Modeling for predicting Leaf Area of Acer rubrum (Red Maple) tree species in the USA.

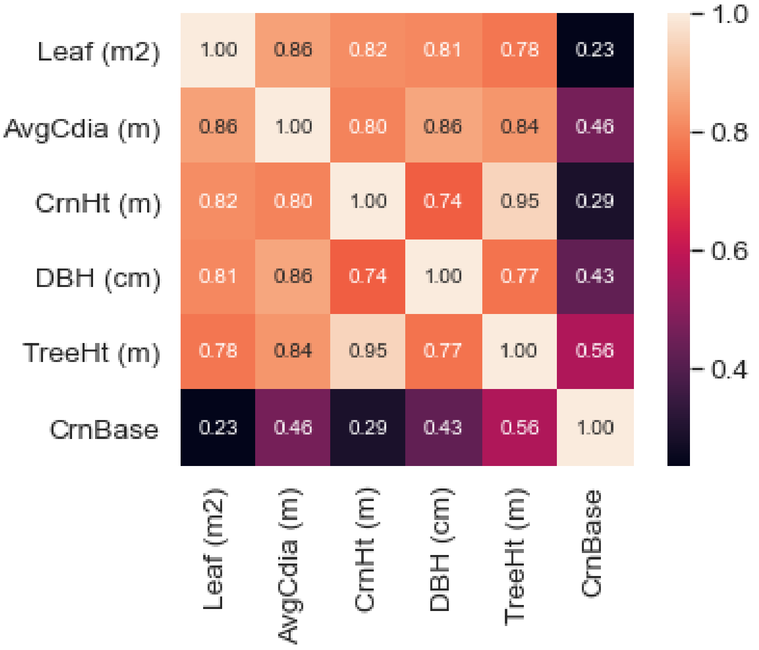
Jatin Mahour

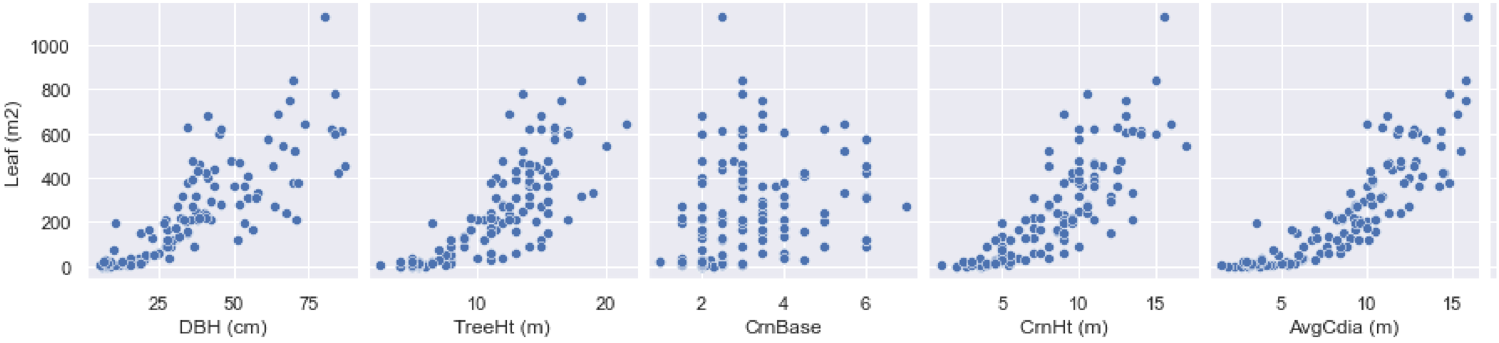
Context: Leaves are the exchange surfaces of plants, absorbing light, taking in carbon dioxide, and releasing oxygen and water vapor. The total leaf area of a tree is a difficult thing to measure, it could be a destructive method to pull all the leaves off a tree, measure them individually, then add the areas together, or a non-destructive digital imaging method. So I have carried out this statistical analysis to estimate the leaf area of a tree from its dbh(Diameter at breast height), tree height, average crown diameter, crown base, and crown height.

Over a period of 14 years, scientists with the U.S. Forest Service Pacific Southwest Research Station recorded data from a consistent set of measurements on over 14,000 trees, with 171 distinct species in 17 U.S. cities. The online database is available at http://dx.doi.org/10.2737/RDS-2016-0005. I have used a subset of this database containing only red maple trees for my study, as it's one of the most common tree species in the USA. My sample size is 129 trees collected from three climate zones namely, Central Florida, Midwest, and Pacific Northwest.

