Penny Silliman

EGR 361 A

04/15/2024

Portland Parks Tree Inventory: Day 1

Any code or formulas used to answer the below questions?:

https://github.com/PennyS8/Portland-Parks-Tree-Inventory

What format did the dataset come in? (e.g. text file, .csv file, etc.):

The dataset came in a .csv (Comma Separated Values) file format.

What type of data is available? Is it continuous or discrete?:

The dataset contains various types of data (most of the numeric data appears to be continuous):

- Date (Inventory_Date)
- String (Species, Condition, CollectedBy, Notes, etc.)
- Numeric (DBH, TreeHeight, CrownWidthNS, etc.)

How was the data generated? By whom? What did they control for?:

The data was generated by the City of Portland. They ensure effective data collection by including details such as species, condition, city employee collecting the data etc.

How many samples/examples/datapoints are in the dataset?:

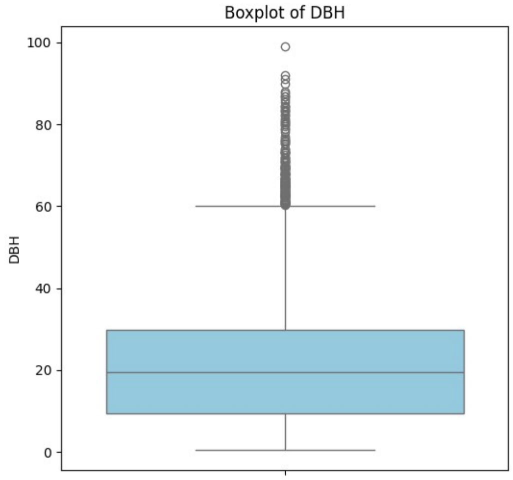
There are 25,740 datapoints in the dataset.

What is the name and data type of the label or "true value" in the dataset?:

The label or "true value" in the dataset is "Total_Annual_Benefits", and its data type is Money Data Type (USD).

What does the distribution of the label look like? What are the descriptive statistics? (Mean, median, range, standard deviation):

Diameter at Breast Height (DBH): Numeric Float (Feet) XX.X



DBH Box-Plot Values:

- Minimum: 0.5
- (Q1) 25th Percentile: 9.5
- (Q2) Median: 19.5
- (Q3) 75th Percentile: 29.8
- Maximum: 99.0

DBH Descriptive Statistics:

- Mean: 20.65
- Median: 19.50
- Range: 98.50
- Std-Dev: 13.38

How many variables are available? What are their names? How are they relevant to the problem?:

There are 42 variables in the dataset, but all except for 2 will be ignored for this analysis. The two variables used will be Diameter at Breast Height (DBH) and Total Annual Benefits (TAB). Other variables that are related but will not be included:

- Categorical variables:

[Species, Genus, Condition, Origin, Size (S/M/L)]

- Size measurements variables:

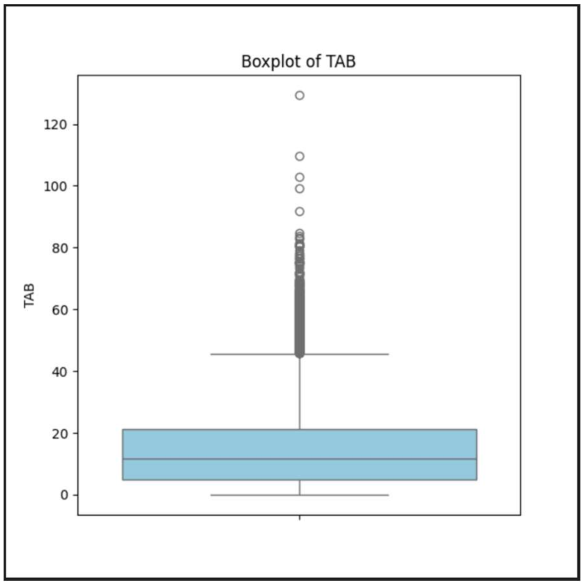
[Tree Height, Crown Width (EW & NS), Crown Base Height]

- Direct cost evaluation related variables:

[Structural Value, Carbon Storage lb, Carbon Storage value, Carbon Sequestration lb, Carbon Sequestration value, Stormwater value, Pollution Removal oz, Pollution Removal value]

What are the descriptive statistics for each variable? (Mean, median, range, standard deviation)
What do the distributions of the variables look like? (Attach screenshots)?:

Total Annual Benefits (TAB): Money Data Type (USD) \$XX.^{XX}



TAB Boxplot Values:

- Minimum: 0.5
- 25th Percentile (Q1): 9.5
- Median (Q2): 19.5
- 75th Percentile (Q3): 29.8
- Maximum: 99.0

TAB Descriptive Statistics:

- Mean: 14.58
- Median: 11.61
- Range: 129.40
- Std-Dev: 12.46

How correlated are each of the variables with the labels?:

