Problem 2 – Ensemble learning (35%)

Using the nutrition dataset included in HW2 / Problem 3, develop a classification model based on bagged classification trees.

15% credit. Train a bagged tree model with 50 trees using the training data (x1). Estimate OOB performance as a function of the number of trees grown. Discuss your findings.

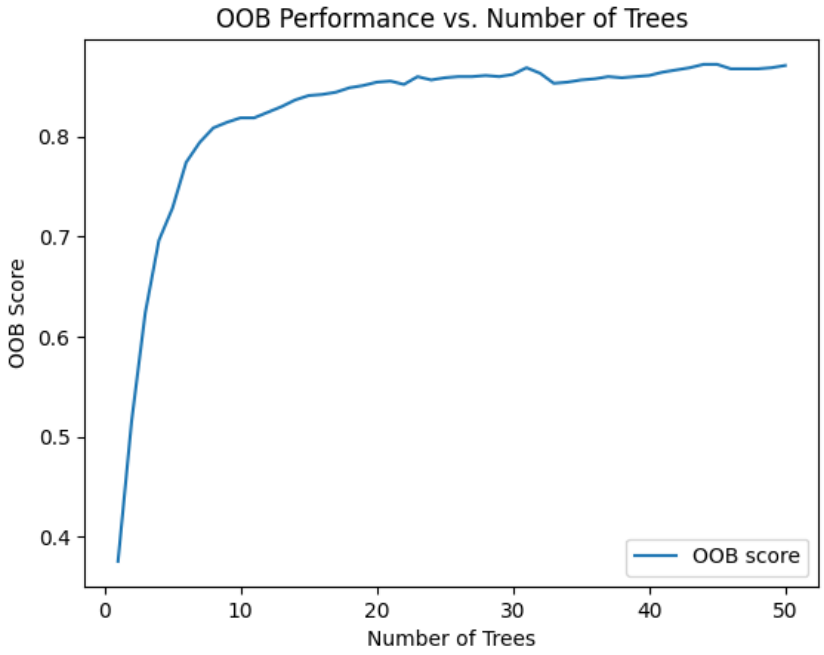

Figure 1. OOB Performance vs. Number of Trees Using Bagging

Discussion of findings:

Initially, it appears that the OOB score for smaller numbers of trees is very low; however, this aligns with our understanding of how Bagging works in terms of Classification Trees. Because Out of Bag error depends on testing the model with the remaining ⅓ of data not used for training in other trees, since Bagging models find success with larger numbers of trees (as it is an ensemble method that requires averaging of different models) increased performance with increased number of trees makes sense.
More trees means that we have more model samples to average from, and thus, outliers in models (or poorly performing ones) don't have as much power in introducing bad interpretations of the training set. Thus, the OOB score for smaller numbers of trees used in bagging does not do as well, as when we have a small number of trees, the OOB score is less reliable because there are limited OOB instances. Since OOB relies on aggregating predictions, as we increase the number of trees, the tree gets more diverse models and we get more OOB instances.

If you have a low number of trees, the OOB score is not a good estimate of error rate as each tree in the bagging model underfits, and this is even highlighted in the warnings when we run the bagging model with a small number of trees. We are warned that when too few trees are used, we don't generate reliable OOB estimates.

10% credit. Generate a variable importance plot for the model you developed in part (a). Interpret the results –this will require you to do some research on the meaning of the various predictors for different food categories.

