EAS 508 HW_3
Rajiv Nagesh 50412150

Q1) Using the four approaches: Random Forest, Decision Trees, Boosting, Bagging; compare the model on these techniques. How sensitive is each of the approach on tree size, number of trees for bagging, etc. What are the most important features?

Answer: After modelling the dataset on the four approaches provided to us, we can infer that the approach of Decision Tree with Bagging is the most preferred one as it has the highest model accuracy among the four. Same can be seen in the table obtained below.

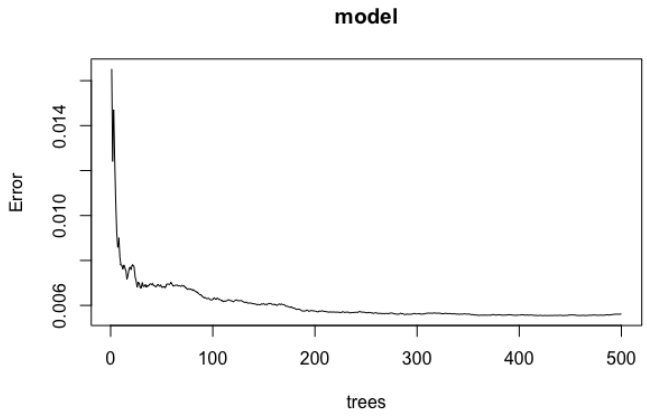
file.R × model_accuracy ×

Filter

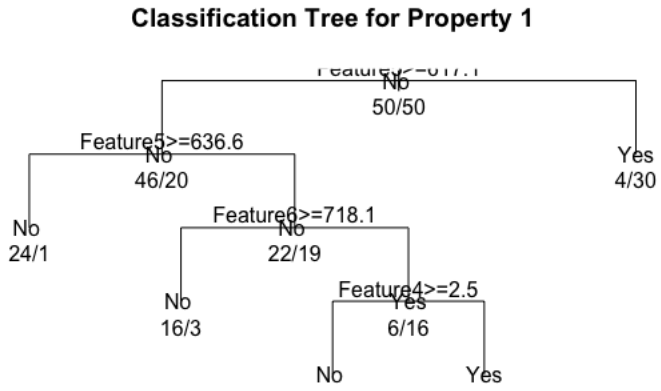
	Technique_Used	acuracy
2	Decision Tree w/ Bagging	91.2103174603175
1	Decision Tree	90
3	Random Forest	88.3928571428572
4	Random Forest Boosting	87.4603174603175

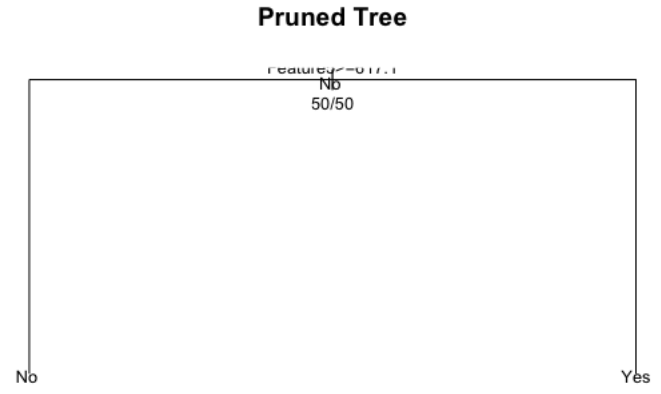
An accuracy of 91.21% is achieved with the Bagging approach. Bagging is basically breaking down the original datasets into a few smaller datasets and then trained on the same. Models are built independently in Bagging. In Bagging, training data subsets are drawn randomly with a replacement for the training dataset. Bagging decreases the variance, and each model has equal weight.

Furthermore, commenting on the tree size, bagging is a 'data hungry' technique. More the better. So, if the tree size increases, the Bagging would have a better accuracy. In the dataset given to us, there are only 100 rows which is comparatively less, hence the Bagging accuracy is more. We could not comment now if increasing the dataset would affect the model accuracy or not. It is a trial-and-error basis.