The correlation coefficient between DBH & TAB: 0.80

Correlation Matrix	DBH	Tree Height	Crown Width NS	Pollution Removal oz	Stormwater ft	Total Annual Benefits
DBH	1.000000	0.800173	0.748444	0.782113	0.782083	0.799200
Tree Height	0.800173	1.000000	0.568684	0.581136	0.581108	0.584098
Crown Width NS	0.748444	0.568684	1.000000	0.899588	0.899576	0.915668
Pollution Removal oz	0.782113	0.581136	0.899588	1.000000	0.999998	0.994936
Stormwater ft	0.782083	0.581108	0.899576	0.999998	1.000000	0.994935
Total Annual Benefits	0.799200	0.584098	0.915668	0.994936	0.994935	1.000000

Are any of the variables highly correlated with each other?:

A correlation greater than 70% is considered a highly correlated variable. The highlighted values are the intersection of the variables that are indeed highly correlated.

Penny Silliman

EGR 361 A

04/18/2024

Portland Parks Tree Inventory: Day 2

Any standardization or normalization procedure used?:

DBH was standardized using sklearn's StandardScaler.

The dimensions of your training data:

The training data has dimensions of (20592, 1), indicating that after removing data points without sufficient data remained 20592 data points, with the 1 feature DBH.

The dimensions of your testing data:

The testing data has dimensions of (5149, 1), indicating 5149 samples were taken from the 20592 data points, and of course still the 1 feature DBH.

Any code or formulas used to answer the above questions?:

https://github.com/PennyS8/Portland-Parks-Tree-Inventory

R^2 and RMSE values on your training and testing sets:

Training set performance:

R^2: 0.641813951540081

RMSE: 7.4432723311796348

Testing set performance:

