Machine Learning Assignment 2 - Analysis of Abalone dataset

Joanna Wojcik

R00169918
Cork Institute of Technology

Contents

Abstract

The purpose of this document is to critically analyze the applicability of linear, decision tree and random forest regression models towards predicting the age of abalone mollusks - a process which is otherwise very time-consuming. Additionally, the document will feature a number of data cleaning and transformation operations as well as research into loss functions, RSS, RMSE, MAE, MSE, and MAPE. The main objective is to achieve the highest accuracy. The accuracy results obtained indicate that all models performed with the optimized Random Forest model performing the best with 87.14%.

1 Introduction

1.1 Analysis motivation and objectives

Abalones are marine snails. Their shells can be used for decoration or jewellery while the body is considered a food source and a delicacy in some regions of the world.[1]. However, the consistent overfishing has lead to the population decline which in turn brought about the commercialized farming where determining the maturity level of a given specimen is critical to the business operations. Accurate defining of the age of a given abalone specimen is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope - a rather time-consuming task that's not suited for farming operation at even the smallest of scales.

Alternatively, physical measurements of the shell and specimen weight can be used for predicting the specimen age, allowing for much faster processing.

Therefore the objective of this analysis is to predict the age of a given specimen, where the age of abalone can be calculated as the number of rings +1.5, as accurately as possible through a number of regression models.

The following models have been evaluated:

1. Linear Regression
2. Lasso Regression
3. Ridge Regression
4. Decision Tree Regression
5. Random Forest Regression

Many papers[2], [3], conference proceedings[4], theses or dissertations[5] have used this dataset in the past. The non exhaustive list has been maintained by at the dataset source at UCI Dataset Repository.

1.2 Dataset Structure

The Abalone dataset has been obtained from UCI Dataset Repository. It consists of 4177 records and 9 columns with each column corresponding to a feature.

Name	Units	Description	Imported Data Type
Sex	M, F, I	M (male) F (female) I (infant)	object
Length	mm	Longest shell measurement	floating point number
Diameter	mm	Perpendicular to length	floating point number
Height	mm	With meat in shell	floating point number
Whole weight	grams	Whole abalone	floating point number
Shucked weight	grams	Weight of meat	floating point number
Viscera weight	grams	Gut weight (after bleeding)	floating point number
Shell weight	grams	After being dried	floating point number
Rings	-	+1.5 gives the age in years	object

Table 1: Dataset Features

2 Research - Error Definition and Accuracy Metrics

Critical to any Machine Learning analysis is the understanding of how to measure the accuracy of a given prediction regardless of the model chosen, therefore, it has been chosen as a focus of the research section. The reasoning behind it is that one may not even be aware of the necessity of data pre-processing in order to attempt to begin building predictive models, however, it will be impossible to asses any model without accuracy measurements and error function definitions.

It should be recognized that there is no omnipotent best accuracy measure to be used when building prediction models[6], therefore all measures selected have been selected in relation to the problem under analysis.

For the purpose of the below equations and measure definitions a residual, or error has been defined by the data point not residing on a given prediction function, where the distance of a given data point from the proposed regression target determining its value.

The simplest definition of an error function denoted by a Greek letter ε(epsilon) is, to sum up all of the residuals as per Formula 1.

$$\epsilon = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{1}$$

However, such a method would be highly inaccurate as any positive residual would be immediately cancelled out, partially or completely, by a negative residual. Therefore following methods to measure model errors are defined instead:

1. Mean Absolute Error (MAE)
2. Mean Squared Error(MSE)
3. Root Mean Square Error (RMSE)
4. Least squares technique (RSS)
5. Mean Absolute Percentage Error (MAPE)

to counter the effects of the bias of the model by taking into the account the likes of the variance of the error.

MAE is defined as the average of the sum of the absolute value difference between predicted and actual values:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \tag{2}$$

Depending on the context of the analysis MAE method's main advantage as well as a disadvantage is giving all the errors the same weight without any penalty applied on account of the magnitude of the error.

Similarly to MAE, the root mean square error can be used to measure the error of the prediction model and some [7] recommend computing both measurements to asses the error of the model.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \tag{3}$$

The main advantage of RMSE, or disadvantage depending on the context of the model under assessment, is applying weight to the magnitude of the errors detected by squaring the difference of y_i from $\hat{y}_i$. RSS sums up squares of any given residuals, therefore, preventing the positive and negative residuals from cancelling out and is defined in Formula 5

There is a direct relationship between RMSE and MSE as the former is the root square of the latter:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = RMSE^2 \tag{4}$$

Standard definition of RSS is:

$$RSS = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{5}$$

From equation 1 we know that the error, epsilon, for a single value can be expressed as:

$$\epsilon = (\hat{y}_i - y_i)^2 \tag{6}$$

therefore RSS can also be defined as the sum of all εvalues:

$$RSS = \sum_{i=1}^{n} \epsilon_i^2 \tag{7}$$

From there another definition indicating the relationship between RMSE and RSS can be induced as per Formula 8.

$$RSS = nRMSE^2 \tag{8}$$

Mean Absolute Percentage Error (MAPE) is defined as the sum of the absolute value of the difference between actual and predicted values multiplied by 100% and subsequently divided by the number of occurrences, as per equation 9.

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{9}$$

Standard MAPE has been selected as the basis for reporting accuracy in simple terms as opposed to any of related measures such as Symmetrical MAPE or Modified MAPE based on the previous research highlighting their issues, such as treating positive and negative sign errors differently[8]. In simple terms, the lower the value of MAPE the better the model is performing.

Finally, the accuracy measure used throughout the analysis is defined as $(100 - MAPE_{\text{mean}})$, where 100 represents 100% accurate prediction.

3 Methodology

3.1 Dataset Cleaning

The abalone dataset has been cleaned of both values impossible for numbers, e.g. NaN or other null value representations as well as of values impossible from the point of view of the domain - values below 0 for measurements, of where the value of the weight component was higher than that of the whole specimen, etc.

The details are as follows:

- row 878 is missing ***viscera weight*** value
- rows 1888 & 3466 are missing ***rings*** value
- row 3093 is missing ***sex*** value
- row 2758 contains negative ***diameter***
- rows 1257 and 3996 contain 0 value for ***height***

Additionally, it is invalid for any specimen to have:

1. higher weight when shucked (stripped of its shell) than whole weight
2. higher viscera weight than the whole weight
3. higher shell weight than the whole weight

and such records can be identified either by a computation method or by drawing a scatter plot dedicated to specific parameters listed above. It has been determined that due to the size of the dataset it is much faster to compute those metrics rather than going with the visual identification method.

- 4 records where ***shucked weight*** is higher than ***whole weight***: 1216, 2627, 2641, 3086
- 1 record where ***diameter*** attribute held a negative value: 2758
- 2 records where ***height*** attribute held a 0 value: 1257, 3996

Row 3996 has been flagged by 2 checks as one containing impossible values: ***height*** and ***shell weight*** higher than ***whole weight***, however, is only contributes towards row-error count once. Given that the grand total of genuinely erroneous records is 11 the records were simply removed from the dataset before further processing. Had the number been more significant the missing data would have had to be substituted.

3.2 Outliers Detection

Outlier detection began with generating a number of boxplots for the given set of features. For the purpose of this analysis, the outliers are defined as a data point lying either $1.5 * IQR$ below first quartile ($Q1$) or $1.5 * IQR$ above third quartile ($Q3$), which is also depicted on the boxplots.

