Machine Learning Engineer Nanodegree

Capstone Project

Aman Aggarwal
April 8th, 2017

I. Definition

Abalone is a shellfish considered a delicacy in many parts of the world. The abalone shell and the meat is of value. The abalone dataset from UCI Machine Learning Archives comes with the goal of attempting to predict abalone age given various descriptive attributes of the abalone. You can find the link to this dataset here. The dataset comes from a 1994 study "The Population Biology of Abalone" in Tasmania. It was donated to the UCI Machine Learning Repository in 1995 by Sam Waugh from the Department of Computer Science at the University of Tasmania (Australia).

In this project, we have attempted to build a regression model that will predict the age of abalones based on their physical measurements. Since this is a regression problem with dependent variable already given to us, the machine learning technique used here is a supervised learning model.

References :

1. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.14.2321&rep=rep1&type=pdf – Helped me to understand the data mining concepts that'll be used in this project like treating missing values, conversion of continuous variable to categorical variable, attributes correlation.
2. http://www.cs.toronto.edu/~delve/data/abalone/abaloneDetail.html- This reference material helped me in understanding the data description and also the benchmark models that have been analyzed in the past.
3. http://shapbio.me/courses/biolB215s13/abalone_cleaning.html - This article helped me in visualizing the dataset and comprehending the relationship between different attributes of the dataset.
4. https://archive.ics.uci.edu/ml/support/Abalone - This is the support site of Abalone dataset from which we can get numerous reference material like research paper, articles to understand the work done in the past.

Problem Statement

Abalone dataset consists of physical information about the abalones like their Weight, Height etc. which can be used to predict the number of rings on the abalone which, in turn, helps us to find out their age. The age of abalone is determined by cutting the shell through the cone and counting the number of rings through a microscope.

We have used Decision Tree Regression to make our predictive model to solve this problem. The process involves cleaning/modification of the continuous data into Boolean or discrete data. Then dividing the data into training and testing set and building a Decision Tree Regressor over it. Then we have tried to optimize our model for precise results.

Metrics

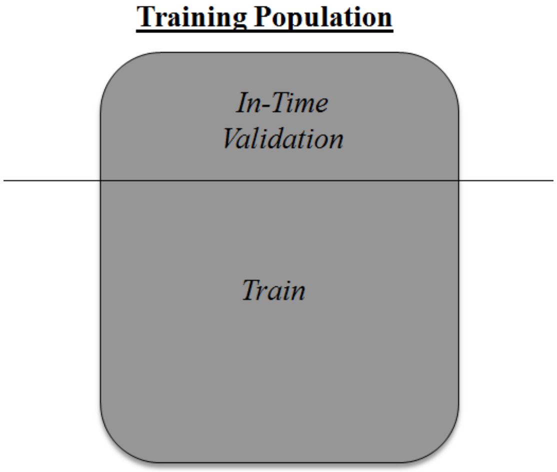
1. **RMSE** is the most popular evaluation metric used in regression problems. It follows an assumption that error are unbiased and follow a normal distribution.

$$RMSE=\sqrt{\frac{\sum_{i=1}^{N}(Predicted_i-Actual_i)^2}{N}}$$

Where, N is Total Number of Observations.

We aim to reduce it as it represents the deviation of our predicted class from the true predicted class

2. **Cross Validation** - Cross Validation is one of the most important concepts in any type of data modelling. It simply says, try to leave a sample on which you do not train the model and test the model on this sample before finalizing it.



