Benjamin Lin

How Random Is Your Forest?

Project Design

My project focused on expanding the National Roosevelt Forest in Colorado. The forest department is looking at the undeveloped land around the forest and wants to develop those areas. With that being said, the objective of this project is to predict the predominant tree type in a section of the forest, in order to provide the proper nutrients and irrigation for that area. I am planning on first taking a look at my data to make sure everything is in a good state. I am then going to play around with base models (Naive Bayes, KNN), then play around with more advanced models (SVC). Finally, I'm going to try various decision tree algorithms to see which gives me the best metrics.

Tools

As for tools, I used pandas and numpy to clean my data and make any necessary changes. For visualization and EDA, I used PCA, matplotlib, seaborn, and standard scaler.

For my models, I used many of sklearn's tools. Specifically, they were: gridsearchCV, pipeline, several metric calculators, and several built in models and algorithms (more information in the Algorithms section).

Data

I came across an interesting dataset that looked at 30mx30m areas in the Roosevelt National Forest in Colorado. For each area, there are several cartographic and environmental features, including the predominant soil and tree type in that area.

Here's a quick overview of the data:

Response Variables:

- Tree Type*: the type of tree that is predominantly in that area
    - 7 different trees
        - 1: Spruce/Fir
        - 2: Lodgepole Pine
        - 3: Ponderosa Pine
        - 4: Cottonwood/Willow
        - 5: Aspen
        - 6: Douglas-fir
        - 7: Krummholz

Predictor Variables:

- Elevation - Elevation in meters
- Aspect - Aspect in degrees azimuth

- Slope - Slope in degrees
- hDist_Water - Horz Dist to nearest surface water features
- vDist_Water - Vert Dist to nearest surface water features
- hDist_Road - Horz Dist to nearest roadway
- shade_9am - (0 to 255 index) - Hillshade index at 9am, summer solstice
- shade_noon (0 to 255 index) - Hillshade index at noon, summer solstice
- shade_3pm (0 to 255 index) - Hillshade index at 3pm, summer solstice
- hDist_Fire - Horz Dist to nearest wildfire ignition points
- Wilderness Area* - Designated wilderness area
    - 4 different areas
- Soil Types* - Designated soil type
    - 40 different soils

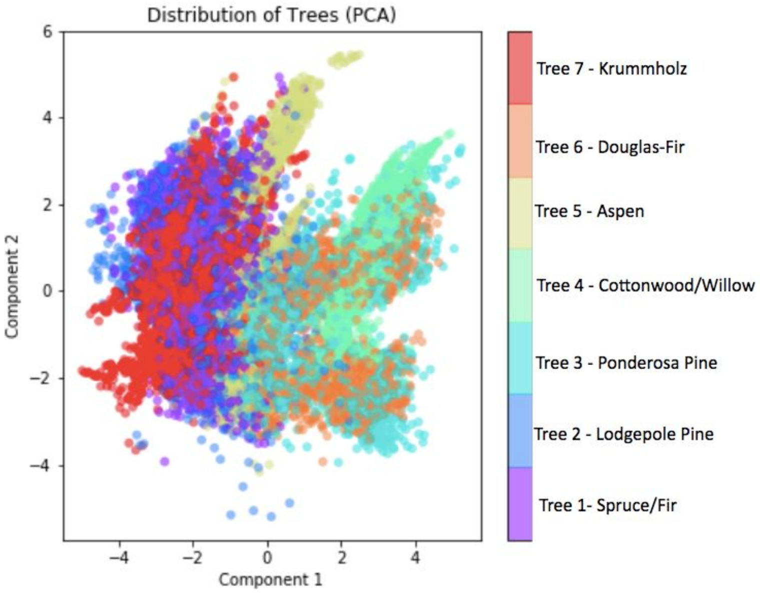
*** = Categorical**

The data cleaning was pretty straightforward. There were two soil types (ST7 and ST15) that did not belong to an area, so I took them out because they were just adding noise.

