Random Forest Regression

This homework uses the same dataset with over 20,000 recipes from the website Epicurious as Homework 5A and 5B.

(a) A step in building a machine learning model that you may have been skipping is the "exploratory data analysis" stage–when you take a look at summary statistics and distributions of the variables in the dataset, as well as exploring relationships between variables. Create a table of the summary statistics (at least the mean and standard deviation) for the numeric variables (`rating`, `calories`, `protein`, `fat`, and `sodium`). In addition, plot the distributions of each variables. Briefly describe the distributions.

(b) Hopefully you've noticed that four of the five distributions are seriously skewed by a handful of extremely large values. The USDA daily recommended intake of these nutrititional variables is as follows:

- Calories: 2,000 per day
- Protein: 60 grams per day (for a 165 lb. person)
- Fat: 66 grams per day (for a 2,000 calorie diet)
- Sodium: 2,300 mg per day

While it is reasonable that some recipes may have larger quantities (for example, if the recipe is for a whole cake, not split into smaller serving sizes), and of course, sometimes we eat more than the daily recommended value, there's no reason to have millions of calories in a recipe. Remove any values above 5 times the daily recommended amount, then redo the histograms and briefly describe them.

(c) Now, create and briefly describe scatterplots for the following pairs of variables: `calories` and `rating`, `calories` and `protein`, `calories` and `fat`, and `calories` and `sodium`.

(d) Remember that some of the extreme nutritional values are likely related, so we might instead be interested in predicting the calorie count for a recipe based off its ingredients. Remove `rating`, `protein`, `fat`, and `sodium`. Take a subset of the data–it can be a random sample, or you can get creative and look at a specific subset (e.g., desserts, salads, etc.). Use 70% of your subset for training and 30% for testing. Using the training set, fit at least five random forest regression models, experimenting with the settings such as the number of trees to grow, minimum and maximum number of nodes, etc. Choose a final model and justify your choice.

Note: I fit a model with the default settings in R–specifically, with the number of trees in my forest being equal to 500. By the time all was said and done, it took me about 8 hours of computation time–but the mean squared error stabilized much earlier. I don't want you to have to take this long... there are a few things you can do to speed things up. First, I recommend training with far fewer trees since we are mostly sure the MSE will converge. Second, you can work with a fairly small subset. Make it small enough such that your model fits are taking no longer than 10 minutes.

(e) Some random forest routines (including the routine in Scikit-Learn) can calculate something called "feature importance". Use this information to describe twenty ingredients most useful for predicting the calories, and twenty ingredients that are least useful for predicting calories. Within the context of food, do these make sense to you?

Note: You can also play with feature importance in your model building step (Part c)!

(f) Research different modules for visualizing your tree. Try to showcase at least the first five levels.

(g) Now, apply your model to the rows you removed from the dataset. Does the random forest help result in more realistic calorie counts?