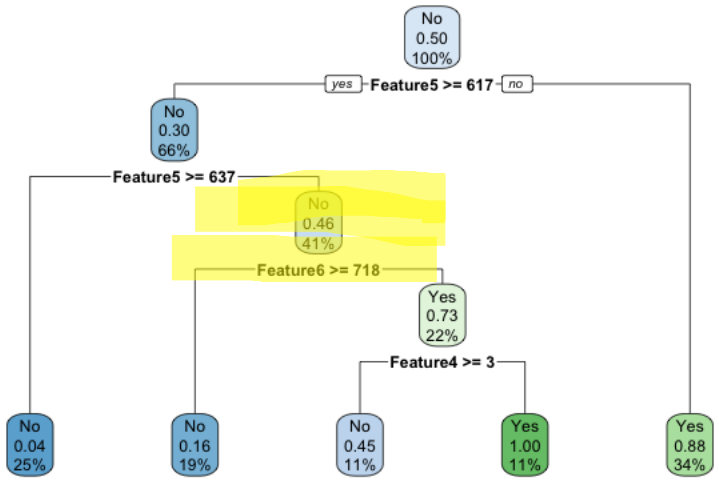


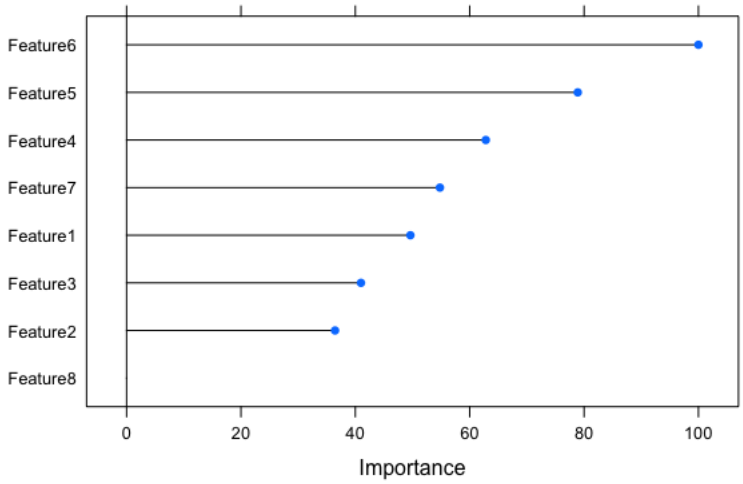
Likewise, adding on to the latter part of trial and error, number of trees used for Bagging also plays a vital role in the accuracy of the same. Here I first tried the 'nbagg value as 20' and got an accuracy like 81. Then I increased the same 'nbagg' value to 100 and got an accuracy of 91. So, on trial-and-error basis, I have concluded that the Bagging approach will give a maximum accuracy for this dataset when the 'nbagg' value is maxed out to 40 which is 91.21 as shown in the table above. Furrther increasing the 'nabgg' value will result in the same accuracy value. 'Nbagg' value means increasing the number of trees used for Bagging on how many iterations the model is going to be trained on a smaller dataset. This concludes that a small increase in the number of trees or the nbagg value, a significant change is seen in the accuracy of the model. Hence, we can infer that this approach is very sensitive to fine tuning of the parameters.





As this is a classification problem, the wiser approach was to choose ‘yes’ or ‘no’ values for property 1 as stated in the question. Breaking down the category was done as - From data set, define Property as categorical: i.e. Yes or No, we select 7.825 as it is in middle.
The most important feature of the model could be concluded as to be “Feature 6” as seen in the image below.