Data - Analysis

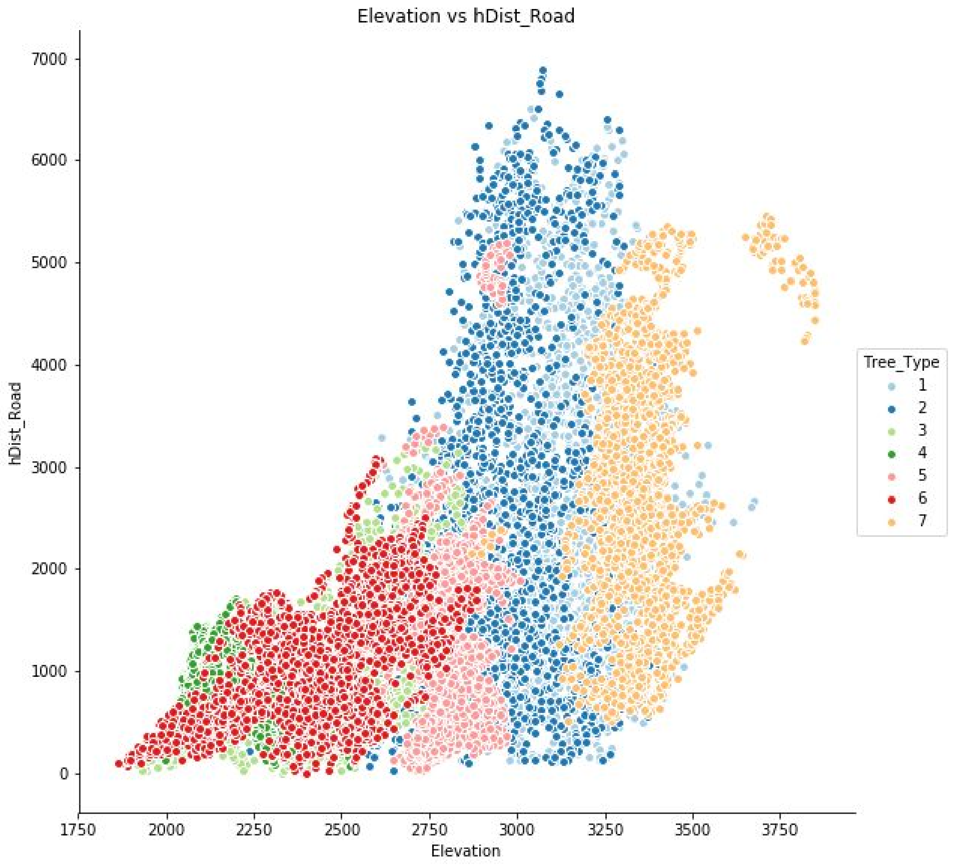
I first looked at the distribution of all 7 tree types in my dataset to make sure all the classes were balanced. Surprisingly, it turned out that all 7 types were perfectly balanced with 2,160 sections each. I then looked at my distribution of the Wilderness Areas and Soil Types (after removing the two), and they were also relatively balanced.

Then, I applied PCA to my features to get a clearer visualization of the distribution of different tree types. Transforming all my features into two components, the graph showed:



From this chart, it’s clear where all 7 tree types are and their varying distributions in respect to the two components. Digging a little deeper, I looked at the transformed value of each feature in both Component 1 and 2. For Component 1, Wilderness Area 4, Slope, and Soil Type 10, captured the most variance. For Component 2, 9am shade, Wilderness Area 1, and Soil Type 30 capture the most variance. So this conveys that the 6 features above are especially important in distinguishing between trees. With PCA, I was able to better visualize my tree type distribution and get a better understanding about my features.

I then looked at the correlation between my features to see if there was possible multicollinearity. None of the features had a high enough correlation with another, so I didn’t have to address that issue. However, I was interested to see the relationship of Tree Types with respect to my highest correlated features, so I plotted some of those scatterplots next. The most interesting one was the relationship between Elevation and Horizontal distance to the road.

