AI Academy Semester 2 Capstone – Forest Coverage

Business Understanding

This capstone is attempting to determined tree coverage type based on given characteristics within Roosevelt National Park and is relative to the forestry and land management industry. The business case for this capstone is the Bureau of Land Management wishes to thin parts of the Roosevelt National Forest so that trees won't have to fight over water and light within the forest. They have determined that there are many Lodgepole Pine trees within the forest and that selling the lumber from these trees will be financially beneficial. Therefore, the Bureau of Land Management wants to know if trees in certain areas are Lodgepole Pines, based on the surrounding environmental features within the forest. Furthermore, it is important that the Bureau not cut down other types of trees as it will not benefit the health of the forest, and the lumber from those trees will be worth less financially.

Data Understanding

The dataset used for this capstone is a csv file that will be downloaded from Kaggle. The dataset includes information on the following:

- Elevation
- Aspect
- Slope
- Horizontal distance to hydrology
- Vertical distance to hydrology
- Distance to roadways
- Distance to fire points
- Shade on the hillside at 9am, noon, and 3pm
- Specific wilderness area within the park
    - The areas are represented with 4 binary columns
- Soil type.
    - The soil type is represented with 4 binary columns
- Type of tree coverage

I will be focusing on discovering the type of tree coverage based on the other factors. While there are 7 different tree types within the park, I will only be determining whether the tree is a Lodgepole Pine or not. I am changing this dataset to a binary problem focused on Lodgepole trees because there are a majority of Lodgepole trees within the data, which will be explained more in depth within the Data Preparation section.

Data Preparation

This is a very easy dataset to work with as it has already been largely cleaned and preprocessed. There are multiple categorical columns within the dataset that have already been encoded. However, I will have to encode the cover type class. There are also no null values within the dataset. There is no class imbalance as there are 283,301 Lodgepole Pine trees and 291,711 other trees, which is only a 2.9% difference. Due to the large majority of Lodgepole Pine trees compared to the other trees, I am changing this data from a multi-class problem to a binary problem to avoid a class imbalance.

Modeling
This capstone is attempting to determine the type of tree coverage depending on multiple factors, it will therefore be a binary classification problem. The target variable is the tree coverage type, specifically if the coverage is a Lodgepole Pine tree. I will be using python, python packages, and GitHub as tools to complete the capstone.

Evaluation
Given that this is a classification problem, I will be using precision, recall, accuracy, and F1 to determine the success of my model. The business case wants to determine the type of tree within an area to know if it needs to cut down that tree for forest health and revenue gain. Therefore, I will be prioritizing precision over recall because it is more important that I cut down the correct type of tree while not harming other types of trees. Essentially, it is more important to have more false negatives (type 2 error) than have more false positives (type 1 error). This is because we don't want or need to cut down all the Lodgepole Pine trees and we don't want to cut down any other type of tree. More false negatives would only leave more Lodgepole pines in the forest, while more false positives would result in unnecessarily cutting down trees that are not Lodgepole Pines.

Tools/Methodologies
I will be using decision trees for this capstone because decision trees maximize information gain while also having parameters that help reduce overfitting of the data. Decision trees are also easy to explain to non-technical stakeholders. Therefore, the Bureau of Land Management will appreciate having a straightforward understanding of the machine learning tool used to create decisions for their business that will help ensure the health of their forest and result in profit gain.