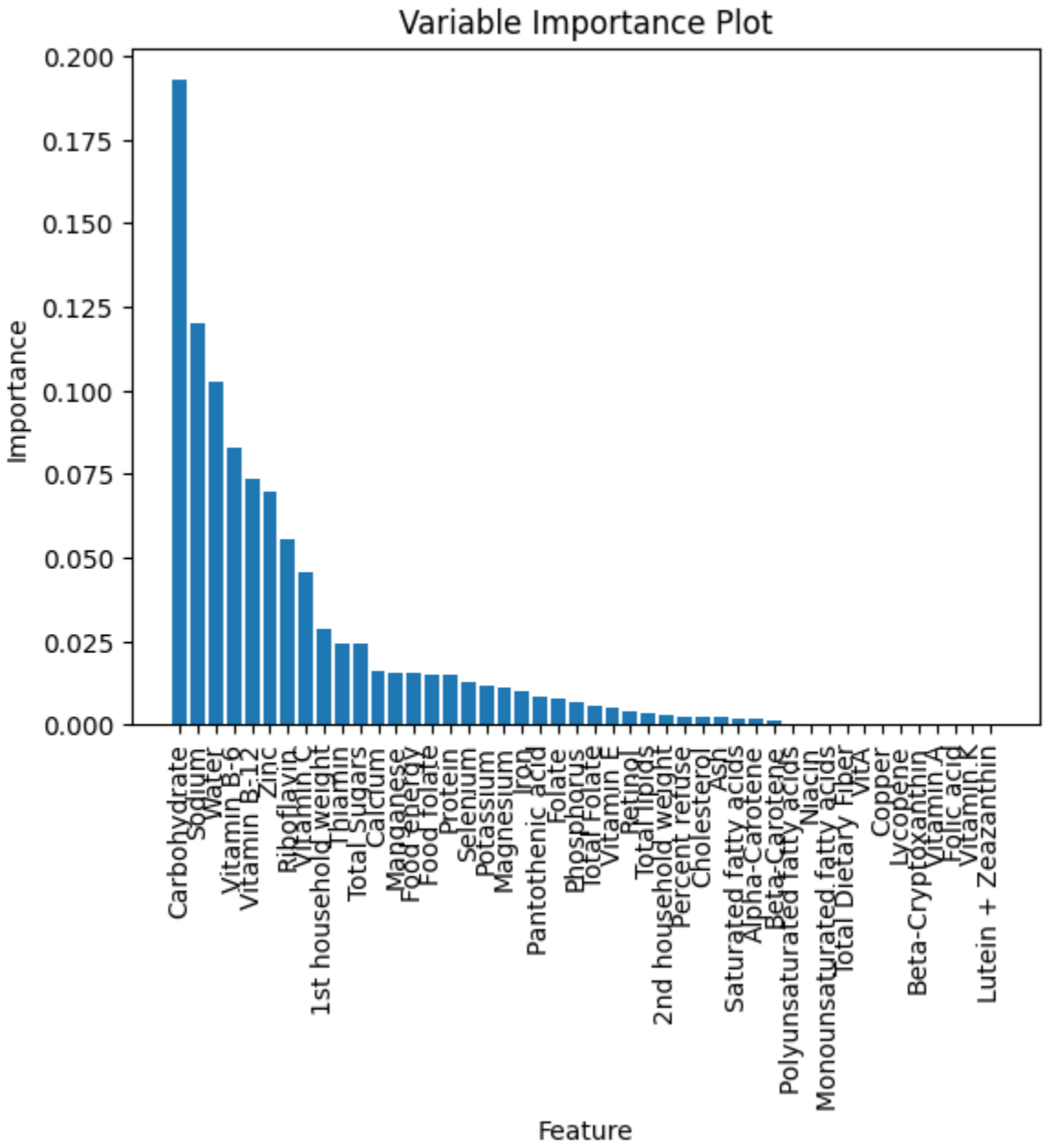

Figure 2. Variable Importance Plot For Bagging Model Using 50 Trees

Discussion of findings:

This model aims to correctly classify the class of a food item based on nutritional facts. The target classes that the model aims to use are “Baked Products", "Vegetables and Vegetable Products", "Soups, Sauces, and Gravies","Sweets","Fast Foods", "Fruits and Fruit Juices", "Breakfast Cereals", "Poultry Products", "Beef Products" , "Lamb, Veal, and Game Products”. Based on the above Variable Importance plot, we can get an aggregated summary of each variable’s importance, or the mean decrease in Gini index expressed relative to maximum. Thus, we interpret the top 6 variables with highest performance.

Carbohydrate

- As seen from Figure 2, Carbohydrate has the highest importance, which tells me that for the bagged tree model using 50 trees, Carbohydrate had high predictive power.
- Carbohydrates are sugar molecules, and can be found in three different forms: sugar, starches, and fiber. Carbohydrates exist in grains, fruits, dairy products, etc. Thus, carbs can be a good predictor of class as it could be used to separate things such as baked products, vegetables and vegetable products, sweets, fast foods, fruits and fruit juices and breakfast cereals, as these different categories have varying levels of sweetness.