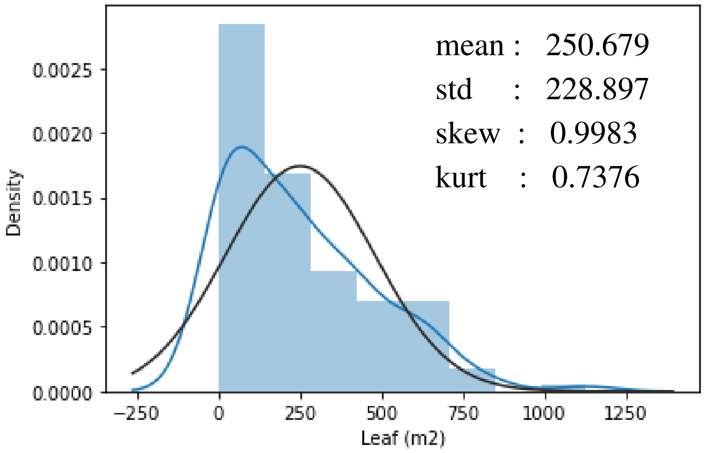
Exploratory data analysis

- There are five independent variables dbh(Diameter at breast height), tree height, average crown diameter, crown base, crown height, and one dependent variable Leaf area.
- Checking for correlation between variables. There is a high correlation between dependent and independent variables and also among dependent variables, resulting in high multicollinearity.
- Exploring the relationship between dependent and independent variables using the pair plot function of python. There is a slight exponential relation between Leaf area and tree height, leaf and crown height, and leaf and average crown diameter.





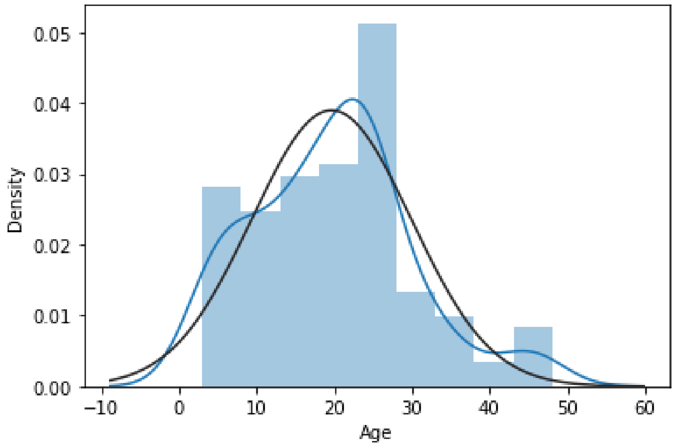
- Checking for normality of target or dependent variable i.e leaf area. Linear regression will make more reliable predictions if your input and output variables have a Gaussian/normal distribution. You may get some benefit using transforms (e.g. log or BoxCox) on your variables to make their distribution more Gaussian looking.
- The target variable is rightly skewed.

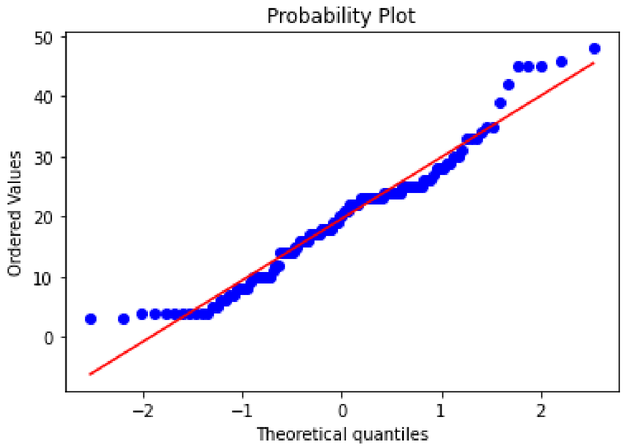


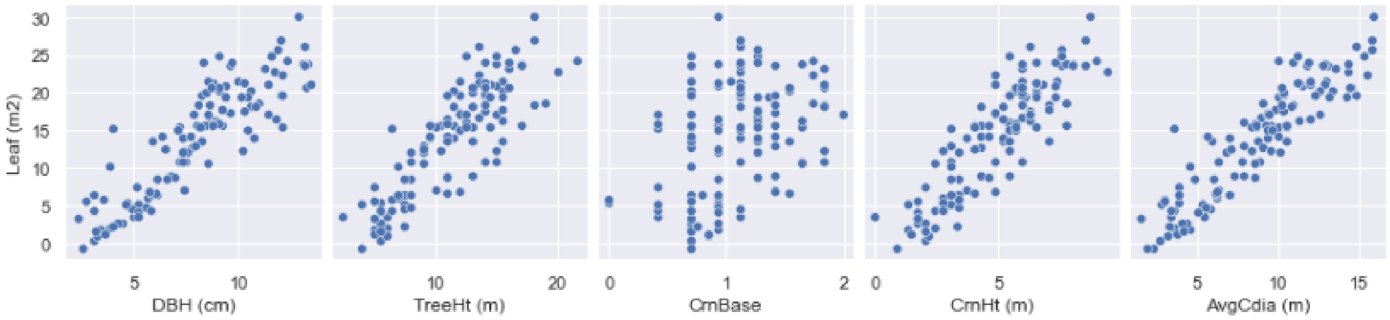
Preparing data for modelling

- There are no significant outliers and missing values in the data.
- Performing box-cox transformation on target and dependent variables to bring them close to normality and transforming exponential relationship to linear.
- After box-cox transformation, the target variable has become more normally distributed. The relationship between variables is more linear after the box-cox transformation.