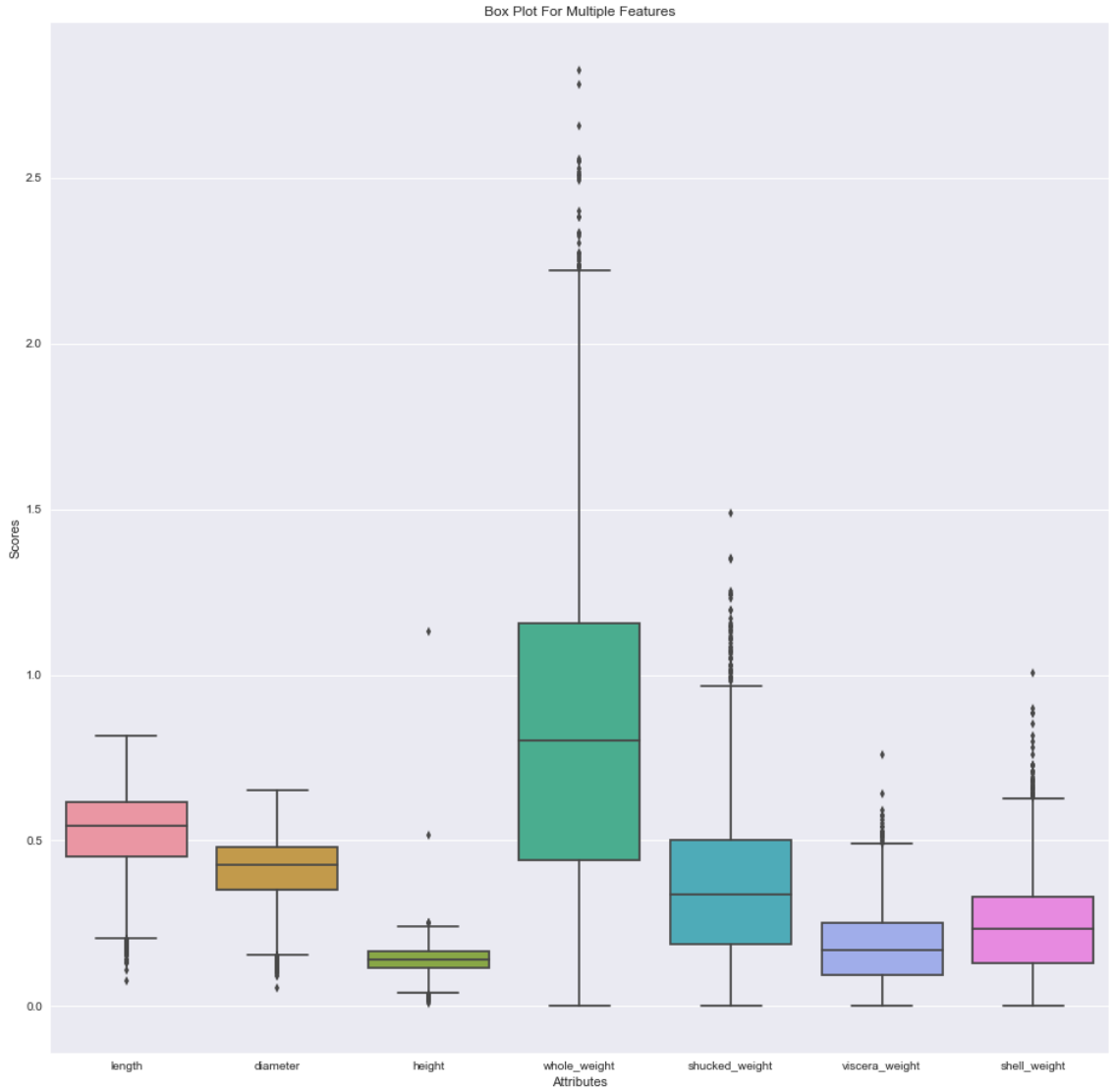

Figure 1: Box Plot of Dataset Attributes

Presence of outliers can be easily observed from Figures 1 and 2, however, it is insufficient a method for determining how many data points are truly affected. Instead, a computational method of determining values contained between upper and lower IQR bounds has been employed and summary can be seen in Table 2. The total came to 403 outlying data points, which constitutes approximately 9.76% of the dataset.

	Rings	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight
Outliers count	278	49	12	6	28	25	16	28

Table 2: Summary of outliers

Again attribute boxplots were generated, Figures 4 and 3 and while graphs indicated the presence of values outside of the IQR for ***shucked weight***, ***viscera weight*** and ***shell weight*** attributes they are so close to the allowed range that it was decided against removing them from the further analysis.

