat + ptratio + black + dis + nox + chas + zn + rad + tax + crim

Then a regression tree is created with a training set of data which is 75% of the full data, the testing data is the other 25% of the data. This tree is shown in figure 1 below, with the pruned tree on the right using a cp of 0.021. These two models are the bases to compare our advanced tree models. The out-of-sample performance for all models and trees are shown in table 1, clearly the advanced trees out perform the linear model and the regression tree by far. The boosting method has the best performance of all methods with an MSPE of 7.1438. The table also shows the in-sample prediction error (MSE) for completion sake.

Figure 1. A regression tree modeled on the training data set.



Table 1. Summary of prediction error of all models.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Linear Model** | **Regression Tree** | **Bagging** | **Random Forest** | **Boosting** |
| MSE | 21.57721 | 17.77735 | 18.98079 | 11.20967 | 0.01416733 |
| MSPE | 24.15419 | 19.53428 | 12.89348 | 7.87853 | 7.143803 |

# Bagging

Bagging is bootstrapping the data and putting it in (n) number of bags and fitting trees to the data (n) times. It then aggregates the predicted values from all the different trees to reduce prediction error. The bagging tree method was performed on the training data with 50 bags, which was determined from figure 2 below. Based on this I decided that 50 bags are sufficient to also get good performance. Keeping the number low is important when there are many variables to conserve computing power. This method did do what is expected, which is to improve prediction accuracy. The MSPE drops significantly to 12.89 from around 17 in a single regression tree. This makes sense, using 50 aggregated trees is better at predicting the response variable than just one regression tree.

Figure 2. Shows how many trees are needed to get a low prediction error while keeping the number of bags low.



# Random Forest

Random forest is a further extension of the bagging method from above and makes significant improvements in prediction accuracy. This method takes the number of predictor variables and divides it by three to use for each split in each tree. This decorrelates the trees and thus reduces variance when the trees are aggregated. We can also use this model to see the importance of each variable. This is shown in table 2. Here the most important variables are lstat and rm, with %IncMSE of 60.79 and 37.19 respectively. In figure 3 the plot shows how the out of bag error decreases as the number of trees increases, however at some point the error no longer decreases, so increasing the number of trees no longer is helpful and could be detrimental. The MSPE also went down further, to 7.87 as shown in table 1 in the executive summary. This is lower than the bagging method, and much lower than a regular single regression tree.

Table 2. Shows the importance of each variable from random forest method.

|  |  |  |
| --- | --- | --- |
|  | % IncMSE | IncNodePurity |
| Crim | 7.9371550 | 2194.6328 |
| Zn | 0.6110028 | 195.4392 |
| Indus | 7.0245312 | 2068.4906 |
| Chas | 1.0554898 | 246.8840 |
| Nox | 10.1495678 | 2209.1796 |
| Rm | 37.1925243 | 9743.4753 |
| Age | 3.3519624 | 840.7240 |
| Dis | 6.4835474 | 1982.1890 |
| Rad | 2.8144004 | 374.7218 |
| Tax | 4.8309851 | 1199.4578 |
| Ptratio | 6.3267491 | 1927.8734 |
| Black | 1.9507974 | 694.0553 |
| lstat | 60.7975809 | 9493.8922 |

Figure 3. Plot that shows how the out of bag error decreases as the number of trees increases.



# Boosting

Boosting builds many small trees and in each new tree the response variable is the residual from the previous tree, this happens sequentially until completed. Below, in figure 4 and table 3, both show the relative influence of each variable on the model. Again, rm and lstat are the most important variables also seen in random forest above. The boosting method also shows the relationships between the variables and response variable medv, as shown in figure 5, I choose to only show the top three influential variables, but all can be shown if wanted. This is the best method for reducing prediction error in the model developed, as seen in Table 1 where the MSPE is 7.14, the lowest of all advanced tree models discussed here. While all three advanced tree methods improved prediction accuracy over a standard regression tree, the boosting method reduced the prediction error the most.

Figure 4. Plot shows relative influence of each variable in the model.

|  |  |
| --- | --- |
| Variable | Relative Influecne |
| Rm | 37.5702366 |
| Lstat | 29.9112046 |
| Dis | 9.1723109 |
| Nox | 4.9927672 |
| Crim | 4.8428674 |
| Age | 3.3710892 |
| Black | 2.8559603 |
| Ptratio | 2.5955702 |
| tax | 1.6454895 |
| Indus | 1.2821980 |
| Chas | 0.9787123 |
| Rad | 0.6653763 |
| zn | 0.1162176 |



Figure 5. Relationships between the top three variables (lstat, rm, and dis) and the response variable (medv).

  