II. Analysis

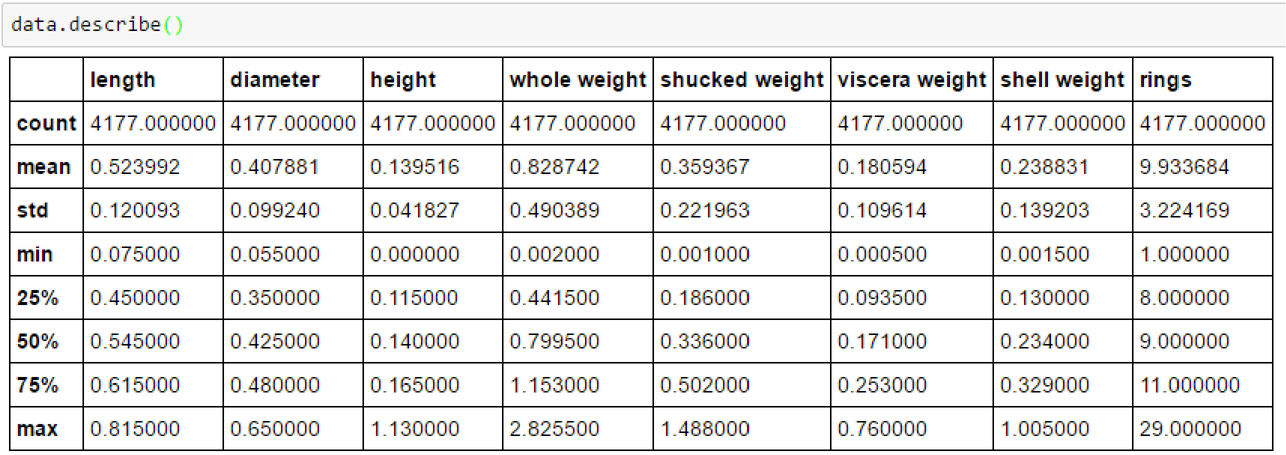
Data Exploration

The Abalone dataset consists of approximately 4000 records and 9 variables that consists of both categorical and continuous variables.

The 9 variables description is as follows:

Attribute	Domain	Variable type
Sex	{M, F, I}	Categorical
Length	[0.075, 0.815]	Continuous
Diameter	[0.055, 0.65]	Continuous
Height	[0.0, 1.13]	Continuous
Whole_weight	[0.0020, 2.8255]	Continuous
Shucked_weight	[0.0010, 1.488]	Continuous
Viscera_weight	[0.0005, 0.76]	Continuous
Shell_weight	[0.0015, 1.005]	Continuous
Rings	{15, 7, 9, 10, 8, 20, 16, 19, 14, 11, 12, 18, 13, 5, 4, 6, 21, 17, 22, 1, 3, 26, 23, 29, 2, 27, 25, 24}	Continuous

The above table shows the range of values (domain) of each variable in the table with their data type.



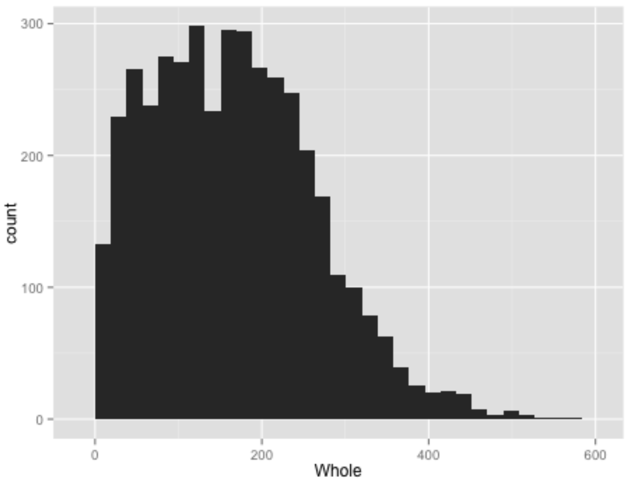
```
data.describe()
```

	length	diameter	height	whole weight	shucked weight	viscera weight	shell weight	rings
count	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000
mean	0.523992	0.407881	0.139516	0.828742	0.359367	0.180594	0.238831	9.933684
std	0.120093	0.099240	0.041827	0.490389	0.221963	0.109614	0.139203	3.224169
min	0.075000	0.055000	0.000000	0.002000	0.001000	0.000500	0.001500	1.000000
25%	0.450000	0.350000	0.115000	0.441500	0.186000	0.093500	0.130000	8.000000
50%	0.545000	0.425000	0.140000	0.799500	0.336000	0.171000	0.234000	9.000000
75%	0.615000	0.480000	0.165000	1.153000	0.502000	0.253000	0.329000	11.000000
max	0.815000	0.650000	1.130000	2.825500	1.488000	0.760000	1.005000	29.000000

