Forest Cover Analysis

TCSS-551 Term Project

Jonathan McFadden

Introduction

The goal of this project is twofold. First, we seek to create an accurate big-data/machine-learning model to predict which type of trees will grow best in a certain area based on the conditions of that area. Finally, we seek to determine which features of a given area are most important for determine which type of tree will grow best there. To do this, we have chosen three different methods:

- Linear/Logistic Regression
- Decision Trees
- Random Forest

We chose to try a Linear/Logistic regression first because they're provide a simple and low computational cost way to see a relationship between the features and outcomes of the provided data does indeed exist. This method will be unlikely produce predictions with good accuracy due to its simplicity, however should these methods show a relationship or relationships within the provided data, that would be a positive indicator to proceed to more complicated and compute intensive algorithms.

METHOD A - Linear/Logistic Regression

As described above, the initial method we tried were Linear/Logistic Regressions. We tried three variations on this method. First, we ran a simple linear regression. This gave a poor result, but *did* indicate that there was a relation between the features and outcomes in the data. This linear model gave an R^2 value of 0.4214 on the test data and 0.4006 on the training data. This is odd as, one would expect this value to be higher for the training data than the

test data. This trend also continued when comparing via MSE[1], as the training data had an MSE of 2.3952 while the test data had an MSE of 2.3257.

[1] mean squared error