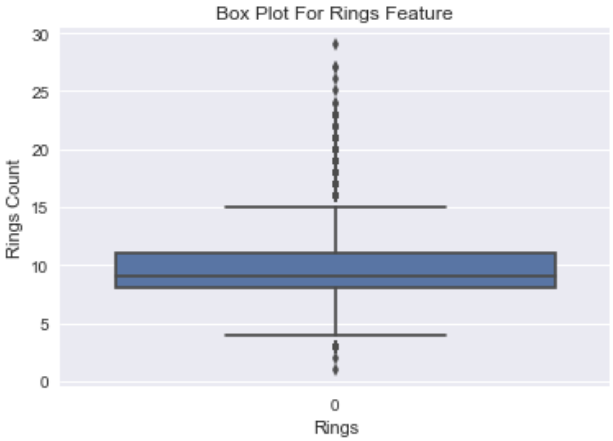

Figure 2: Box Plot of Rings Attribute

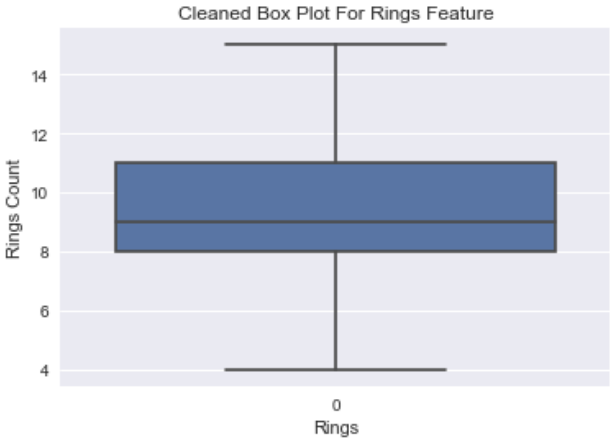

Figure 3: Cleaned Box Plot of Rings Attribute

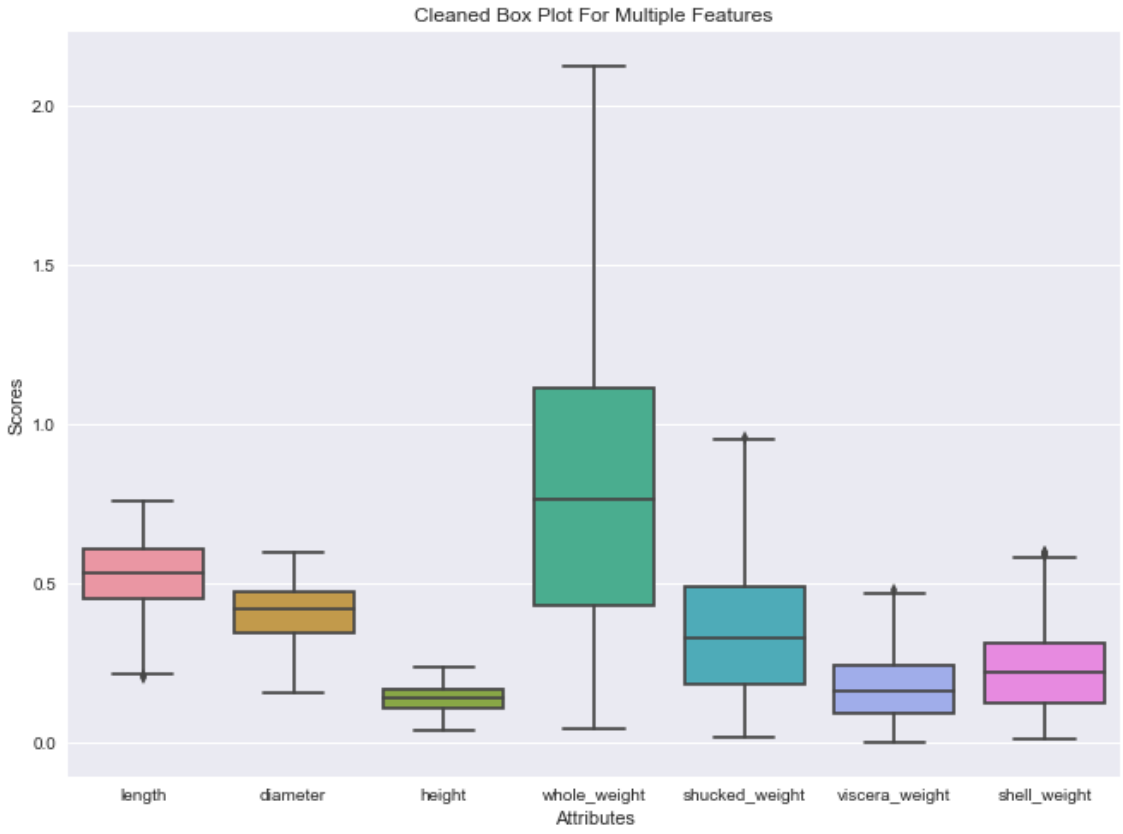

Figure 4: Cleaned Box Plot of Dataset Attributes

3.3 Dataset Balance

The feature balance checking has been carried out using both computational method, with results collected in Table 3 as well via generating histograms for each of the 8 numerical attributes, with results presented on Figures 5 through 12. Furthermore, as a measure of double-confirming the observations Q1, Median and Q3 can be read from the boxplot graphs in Figures 1 and 2.

Starting with ***length***and ***diameter*** attributes, both appear approximately symmetrical with the magnitude of the left-skew considered insignificant towards the prediction results. The ***height*** attribute's mean and median almost matching it is decidedly the most symmetrically distributed attribute in the dataset. From all sources, it can be determined that the distribution of all of the weight attributes follows the same pattern of even distribution. The appearance of right skew of data comes from the fact that the graphs terminate at 0 value, where due to domain reasons none

	Mean	Std Deviation	Minimum	Q1	Median	Q3	Maximum
Length	0.520708	0.111227	0.205	0.45	0.535	0.61	0.76
Diameter	0.404721	0.092058	0.155	0.345	0.42	0.475	0.6
Height	0.137218	0.035185	0.04	0.11	0.14	0.165	0.24
Whole Weight	0.79	0.441938	0.0425	0.433	0.766	1.1155	2.1275
Shucked Weight	0.346427	0.202901	0.017	0.181	0.3265	0.4915	0.96
Viscera Weight	0.17324	0.100376	0.0005	0.0905	0.1635	0.2435	0.478
Shell Weight	0.225437	0.121926	0.013	0.125	0.22	0.3125	0.6
Rings	9.427053	2.328129	4	8	9	11	15

Table 3: Distribution Metrics for Cleaned Dataset

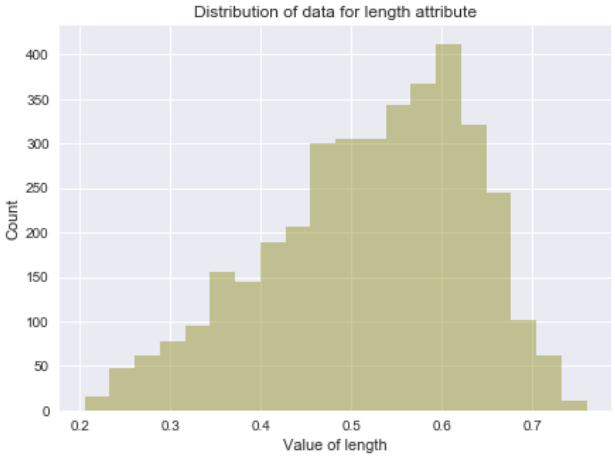

Figure 5: Histogram of Length Attribute

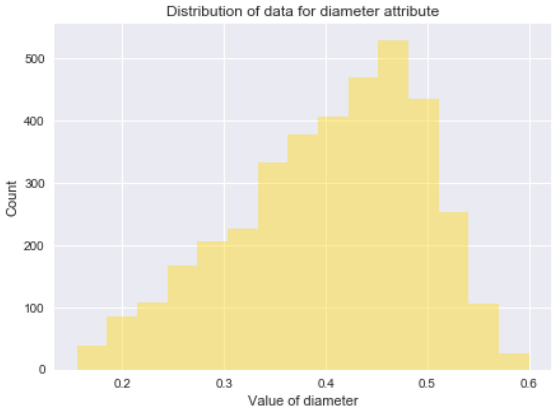

Figure 6: Histogram of Diameter Attribute

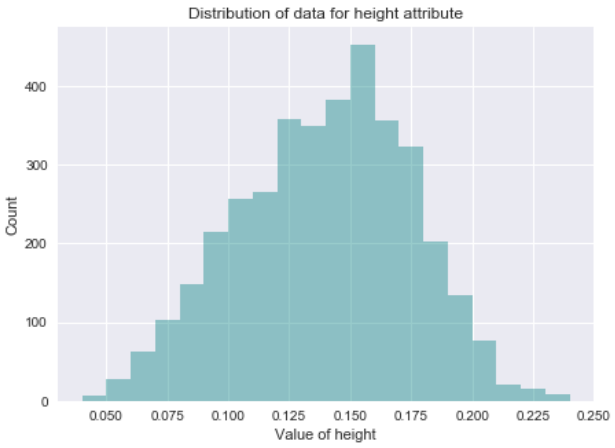

Figure 7: Histogram of Height Attribute

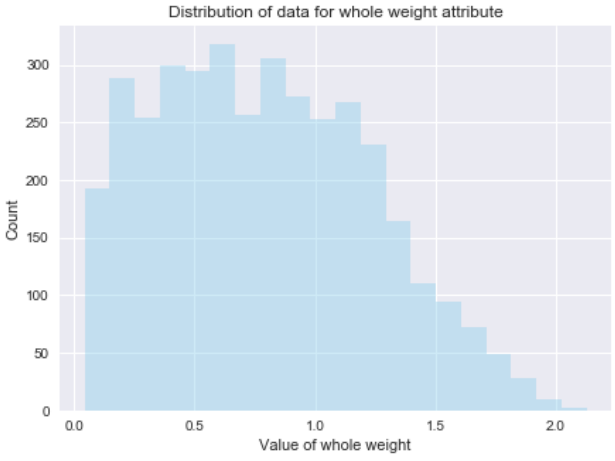

Figure 8: Histogram of Whole Weight Attribute

of the attributes can hold negative values. It should be noted that the visual cues are backed up by the numerical data from Table 3 where the difference between mean and median, where while mean is indeed greater of the two numbers, is too small to conclude skewness of the data.

The final attribute, ***rings***, also appears to be symmetrically distributed, Figure 12 support that observation.

To conclude the features of the dataset are distributed in an approximately symmetrical fashion allowing for reliable prediction of typical results for a given value of a given attribute

