Abalone Case Study /Blog Submission

The purpose of this article is to summarize the predictive modeling process, from exploring the data to deploying the prediction. This includes the thought process behind various decisions during the process, be it selecting the features that could help with better prediction accuracy or engineering certain features to help get more details out of the data.  In this analysis we seek to understand the distribution of the dataset attributes, as well as the relationship between them.

The analysis is divided in six sections: on section 1 we briefly present what an abalone is. On section 2 we present the Abalone Dataset and their attributes. On section 3 we will perform the analysis of each attribute individually. On section 4 we seek for correlations between the attributes and how the segmentation affects the results. On section 5 we will build our Model based on observation & analysis done. Finally, on section 6 we present our conclusions.

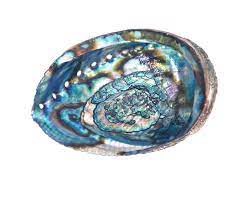
Table of Contents:

1.      Problem Definition  
2.      Data Analysis  
3.      EDA Concluding Remarks  
4.      Pre-processing Pipeline  
5.      Building Machine Learning Models  
6.     Concluding Remarks

Problem Definition: The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope -- a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem.Based on the information given in our dataset we have to predict the rings of each abalone which will lead us to the age of that abalone.

What is an Abalone?

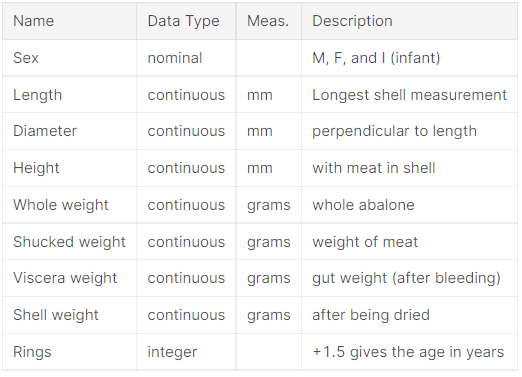
Abalones are marine snails. Their taxonomy puts them in the family Haliotidae which contains only one genus, Haliotis, which once contained six subgenera. These subgenera have become alternate representations of Haliotis. Abalones can be found along coasts of almost every continent. Usually, abalones are consumed as food all around the world, by different cultures. However, the bright and variety of colors of the interior side of their shells makes them an valuable object of adornment and decoration.



**Data** Analysis:We will be using the already collected Abalone dataset to see the algorithms in action. The first step in knowing the data is to know what it contains. This means understanding the type (continuous numeric, discrete numeric or categorical) and meaning of each feature and noting down the number of instances and features in the dataset.

*A brief aside on the motivation behind collecting the dataset. Abalone is a type of consumable snail whose price varies as per its age and as mentioned here: The aim is to predict the age of abalone from physical measurements. The age of abalone is traditionally determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age.* *Thus, one possible solution is predict the number of rings of an abalone from characteristics like height, diameter, lenght and weight measurements.*

The Abalone Dataset is composed of the following attributes:



In this study, each one of these attributes were analyzed, as well as the relationships between them.

**Exploratory Data Analysis** is a set of techniques that were developed by Tukey, John Wilder in 1970. We use this approach to examine the data before building a model. John Tukey encouraged statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. Data scientists and analysts spend most of their time in Data Wrangling and Exploratory Data Analysis also known as EDA For categorical features, a useful Exploratory Data Analysis (EDA) step is the frequency distribution of categories within the feature, which can be done with the .value\_counts(). The distribution of Males, Females, and Infants are not too different, Males are about 37%, Females and Infants about 32%.

**Univariate Analysis**

Univariate Analysis: Univariate analysis explores each variable in a data set, separately. It looks at the range of values, as well as the central tendency of the values. It describes the pattern of response to the variable. It describes each variable on its own. Descriptive statistics describe and summarize data.

In this section the distribution of each attribute is analyzed individually. We start analyzing the distribution of the target attribute *Rings*. The rest of the attributes are divided in groups for convenience of the analysis: a group called *Size*, containing attributes that represents the dimensions of an abalone, a group *Weight*, containing the different weight attributes and a third group composed only of the *Sex* attribute. The continous or quantitative attributes were analyzed using histograms and boxplots, while categorical attributes were analyzed using barplots.

**The Target Attribute**

The analysis shows that the Ring attribute values ranges from 1 to 29 rings on an abalone specimen. However, the most frequent values of Rings are highly concentrated around the median of the distribution, so that, the 2nd and 3rd quartiles are defined in a range of less than 1 std deviation. We observe that its possible to approximate the distribution of this attribute to a normal curve.