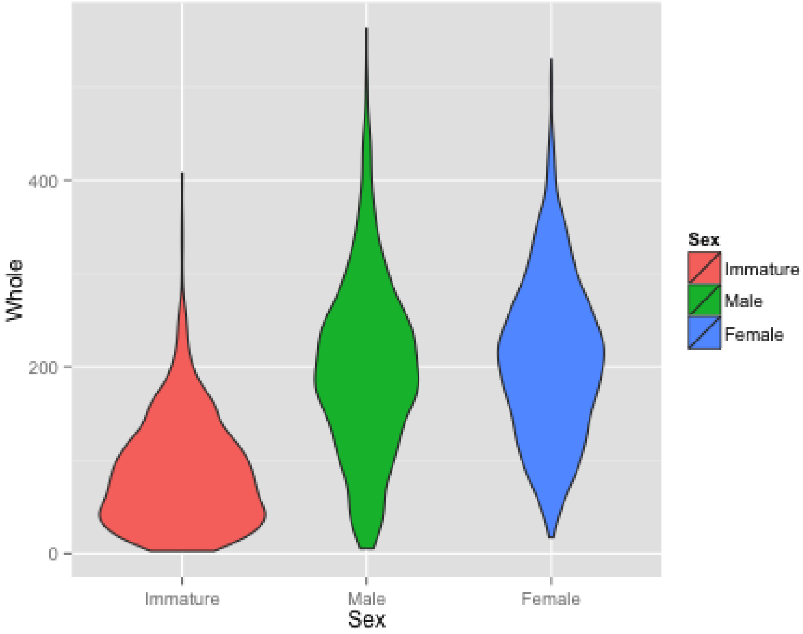
The above table shows the statistical description of the data of the dataset.

Exploratory Visualization

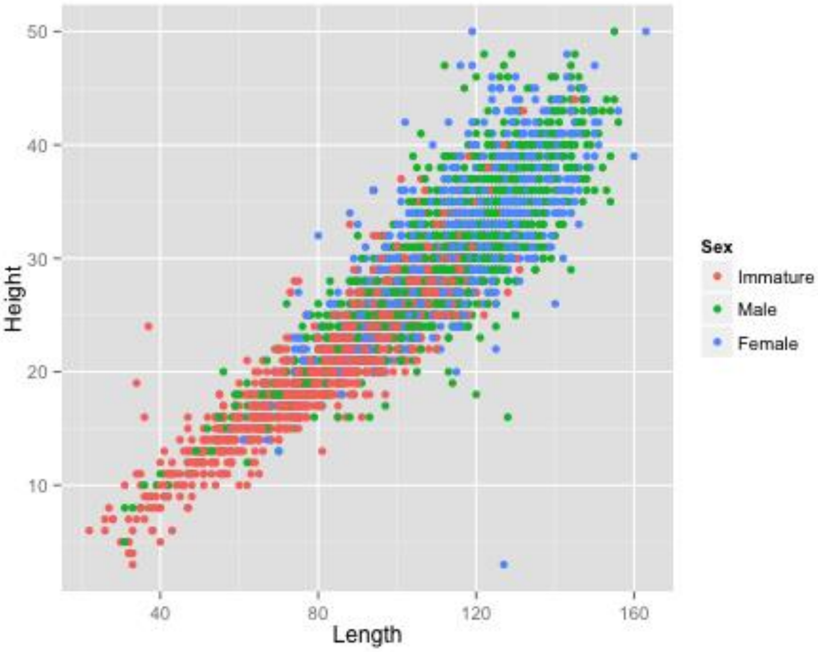
There are many types of plots and graphs that can be used to visualize the data. Here, we have used violin plots, scatterplots and histograms to explore the data.



This is a histogram showing the frequency of abalones with different “whole weights”. The graph looks positively skewed which means that mean > mode.



This is a violin plot showing the “sex” vs “whole weight” plot. It shows the probability density of the data at different values. For the above plot, the count of ‘Immature’ abalones is maximum in the range 0-200 and it’s almost zero beyond that.



.

This is a scatterplot showing the “height” and “length” of abalones in different “sexes”. As the datapointsreally close to each other, we do not see any clear boundary of distinction which makes the attributes highly correlated. The above mentioned graphs and plots are plotted using **matplotlib** library.

Algorithms and Techniques

In this project, we have used a Decision Tree Regressor to model our estimator of the number of rings of abalone. Decision Tree Regression seems a valid algorithm here because the dependent variable is of continuous type. Other options could have been Polynomial Regression, Lasso Regression, K-NN but I’ve chosen Decision Tree Regression because they are simple in nature and easy to understand. It gives a start-off to our project and also provides us with a scope of improvement and optimization of our model.

The method starts by searching for every distinct values of all its predictors, and splitting the value of a predictor that minimizes our statistic RMSE. Unlike other linear regression models that calculate the coefficients of predictors, tree regression models calculate the relative importance of predictors such as overall reduction of optimization criteria like Sum of Squared errors(SSE). The Random Forest being more advanced technique, uses many such decision trees and chooses their average for the resutls.

