Alexander Pastora

Final Project Write Up

For my final project, I chose to use the Forest Cover Type Prediction data set. Essentially, the goal of this set is the create a model that predicts whether the forest covert type is Spruce/Fir, Lodgepole Pine, Ponderosa Pine, Cottonwood/Willow, Aspen, Douglas-fir, or Krummholz. Even though this output is categorical, it is represented as integers from 1-7, respectively. The predictor variables, for the most part, are either numerical or binary variables. On this data set, I simplified a few variables. One such variable was Distance to Hydrology, where I used the change in distance formula to combine both Horizontal and Vertical Distance to Hydrology. I also took the wilderness area binaries, and the soil type area binaries, and made them into two variables called wilderness area and soil type. I then used the fct_recode function in the for cats package in conjunction with the as.numerical function to represent all of these categories as a numerical value. Using Exploratory Data Analysis, I then determined that Elevation, Horizontal Distance to Roadways, Distance to Hydrology, Wilderness Area, and Soil Type would be decent variables to use for the model. I did do EDA on all of the other variables, but their trends were not that interesting, so these plots are only included in the Extras Section.

For completeness, I considered linear regression, ridge regression, LASSO, and K nearest neighbors as potential models. I first determined lambda star values for ridge regression and LASSO. I attempted to determine a k-star value for K nearest neighbors, but the run time for that took way to long, because the training data had 15,120 observations. Instead, I arbitrarily made k-star equal 450. I then did cross validation on all of these models using % correctness, since this is how the Kaggle competition scores submissions. Based on the results, K nearest neighbors was the best model, so I trained the model to the training data set, and then made predictions on the test set. Based on my submission, the model has a success rate of 34.885%, which is not the greatest, but it is still higher than the CV values for any of the other models. However, the CV for K-Nearest Neighbors was quite a bit higher (50%-ish) than the Kaggle score. I'm thinking that this may have to do with the K-star value, since I literally just made that up. Overall, I feel that this model is a good start, and that with some tweaking (possibly in the soil data), that a better model could be made.