Data Mining Coursework: Predicting Age as a Target Variable in the Abalone Dataset

39612295
School of Computing and Communications
MSc. Data Science
Lancaster University Leipzig
Leipzig, Germany

***Abstract*—This document aims to dissect, explore, and evaluate the Abalone dataset and different patterns throughout to understand, evaluate, and enhance a Neural Network model that attempts to predict the age of Abalone. Different techniques are explored in order to derive the optimal pre-processing methods, hyper-parameters, and model structure as an attempt to provide a meaningful model for prediction accuracy. An Extreme Gradient Boost model is then derived as a second model to compare the performances and provide a fair scrutinization of the dataset.**
NOTE:
In the beginning of each file, the first line of code after importing all the necessary libraries is where the dataset is imported and loaded. To replace or modify this dataset, replace the "abalone.head" with the desired file name (the relative path), and it should run without any issues if the dataset is in the dame folder. The codes for the two models are in the two files model1.py and model2.py respectively. These codes run on the Linux operating system using the following commands "pip3 install –upgrade pandas" , "pip3 install numpy pandas scikit-learn" , "pip3 install tensorflow" , "pip3 install seaborn" , "pip3 install xgboost", and then python3 -filename-.py

I. Introduction

Abalones are a type of sea snail or molluscs that are an endangered species typically found in cold coastal waters. The meat is considered to be a highly nutritious food, while the shells are commonly used for decorative purposes, sought out after and enjoyed in multiple parts of the world, and thus, are in high demand. The age of the abalone is used to derive the price of the abalone; however, determining the age is a time-consuming and tedious process. The shells must be polished and stained, and then examined in a lab in order to obtain the number of rings under a microscope. Because of this, it is of significance to farmers and consumers to achieve a way to predict the age of abalones to ensure the correct price, as well as the environmental importance of ensuring this endangered species is protected to the best of our abilities [1].

Due to this tedious and complex process, it would seem counterproductive to develop a predictive algorithm which utilizes the rings as one of the predictive variables. As the stakeholders involved would benefit from a model that is able to predict the age without the number of rings, the choice of eliminating the rings from the predictive variables is almost imperative. Thus, the rings will not be considered in predictions. A neural network model was chosen to complete the task of predicting the age of abalones using the abalone dataset. Since this is a regression task, and accuracy is somewhat difficult to assess here, the metrics that will be measured and used to assess model performance and accuracy are R^2, MAE and MSE. R^2 is a measurement of how much of the variance is captured by the data in the predictions. The closer the value to 1, the better your model can predict; hence, this value is used to describe the 'accuracy' of the model. The MAE is an average of the absolute difference between the predicted and actual values by the model, and is less sensitive to outliers, making it a suitable measurement for this dataset. The MSE is the average squared difference between the predicted and actual value by the model, and thus, since it is a square, it accentuates larger errors [2]. Different pre-processing techniques and applications will also be applied, assessed and compared to find and construct the optimal predictive model for this task. Other feature extraction techniques will also be examined, as well as hyper-parameter tuning.

December 13, 2024

A. Methodology

1) Understanding the Dataset: The dataset contains one categorical variable, Sex, and eight numerical features: length, diameter, and height. These are all of type float and in millimeters, as well as whole weight, shucked weight, viscera weight, shell weight, all in grams and of type float. Lastly, there is a rings feature which is an integer number in the range 1-29. The age of the abalone can be calculated using from the rings value, where the age is equal to the number of rings plus 1.5. The sex of abalone contains three labels: M, F, and I, representing "Male", "Female", and "Infant". The length feature which is a continuous value, is equal to the abalone's longest shell measurement. The diameter and height which are also continuous features, refer to the measurement that is perpendicular to the length, and the length of the abalone with meat in the shell, respectively. The whole, shucked, viscera, and shell weights are also continuous variables where the whole weight refers to the entire weight of the abalone. The shucked weight refers to the weight of the meat, the viscera weight refers to the gut weight after bleeding, and the shell weight refers to the weight after being dried. If attempting to train a model to predict the age of an abalone based on its features, it is essential to test and understand which features contribute in determining and predicting the age.