**Size Attributes**

In this section, we analyze the attributes that represents the dimensions of an abalone. These attributes are Lenght, Diameter and Height. For each of these attributes we plotted two histograms and their respective boxplot. The first histogram is a density histogram and a kernel density estimate, and the second is the absolute frquencies of the attribute, with some adjustments to ticks and bins. Once more, we observe an approximate normal distribution. However we notice the high peak formed by the Height histogram. Analyzing the Height boxplot we conclude that the high peak is formed due the presence of two observations that lie far beyond the central positions of the distribution.

Weight Attributes

The weight attributes were analyzed following a similar approach to the Size attributes analysis. A similar distributions were observed, however, for the weight attributes the bell curve is a little larger.

**Sex**

The Sex attribute is a categorical variable for which the possibles values are: M for Male, F for Female and I of Infant (an abalone which is not adult). We analyzed the count of each category with a bar plot, and concluded that relative to this attribute, the dataset is balanced.

**Multivariate Analysis**

In this section we analyzed how the dataset attributes are related and how the independent variables influences the target variable. Our first step in the multivariate analysis was to visualize the correlation matrix in a heatmap:

How the independent variables influences the target variable.We need to visualize the correlation matrix in a heatmap

Analyzing the correlation matrix, we notice that *Height* and *Shell weight* are the attributes that most correlates to *Rings*. Therefore, we concentrated the multivariate analysis on the correlation of these two attributes with *Rings*:Height and Shell weight are the attributes that most correlates to Rings

Correlation of these two attributes with Rings

How correlation varies with the number of rings?

Based analysis, we decided to investigate the variation of the correlation regarding the number of rings in more detail. We tested for many values, and found that the region delimited by Rings < 10 has greater correlation between the independent attributes and the target variable.

We will see the correlation between size attributes and Rings.

The violin plots bellow show that the median of *Size* attributes increases as instances are grouped by *Rings*:

Comparing Height and Shell weight to Rings:

W With more hen we will compare we found than 10 Rings we observe that correlation decays drastically to near 0 (zero).These results suggest that abalones grows in size and weight until a certain age, near 10 years. After this age their size and age remains stable.

Influence of Sex on attributes

Finally, we analyze how the Sex category influences the distribution of variables Rings, Height and Shell weight. Our objective is learn if the different categories of abalones have different distribution parameters or even form. To accomplish this objective, we visualize the distribution of each one of these parameters in relation to Sex. Finally, we analyze how the different Sex categories of abalones influence the correlation of Rings, Height and Shell weight.

Our first step is to analyze how *Sex* categories influence the number of *Rings*. We observe that the median of *Rings* for the *I* category is lower than the median for *M* and *F* categories.Infants have lower number of rings as compared to males and females

**For shell height & weight**

Finally, we analyze how the categories influence the correlations *Ring* x *Height* and *Ring* x *Shell weight*. We already concluded that these attributes has stronger correlation for lower values of *Rings*. Because *Infant* abalones have lower values of *Rings*, the consequence is that *Height* and *Shell weight* have stronger correlation to rings. Observing the regression curve for the *Infant* category, we notice that its inclination is closer to 45°.

Below process we use :

* Outlier detection and removal
* Z score method & Separating Input Features and Output Features
* Splitting the data into Train and Validation Set

**Model Evaluation:**

Once we trained the data, we need to find out if our model generalizes well with unseen data. We need to find out if our model works and if we can trust the predictions. For which we need to evaluate the model.

There are many ways to evaluate the model. I am going to use Root Mean Square Error measure accuracy, for the regression problem Mean squared error loss will be able to give better prediction accuracy.

**Model Validation**

Model Validation is the process of checking the behavior of a model before it is used in production. Different organizations and Data Scientists might employ different methods to validate the model that they have built. The steps involved will vary across both horizontals and verticals to ensure that the model performs as expected and suits the specific needs of the team in question. Some of the common steps that are involved during the validation are as follows:

**Conceptual Design —**In this phase, the underlying principle behind the model development is questioned and verified. For a financial institution, this step may involve ensuring that the granting of loans is not subject to racial bias while in Human Resource tools, analytical models might want to ensure that the model does not prefer men, more than women.

