Andrew Rucker
Analytics Early Talent Case Study

Predicting forest cover type from cartographic variables

Exploratory Data Analysis (EDA)

This dataset is comprised of tree observations from four areas of wilderness located in Roosevelt National Forest of Colorado. There are 581,012 instances with each one being a 30 x 30-meter cell of forest. Each observation is comprised of information on tree type, area topography, shadow coverage, proximity to nearby landmarks (water, roads, etc.), and soil type. I will be using the first 15,120 observations as my training and validation subset as they contain all attributes along with the cover type of the tree. Once the model is trained, I will then test it against the validation data subset and report its accuracy doing so. Also, the dataset has zero missing values. If there was a small amount of missing values, I would drop those observations. If there was a larger amount, I would replace the missing values with the mean/median values of that area.

Here are the data fields and forest cover types:

Forest cover type: Spruce/Fir (1), Lodgepole Pine (2), Ponderosa pine (3), Cottonwood/Willow (4), Aspen (5), Douglas-Fir (6), Krummholz (7).

Quantitative data fields: Elevation (meters), Aspect (in degrees azimuth), Slope (in degrees), Horizontal/Vertical distance to ‘hydrology’, roadways, and fire points, and Hillshade (9am, noon, 3 pm on a 0 to 255 index).

Binary data fields: 4 types of wilderness area and 40 soil types.

The elevation and hillshade variables seem important for predicting which type of tree will be in what area. As a study in the *Journal of Ecology* found that basal area (area occupied by tree stems) and composition of tree species is most affected by altitude.[1] Another thing to consider is that the cover type is heavily dominated by 1. Spruce-Fir (211,840) and 2. Lodgepole Pine (283,301), and the standard deviation in horizontal distance to roadways (1,559m) in the original dataset.

Insights from training set

As found in the 2007 study, tree density is an important variable as well in forest composition. To show the cover type density within each wilderness area is figure 1, [2] which is a kernel density estimation plot (KDE). The Cache la Poudre area has the highest levels of density in any area. As noted in the ‘covtype.info’ file, it is also the area

with the lowest mean elevation. Showcasing the importance of low elevation for tree density due to the levels of effective oxygen lowering by around 1% per 1,000ft of altitude.[3]

To understand the relationships between cover type, elevation, and the other variables a correlation heatmap was created. [4] This type of matrix requires continuous data; therefore, all binary values were left out.