2) Exploring the Variables and Their Relationships: After creating scatter-plots of the variables against the target variable (Age), and creating a heat-map of the Pearson's coefficients between the variables, it is noticeable that there is not any significantly linear relationships between any of the variables with the target feature, as the highest Pearson coefficient value is 0.63, for the shell weight. However, it can be observed that the variables have very high linear dependencies with one another, indicating that the variables have a high correlation with each other, but not with the target variable. The height variable having the lowest correlation with the other variables, where it's lowest coefficient is with the shucked weight, being 0.77, which is still relatively high. In attempt to further understand the dataset, simple models for Random Forests and LASSO regression were constructed in order to explore the Gini Importance for each feature and the LASSO coefficients, to understand which features were contributing the most in making predictions.

	1st Feature	2nd Feature	Importance
Gini Import.	Shell wt	Shucked wt	0.497
LASSO Coeffs.	Shucked wt	Whole wt	-3.558

The table above highlights the most important features according to the corresponding measurements. Since the Gini Importance accounts better for nonlinear relationships and splits, while LASSO focuses more on linear relationships, it is reasonable to assume that the weight features, particularly the shell, shucked, and whole weights, are the most contributing variables to the age. The model scores for the simple random forest and LASSO regression models were 0.58 and 0.52 respectively.

After this, boxplots were plotted for each variable against the target in order to investigate the outliers, if any present. As displayed by the plots, there are a significant amount of outliers, with high amounts of outliers in the lower ranges for the length, diameter, and height variables. The length and diameter variables are pretty symmetrical with no apparent skewness. The height variable has a significantly lower inter-quartile range and but obtains similar symmetry to the length and diameter. The weight variables exhibit very similar patterns, indicating that there is a high correlation between these variables, as emphasized in the heat-map. The four features have notably great amounts of outliers in the higher ranges. This great amount of anomalies signifies the necessity of investigating these outliers further.

The amount of outliers present denotes that it is both in-feasible and irrational to remove the outliers or perform mean imputation. For the purpose of focusing on age prediction, there are too many outliers present, and consistent within the feature groups–either the dimensions or the weights–to neglect them and their contributions. This prompted the investigation of what would be a logical and fair way to handle them, or at least mitigate the potential noise that may arise? The clearly erroneous samples could be removed, but the groups allow for a more specialized way of being handled.