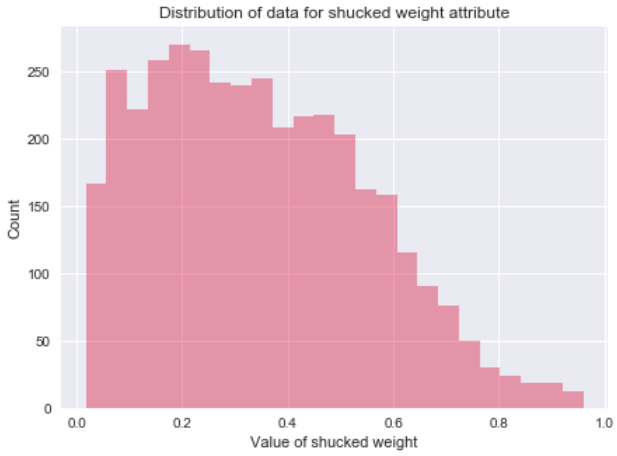

Figure 9: Histogram of Shucked Weight Attribute

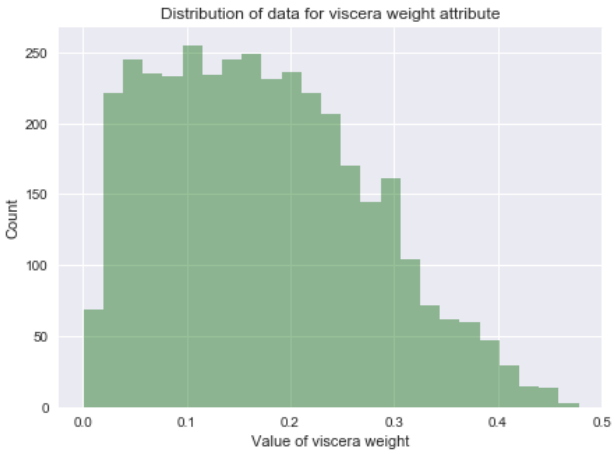

Figure 10: Histogram of Viscera Weight Attribute

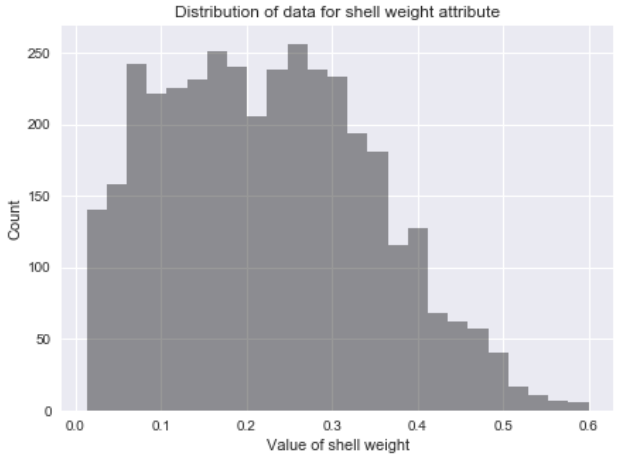

Figure 11: Histogram of Shell Weight Attribute

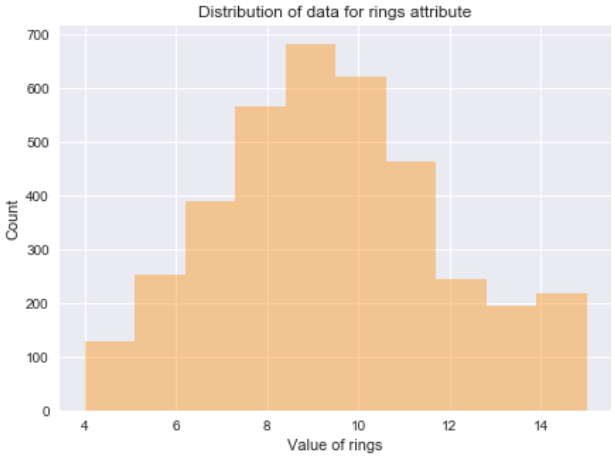

Figure 12: Histogram of Rings Attribute

3.4 Feature Encoding

Only 1 value encoding option has been experimented with and subsequently implemented in the analysis - one hot encoding where non-numerical, non-boolean values are encoded into quantitative data. The method has been utilized to encode ***sex*** attribute values into 3 new, mutually exclusive, features, Female (F), Male (F) and Infant (I), where only 1 of the listed can hold 1, or true value, per record. At the same time, sex feature is removed from the dataset. Unfortunately, one hot encoding comes at a significant disadvantage to the state of the data as newly created F, M and I features are highly correlated with one another.

3.5 Feature Correlation, Transformation and Relationship

The next step in the analysis process, before building out the predictive models there are 3 distinctive steps that will be taken:

1. computing feature correlation metrics - as it will impact on the non-regularized linear regression model
2. transforming existing highly correlated features - to attempt to alleviate the impact of high correlation on prediction target in the non-regularized linear regression model
3. determining if the relationship between features selected can be linear - as otherwise Linear Regression models, regularized or not, could not be used

To identify the features with the most impact towards prediction model feature correlation metrics have been computed, available in Table 4, however, given the number of features, albeit

	Length	Diameter	Height	Whole Weight	Shucked Weight	Viscera Weight	Shell Weight	Rings
Length	1.00	0.99	0.89	0.94	0.92	0.91	0.92	0.58
Diameter	0.99	1.00	0.90	0.94	0.91	0.91	0.93	0.60
Height	0.89	0.90	1.00	0.89	0.85	0.87	0.90	0.61
Whole Weight	0.94	0.94	0.89	1.00	0.97	0.97	0.96	0.56
Shucked Weight	0.92	0.91	0.85	0.97	1.00	0.93	0.90	0.47
Viscera Weight	0.91	0.91	0.87	0.97	0.93	1.00	0.92	0.54
Shell Weight	0.92	0.93	0.90	0.96	0.90	0.92	1.00	0.62
Rings	0.58	0.60	0.61	0.56	0.47	0.54	0.62	1.00

Table 4: Feature correlation metrics prior to any transformation

still very small as far as machine learning datasets can be, it's difficult to immediately determine which features are correlated and by what factor. Therefore as an additional measure, a heat map has been generated, Figure 13. From the figure it can be observed that all features are correlated, however, the ***diameter***, ***length*** and ***height*** are showing correlation values of over 85%, making them the first target for feature transformation. From the correlation heatmap available in Figure 14 it can be surmised that the feature showing the least value of correlation with the target variable is ***shucked weight***, however, it does not provide an answer

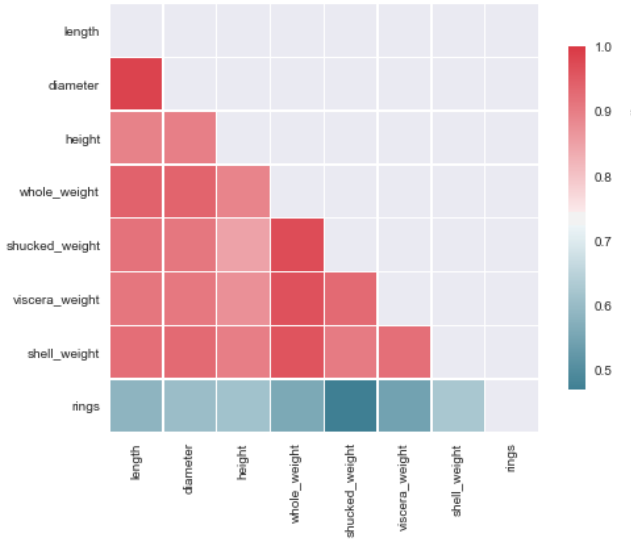

Figure 13: Feature Correlation Heat Map

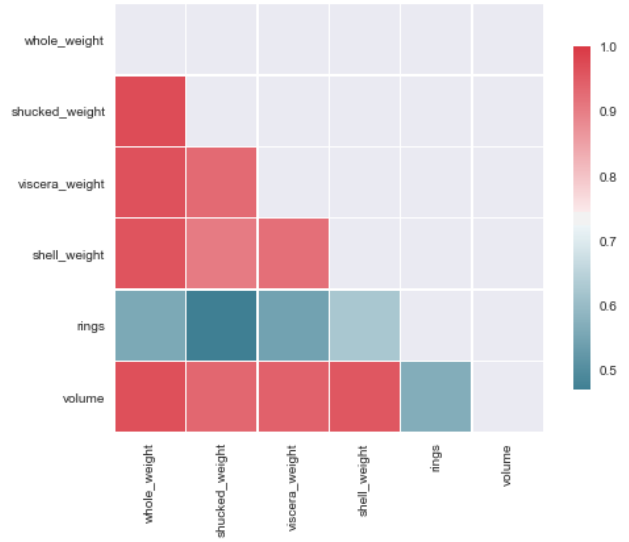

Figure 14: Transformed Feature Correlation Heatmap

The last step is the verify that the Linear Regression model can be used as part of the analysis, by checking if the relationship between the features can be considered linear. To that end a number of scatterplot graph have been generated, Figures 15, 16 and 17.

In Figure 15 while it seems that there is a linear relationship between all features, Figures 16 and 17 depict that relationship in more detail. Additionally, it can be observed that the target regression function will differ by the gender of a given abalone specimen, with male and female adults following the same function while the infants following a significantly different one, which will no doubt affect the results of the overall linear regression model built. While it is beyond the scope of this analysis, it is feasible to assume that if all infant specimens could be pre-filtered then most likely non-regularized linear regression model would perform significantly better than when the whole dataset with all data points is taken into the account.

