Capstone Project Approach Tom Thorpe

The Forest Cover Capstone Project will apply techniques learned in the Springboard Data Science Foundation Course to predict the type of tree coverage in four different wilderness areas in Colorado.

The data comes from the Ph.D. dissertation by Jock Blackard, 1998, “Comparison of Neural Networks and Discriminant Analysis in Predicting Forest Cover Types.”, Department of Forest Sciences, Colorado State University, Fort Collins, Colorado. The data can be found at: https://archive.ics.uci.edu/ml/machine-learning-databases/covtype/.

The Forest Coverage predictor is used by the US Forest Service to “support the decision-making processes for developing ecosystem management strategies.” Improving the accuracy of this predictor could help the Forest Service improve their planning.
Dr. Blackard found that neural networks were able to predict forest coverage with 70% accuracy. This was a 12% improvement over the Discriminant Analysis methods currently in use which had an accuracy of 58%. This project will see how logistic regression prediction compares to the neural network.

The data consists of 581,012 rows with each row representing a 30 meter by 30 meter cell/area of land. The main features of the data include the outcome, tree type and environmental data consisting of

- Elevation of the cell in meters
- Slope of the cell from 0 to 90 degrees
- Aspect: the compass direction the cell is facing
- Distance of cell to nearest water, fire and road features
- Soil Type which aggregates
    - Geologic Zone
    - Climate Zone
    - Soil Families
    - Rock Types within the soil

Logistic Regression will be used to determine the tree type. Since logistic regression can only provide a true / false type of response, logistic regression models will be created for each tree type and then applied in a specific order to assign tree types if the probability exceeds the threshold for each model and the tree type has not already been assigned. The thresholds will be determined for each model individually and then tuned for the combined final model.

The Soil Type will be split into it's constituent components to see if the constituent components will allow for finer tuning and better performance compared to the aggregated Soil Type value. The two different approaches will be analyzed.

The overall accuracy of the combined model will be used to evaluate the performance. The sensitivity and specificity will also be used to help refine the thresholds. Confusion matrices, 7x7 in size, will also be used to evaluate the effectiveness of the combined models.