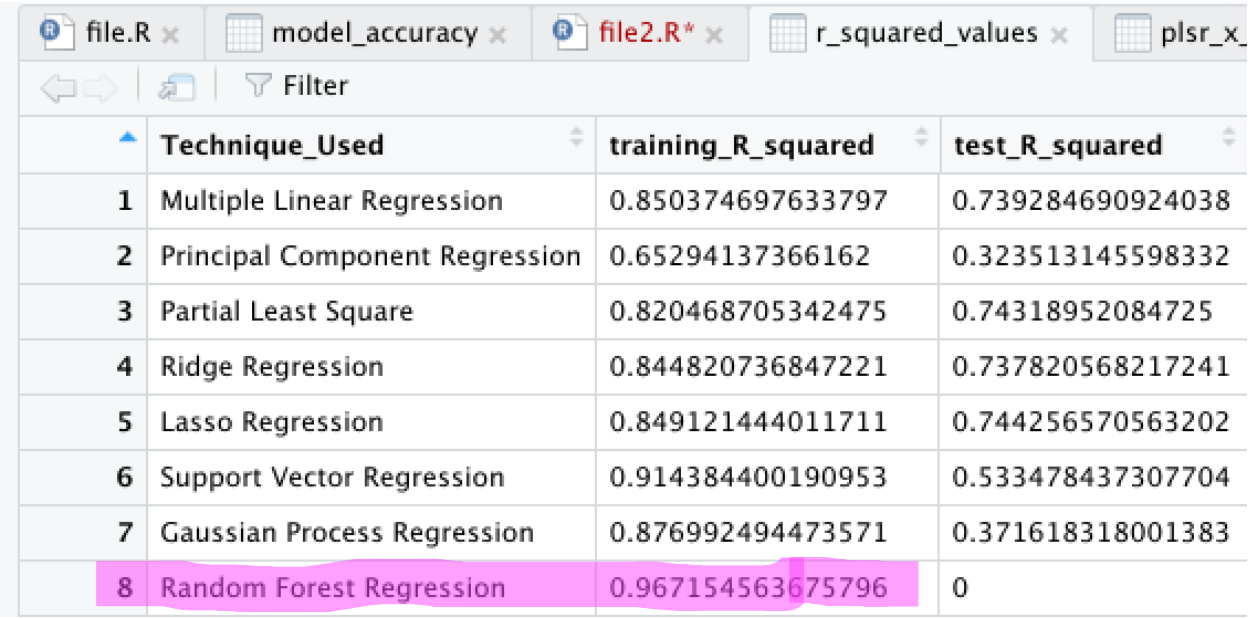




From the images we can see that Feature 6 is the most important feature. The other contributing features to the model accuracy and that could also be said as important are features 4 and 5 which have an importance of greater than or equal to 60.

Q2) Perform Random Forest Regression and compare the values to the values obtained from HW_2. Comment on the differences, if any.

Answer:



	Technique_Used	training_R_squared	test_R_squared
1	Multiple Linear Regression	0.850374697633797	0.739284690924038
2	Principal Component Regression	0.65294137366162	0.323513145598332
3	Partial Least Square	0.820468705342475	0.74318952084725
4	Ridge Regression	0.844820736847221	0.737820568217241
5	Lasso Regression	0.849121444011711	0.744256570563202
6	Support Vector Regression	0.914384400190953	0.533478437307704
7	Gaussian Process Regression	0.876992494473571	0.371618318001383
8	Random Forest Regression	0.967154563675796	0

From the table above we can infer that the highlighted portion, Random Forest has the highest model accuracy compared to the other model approaches used. This is because – in the earlier homework we were asked to split the data into training and testing sets and then model it on that basis. This accuracy would depend on how an individual splits the dataset. Whether it is a, 80-20, 60-40, 70-30 split, etc. Every split, would have different accuracy percentage for every model. Split used for HW2 was 80-20 split.

In the Random Forest technique, the approach implicitly splits the data into training and testing sets iteratively and models accordingly. It then throws out the accuracy for the best possible split chosen after many iterations. We do not have to specify the split portion or the number of iterations. Hence, the technique does the best possible split for us and provides with the model accuracy.