Solutions to Assignment 2

Valentin Zulj & Vilgot Österlund

October 10, 2018

1 Introduction

In this project, we attemt to solve one of the competition problems published on Kaggle. The purpose of the competition is to use machine learning methods – random forests – in order to classify the forest cover type of a given 30 x 30 meter cell. In other words, we want to use the data supplied by Kaggle to determine which type of tree is most common on a specific plot of land. It is an important subject because it would save a lot of time and money if you could classify the tree type based on geographical data. If so, you would not have to fly over the specific area but instead classify the tree type from other – already collected – data. We will mainly focus on the random forest classification method, but we will compare it with the performance of a classification tree and a multinomial logit regression model. Not only will we compare the random forest model, but we will also do a small simulation study where we check that the models works the way it should.

2 Data

2.1 Original data

The data used in the project can be found on found on http://www.kaggle.com/c/forest-cover-type-kernels-only/. The data consists of 15 120 observations of 55 variables. The response variable is, of course, the type of tree, and consists of discrete, interger values on the interval [1, 7]. These types are: Spruce/Fir, Lodgepole Pine, Ponderosa Pine, Cottonwood/Willow, Aspen, Douglas-fir and Krummholz. In total, the data consists of 2160 observations of each Cover type.

The data set consists of 55 variables, most of which – 40 to be precise – are dummies regarding different types of soil. As for the other variables, we present a brief summary in the bullet point list that follows:

- Elevation: Elevation – in metres – of the plot of land
- Aspect: Aspect in degrees azimuth
- Slope: Slope – in degrees – of the plot of land
- Horizontal: The horizontal distance to the nearest water source
- Vertical: The vertical distance to the nearest source of water
- Horizontal roadways: Horizontal distance to the nearest roadway
- Hillshade (9am, noon, 3pm): Index showing degree of shade from hills at different times of day

- Horizontal: Horizontal distance to the nearest point where wildfire ignition is allowed

The variables themselves are quite straight forward. Perhaps we should mention that the hillshade index is measured on a scale from 0 to 255. Furthermore, there are some binary columns specifying whether the plot of land is located within one of four wilderness areas, as well as the 40 binary columns that indicate the type of soil in which the trees grow. However, two type of soils, coded as 7 and 15, does not show up in any observation. Problably the data has been split in a way so that there is observations in the test set with theese soil types.

2.2 Data handling

To get the data tidy and more easily foreseeable, we merged the dummy variables associated with soil type and wilderness area into one column each. This reduced the number of covariates to 12 and means that soil type is now a factor variable with 38 levels (due to the fact that there where no observations of type 7 and 15) and that wilderness area is a factor variable with four levels. After getting to know the data with different plots, which you can read about in the next section, we split the data in to a training- and test set. We sampled half of the data in to the training set and the other half to the test set. However, because of the fact that there where two soil types with only one observation each, we had to force these two observations to the training set. This is because our models can not predict or classify a response variable from covariates that it has not been trained on. The training set consists of 7562 observations and the test set of 7558 observations.

2.3 Graphical presentation

To get an understanding of the data, we explore it by several different plots. Figure 1 shows the frequency of each Soil type, and it becomes clear that some soil types are very rare. As stated earlier, the data set does not consists of any observation of an area where soil type 7 and 15 is the primary soil type. We can also see the observations mentioned in the previous section, where soil type 8 and 25 are the types with only one observation each.

In figure 2 it is clear that different trees grow at different elevation. For a decision tree implementation where elevation is the only explaining variable, it would be very unlikely to predict a tree as Krummholz if the elevation was less than 3000 metres. The plot tells us that the elevation of a specific area must be of great importance when it comes to what type of trees that grows there.

From figure 3 we can tell that there seems to be no correlation between elevation and the hillshade at 3 pm. Noor does it seem like the hillshade affects the type of tree. This we can tell from the fact that all types of trees grow at all levels of hillshade. We can see that the points in 3 tend to get bigger when the elevation is higher. This is intuitve, since the size of the points is determined by the horizontal distance to the nearest roadway.

Figure 4 shows a boxplot of cover type and horizontal distance to water for the corresponding plot of land. It seems like the horizontal distance to water have a very small, if any, impact on which type of three that grows in a specific area.

We can tell from figure 5 that the soil type depends on the elevation of the specific area. Some soil types exists in a wide range of elevation and other types exists only in a more narrow range of elevation. However, no soil type span the whole elevation span and for example we can see in the plot that soil type 40 is never found below 3000 metres.