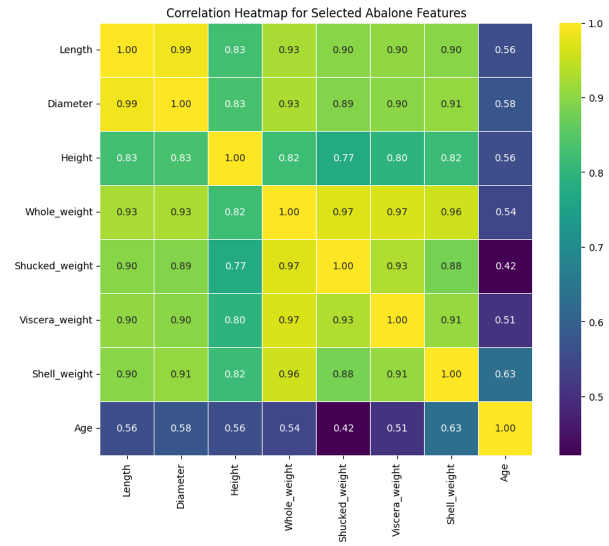

Fig. 1. Heatmap With Selected Abalone Features

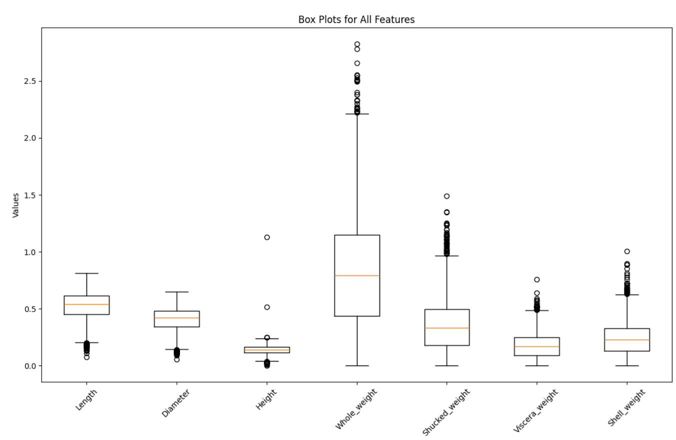

Fig. 2. Boxplots Of Selected Abalone Features

One technique is to adjust or adapt the interquartile range to 'accept' more data samples by having more flexible and wider boundaries, and then omit the sample outside of the bounds. Looking at the diagram, it is evident that are some outliers which are not clustered with the others. Looking at height, it is very clear that some samples could be causing some unnecessary noise, at around 0.5 and 1.2. There are similar visible outliers for all of the weight features, especially the whole weight. Thus, one way to handle this is to remove the outliers outside of the ”outlier” range; this helps both with the noise, and with ensuring that necessary patterns that may come from the outliers will not be affected or negated. In order to extend or increase the boundaries of the IQR, the IQR can be modified. The IQR is calculated as:

$$IQR = Q3 - Q1$$

The lower bound and upper bounds are found through the following equations, where $k = 1.5$:

$$LowerBound = Q1 - k * IQR$$

$$UpperBound = Q3 + k * IQR$$

Thus, in order to extend the boundaries, k needs to be a number greater than 1.5 [3]. K was changed to the values 2, 2.5, 3, 3.5, and 4, where the lower and upper bounds were computed, and the sampled lying outside of these boundaries were discarded. The simple, not yet tuned or adjusted neural network was trained on these datasets, where the highest amount of sampled discarded was 36, meaning the sample sizes ranged from 3964 to 4000 in this measurement.

	k = 2.0	k = 2.5	k = 3.0	k = 3.5	k = 4.0
Samples	3964	3989	3994	3997	4000
R^2	0.57	0.57	0.57	0.63	0.66
MAE	1.49	1.57	1.57	1.43	1.37
MSE	4.96	5.20	5.20	4.35	3.96

After completing this, shockingly, it was noticeable that the model performs the best and has the lowest error when none of the outliers are discarded or modified. The outliers thus can be assumed to provide patterns that are necessary for the predictions. Therefore, none of the outliers will be removed or modified in the dataset.

B. Model Induction

The process of model selection, particularly within a task to predict values on a dataset with complex relationships, is a critical choice to ensure that the model successfully captures these patterns and relationships. The decision here can come from different factors; would a statistical model make more sense here, or an algorithmic model? Would an eager model be more suitable for this task, or a lazy one? Since the choice to remove the rings from the predictive variables was made, a robust model is needed to capture the nonlinear and assumably redundant patterns in the data, as many features appear similarly distributed, and are highly linearly correlated. Thus, a statistical model might be decent after tuning and adjusting, but statistical models assume linearity and independence in the data distribution. An eager model seems like a more suitable choice here as well, as it makes sense for the model to use all of the data for training and provide a generalization of it. Also, since this is a regression task, it is sensible to choose a model that performs well for regression tasks. Additionally, considering there is a significant amount of outliers, the model selected needs to be fairly robust to outliers. Considering these factors, the decision to use a neural network model in attempt to predict the age of abalone seemed sensible and reasonable. Running a basic neural network on the data, with no hyper-tuning also gave a model score of 0.67 for the R^2, and a mean average (MAE) value of 1.41, which gave significantly better metrics than both the LASSO regression and the Random Forest models.

1) Pre-processing and Transforming Data Alongside Feature Engineering: The dataset contains a categorical variable, which means one-hot encoding will be used here to transform this feature into a numerical one. Also, the dataset was provided without an ”Age” feature. Thus, this column was added to ensure its availability as the target variable. This was carried out by adding 1.5 to the value of the rings for each row to get the corresponding age. A check was performed to ensure there were no missing or NA values present in the dataset. Lastly, after examining the dataset, it was noted that the data is in different units (mm, g), as well as counts for the rings, and thus, for the age. Therefore, the data, with the exceptions of the Sex, and Age features, were standardized before performing any data modeling. This was done to ensure features do not obtain any advantage of contributing unnecessarily more and thus contributing falsely in prediction making.

