FINAL PROJECT REPORT: Predicting Forest Coverage

Anushri Bhagwath, Prathamesh Patil, Rashmi Chhabria, Yash Nanda

The main motivation behind choosing this topic was to answer the question "How are the nation's forests doing?" Since the advent of global warming and the ever-increasing demand for wood and other forest products has grown, understanding this question has become very imperative. There are numerous techniques which are being developed by researchers so that people have a better understanding regarding the forest conditions and the effects of changing climate, human intervention and other management practices. There are various scientific techniques available now which can be used in order to study and monitor the forests closely.

For our project, we have used a dataset from the Roosevelt National Forest of Northern Colorado in order to predict seven different cover types in four different wilderness areas.

The four wilderness areas are:

1: Rawah

2: Neota

3: Comanche Peak

4: Cache la Poudre

This dataset was picked up from Kaggle and it consists of 12 real features, 54 variables and 581k observations. The 54 variables were divided into 10 quantitative variables, 4 binary wilderness areas and 40 binary soil type variables.

The different variables in the dataset are described below:

- Elevation: It represents the elevation of the cover type in meters.
- Aspect: It represents the aspect in degrees azimuth.

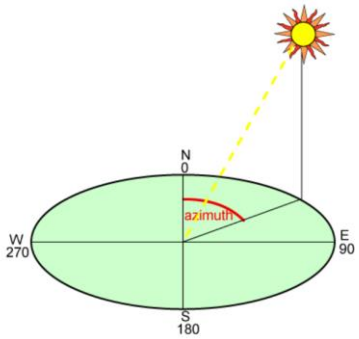

FIGURE 1: Pictorial representation of the azimuth angle. From Azimuth Angle, n.d., (Link) Copyright 2022 by pveducation.org

In the figure shown, the azimuth angle is the compass north and the direction of the sunlight. It keeps on changing as the position of the sun changes during the day.

- Slope: It provides information regarding the slope of the hill in degrees.
- Vertical distance to hydrology: Indicates Vertical distance to the nearest surface water features
- Horizontal distance to hydrology: Indicates Horizontal distance to the nearest surface water features.
- Horizontal distance to roadways: This indicates the horizontal distance to the nearest highway or any other road.
- Hillshade 9am/noon/3pm: Indicates the hill shade index at 9am/noon/3pm, summer solstice (0 to 255 index, here 0 indicates the shadiest area and 255 indicates brightest area).
- Wilderness Area: Indicates the wilderness area designation (4 binary columns).
- Soil Type: Indicates the designated soil type (40 binary columns).
- Cover Type: This column indicates the 7 different cover types.

The 7 different cover types are numbered 1 to 7 in the column Cover Type in the dataset. The cover types are classified as mentioned below:

1: Spruce/Fir

2: Lodgepole Pine

3: Ponderosa Pine

4: Cottonwood/Willow

5: Aspen

6: Douglas-fir

7: Krummholz

Research Questions:

Our research surrounds forest and tree types, with cover type, which is nothing but types of trees, as the outcome. Predictors or dependent variables can be wilderness area, soil type, hillshade index, distance to roadways, fire points, water bodies, slope, aspect, elevation of ground, etc. These multiple factors are the ones that affect the environmental conditions for trees in every area. Hence our main research question helps in determining the type of predominant tree that will develop in every location based on the environmental conditions. To answer this question, we will be fitting various models and identifying the model that is best suited to classify the cover types of the trees based on existing dataset. Therefore, the primary question is framed as: “Which model will be best suited to classify the type of predominant tree that will develop in each location based on the environment?”

Our secondary questions further analyze factors such as “What are the most prevalent tree species in the Roosevelt National Forest?”, “Which tree species can thrive in a wider range of conditions?”, and

“Are there any tree species that are more susceptible to environmental factors like elevation or soil type than others?”

Exploratory Data Analysis (EDA):

The EDA was performed in order to find the influence of different variables such as soil type, elevation, cover type, wilderness area, aspect, slope, distance to hydrology and roadways and horizontal distance to fire points have on the tree species. We chose “violin charts” as well as “bar graphs” to represent these relationships. Bar charts were used to count the number of tree species in each wilderness area and to find the correlation between cover type and the different soil types. The violin charts were chosen as they help in representing numeric data distribution and depict the median, standard deviation and density of the data points. They are specifically used when we need to make a comparison between distribution of different groups. In our case, we were comparing the distribution for different cover types.