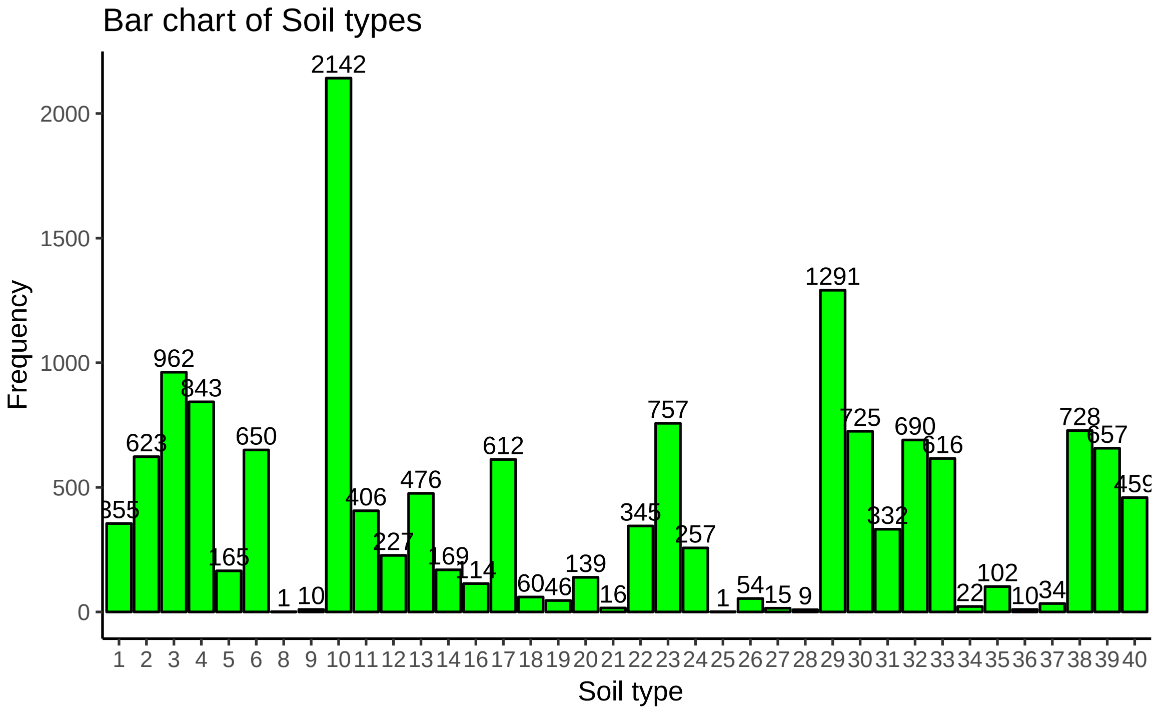

Figure 1: Bar chart of the frequency of different soil types.

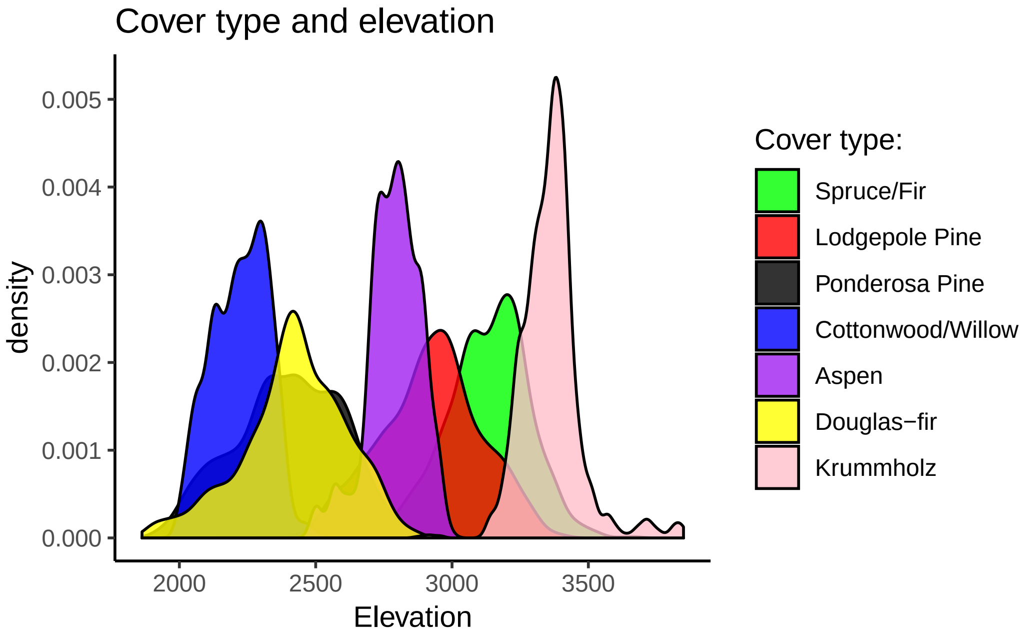

Figure 2: Densityplot of Cover type and elevation.