Principal Component Analysis (PCA) and Autoencoders will also be investigated for the purposes of dimensionality reduction and feature extraction as well as feature transformations. Feature transformations may aid the model in performing better because certain features which may be redundant or highly correlated can be combined in order to provide less ”confusion” to the model. Similarly, features that are exponential, logarithmic, or even have a skewed distribution can be transformed in order to better feed the model a specific pattern which might be necessary for training. Since the data has relatively high positive linear correlations between each other, it is sensible to apply PCA, considering that the features are similar also [4]. Autoencoders are another type of data compression that can be applied here. They are used for feature extraction and dimensionality reduction, which is sensible to attempt here considering that the features are highly correlated, meaning the model could be learning unnecessary redundant patterns, providing more confusion and difficulty in choice [5]. Surprisingly however, applying PCA after standardizing the data from the X variables, with both 0.95 and 0.99 variance levels, significantly decreased the model performance. Also, after adding autoencoders and tuning them, the performance of the model also decreased. For this reason, neither were used, and this signifies that the dimensionality is important for predictions, at least for neural networks. The dimensionality helps the model learn different patterns, even between features that might have high correlations, but could be significant in relationship to other features. The dataset containing eight features is also relatively low for a neural network, denoting that the dimensionality is not an issue here, and does not need to be reduced.

2) Hyper-parameter Tuning and Improving the Model: Selecting, enhancing and tailoring the model to suit the data as well as the needed outcome is extremely pivotal in the success of the feature prediction. The first attempt at improving the model was adjusting the hyper-parameters. A seed was set to 42, to ensure reproducibility and so that the same results would be yielded each time. It was noticeable that the model converged in the first 100 epochs. Also, the batch size was tested for 16, 32, and 64, and 16 yielded the lowest error and R^2 values and thus, the number of epochs was set to 100 and the batch size to 16. When defining the model, for the activation function, since the predicted variable is age–which is always a positive number–it is perceptible to use the ReLU or LeakyReLU activation functions. The model performed slightly better with the ReLU activation function–a lower MSE by 0.04 and a lower MAE by 0.03–and thus, as per convention and these metrics, ReLU was chosen as the activation function. It is also less computationally expensive and efficient for computation. [6] For the model optimizer, SGD and Adam were both tested, as these are the most suitable for robust regression tasks. It was apparent that Adam performed better, and thus, Adam was chosen as the optimizer.

The loss function was tested for different functions, including MSE, MAE, Minkowski's loss, which is a generalization of MSE and MAE, however, MAE was chosen as it is the least sensitive to outliers and is the most robust. Also, regularization techniques to penalize larger coefficients in attempt to prohibit the model from overfitting, and learning the "noise" of the outliers could also provide some benefit to the model. Thus, an L1 regularizer was added to the first hidden layer of the neural network, and served benefit to the model, even slightly. It was very interesting to find that the model was tuned using different learning rates and learning schedulers, as well as experimenting with different amounts of neurons in the layers as well as the amount of hidden layers. Three hidden layers with the neurons halved at each layer showed to yield the best results. However, when the training and validation losses were inspected, it was apparent that the model was overfitting. This was the same case when the batch size was decreased, and it also took almost double the amount of epochs to converge. Hence, even after fine-tuning the model, the model that was untuned with these metrics, produced better metrics for the model, as well as better training and validation loss results. Lastly, it was noticed that manually standardizing the data, instead of using the StandardScaler() function worked better, and thus, the data was standardized manually.

It was noticed from the box-plots of the Sex that the Male and Female attributes were almost identical. The Infant category was directly below, meaning the 25th quartile of the Male and Female abalones was the 75th quartile for the Infants.