Benchmark

Many attempts have been made on improving the prediction quality of the model for this dataset. Previous works have shown that on using:

a. Polynomial Regression technique with RMSE = 2.5
b. Lasso Regression technique with RMSE = 2.2

We can see that the prediction models are quite successful with such a small dataset. We have tried to reduce the RMSE with our prediction model and understand the underlying complexity of this dataset empirically. The results have shown that the Random Forest Regressor also reduces the RMSE to 2.2

This problem has also been seen as a classification problem and has been solved with the techniques such as SVM, Logistic Regression but the accuracy was quite low. Here, we only take our dig at linear regression to make a better estimator of the rings.

III. Methodology

Data Preprocessing

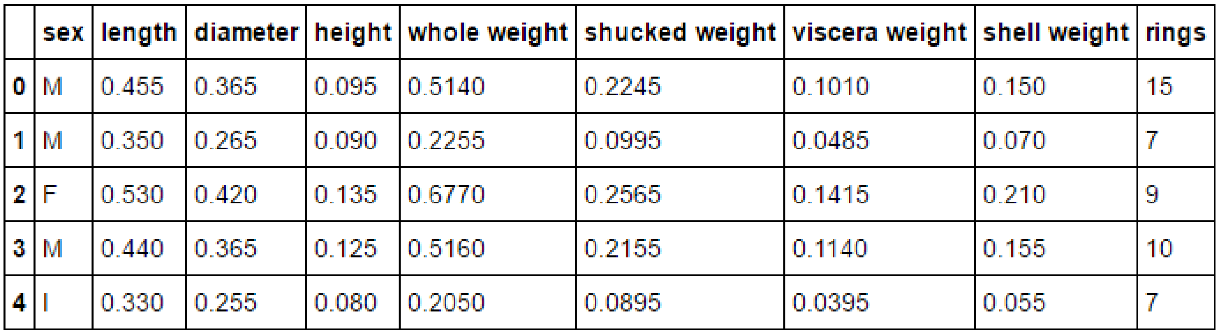
There are 2 files that are downloaded in this project: abalone.data and abalone.names. The data preprocessing steps required adding columns names to the data and treatment of categorical attributes of the data.

The data pre-processing steps are as follows:

a. Adding the column names to the abalone.data file from abalone.names file.

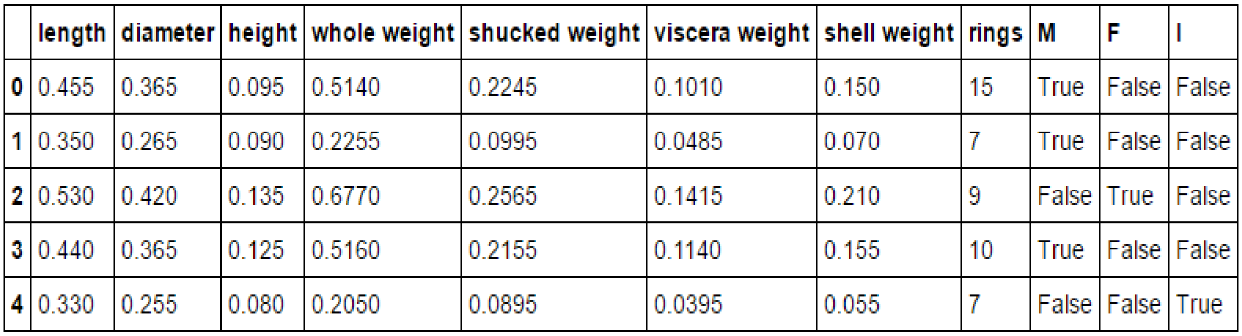
```
column_names = ["sex", "length", "diameter", "height", "whole weight",
                "shucked weight", "viscera weight", "shell weight", "ri
ngs"]
data = pd.read_csv("abalone.data", names=column_names)
```

b. One-hot encoding of 'Sex' variable:
So before one-hot encoding, the data looked like:



	sex	length	diameter	height	whole weight	shucked weight	viscera weight	shell weight	rings
0	M	0.455	0.365	0.095	0.5140	0.2245	0.1010	0.150	15
1	M	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.070	7
2	F	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.210	9
3	M	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.155	10
4	I	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.055	7

c. After one-hot encoding, the data looked like:



	length	diameter	height	whole weight	shucked weight	viscera weight	shell weight	rings	M	F	I
0	0.455	0.365	0.095	0.5140	0.2245	0.1010	0.150	15	True	False	False
1	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.070	7	True	False	False
2	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.210	9	False	True	False
3	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.155	10	True	False	False
4	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.055	7	False	False	True

After the data pre-processing phase, our data has become suitable for further analysis and ready to enter the model building and implementation phase.