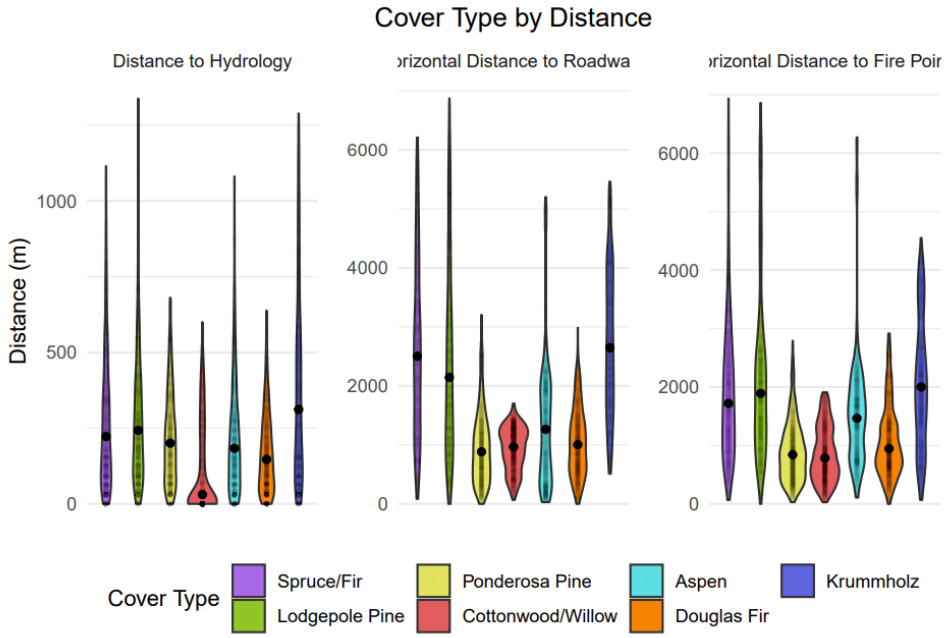

FIGURE 2: Violin charts depicting the correlation between different cover types and distance to hydrology/ horizontal distance to roadways/ horizontal distance to fire points.

The dark black dots in the violin charts shown in Figure 2. are the median. The central line inside each leaf represents interquartile range. The thin elongated line represents the distribution of the data points. The thick section in each leaf of the violin chart indicates that there are more data points having that data value. Looking at the first graph in Figure 2, we can say that as the Cottonwood/Willow has the median value close to 0 meters, that tree species can be found closest to water bodies. Also, as the median of the tree species Krummholz has the highest value compared to other tree species, they can be seen growing in areas farther away from water bodies. We can read the other two charts in Figure 2 in a similar way. The violin charts were used to represent the correlation between cover type and elevation, cover type and wilderness area and cover type and aspect.

Models:

Initially, we had planned to fit three models: Support Vector Machine, K Nearest Neighbours, and Random Forest. During the actual implementation of these models, we faced incorrect classifications, some did not logically make sense, such as k=1 giving the best accuracy for KNN. Therefore, we decided to try other

additional models and ended up performing bagging, boosting, random forest, support vector machine and support vector classifier. We plan on implementing decision trees as well, in the future scope.

- Planned: SVM, KNN & Random Forest
- Actual: Bagging, Random Forest, Boosting, SVM, SVC
- Future Scope: Decision Trees

Results & Findings:

Predicting the most predominant tree:

The model best suited to predict the types of trees that will develop in each location based on the environmental factors are boosting, giving an accuracy of 93%, which is much higher compared to bagging, random forest, SVM and SVC.

Models <chr>	Accuracy <dbl>
Bagging	0.7228571
Random Forest	0.6833333
Boosting	0.9323810
Support Vector Classifier	0.2719048
SVM Classifier	0.6342857

FIGURE 3: Different models used along with their accuracy.

Most prevalent tree species:

The most prevalent tree species in the Roosevelt National Forest with the count of their tree types ordered in descending order are listed below.