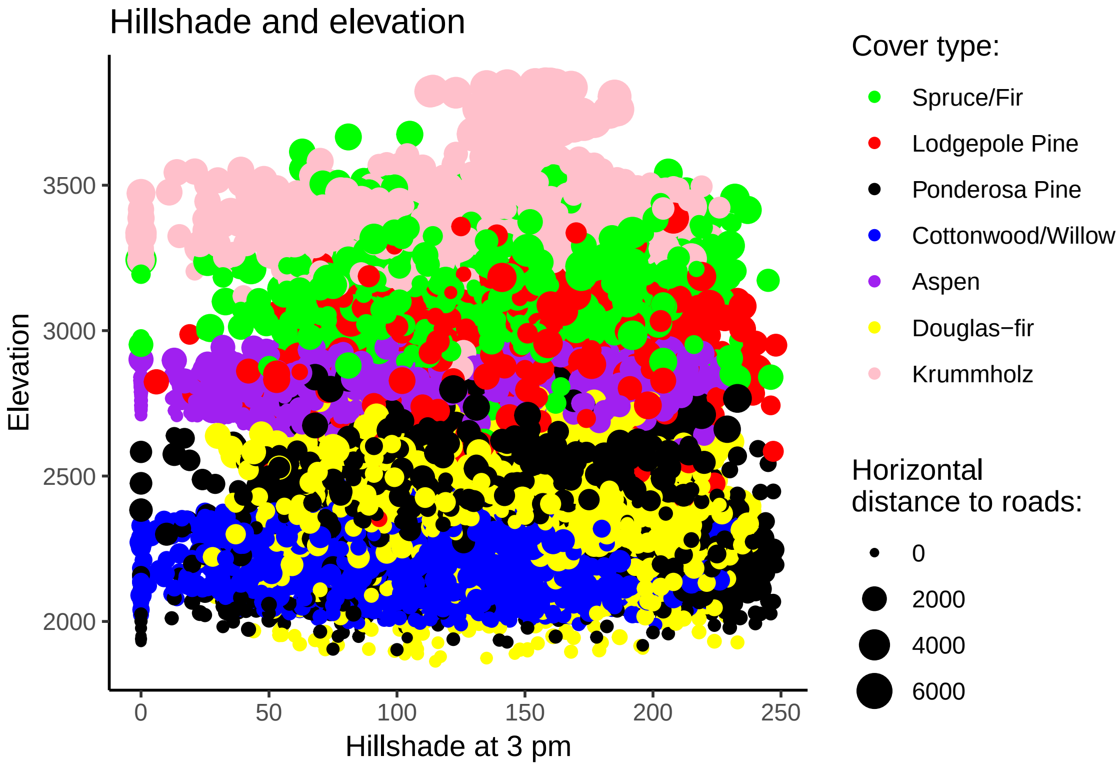

Figure 3: Scatterplot of Hillshade and Elevation.

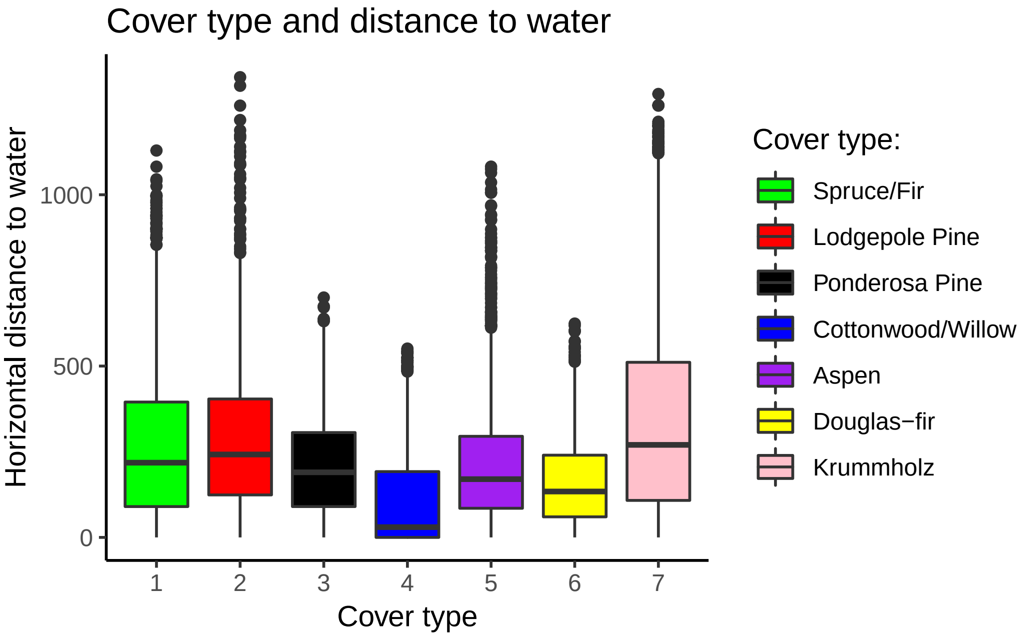

Figure 4: Boxplot of Cover type and horizontal distance to water.