Implementation

In this project, we have used Python libraries like **URLretrieve** to download the dataset from the UCI repository website, **numpy** and **pandas** to shapen and describe the data, **matplotlib** to plot the graphs and scatterplots, **sklearn** to make use of regression and evaluate performance metrics and cross-validate the data.

The project design and implementation is as follows:

1. Download the dataset and explore it. Describe the data using statistical tools like mean, standard deviation, range etc.
2. Data wrangling – Conversion of continuous variables to categorical/Boolean variables for analysis.
3. Check for attribute usefulness and correlation between the attributes by plotting graphs.
4. Split data into training (75%) and testing set (25%).
5. Build a prediction model using decision tree regressor and explore the results on scatterplot and check for RSME.
6. Use cross validation and run the prediction model several times on the cross-validated datasets.
7. Plot learning curves to check if there is a scope of improvement.
8. We employed advanced methods like RandomForestRegressor and optimized our model by tuning it’s parameters.

Refinement

Initially we employed Decision Tree Regressor without using any of it's parameters and found the RMSE to be 3. Then, we tuned one of its parameters 'max_depth' and equaled it to 10 (default = None). By doing so, we explicitly expanded the tree until the tenth node and not beyond it. The result got better with RMSE = 2.7. Next, we used a more sophisticated model called Random Forest which gave RMSE = 2.4. All parameters that influence the learning are searched simultaneously and fitted to the model. It's performance is slower but it's better than GridSearch as GridSearch finds a particular number of parameters that influence the learning giver by n_iter. Yet, there was scope of improvement by tuning it's parameters as well. So we used 'n_estimators' parameter to be 100, which is 10 by default, and is the number of trees in used in the model.
Now this model gave us an optimum result of RMSE = 2.1

IV. Results

Model Evaluation and Validation

The final model has been achieved by a systematic approach to solving this problem. It seems quite reasonable and totally aligned with the expectation. With such a few data points at hand, the results seem to be good enough.
The final model has been tested on the test data which has been cross-validated already. This ensures that the results of a statistical analysis generalizes to an independent data set. Cross validation also ensures that the small perturbations in training data or the input space does not affect the results.
The results have been cross checked several times in cross-validation and different models used in the approach in this problem. Thus, they can be trusted.

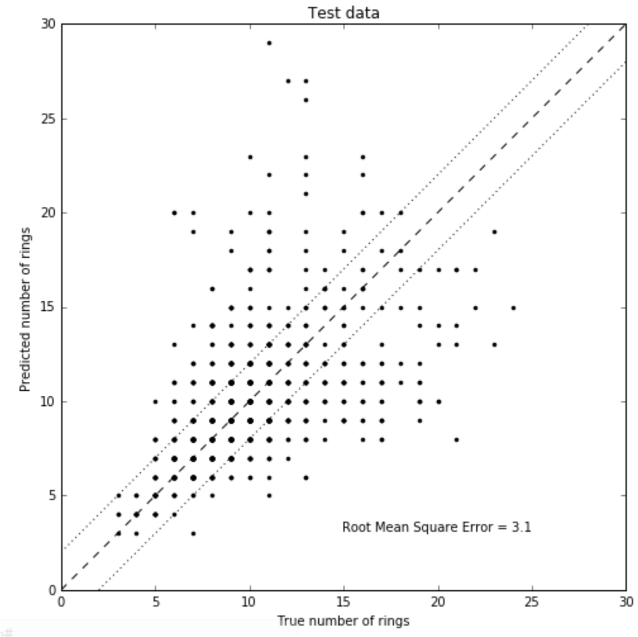
Justification

Our attempt to solve this problem is of significant value as it outperforms one of the benchmark models polynomial regression that gives the RMSE to be 2.5 while our model gives an RMSE of 2.2. It remains in tie with the Lasso regression technique that achieved the same RMSE as our model.