Cover_Type	Count
Lodgepole Pine	283301
Spruce/Fir	211840
Ponderosa Pine	35754
Krummholz	20510
Douglas Fir	17367
Aspen	9493
Cottonwood/Willow	2747

FIGURE 4: List of most prevalent tree species in the Roosevelt National Forest along with their count arranged in descending order of their count.

Tree species that can thrive, tree types that can grow in most diverse environments:

After looking at the EDA, we can say that Krummholz seems to grow in much diverse environments like widespread elevation, distance to hydrology and soil type.

Trees more susceptible to environmental factors:

Cottonwood/Willow has the lowest count of trees in the Roosevelt National Forest and the EDA also confirms that this tree type is the most susceptible to all the factors.

Additional Topics:

Are there other previously published solutions to this problem? That is OK! If so, how does your solution differ or compare?

There are previously published solutions to this problem since this dataset and analysis were part of a Kaggle competition. Although there are various solutions to this in Kaggle, we were captivated by the dataset and curious to explore it ourselves to understand how the predictors come into play with nature. Our solution differs from others mainly because of the feature engineering transformations performed on the columns, wherein we reduced the multiple columns for soil type, wilderness area, hill shade index, and distance to hydrology, and aggregated them into one column each. This considerably made our classification easier and more precise. We also used a different set of models to test them out compared to the ones picked in other solutions. Out of the ones we implemented, boosting seemed to be a great fit with 93% accuracy which was honestly a surprise.

Did you have expectations going into the project that were proved correct or incorrect?

The team was astonished by the first glance at the dataset since we learned about various factors that in fact could possibly affect the tree types. Predictors such as distance to fire points or roadways seemed strange initially and hence, we wanted to explore their relationship with the cover type. Our expectations were that these predictors might not be of great importance to the response variable, but we were proved incorrect and found out that although it does not have a direct impact on the response variable, these predictors are strong factors for the wilderness area, which in turn is a predominant predictor for cover type.

How did your final analysis compare to your proposal? Did you have to make changes to your analysis as you encountered problems along the way?

As per our proposal, we thought it would be easy to fit various models on the dataset once cleaned and transformed, that we would be able to explore many more models and compare them to find the best fit, however, when we actually started cleaning the dataset, we realized the complexity of it, especially with having 7 different types of possible outcomes in the response variable made it quite difficult to fit models. Another issue we faced while fitting the initially proposed models is the memory complexity due to the size of the dataset, which we later reduced by picking a sample of 1000 records from every cover type. At the end, although our analysis was proposed to be based on Support Vector Machine, K-Nearest Neighbours & Random Forest, we ended up implementing Bagging, Boosting & Random Forest as we continued to face problems with the other models. For example, with KNN, k=1 was giving the best efficiency which wasn't an ideal case scenario and so we understood that it wasn't classifying the data correctly as our response variable is not binary.

How did your team work together? Did team members work together on pieces of the project, or split the project into different topics for different team members? What did you do?

The team worked great together since it was easy for us to physically get together, sitting down to work together made it easier to divide and distribute as well as help each other. We all complimented each other and worked together on pieces of the project. At first, we started working on cleaning the dataset together, after some progress, we divided into teams of two for the transformation and feature engineering. As for EDA and model fittings, we split the work amongst all of us with all of us trying various models, discussing the setbacks, and then helping each other before deciding to discard that model and trying another one (since that happened a lot). Yash and Prathamesh worked on EDA while Anushri and Rashmi tried implementing different models. We faced hurdles in both parts and then worked together with 4 of us trying to fix things and swapping sections. We tried implementing many other models, however, we learned a lot in the process before settling on bagging, boosting, and random forest. The team was constantly supporting and troubleshooting issues together, which was commendable.

What would you like to do with this project in the future if you had more time?

If we did indeed have more time in the future, we would love to work on decision trees and classify the most predominant trees based on the environmental conditions, that are nothing but our predictor variables in this case. We understand that it would be a complicated process since the predictors classify it down to multiple outcomes and there isn't a set of unique outcomes for different combinations of predictor variables. However, we would love to classify and try to bring it down to the most preferable cover type as a unique outcome for every different set of predictor variable values.

References:

[1] https://towardsdatascience.com/predicting-forest-cover-types-with-the-machine-learning-workflow-1f6f049bf4df