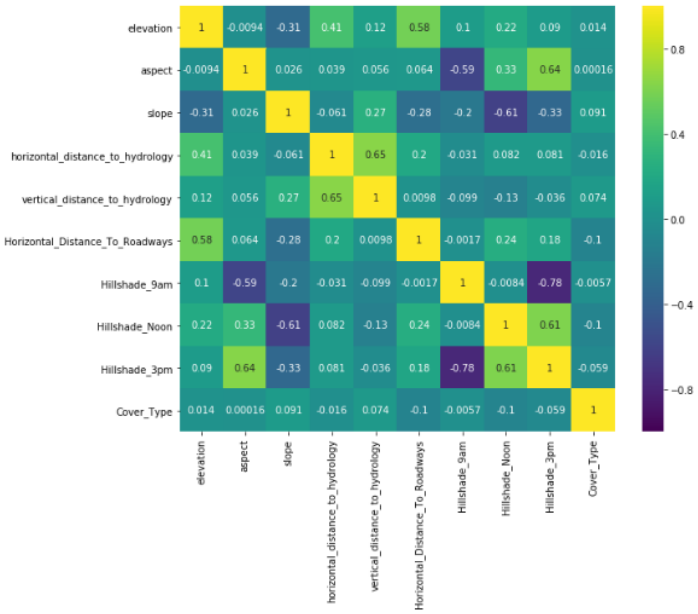

Figure 2 – Correlation heatmap

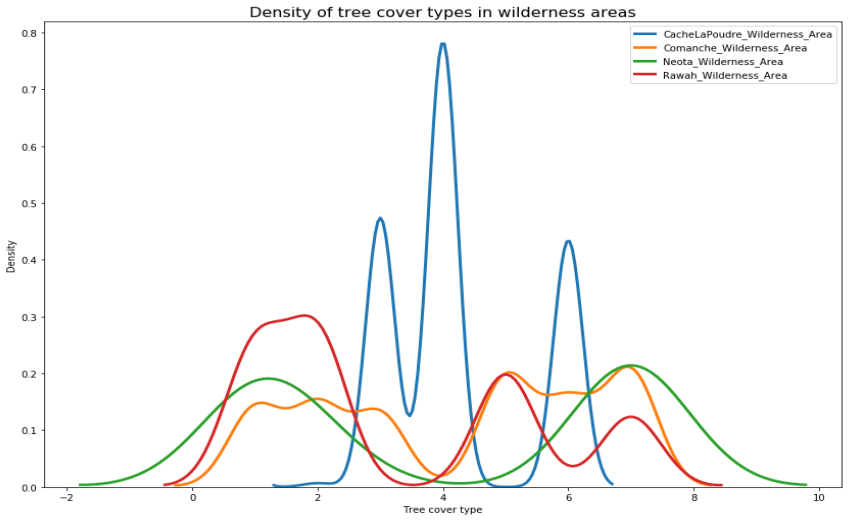

Figure 1 – Kernel density estimation plot

This matrix tells us a lot about the relationships between continuous variables and cover type. As expected, elevation is linearly related to cover type, along with aspect and slope. Another thing to note is that all hillshade variables are similar in their linear relationship with cover type. This aligns with what was found in the study [5], which found that light availability and leaf mass per unit area (similar to our hillshade variable and findings on density) were the two variables that impacted tree species in the selected area of canopy.

Modeling Strategy

As we are looking to classify the cover type within the given area, a multinomial logistic regression model was pursued. This model was pursued because we have a categorically dependent (shown by our insights from the EDA) variable in cover type. With this model, the implemented algorithm will run through our training data set and test the relationship between cover type and our cartographic variables. This model utilizes the softmax function which when giving each cover type returns a ‘high probablitity’ value for a cover type and returns a ‘fewer probability’ for the remaining cover types. The sum of probabilities for cover type of each observation will be equal to one, and its prediction will then be the highest value of those probabilities. [6] The model was created with the ‘train_test_split’ function from the sklearn python library, training the model to the first 11,340 records and testing it against the next 3,780 records which are labeled as the validation data subset in covtype.info. Along with the ‘TTS’ function, the model was built utilizing the sklearn functions ‘LogisitcRegression’,

‘StandardScaler’ (standardizes our variables), and ‘accuracy_score’ (computes given subset accuracy).

Due to the significant difference in correlation between our cartographic variables and cover type, an algorithm that utilizes decision trees could improve accuracy. The random tree algorithm is perfect for this problem as it creates a large number of individual decision trees that operate as an ‘ensemble’. For our problem, each decision tree will give a cover type prediction for each observation based on a series of basic logical decisions (True/False), and the cover type with the most total predictions will become the model’s prediction for each observation. [7]

Again, utilizing the sklearn python library, a second model was built with the ‘RandomForestClassifier’ function. It used the same training split to train the random forest model and would then test the model against the validation subset.