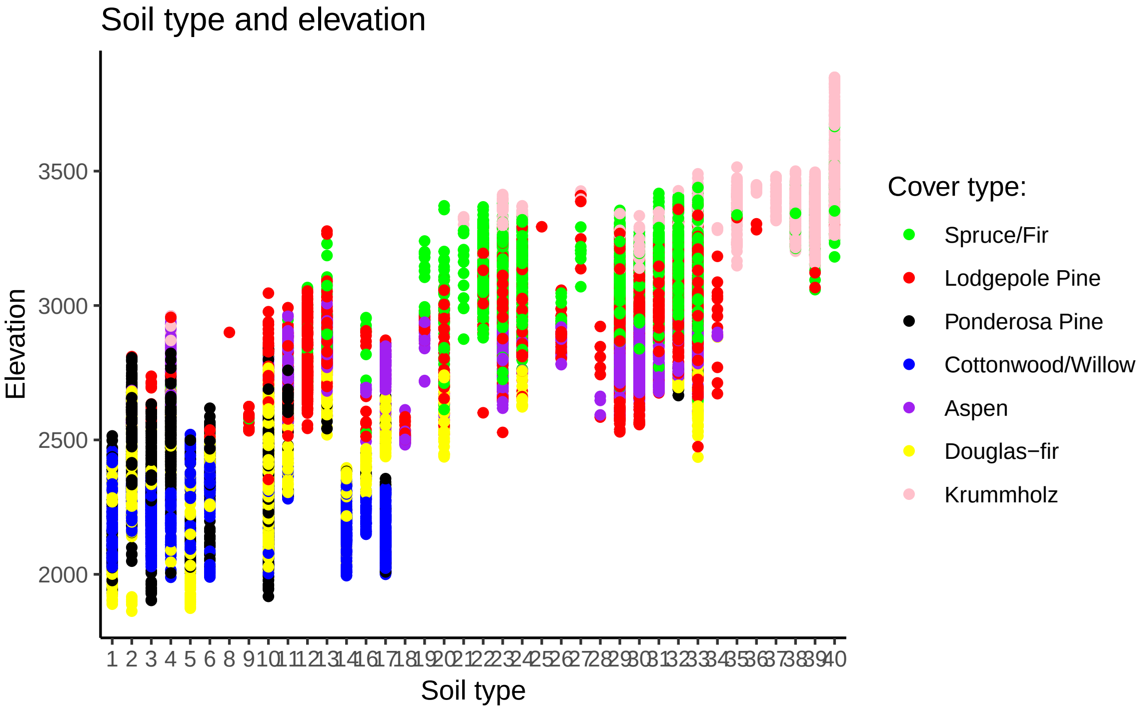

Figure 5: Scatterplot of soil type and elevtion.

3 Simulation study

The simulation study can be described in two parts, the first part is about generating new data and the second part is about the simulation. The idea is to use the new data to examine how different settings in the random forest model affect the performance of the model.

3.1 Generating data

Our first plan to generate a new data set was to use leave one out cross validation (LOOCV) together with linear- and multinomial logit models. The idea was to start with the original data set and use a double for loop to predict new vales of the covariates and response variable. Excluding the i:th row, we wanted to estimate a model with the j:th column as the response variable. Depending on if the j:th column is a factor or numerical variable, the model would either be a multinomial logit model or a linear model. After estimating the model, we use the model and the values of the covariates on the i:th row to predict a new value. Moving on to the next value of j, and later on to the next value of i, this would give us a new data set with the same size as the original set. However, when starting the for loop we quickly realised that it would take about two weeks for the loop to go trough all of the rows of the original data. It was mainly the creation of the multinomial models that where too time consuming. Therefore we abandoned the LOOCV idea and for some extent changed the generating plan. Instead of creating new models for each row, we created just one model for each variabel with all of the observations included. This means that in the first model the first column is the response variable and all other columns are the regressors. In the second model the second column is the response variable and the rest of the columns are the regressors, and so on. Using theese 13 models, we predicted new values based on the original data. This new data set should resemble the original data and it is on this data we will perform our simulation study.

3.2 Simulation