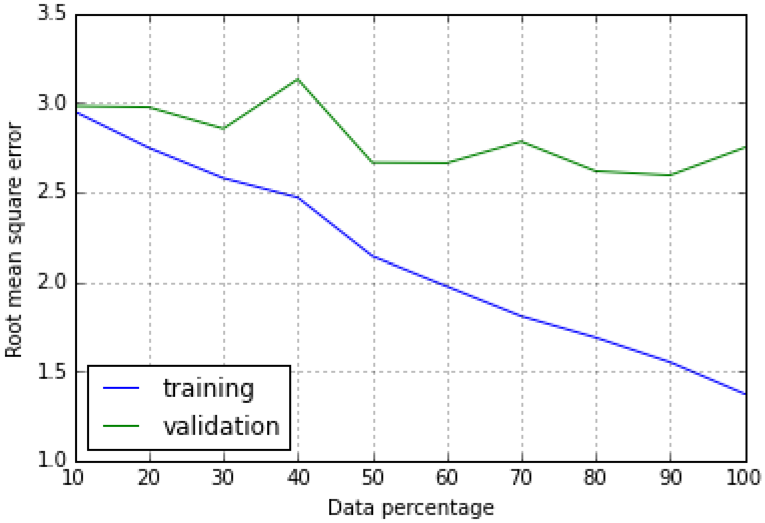
V. Conclusion

Free-Form Visualization

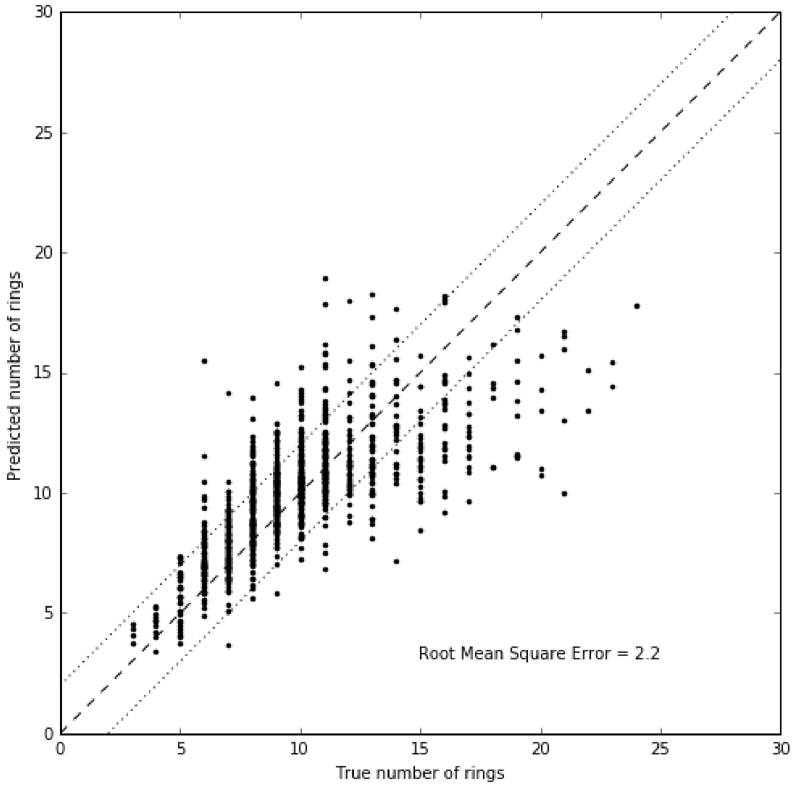
1. The initial scatterplot using Decision Tree Model with un-tuned parameters.



2. The RMSE of training and validation dataset. It shows a large gap in the end showing that the data is not fitting that perfectly in the Decision Tree model. This is when we realize the need of a more sophisticated model.



3. The final scatterplot, shown below, using Random Forest Model with tuned parameters shows that the data points have come closer to the line of regression thus decreasing the Root Mean Square Error.



Reflection

Abalone age prediction is an interesting problem to solve in terms of how we can use its physical measurements to find its age. It’s actually really cool! Bust as interesting it was, we faced a few challenges in the process of solving it like choosing a good, but, easy to implement and understand algorithm. Then, choosing a suitable metric to justify our model like Absolute error or Squared error or R2 score. After hustling with all these hurdles, we have finally created a model that sets a benchmark for others.

Improvement

The final model is a rational one, but there is always a scope of improvement. Other attempts made in the past suggest that although the physical measurements can be used to predict the number of rings (and thus its age) with some accuracy, it is noted that information not present in the dataset (weather patterns and location, food availability) could be used to improve the prediction quality of our model.