This highlighted a clear distinction, and separation in the 2 'groups'. In attempt to further investigate different ideas on how the data could be handled, separated, or pre-processed, a model where the Sex divided the data into two subsets, and two smaller models were trained on each data subset was investigated. It was as follows: the data was split into 2 subsets, and a neural network was trained on each subset,

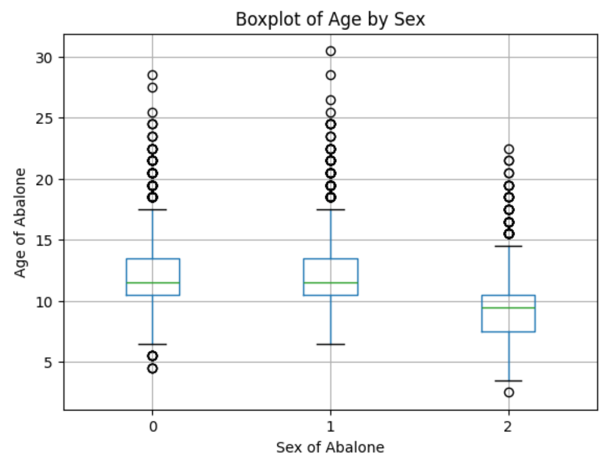

Fig. 3. Boxplot of Age by Sex Feature in Abalones

and predictions were made based on the combining of both models–similar to a meta-model. This was done in order to seek for an optimized model [7]. The results for this, however, were notably poor, as the results for the Female-Male model yielded approximately an R^2 of 0.37 and 0.65 for the Infant model, despite the Infant model having significantly less data in the subset, with less than 1000 samples. It can be inferred that since the characteristics for female and male overlap heavily, it is very difficult for the model to distinguish between the two, and finds difficulty in detecting patterns related to the Sex. Thus, this idea was investigated, but not considered for the model due to its incompetent performance and aid. At first, it was suspected that the model performed better without the Sex as a predictive variable, but the model performed slightly better with it compared to without it, and hence, it is used. This indicates that the neural network does in fact favor feature increased dimensionality for possible patterns and correlations with other variables, in addition to everything else, over reduced dimensionality. This is not counterintuitive considering how complex and powerful neural networks are.

Another idea is to perform a k-means clustering technique on the data in order to observe the impact of clustering on a model. This is also an intuitive approach, considering that the variables are very groups similarly–where abalones belonging to a given age range have very similar dimension and weight ranges. In order to find the optimal number of clusters, the elbow method was used.

From the graph it can be deduced that the "elbow" is present at 3, indicating that the optimal amount of clusters is 3 [8]. Thus, k-means clustering was investigated here as a pre-processing step. However, this technique was tested two ways. The first way, was that the algorithm ran through the dataset, and divided the data based on 3 clusters that it found optimal, based on the predictive variables. Then, three variables were added to the dataset, cluster1, cluster2, and cluster3. Depending on which cluster the data point fell into, one hot encoding was used as True and False here, corresponding to 1 and 0. Though this is a dimensional

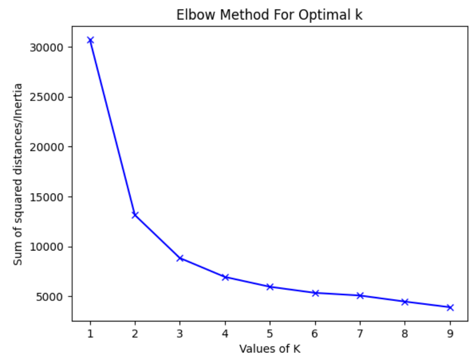

Fig. 4. Elbow Method For Optimal K for Predictive Variables

increase, hypothetically this might help predictions, as the data points were classified into three different clusters depending on their dimensions, giving the model a new pattern. Because the k-means projects the data into a multi-dimensional space, it is understandable that applying such a technique could potentially add an unseen or not easily detected trend of pattern in the data. The second method was applying dividing the dataset into subsets depending on the cluster. So clusters 1, 2, and 3 would each obtain their respective dataset, then the model would train on each subset, and a meta-model is trained on the predictions of the cluster models.