Variables	Before box-cox	After box-cox
skew(Leaf)	0.99838	-0.22838
skew(TreeHt)	-0.07621	-0.07621
skew(DBH)	0.57492	-0.08686
skew(CrnBase)	0.87187	-0.00068
skew(AvgCdia)	-0.06983	-0.06983
skew(CrnHt)	0.14240	-0.10233

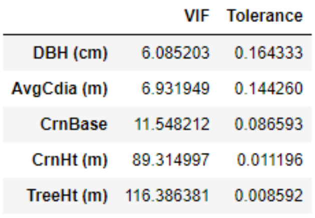




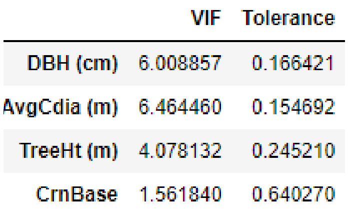


- Checking for multicollinearity using Variation Inflation Factor and Tolerance values. So either a high VIF or a low tolerance is indicative of multicollinearity. VIF is a direct measure of how much the variance of the coefficient (ie. its standard error) is being inflated due to multicollinearity.

- I have removed crown height from our sample to bring down the multicollinearity.



	VIF	Tolerance
DBH (cm)	6.085203	0.164333
AvgCdia (m)	6.931949	0.144260
CrnBase	11.548212	0.086593
CrnHt (m)	89.314997	0.011196
TreeHt (m)	116.386381	0.008592

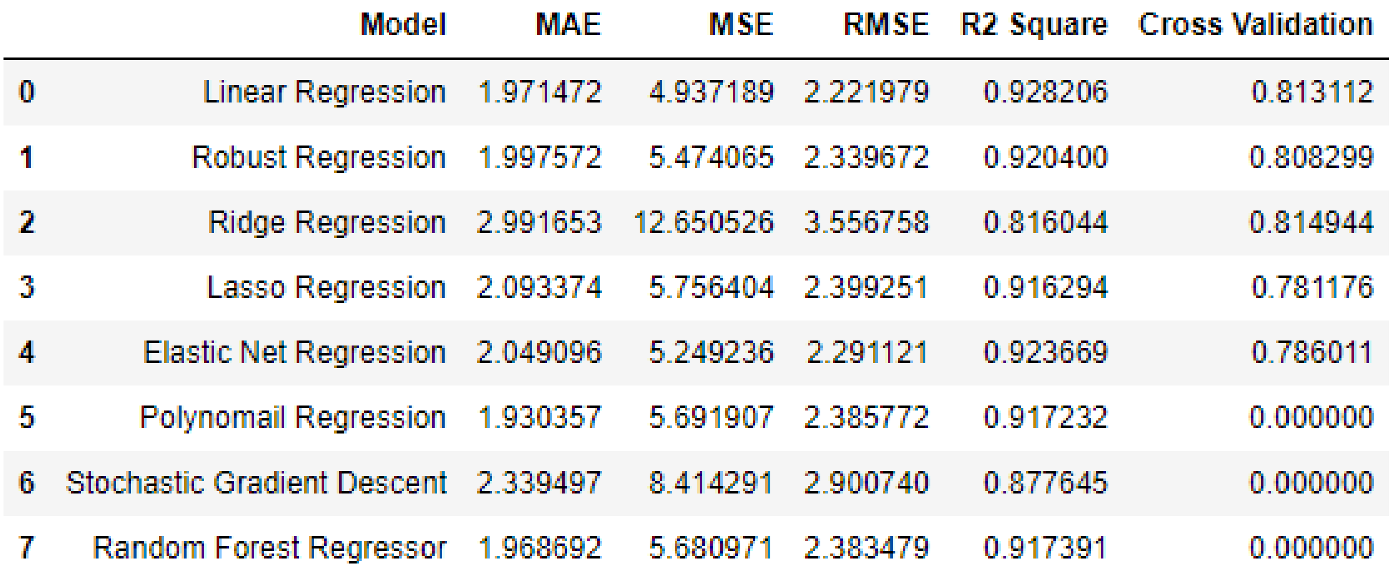


	VIF	Tolerance
DBH (cm)	6.008857	0.166421
AvgCdia (m)	6.464460	0.154692
TreeHt (m)	4.078132	0.245210
CrnBase	1.561840	0.640270

Building Models

- I have split the data into 7:3 ratio, i.e. I have used 70% of the data to train our model and 30% of the data to test our model's validity.
- I have used three independent variables, i.e tree height, dbh(diameter at breast height), and average crown diameter to predict the leaf area.