**Data Quality Assessment —**In this phase, the quality of the data going into the model as training data is checked and analyzed. This will normally involve the following steps:

Outlier Detection

Data Imbalance

Noise Removal

Data Integrity Checks

Diversity Checks (Checking if people are from all parts of the society are represented equally)

**Model Performance Assessment —**This step deals with evaluating the performance of the model by performing various checks. The model might be run again on specific sets of data and the results that are output by the model will be compared and analyzed. If the model is churning out results that are inconsistent, then the model won’t go into production and could be sent back for further understanding of issues in the model, if any

**Infrastructure Assessment —**This step deals with evaluating the infrastructure that the model will be deployed to. The infrastructure will be evaluated differently, based on the business needs of the model and the type of the model. For the unversed, there are primarily 2 types of models — online models and batch models. Online models serve real-time requests and provide predictions in real-time while for batch models, the predictions are made as a ‘batch’, at one go. For such models, there might be scheduled runs of the model where the data is aggregated and fed into the model at one go.

On-line models: For these models, the infrastructure should be available at all times to serve requests. The latency should be at a bare minimum and any downtime could cause complications from a business standpoint. Hence, for such models, there will usually be back-ups that can take over at all times to reduce downtime

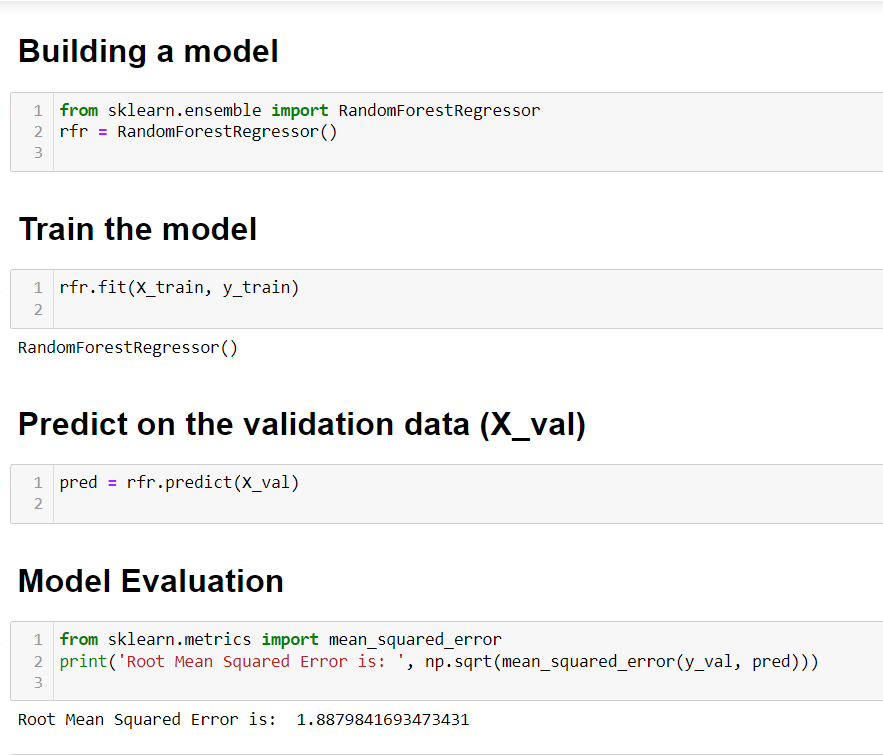
Batch models: For these models, the infrastructure should have enough computing power to be able to make predictions in a timely manner. If the infrastructure doesn’t have enough computing power to handle the volume of data that will be fed into it, then either the infrastructure should get upgraded for more computing power or the model will have to be optimized more to adapt to the constraints in the infrastructure.

As can be seen, the model validation is a series of steps that involve checks from both the business side and the technical side of things. On the technical side, care should always be taken to ensure that the model should only be validated on data that was not used to train it, ie. training data.

Model, not a one-time process and is instead a series of steps that should be repeated whenever modifications are made to a model. The above steps will vary from organization to organization to adapt to individual needs but these are the most common steps that need to be included in any pipeline of steps for validating a model.

**Modelling Phase** : In this stage of the process one has to apply mathematical, computer science, and business knowledge to train a Machine Learning algorithm that will make predictions based on the provided data. It is a crucial step that will determine the quality and accuracy of future predictions in new situations.

**Finalizing the best Model**:Once you have an accurate model on your test harness you are nearly, done. But not yet.There are still a number of tasks to do to finalize your model. The whole idea of creating an accurate model for your dataset was to make predictions on unseen data.



**Conclusion**

By observing the correlation between the target attribute Rings and the indepent variables, we conclude that it is possible to build a model to predict the target value in function of the independent attributes.The weight of the Abalones varies proportional to their sizes

There's no significant differences in size, weight and numbers of rigns between male/female abalones.The Infant Abalones groups presents lower mean values of size, weight and number of rings.The weight and height of abalones varies accordingly to age until the adult age, after adult life size and weight stops varying, and after 16.5 years (15 rings) these measurements aren't correlated.