	R^2	MAE	MSE
Clusters as added Features	0.67	1.36	3.85

	R^2	MAE	MSE
Cluster 1 model	0.34	1.82	6.18
Cluster 2 model	0.42	1.42	3.93
Cluster 3 model	0.45	1.62	4.90
Meta-Model	0.52	1.55	5.15

Looking at the results, providing cluster features to the dataset, and concatenating them to the set of predicted variables provides the best metrics of the model so far. The R^2, MAE, and MSE are all the lowest values achieved yet. Thus, it is fair to claim that this does help the model make predictions, and thus, is added to the model as a pre-processing step.

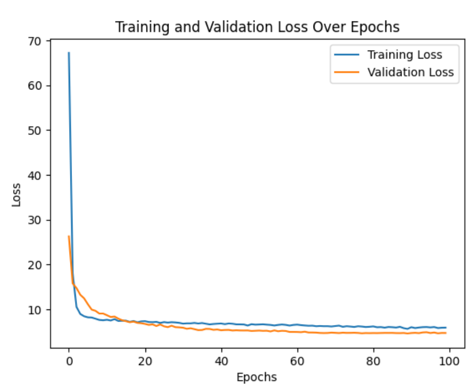

Fig. 5. Training and Validation Loss Over Epochs Graph with K-Means Clustering as a Pre-Processing Step

In attempt to further confirm adding the clusters is a fair and reasonable pre-processing step, the training and validation losses over the epochs was plotted. Here, the training loss indicates to the error amount on the data that the model was trained on. The validation loss involves the error amount on the unseen or test data, and this is used to provide insight on the model's performance on the test set. Tracking these losses is crucial in understanding how well the model generalizes and learns. Looking at Fig. 4, the losses both decrease, and the validation loss being slightly below the training indicates that the model is very slightly underfitting, which is understandable considering the model has not yet had been fully hyper-tuned, but the model showing no signs of overfitting is a good sign. The plot also indicates that the model converges before 100 epochs, meaning the the number of epochs can be reduced, and the learning rate scheduling could be optimized [9].

Another method to investigate for pre-processing this data follows investigating different transformations on the features to explore if reducing and transforming the features has any effect on the model performance, or on the presence of the outliers. After plotting the scatter-plots and visualizing the relationships with the target variable ”Age”, different patterns and similarities can be detected.

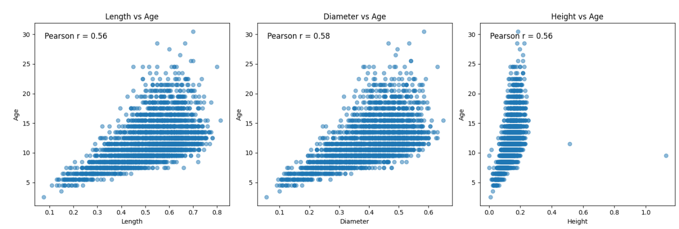

Fig. 6. Scatter-plots for the Length, Diameter, and Height Respectively Against the Age

For the Length and Diameter features, it is very distinct that the distributions are highly similar. For this reason, an attempt to combine the two was examined to check if this had any positive effect on the model. First, the mean of these two variables was obtained–at first, it was placed in the predictive variables set with the Length and Diameter features removed, and then once with it alongside the Length and Diameter in the predictive variable set. It evidently made the model worse in both cases. Then, the same was attempted for the Diameter and the Height variables. This was done in attempt to provide a feature with a more linearly dependent relationship to the x values. This also did not prove beneficial.

For the weight variables, since they are all right-skewed, a logarithmic transformation was applied on each of the weight variables to investigate if this transformation serves any benefit for the model [10]. This was done once for the weight values alone, and one for the weights and height features, and the height is right-skewed as well. Both attempts served no benefit, and were thus not considered for the model.