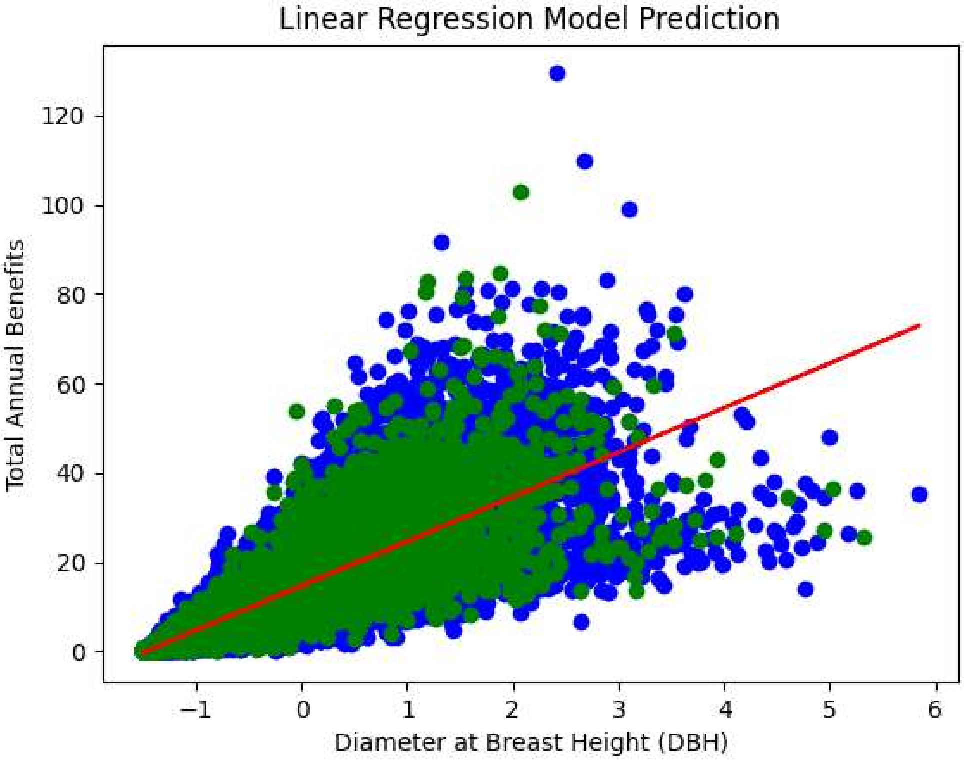
R^2: 0.6357081001919545

RMSE: 7.48831451989267

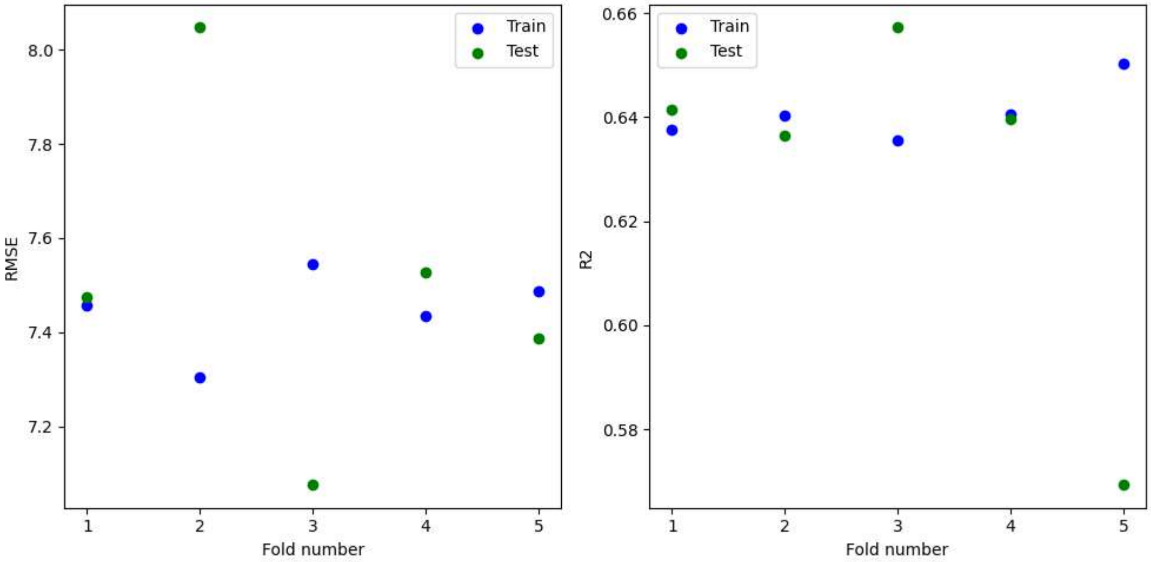
Which features were most important in creating your model? Does it make sense that those features were important?;

The most important feature is DBH (Diameter at Breast Height). Which is sensible as the DBH is indicative of the age and size of the tree, which naturally corresponds to the value of the tree and thus its Total Annual Benefits. The Total Annual Benefits represent the combined monetary worth of the various environmental advantages that trees offer to the community annually, including carbon sequestration, air pollution reduction, and mitigation of stormwater runoff, alongside other ecological benefits.

Plot your model over a scatterplot of the most important feature. Use a different color to plot points used for training and testing. Attach a screenshot of this plot to your report:



Did any of your test datapoints fall outside of the range used in the training data? If so, how good was the prediction on those points?



Describe qualitatively the performance of your model:

For the RMSE (Root Mean Squared Error) values, the closer to 0 indicate better model performance. Since the target variable is “Total Annual Benefits” the units are the dollar value. So the 7.5 that the RMSE hovers near is $7.50. Which is quite reasonably low value.

The R^2 ranges from 0 to 1 as they are a percentage of how predictive the regression is; 1 meaning that the DBH is 100% indicative of the Total Annual Benefits. Thus we see in our table that the BDH is about 64% indicative of the Total Annual Benefits.

There is some difference between the Test and Training points on both tables. This may indicate a possible overfitting of the testing data. However, while the outliers seem dramatic if you recognize the scale of the Y-axis it appears to still be within a reasonable range for the table.

Penny Silliman

EGR 361 A

04/20/2024

Portland Parks Tree Inventory: Day 3

Who you discussed your setup with:

Jonathan Nerenberg commented:

> *"Day 1 looks great, all questions are adequately answered, and I like your choices for the label and variable. It is cool to see how many of the variables are highly correlated."*
>
> *"Day 2 also looks very good. You have plenty of data points for both training and test data. I like your explanation for why you chose DBH. Your Lin Reg Model looks beautiful. I think you could add a few words explaining your graphs of the outliers. Lastly, in your conclusion, you could explain exactly what TAB is used for- is it like how much value the tree gains per year per some volume of the tree? Overall great analysis."*

What conclusions can you draw based off your analysis?:

The results of the analysis align with expectations, as DBH is often correlated with the age and size of a tree, which in turn influences its environmental benefits and value to the community. The R^2 values indicate that DBH explains about 64% of the variation in TAB. Additionally, the RMSE suggests that the model's predictions are within a reasonably low error margin indicating its effectiveness in estimating TAB based on DBH.

What have you learned from this whole process?:

The feedback I was given made me realize that my deep experience with the data and my code made me forget about the context, which is very important but very clear to me because I have been working with the dataset this whole time. But from the perspective of the reviewer it was not so clear. It is important to take that step back and reevaluate my reports and comment, as well as make sure my reports are peer-reviewed so those sorts of issues don't remain in the published version of the report.

Exploring the dataset's variables, distributions, and correlations is important for understanding the relationships between features and the target variable. Initially, I overlooked the other tree features/variables. However, the correlation analysis was insightful, especially regarding the multicollinearity of certain variables. Notably, Tree Height stood out as a variables with significant lack of correlation, contrasting with the interdependencies of most of the other variables.