Results

	precision	recall	f1-score	support
1	0.67	0.69	0.68	721
2	0.63	0.53	0.58	715
3	0.62	0.56	0.59	699
4	0.80	0.89	0.84	703
5	0.74	0.81	0.77	710
6	0.62	0.63	0.62	744
7	0.89	0.88	0.88	698
accuracy			0.71	4990
macro avg	0.71	0.71	0.71	4990
weighted avg	0.71	0.71	0.71	4990

Figure 3 – LogisticRegression model results

	precision	recall	f1-score	support
1	0.79	0.76	0.78	721
2	0.79	0.70	0.74	715
3	0.84	0.82	0.83	699
4	0.93	0.98	0.95	703
5	0.91	0.95	0.93	710
6	0.84	0.87	0.85	744
7	0.92	0.96	0.94	698
accuracy			0.86	4990
macro avg	0.86	0.86	0.86	4990
weighted avg	0.86	0.86	0.86	4990

Figure 4 – RandomForest model results

Shown by figure 3, our logistic regression model was able to predict the cover type within the validation dataset (30 x 30-meter cell) with an accuracy of 71%. Whereas the second model (figure 4) utilizing the random forest algorithm was able to obtain a reported accuracy of 86%. The classification report for the logistic regression model showed that the model had difficulty (less than 70%) with the Spruce/Fir, Lodgepole Pine, Douglas Fir, and Ponderosa Pine cover types. The random forest model had similar issues with predicting the first two with less than its overall 86% accuracy. [F5]

In terms of deciding on which model to use, the random forest model should be used, but tweaked. It is an improvement on the LogReg model because of the number of variables we have. LogReg models tend to perform better when there are fewer noise variables, in this case our cartographic variables (noise variables) create an unbalanced dataset, because of this the linearity of the LogReg model doesn’t perform as well. [8]

With more time, we could have tweaked the weight that our randomForest model puts on certain features, such as elevation, soil types, etc. This way our model takes into consideration the importance of elevation and its impact on tree species shown in our correlation heatmap and the 2007 study on altitude, species competition, and growth. We also could have developed other decision tree models such as LightGBM, XGBoost, or a distance-based model like K-Nearest Neighbor and compared them to the performance of our randomforest model.

References

1. Coomes, D. A., & Allen, R. B. (2007). Effects of size, competition and altitude on tree growth. Journal of Ecology, 95(5), 1084–1097. doi: 10.1111/j.1365-2745.2007.01280.x

2. Visualizing the distribution of a dataset¶. (n.d.). Retrieved from https://seaborn.pydata.org/tutorial/distributions.html

3. Oxygen Levels at High Altitudes - Altitude Safety 101. (n.d.). Retrieved from https://www.wildsafe.org/resources/outdoor-safety-101/altitude-safety-101/high-altitude-oxygen-levels/

4. Holt, A. J. (n.d.). Generating Correlation Heat Maps in Seaborn. Retrieved from https://ajh1143.github.io/Corr/

5. Rijkers, T., Pons, T. L., & Bongers, F. (2000). The effect of tree height and light availability on photosynthetic leaf traits of four neotropical species differing in shade tolerance. Functional Ecology, 14(1), 77–86. doi: 10.1046/j.1365-2435.2000.00395.x

6. How Multinomial Logistic Regression Model Works In Machine Learning. (2017, September 14). Retrieved from https://dataaspirant.com/2017/03/14/multinomial-logistic-regression-model-works-machine-learning/

7. Yiu, T. (2019, August 14). Understanding Random Forest. Retrieved from https://towardsdatascience.com/understanding-random-forest-58381e0602d2

8. Kirasich, K., & Smith, T. (2018). Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets . SMU Data Science Review, 1(3), 1–25. Retrieved from https://scholar.smu.edu/cgi/viewcontent.cgi?article=1041&context=datasciencereview

Extra Visual to explain results (mainly out of curiosity)

Was curious as to why the models (more so logistic regression) that I made had difficulty predicting cover types 1, 2, 3, and 6. So, I wanted to visualize the distribution of cover type and elevation. As we can see, the difficult cover types to predict were not 4, 5, or 7 which are the cover types with the tightest distribution of elevation it was the cover types with the widest distribution of elevation. Suggesting that the elevation variable being distributed is the one that makes cover type hard to predict. This is true with the Random Forest algorithm as well. The tight distribution of elevation with cover types 4,5, and 7 results in a nearly perfect prediction of cover type (all 90%+ in recall).