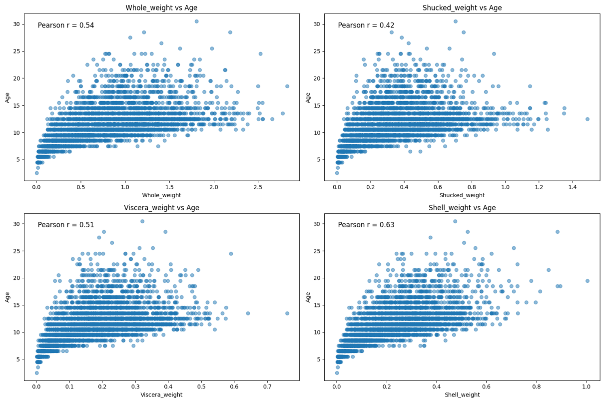

Fig. 7. Scatter-plots for the Whole, Shucked, Viscera, and Shell Weights Respectively Against the Age

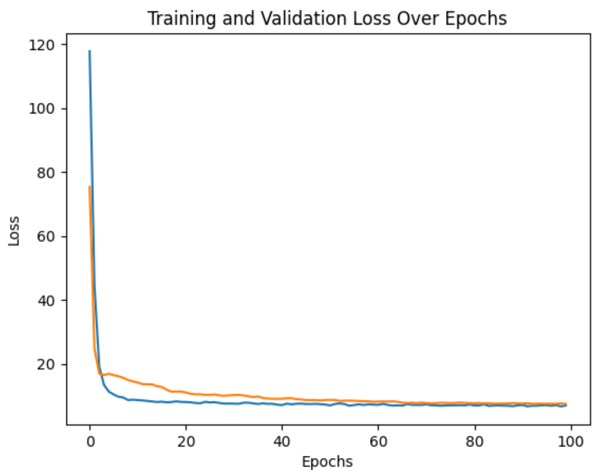

Fig. 8. Training and Validation Losses Over Epochs for Final Pre-Processing and Tuning of the Model

C. Results Analysis

1) Model Two: Extreme Gradient Descent (XGBoost): Considering the nature of the data at hand, it seems sensible to compare this model with a model that is also fairly powerful and robust. Considering the multi-dimensional aspect of the data, it is also reasonable to look at gradient descent algorithms [11]. Consequentially, Extreme Gradient Boost appears to be a suitable algorithm to investigate this task further, and to explore it against the neural network. Extreme Gradient Boost models are very suitable for tabular datasets, such as the abalone dataset [12]. It is considered a tree boosting machine learning method, where it performs gradient boosting on trees. It is noted for its high scalability due to its algorithmic optimizations. It is also well-known for its computational efficiency, allowing one to process millions of data samples on a simple desktop, and is popular because of its performative abilities. This algorithm is also fairly robust to outliers, but surpasses neural networks by far on both a computational level and a interpretability level, as XGBoost algorithms provide some insight into the feature importance [13]. After completing the model and hyper-tuning the parameters, it was evident that the best hyper-parameters for this model and dataset are: `n_estimators = 175`, `learning_rate = 0.07`, `max_depth = 4` and the test set is taken as 0.1. After running the model, the following metrics were yielded, as compared to the neural network.

2) Model Comparison and Evaluation: The metrics indicate that surprisingly, the neural network still outperforms the XGBoost model both in terms of capturing variations and in the error values. XGBoosting.com used an XGBRegressor to predict the age of abalone, using the same predictive variables: 'Sex', 'Length', 'Diameter', 'Height', 'Whole weight', 'Shucked weight', 'Viscera weight', 'Shell weight', and MSE of 4.95, claiming that the best MSE score value is 4.61 [13]. Thus, though these metrics can be highly improved, as R^2 can most definitely be increased, and the error values can definitely be decreased, the models and the research indicate that the data is fallable to a fair degree. Neural networks are extremely powerful and great at detecting complex patterns, however are computationally expensive. XGBoost models are efficient, but performed not as well as expected on this dataset, also considering how powerful they are. Improving the data in different ways essentially could help improve the prediction accuracy of different models, and help derive a stronger and more accurate model in which farmers and consumers can serve benefit from. The patterns in the data suggest that something more needs to be fed to the model for it capture all the patterns effectively–or maybe multiple things.