- Seven different models were created and the best model was determined using the evaluation metrics.
- Out of the seven models, linear regression has the lowest RSME value of 2.22 and best R-Squared value of 0.928.



	Model	MAE	MSE	RMSE	R2 Square	Cross Validation
0	Linear Regression	1.971472	4.937189	2.221979	0.928206	0.813112
1	Robust Regression	1.997572	5.474065	2.339672	0.920400	0.808299
2	Ridge Regression	2.991653	12.650526	3.556758	0.816044	0.814944
3	Lasso Regression	2.093374	5.756404	2.399251	0.916294	0.781176
4	Elastic Net Regression	2.049096	5.249236	2.291121	0.923669	0.786011
5	Polynomail Regression	1.930357	5.691907	2.385772	0.917232	0.000000
6	Stochastic Gradient Descent	2.339497	8.414291	2.900740	0.877645	0.000000
7	Random Forest Regressor	1.968692	5.680971	2.383479	0.917391	0.000000

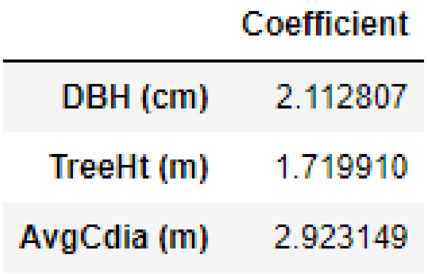
Here are three common evaluation metrics for regression problems:

- **Mean Absolute Error (MAE)** is the mean of the absolute value of the errors. **MAE** is the easiest to understand because it's the average error.
- **Mean Squared Error (MSE)** is the mean of the squared errors. **MSE** is more popular than MAE because MSE "punishes" larger errors, which tends to be useful in the real world.
- **Root Mean Squared Error (RMSE)** is the square root of the mean of the squared errors. **RMSE** is even more popular than **MSE** because **RMSE** is interpretable in the "y" units.
- All of these are loss functions because we want to minimize them.

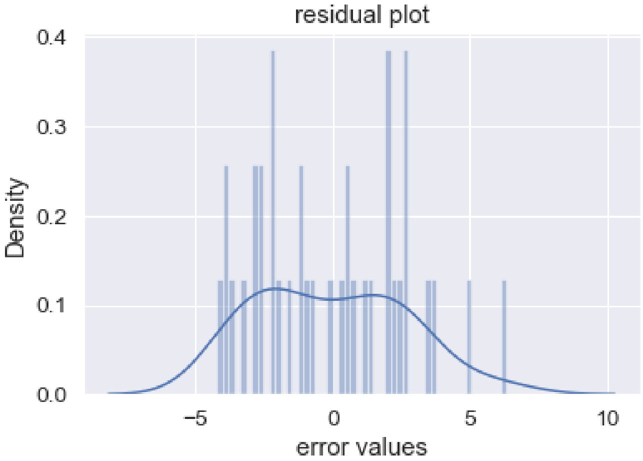
Interpreting the coefficients:

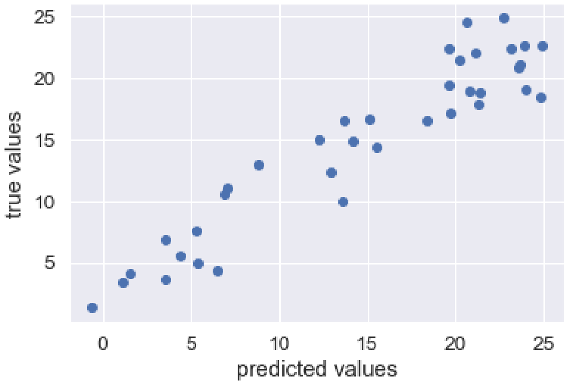
- Holding all other variables fixed, a **1 unit** increase in **DBH** is associated with an increase of **2.11 unit** increase in **leaf area.**
- Holding all other features fixed, a **1 unit** increase in **tree height** is associated with an increase of **1.71 unit** increase in **leaf area.**
- Holding all other features fixed, a **1 unit** increase in **average crown diameter** is associated with an increase of **2.92 unit** increase in **leaf area.**

Intercept :13.7054



	Coefficient
DBH (cm)	2.112807
TreeHt (m)	1.719910
AvgCdia (m)	2.923149





Equation to predict leaf area :

Y = 13.7054 + 2.11*DBH(cm) + 1.712*TreeHt(m) + 2.923*AvgCdia(m)