Sodium

- Sodium has many forms, often taking the form of salt when talking about nutritional values. Sodium typically is in higher amounts in processed foods, but does naturally occur in items like fish, some vegetables (like carrots and cabbage), and some fruits (like avocado and mango). Thus, sodium can be a good predictor of class as it can separate items such as Fruit and Vegetables, as they have low sodium levels. Processed foods typically have higher sodium levels too, thus, this aligns with the information in Figure 2.

Water

- Water is in a lot of foods, and has different percentages/concentrations in foods. For example, processed foods may not have as high water content then that of fruits and vegetables. Thus, water can be a good predictor of class as it can separate things such as fruits and vegetables (which have higher water contents) from that of processed foods. Each class appears to have a range of “acceptable” values, and nodes can be further separated based on other variables as well.

Vitamin B-6

- Vitamin B-6 is a water soluble vitamin that is present naturally in many foods. It often is a generic name used for six compounds with the vitamin B6 activity. Examples of foods that B6 is high in include beef liver, tuna, salmon, chickpeas, bananas, and papayas. This potentially could be a good predictor of Poultry, Beef and Lamb products as all these items are high in vitamin B6.
- B6 can be a good predictor of class as it can provide separability between meat and plant products. Though there are some exceptions to plant products that have high concentrations of B6, later splits can separate the meat products from that of the plant products, reducing the Gini-Index of resulting nodes.

Vitamin B-12

- Vitamin B-12 is naturally present in foods of animal origin (i.e. fish, meat, poultry, eggs and dairy products).

- What is interesting is that some vegans can be at risk of developing vitamin B-12 deficiency should they not eat vegetables high in B-12 or take supplements, as B-12 is commonly found in items such as meat, dairy, and eggs. Thus, B12 could be a good predictor of class, as it could be used to separate items that could potentially be classified as meat products or vegetables.

Zinc

- Zinc is a mineral that is rich in food items such as meat, fish, and seafood. Breakfast cereals can also be high in zinc due to fortification. Thus, Zinc could be a good predictor of class separation due to the higher concentrations of zinc in such food classes and provide separation from plant based products.

10% credit. Train a sequence of bagged tree models with increasing numbers of trees (from 1 to 50). Estimate model performance on the test set (x2). Compare your results against the OOB results in part (a). Discuss your findings