An alternative approach and precisely the one taken in the due course of this analysis would be to build a model that accommodates such significant discrepancies.

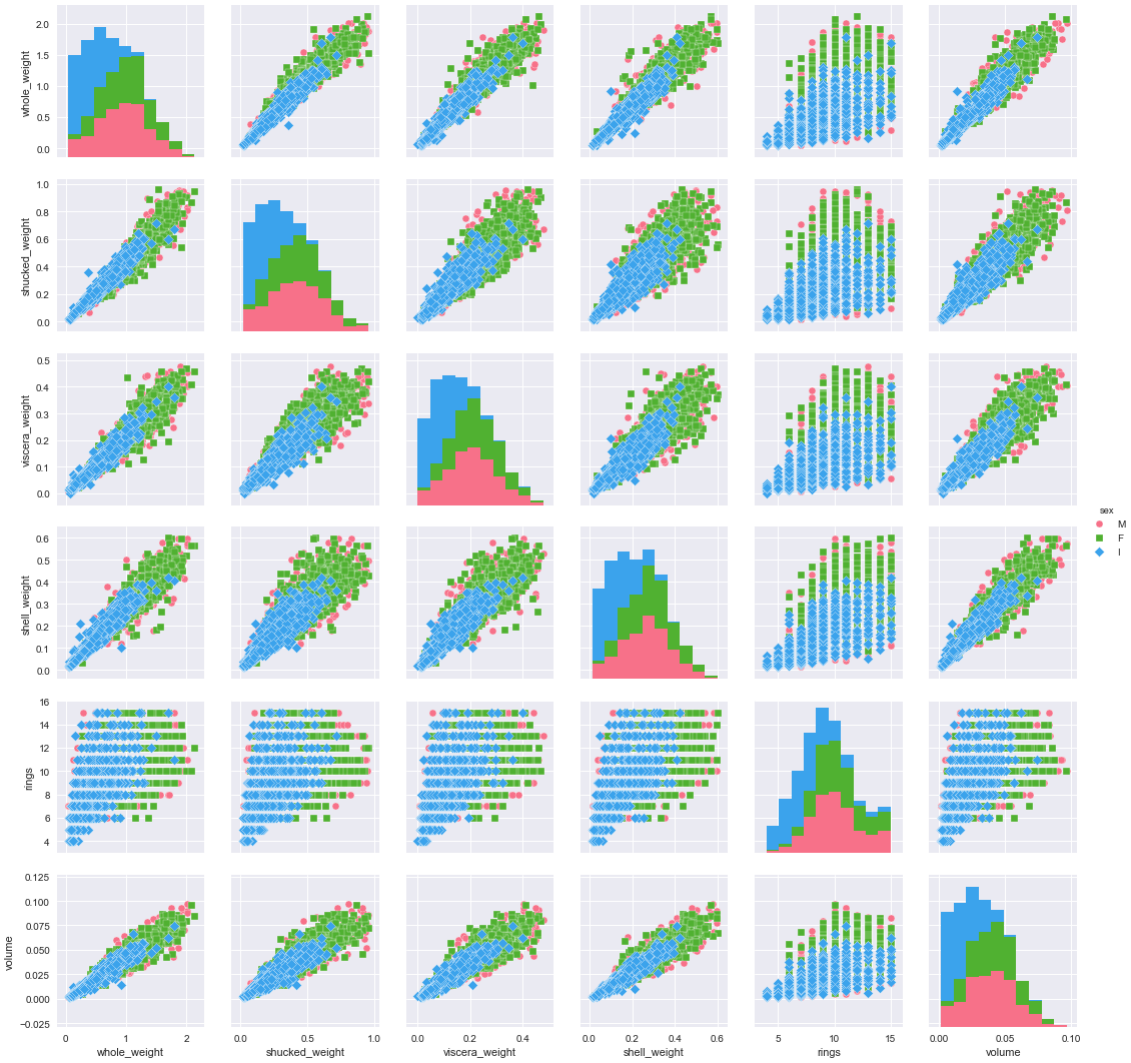

Figure 15: Feature scatterplot

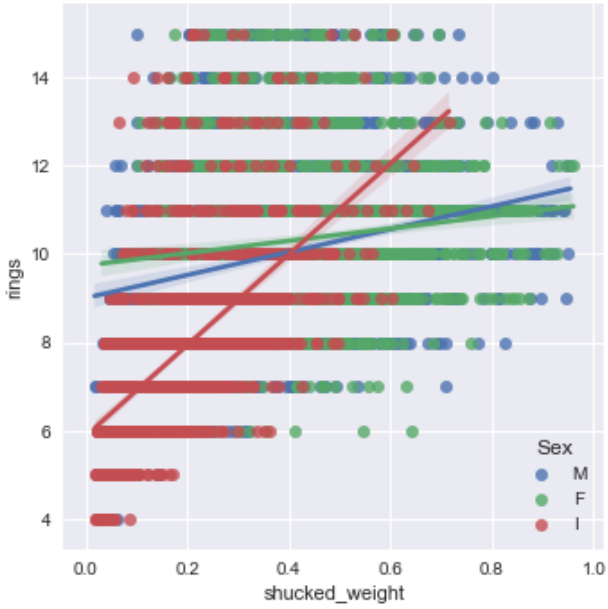

Figure 16: Shucked Weight to Rings relationship

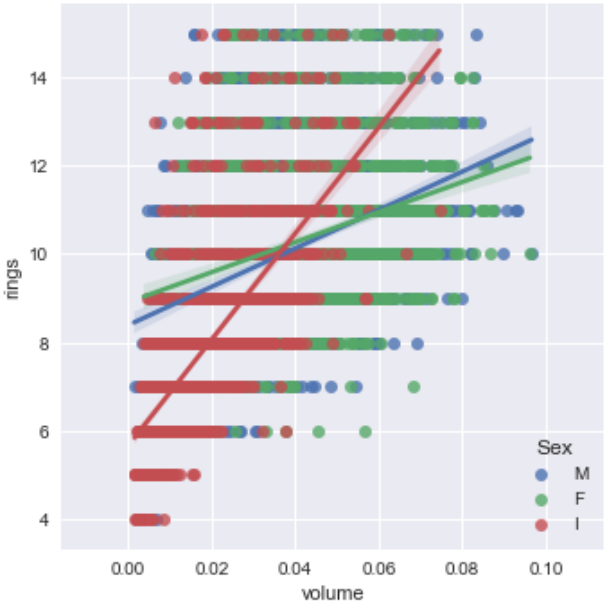

Figure 17: Volume to Rings relationship

3.6 Hyperparameter Optimization

3.6.1 Decision Tree and Random Forest models

Hyper Parameter Name	Search Space	Step
max_features	auto or sqrt	
max_depth	1 to 300	10
min_samples_split	2 to 300	10
min_samples_leaf	2 to 300	10
min_weight_fraction_leaf	0 to 0.5	
criterion	mse or mae or friedman_mse	

Table 5: Hyper Parameter search space

3.6.2 Lasso and Ridge Models

Both Lasso and Ridge regression models are part of the linear regression family and both are regularized.

