Indigo Catton

CS 777 Final Project

Estimating Biomass with Forest Attributes

Project Introduction

There is increasing interest in biomass estimation for forest monitoring, carbon sequestration, and renewable material production. However measuring biomass is time consuming and expensive to collect and calculate while other attributes like tree count, elevation and height are much easier to measure. The goal of this project is to use a linear regression to predict plot level total above and below ground biomass from input predictor variables. This information could reduce the amount of time spent collecting data and could be used to estimate biomass for given forest attributes without collecting on the ground data. This information could be used in combination with satellite imagery products to more accurately estimate biomass across the country which is valuable information for forest protection, forest productivity, and combating climate change through carbon sequestration.

Dataset

The US Forest Service collects information on forests using forest plots across the country. They collect data on forest attributes including tree size, height, count, age, canopy cover and many other attributes such as biomass. The data for one state contains 10,000 plots but this could be expanded to include the whole country. The data can be downloaded in a sqlite database which is easy to read and work with in python.

Methods

Data Preparation

After reviewing the database guide I selected and summarized a combination of 27 modeled and measured features that are easy to measure and could be indicators of biomass. I used a subset of features here because my limited knowledge of the data made it challenging to summarize all of the features in a meaningful way. I initially chose twice as many attributes but many of these were sparse so I removed attributes with > 2500 null values. One attribute produced a null value with a join but the lack of data indicated 0 so I filled the data with 0's. To handle data with null values, I filtered out data with null values so every record had all values.

The features included a mixture of categorical and continuous data. I used one hot encoding to convert the categorical data to 0 and 1s and split the data 80/20 for train and testing data. I

scaled the data using a standard scalar that was trained on only the training dataset to prevent data leakage.

Linear Regressions

Next I used a linear regression with all of the features as a baseline to assess the impact of feature selection. After reading different methods for feature selection I found that lasso regression for feature selection can be used on both categorical and continuous variables. This can be implemented with pyspark ml with LinearRegression(elasticNetParam=1.0). The L1 (Lasso) regularization drives the coefficients of unimportant features to 0. To select the important features I selected features where the coefficients are not equal to 0.
Then I reran the linear regression with the subset of selected features but saw no changes in the coefficients even after increasing the regularization parameter. The lasso regression coefficients were not converging to 0 so all of the features were being used and the regression was producing the same results.
To continue trying to reduce the number of features needed to estimate biomass I used the pyspark univariant feature selector with f-score to select the top 5 features and re-ran a linear regression with these five features. All of the linear regressions had a regularization parameter of 0.3 and ran for 10 iterations.

Results

The initial linear regression performed well with a R^2 value of 0.91. The regression with the top 5 features decreased but still performed well with an R^2 of 0.81. The linear regression with lasso feature selection was exactly the same as the initial regression.

Linear Regression	RMSE	R^2
All Features	3865	0.9124
Lasso Feature Selection	3865	0.9124
F-score 5 Features	5625	0.8145

Conclusions

While this is just an initial exploration into the dataset the initial results suggest that a combination of modeled and easy to measure attributes can be used to predict biomass. This

was a simple analysis but a more complex analysis including stratification by specific attributes could improve the model. Since this analysis was only run on the data for one state it would be interesting to see how much the model varies by state and if similar features are important in other geographic areas. Further learning and testing is required for me to understand why the lasso feature selection did not change the results. This project was an excellent opportunity to implement the skills and knowledge gained in this class.

References

https://medium.com/@agrawalsam1997/feature-selection-using-lasso-regression-10f49c973f08
https://spark.apache.org/docs/latest/ml-guide.html
https://research.fs.usda.gov/understory/forest-inventory-and-analysis-database-user-guide-nfi
https://apps.fs.usda.gov/fia/datamart/datamart.html