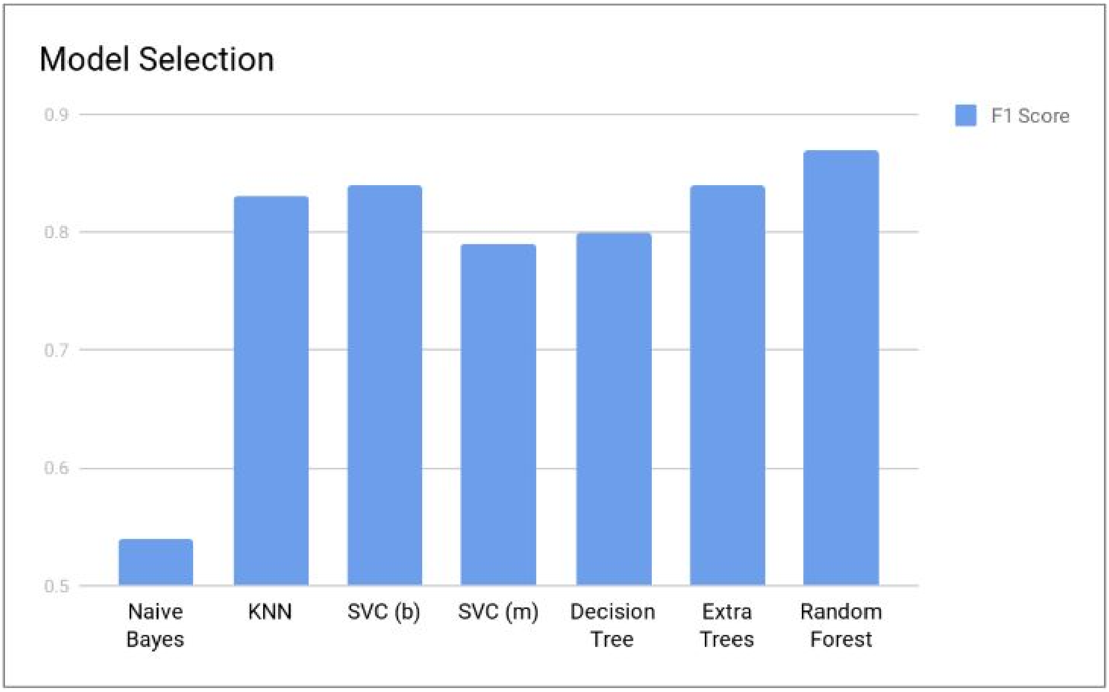


As you can see from the graph, it's pretty easy to distinguish the 7 tree types by these two features. If we only look at Tree 2 (dark blue dots), we can tell that Tree 2 is mainly within a small elevation range of (2875 and 3200) and within a large range of HDistance to Road (0 to 6000).

Algorithms

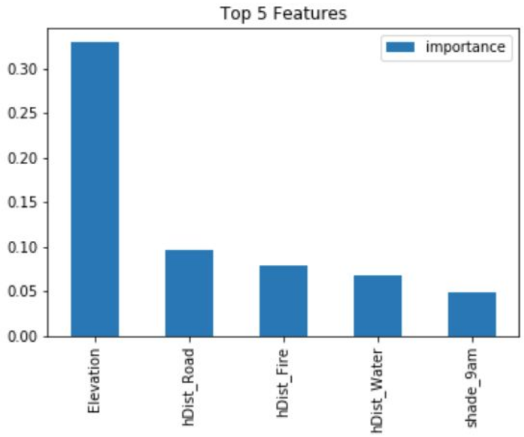
To start, I split my dataset into a 80% training and 20% test set. I validated my models using the CV score on the training data, and chose the best model based on the F1 score of the test set.

I ran 6 different models - optimizing for each model's specific parameters. I first started with Naive Bayes just to get a baseline accuracy and F1 score. I then moved on to KNN and SVM which gave me relatively good metrics. I was curious to dive in a little deeper so I experimented with Trees - Decision Tree Classifier to Extra Trees Classifier to Random Forest Classifier. Here's a snapshot of the F1 scores for each model:



Random Forest did the best, with an F1 score of 0.87.

I then took a look at the most important features for my Random Forest Model:



Elevation makes sense, as we saw from the correlation plot earlier. I found it interesting how important the distance to the road, fire, and water was. But this intuitively makes sense, as all 3 measurements technically measure how deep into the forest one area is, which would be very important in determining tree type.

Conclusion

To wrap everything together, let’s revisit the initial objectives of this project: 1) predict the predominant tree type in a section of the Roosevelt National Forest to expand the forest.

With my random forest model, I am able to accurately predict the tree type in a section of the forest. Now, the forest department can go into any damaged area with the proper soil and nutrients to expedite the restoration process!