Regularization is defined as introducing an additional term to the loss function, the prediction model, in order to prevent overfitting of the same α - note that more on regularization and its impact is defined in Section 4.2

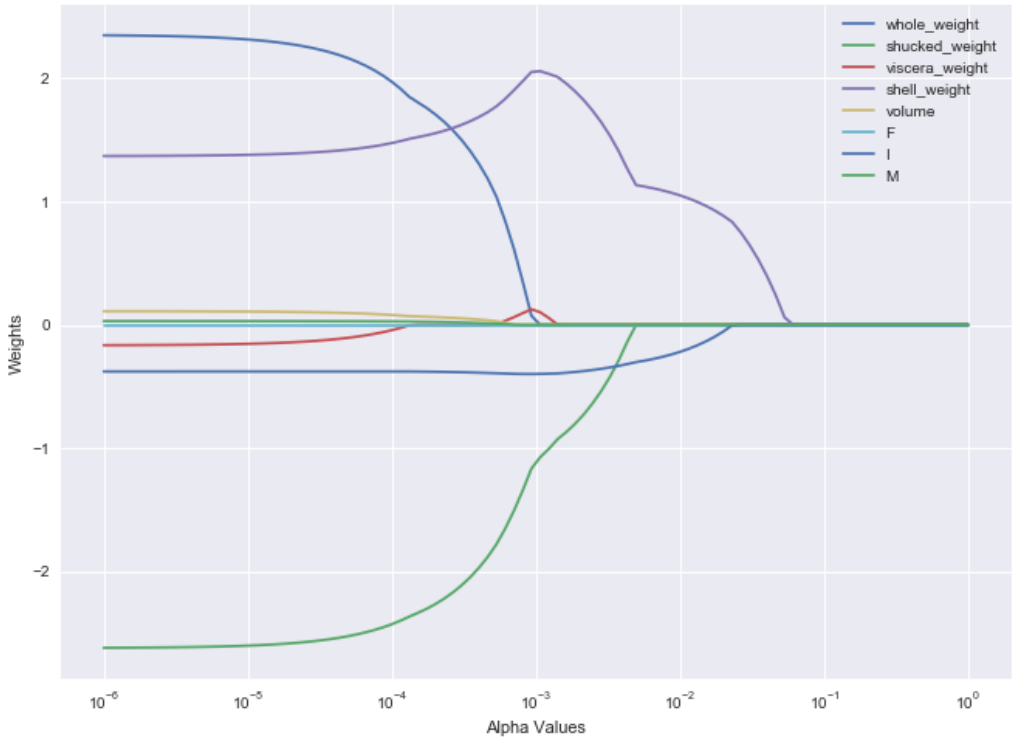

Figure 18: Lasso: Plot of Changes of Weights for Given Alpha Factor

α value ranges used in this analysis are in Table 6, with results visible in Figures 18 and 19.

Since it would be challenging to obtain the best value of α parameter from the graphs, the appropriate ranges defined in Table 6 have been used in LassoCV and RidgeCV for same.

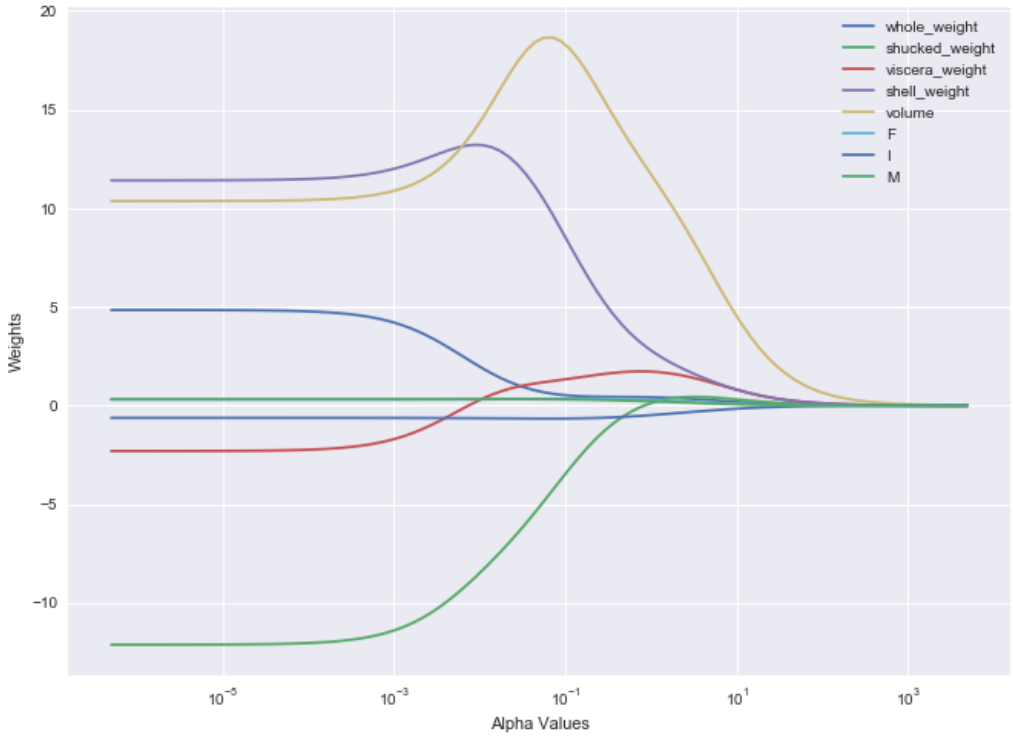

Figure 19: Plot of Changes of Weights for Given Alpha Factor

	Lower Bound	Upper Bound
Lasso	10^{-6}	10^{4}
Ridge	10^{-6}	10^{0}

Table 6: Alpha parameter value ranges

4 Evaluation and Conclusions

4.1 Non Regularized Linear Regression

A Linear Regression, using 5 K-fold cross-validation was performed on the dataset with ***length***, ***height*** and ***diameter*** combined into ***volume***.

It should be explained that K-fold cross-validation is a technique where a given dataset is divided into an X amount of equally sized chunks called folds and then the model is trained iteratively X amount of times. During each iteration 1 fold is reserved for validation set while remaining ones are used as a training set.

Distribution of errors can be observed in Figure 20 with most errors centered around the mean and median of the graph, with computed error mean and median provided in Table 7. However, as suspected the RSS score which could be interpreted as a measure of prediction accuracy - the less the value of residuals the better the prediction - is significant, implying that the model is indeed suffering from the problems outlined before.

However, since it has been observed that the infants abalone follow a different regression function as depicted in Figures 16 and 17 and additional experiment data with Infant specimens removed has been conducted and as per Table 7 it should be noted that the RSS score improved dramatically, providing the best RSS score across all experiments, while MAPE improved by less than 1%.

	Linear Regression	Linear Regression no Infants
MAPE mean (%)	14.42802	13.88254
Accuracy (1-MAPE %)	85.5720	86.1175
Mean absolute error	1.3061	1.3655
Mean squared error	2.8585	3.0017
RMSE	1.6907	1.7326
RSS score	0.4750	0.2918
Error mean	0.0003	-0.0022
Error median	-0.2119	-0.1891