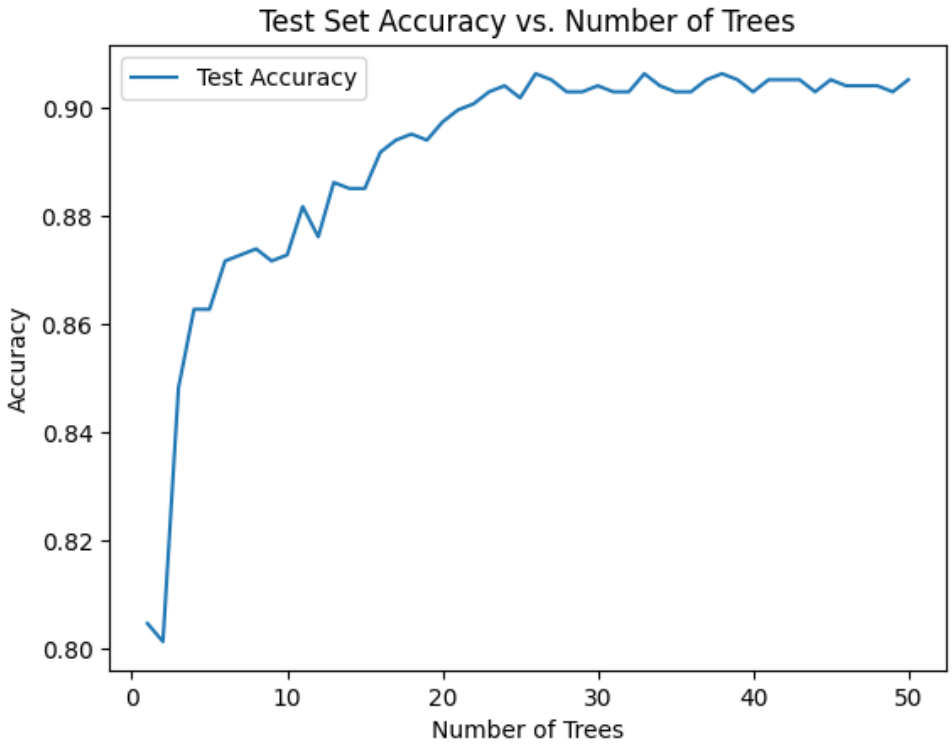

Figure 3. Test Set Accuracy Against Number of Trees

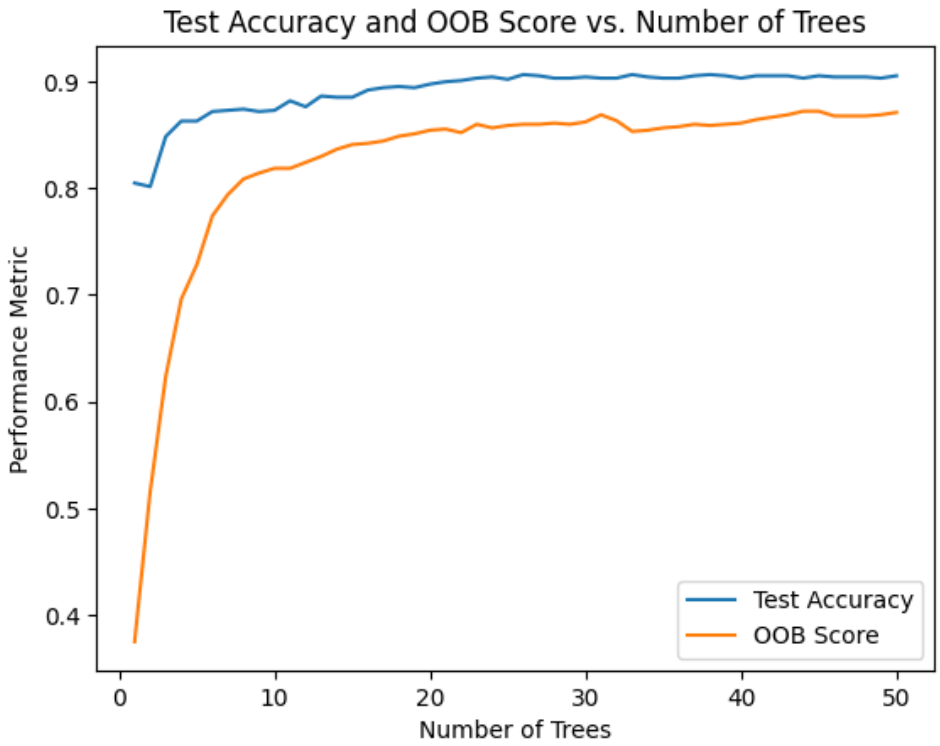

Figure 4. Test Set Accuracy and OOB Score Against No. Trees

Discussion of findings:

What was interesting in terms of model performance on a test set against that of OOB score, testing accuracy always seemed to be higher than that of the training. Initially, when tree size is small, we get a small plateau in test accuracy, as seen when we use 1 tree for bagging. As we increase the number of trees in the bagging model, performance for both OOB and Test scores increase. This makes sense, as increasing the number of trees used in the bagging model reduces variance, and allows for creation of a better model that averages the generalizations of the individual trees.

This observation also aligns with our understanding of how the model interprets Testing Accuracy, as OOB scores are an internal training method used during training that does fully represent model performance on a test set. Because the test set includes data that is unseen by the models, and since OOB simply is an estimate of how the model performs on true unseen data (i.e. the testing), having test accuracy perform better than that of OOB aligns with understanding of bagging.

References:

- *OOB Errors for Random Forests in Scikit Learn - GeeksforGeeks*
- *Feature importances - Bagging, scikit-learn - Stack Overflow*
- How to interpret OOB Error in a Random Forest model - Cross Validated