	R^2	MAE	MSE
XGBoost Model	0.61	1.49	4.17
Neural Network	0.67	1.36	3.85

The neural network did outperform the XGBoost model in all metrics–having a higher R^2 value and lower error values. Thus, the neural network might be a better option for this dataset, despite the metrics and model performance still being relatively subpar.

II. Conclusion

Overall, the attempt to provide a predictive model for predicting the age of abalone is both important, but constantly proves as a challenge considering the data at hand. The idea is that the variables, though very similar–such as the dimensional features and the weights–follow different patterns, however, appear to have underlying or discrete patterns that models are generally confused with. Also, the weak correlation between the variables and the Age, does not aid this. Further pre-processing and a more meticulous understanding of the data would significantly improve the predictive accuracy. Though multiple attempts of predicting the model accuracy have been made, there could still be more research, models, and feature engineering that could be tested in order to find an optimal combination. The data clearly does form clusters in the multi-dimensional space, as seen through the k-means clustering being successful here, as well as the features being very similarly distributed for samples of simple features and aspects.

References

[1] K. Mehta, "Abalone Age Prediction Problem: A Review", International Journal of Computer Applications, vol. 178, no. 50, pp. 1-5, 2019. [Online]. Available: https://www.ijcaonline.org/archives/volume178/number50/mehta-2019-ijca-919425.pdf

[2] A. Kumar, "MSE vs RMSE vs MAE vs MAPE vs R-Squared: When to Use?" Vitalflux, Oct. 8, 2021. [Online]. Available: https://vitalflux.com/mse-vs-rmse-vs-mae-vs-mape-vs-r-squared-when-to-use/

[3] S. Hleap, "Interquartile Range Method for Reliable Data Analysis," ProCogia Blog, Apr. 1, 2020. [Online]. Available: https://procogia.com/interquartile-range-method-for-reliable-data-analysis/

[4] Scikit-Learn Developers, "sklearn.decomposition.PCA," Scikit-Learn Documentation, 2023. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

[5] F. Chollet, "Building Autoencoders in Keras," The Keras Blog, Dec. 2016. [Online]. Available: https://blog.keras.io/building-autoencoders-in-keras.html

[6] GeeksforGeeks, "Activation Functions in Neural Networks," GeeksforGeeks, 2021. [Online]. Available: https://www.geeksforgeeks.org/activation-functions-neural-networks/

[7] K. Wang, L. Lozano, C. Cardonha, and D. Bergman, "Optimizing over an ensemble of neural networks", ArXiv preprint, arXiv:2112.07007, Dec. 2021. [Online]. Available: https://arxiv.org/pdf/2112.07007

[8] Ankita, "K-Means: Getting the Optimal Number of Clusters," Analytics Vidhya, May 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/

[9] GeeksforGeeks, "Training and Validation Loss in Deep Learning," GeeksforGeeks, 2021. [Online]. Available: https://www.geeksforgeeks.org/training-and-validation-loss-in-deep-learning/

[10] S. Khan, "Enhancing Machine Learning Models: A Deep Dive into Feature Engineering," Dev.to, Jul. 2023. [Online]. Available: https://dev.to/salmank/enhancing-machine-learning-models-a-deep-dive-into-feature-engineering-2dh7

[11] XGBoosting.com, "XGBoost for the Abalone Age Dataset," XGBoosting Blog, 2022. [Online]. Available: https://xgboosting.com/xgboost-for-the-abalone-age-dataset/

[12] A. Moore and M. Bell, "XGBoost, A Novel Explainable AI Technique, in the Prediction of Myocardial Infarction: A UK Biobank Cohort Study", *SAGE Open*, vol. XX, pp. XX–XX, Oct. 2022. [Online]. Available: https://journals.sagepub.com/doi/pdf/10.1177/11795468221133611

[13] XGBoosting.com, "XGBoost for the Abalone Age Dataset," XGBoosting Blog, 2022. [Online]. Available: https://xgboosting.com/xgboost-for-the-abalone-age-dataset/