Practical Data Science: Assignment 4

Setup

Welcome to the fourth and final assignment for this course. The objective of this assignment is to once again complete a full analysis of data sets, in this case one for classification and one for regression. For this assignment, you must use **scikit-learn** for training and evaluating any models on the data provided. Consequently, you might want to review the material in the final two lectures for this assignment.

Questions

There are **two** questions in this assignment and you must complete both. The style of this assignment is similar to assignment three in that you will be given a dataset and asked to investigate a particular problem on that data set. You are expected to use the `scikit-learn` module for this assignment. In contrast to assignment three we are looking for best practices in model building so we are looking to see ways of evaluating the models that you build beyond simply measuring the performance on a training data set.

Question 1 (15 points)

The data set for this problem is hosted on Kaggle, and it concerns the prediction of a house's price based on a number of different variables such as the type of house, the zoning classification and the lot size. You can find this data set here:

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Your object is to build the best model you possibly can for predicting house price using the `scikit-learn` module. As price is a numeric quantity, this is a regression-type problem. You should use the `train.csv` file for training your models. **After** you have picked the best, evaluate that on `test.csv`.

To receive full credit for this question, I am not looking so much for final accuracy as I am looking to see the process you followed in order to arrive at your chosen model as well as how you evaluated the different models that you trained. Consequently, I want to see how you picked which model works best. Please make it clear in your submission which model of all those that you trained you consider best and **why**. I am also looking for:

- How you pre-process the data you are given
- How you used your training data to train and evaluate each model
- How you chose the relevant parameters for each model
- How you evaluated and compared each model to the rest

I'd like to see a minimum of three different model types that you compared. Once you pick your final model from these, I'd also like a short discussion (up to 5 lines) explaining which features in the data set you found were most useful and whether your model excluded any variables from those given.

Question 2 (15 points)

The data set for this problem is also hosted on Kaggle. This time, we have a classification problem. More specifically, our objective here is to use different geographic and geological variables to predict the type of tree found in a particular area of the Roosevelt National Forest in Colorado. The data can be found here:

https://www.kaggle.com/uciml/forest-cover-type-dataset

The variable `Cover_Type` is what you are trying to predict. Notice that we have a multi-class classification problem here, as this variable has 7 levels. Make sure you do not accidentally treat your output as numeric – this is not a regression problem.

If you load this dataset, you'll see that it is quite large. To that end, if you run into performance issues I propose you sample the dataset to create a training dataset with 20,000 examples and a test set with 20,000 examples. Make sure the distribution of the 7 different cover types is roughly the same in your train, test and global data set.

Just as with the first question, we want you to go through the process of training at least 3 different classifiers to solve this problem and to then pick the best one. The criteria for this question are the same as those for the first question so the same questions asked there apply here e.g. we want to once again see how you evaluated and compare each model to the rest.

Once you pick your final model from these, I'd also like a short discussion (up to 5 lines) explaining which features in the data set you found were most useful and whether your model excluded any variables from those given. For this question I am also interested to find out which pair of cover types your model finds it hardest to distinguish.

IMPORTANT NOTE FOR BOTH QUESTIONS:

I appreciate that parameter tuning for some types of models can be quite time consuming. I am not looking for solutions where you search across a vast array of different parameter values, as I do want to be able to run your entire assignment within an hour.