Table 7: Metrics for Non-Regularized Linear Regression

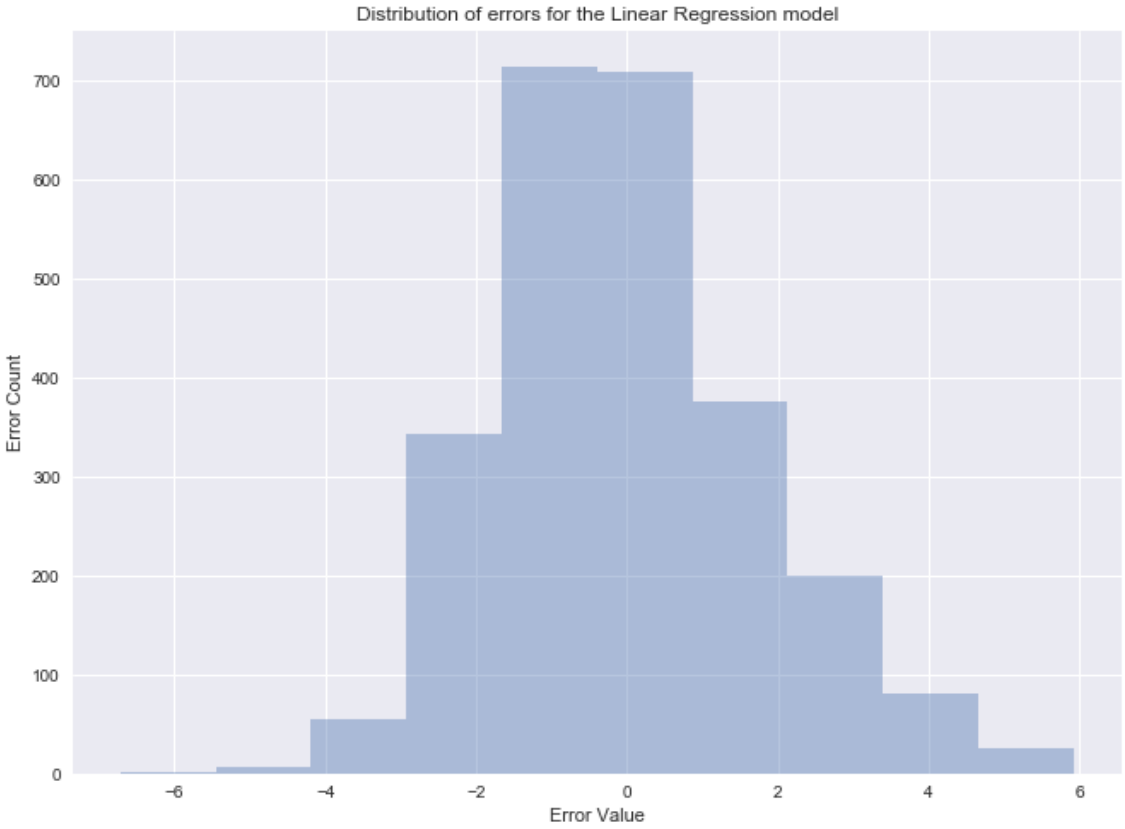

Figure 20: Distribution of errors for the linear regression model using transformed dataset

4.2 Regularized Linear Regression

Both Lasso and Ridge regression models are part of the linear regression family and both are regularized.

Regularization is defined as introducing an additional term to the loss function, the prediction model, in order to prevent overfitting of same. Lasso and Ridge are representatives of L1 and L2 regularization respectively. L1 and L2 techniques both focus on coefficient shrinking, however, Table 8 outlines some key differences between the 2 methods starting with the metrics measured by each.

Both methods perform well in spite the presence of correlated features: L1 by picking the most significant feature and zeroing the coefficient of the related features effectively excluding them from the prediction model, L2 ensures even distribution of the coefficients of the correlated features.

Lasso stands for Least Absolute Shrinkage and it is defined as $RSS + \alpha *$ (sum of absolute value of weights) or:

$$\sum_{i=1}^{n}(\hat{y_i} - y_i)^2 + \alpha \sum_{j=0}^{n}|w_j| \tag{10}$$

L1	L2
Sum of weight	Sum of square of weights
Sparse outputs	Non-sparse output
Built-in feature selection	No feature selection
High sparsity for highly correlated features	Even coefficient distribution for highly correlated features
Ability to interpret models with large feature sets	Main use case is preventing overfitting

Table 8: Summary of L1 vs L2 Regularization Techniques

where α refers to the factor of the penalty applied to a feature and w_i refers to the weight of the feature[9].

Ridge regression works by adding a penalty factor to square of the magnitude of coefficients[10]. It can be represented as:

$$\sum_{i=1}^{n}(\hat{y_i} - y_i)^2 + \alpha \sum_{j=0}^{n} w_j^2 \tag{11}$$

From the same definition it can be deducted that impact of α will be the same is both methods, namely:

1. $\alpha = 0$, is the same as linear regression,
2. $\alpha = \infty$, the coefficients will be zero due to the infinite weighting on square coefficients anything less than 0 will make the objective infinite
3. $0 < \alpha < \infty$, the coefficients will be found between 0 and the ones obtained from simple linear regression

4.2.1 Lasso Regression and Ridge Regression

Both Lasso and Ridge regression models are an improvement on the non regularized Linear Regression discussed in Section 4.1.

	Lasso	Ridge
MAPE mean (%)	14.06769	14.06828
Accuracy (%)	85.9323	85.1575
Mean absolute error	1.2689	1.3261
Mean squared error	2.7304	2.9001
RMSE	1.6524	1.7030
RSS score	0.4746	0.4743
Error mean	0.0267	0.0393
Error median	-0.1573	-0.1609

Table 9: Metrics for Regularized Linear Regression

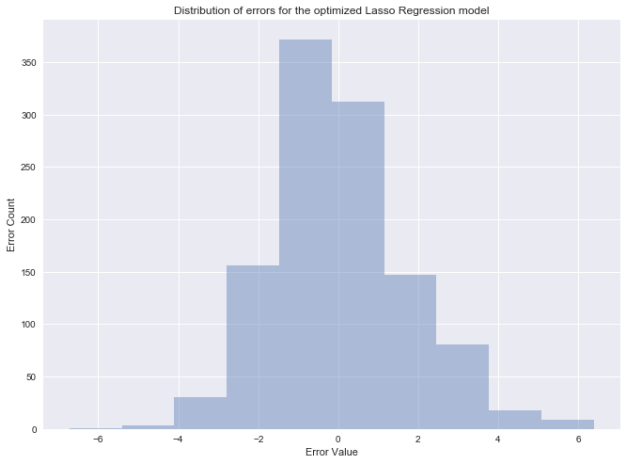

Figure 21: Distribution of errors for the Lasso Regression

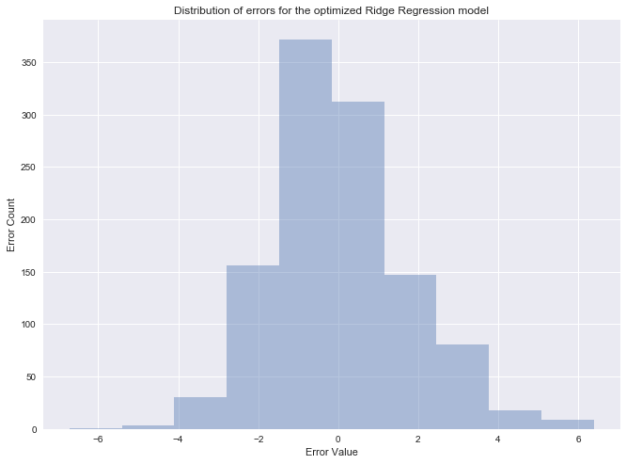

Figure 22: Distribution of errors for the Ridge Regression

4.3 Decision Tree and Random Forest Regression

Decision Tree and Random Forest models have been evaluated with 2 iterations per model, one where no hyperparameters were optimized and one where best parameters from ranges defined in Section 3.6.1 have been identified via Random Grid Search.

From Table 10 it can be observed that the Decision Tree model when unoptimized performed worse than even non-regularized Linear Regression model. It indicated that significant improvements can be made post optimization, with the same table providing data about approximately 5% improvement.

As for the distribution of the errors, it can be observed from Figure 23 depicting the baseline for further model optimization with no additional parameters specified that the range of errors is greater than that of optimized decision trees, Figure 24. From the graphs as well as from computed error mean and median available in Table 10 it can be concluded that the distribution is mostly symmetrical with little to no skew observed.

	Baseline Decision Tree	Optimized Decision Tree	Baseline Random Forest	Optimized Random Forest
MAPE mean (%)	18.0523	13.89625	13.98651	12.80984
Accuracy (%)	81.9652	86.1038	86.0135	87.1423
Mean absolute error	1.6766	1.2833	1.2927	1.2227
Mean squared error	5.1387	2.8217	2.8536	2.6918
RMSE	2.2669	1.6798	1.6893	1.6407
RSS score	1.0000	0.5280	0.9103	0.5490
Error mean	-0.0481	-0.0030	-0.0209	0.2269
Error median	0.0000	-0.1164	-0.2000	0.0000

Table 10: Decision Tree and Random Forest metrics

The Random Forest model performs extremely well even without hyperparameter optimization, on par with the best result so far obtained by optimizing Decision Tree, details in Table 10. After optimizing the hyperparameters as per ranges in Section 3.6.1, the best MAPE results throughout the models.

The distribution of errors, depicted in Figures 25 and 26, as well as metrics Table 10, similarly as in the case of Decision Tree model that the errors in an almost symmetrical manner. It should be noted, however, that the unoptimized model does make some higher value errors, as per graph ranges, than the optimized model. This has been anticipated and considered normal as the range of errors indicative of the MAPE and accuracy metrics.

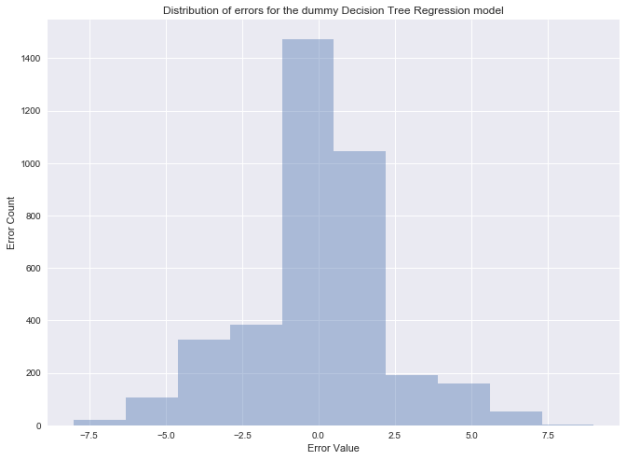

Figure 23: Distribution of Errors in Unoptimized Decision Tree

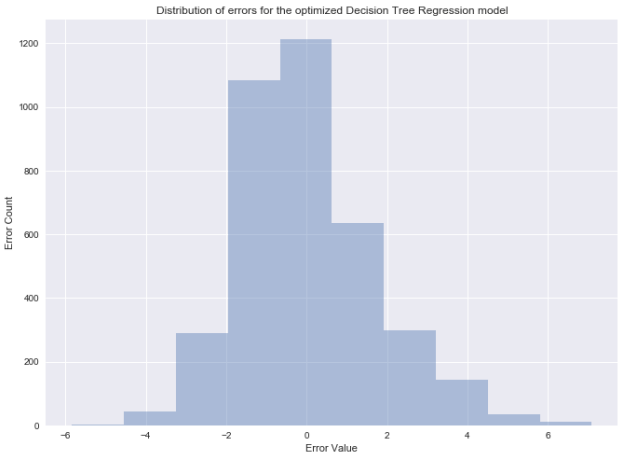

Figure 24: Distribution of Errors in Optimized Decision Tree

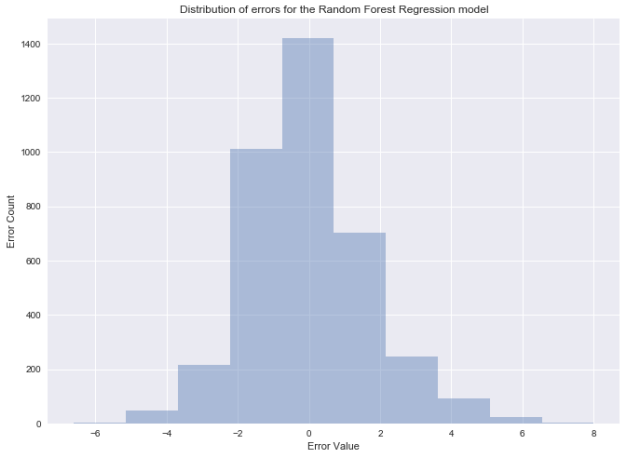

Figure 25: Distribution of Errors in Unoptimized Random Forest

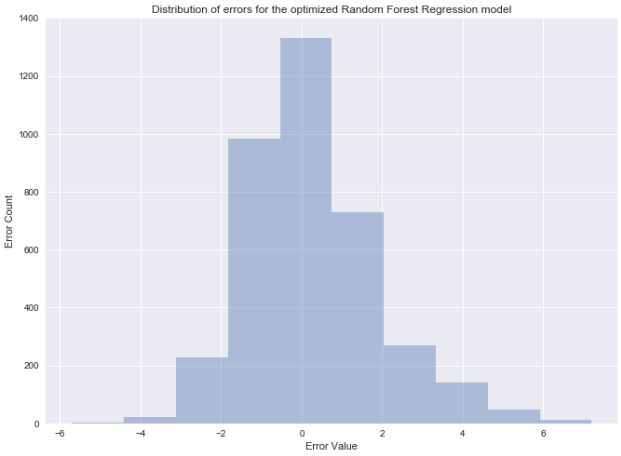

Figure 26: Distribution of Errors in Optimized Random Forest

4.4 Conclusions

Out of all the models constructed the optimized Random Forest performed the best achieving the lowest MAPE and conversely the highest accuracy, as MAPE and accuracy metrics are in direct relationship with one another - the lower the MAPE the higher the accuracy.

	MAPE
Optimized Random Forest	12.85772
Linear Regression no Infants	13.88254
Optimized Decision Tree	13.89625
Baseline Random Forest	13.98651
Lasso	14.06769
Linear Regression	14.42802
Ridge	14.84252
Baseline Decision Tree	18.03483

Table 11: Summary of MAPE metrics indicating accuracy per model

4.5 Future Work

Only in one instance has a prediction model been designed without infant specimens included. It is probable, that it would significantly improve scores obtained from Lasso and Ridge linear regression models. Similarly the effect of choosing a linear model encompassing both L1 and L2 regularization such as Elastic Net method.

References

[1] https://en.wikipedia.org/wiki/Abalone.

[2] F. Gasir, Z. Bandar, and K. Crockett, "An architecture for constructing fuzzy regression tree forests using opt-ainet," in *2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011)*, pp. 283–289, June 2011.

[3] Y. Chen, B. Song, and Y. Ren, "Perpendicular bisector constraint on artificial neural network," in *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 314–321, Nov 2017.

[4] http://users.monash.edu/~dld/Publications/2003/Tan+Dowe2003_MMLDecisionGraphs.pdf.

[5] http://www.cs.bilkent.edu.tr/tech-reports/2000/BU-CE-0009.pdf.

[6] S. Makridakis, "Accuracy measures: theoretical and practical concerns," *International Journal of Forecasting*, vol. 9, no. 4, pp. 527 – 529, 1993.

[7] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, 2014.

[8] P. Goodwin and R. Lawton, "On the asymmetry of the symmetric mape," *International Journal of Forecasting*, vol. 15, no. 4, pp. 405 – 408, 1999.

[9] T. Hastie, R. Tibsharani, and J. Friedman, "Springer Series in Statistics The Elements of," *The Mathematical Intelligencer*, vol. 27, no. 2, pp. 68–69, 2009.

[10] T. Hastie, R. Tibsharani, and J. Friedman, "Springer Series in Statistics The Elements of," *The Mathematical Intelligencer*, vol. 27, no. 2, pp. 61–68, 2009.

5 Appendix

	Linear Regression	Linear Regression without Infants	Lasso	Ridge
MAPE mean (%)	14.42802	13.88254	14.06769	14.06828
Accuracy (%)	85.5720	86.1175	85.9323	85.1575
Mean absolute error	1.3061	1.3655	1.2689	1.3261
Standard deviation of the error	1.6909	1.7329	1.6529	1.7033
Mean squared error	2.8585	3.0017	2.7304	2.9001
RMSE	1.6907	1.7326	1.6524	1.7030
RSS score	0.4750	0.2918	0.4746	0.4743
Error mean	0.0003	-0.0022	0.0267	0.0393
Error median	-0.2119	-0.1891	-0.1573	-0.1609

Table 12: Complete set of metrics computed for Linear Regression models

	Baseline Decision Tree	Optimized Decision Tree	Baseline Random Forest	Optimized Random Forest
MAPE mean (%)	18.0523	13.89625	13.98651	12.80984
Accuracy (%)	81.9652	86.1038	86.0135	87.1423
Mean absolute error	1.6766	1.2833	1.2927	1.2227
Standard deviation of the error	2.2667	1.6800	1.6893	1.6251
Mean squared error	5.1387	2.8217	2.8536	2.6918
RMSE	2.2669	1.6798	1.6893	1.6407
RSS score	1.0000	0.5280	0.9103	0.5490
Error mean	-0.0481	-0.0030	-0.0209	0.2269
Error median	0.0000	-0.1164	-0.2000	0.0000

Table 13: Complete set of metrics computed for Decision Tree and Random Forest models

List of Tables

List of Figures