Abalone Age Prediction

Problem Statement:-

Abalones are endangered marine shells that are found in the cold coastal water world wide, being distributed off the coasts of New Zealand, South Africa, Australia, Western North America, and Japan. They are highly nutritious , elegance, rare food which is good to eat is extensively consumed in France, New Zealand, Certain parts of Latin America, Japan and Korea. The shells of abalone are used for decorative purpose. Therefore , abalone is economically significant. The price of Abalone is positively correlated to its Age. However determining the age of abalone is very difficult process. Rings are formed in the inner shell of the abalone. One ring will be developed per year. The age of Abalone is determined by cutting the shell through the cone and staining it and counting the number of rings through a microscope.

Since some rings are hard to detect in this method,1.5 is traditionally added to ring count as the reasonable approximation of the age of abalone. Knowing the absolute price of the abalone is important to both the farmers and consumers while knowing the correct age is important to Environmentalists who seek to protect this endangered species. Due to the inherent accuracy in the manual method of counting the rings and thus calculating the age. Researchers have tried to employ physical characteristics of abalone such as sex, weight, height, length to determine its age. The corresponding dataset is found on UCI’s repository.

Most of the research on the dataset has seen the abalone age prediction problem being categorized as Classification Problem that is assigning a label to each parameter in the dataset. The label in this case is number of rings of abalone, which is the real number. This leads the classifier to distinguish among classes and is thus bound to poor results. To improve upon this approach, the number of classes is reduced. . For instance, two ages belonging to one of the reduced class but nonetheless causing a large variation in price would render the reduced class model useless. To overcome the problems associated with the classification model, this paper experiments with regression models and analyses the performance.

Data Analysis:-

The abalone dataset is a dataset that contains measurements of physical characteristics of different abalones. It has 4177 instances.

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.

3.1. 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.

3.2 Boxplot:-

The minimum, maximum, mean, median, standard deviation and interquartile range of all the numeric attributes along with dependent variable of the dataset is calculated and plotted using a boxplot for easy visualization of outliers. Due to the larger range of “Rings” variable, an un normalized boxplot renders the other variables’ boxplots incomprehensible by squeezing their ranges. To bring all the variables on the same scale, they are normalized such that they all have zero mean and standard deviation 1. . The attributes Length and Diameter have almost the same normalized range while there are a few outlying values for the Height attribute which might make the task of regression difficult. All the Weight attributes also have almost the same normalized range. The Rings label is not analyzed since it will be used in an un normalized form for the regression to obtain a proper value of MAE

Figure specifically looks at the values of correlation among the numeric attributes and between the numeric attributes and label. Apart from the index starting at zero, the order of attributes is as shown in Table 1 (the Rings label forms the last column and last row). It can be clearly seen that the different attributes have a strong correlation with each other.. Based on the correlation of the attributes with the label, it can be concluded that Shell Weight is the most important attribute for prediction.

Methodology:-

The first step towards applying linear regression to predict the age of abalone was to numerically code the sex variable in the dataset. For this, the following approach was taken:

1.two attributes were created in place of sex. Let them be S1 and S2.

2.The sex value of each sample was checked

3.if it is a Male, then S1 is equated to 1 and S2 is equated to 0.

4. if it is female, then S1 is equated to 0 and S2 is equated to 1.

5.if it is Indeterminate, both are kept 0

The attribute were not normalized before training because most of them distributed in range (0,1). A dataset with normalized attributes was trained on the best performing Linear regression model later to validate the claim. The label was never normalized as we done in the dataset Analysis to ensure proper calculation of Mean Absolute Error.

OLS Model:-

This is the most basic regression model. It was selected so that other models performance could be compared with OLS

Ridge:

Ridge was used to see the effects of penalties on the abalone dataset . other models like lasso, Lasso lars and Elastic Net was tried

All the models were run several times till the performance no longer improved and best performance of the model is reported. For all models techniques like SMOTHE and 10-fold Cross Validation were used to improve the quality of dataset and the performance of the model. CGANs were not used to synthesize samples because the author believes that CGANs amplifies the noise in the dataset too which results in deterioration of performance. For the models without CV and without synthesized data, the dataset was divided into 80% of training data and 20% of testing data. SMOTHE has limitations due to which it cannot synthesize samples for labels having only one sample in the orginal data Hence, parameter having labels between 3 and 19 were synthesized and only labels within this range in the orginal as well as SMOTHE data were used for training models. The data imbalance was removed by bringing the total parameters for each label having less than 600 parameters initially to 600 which is comparable to the majority class. The training and testing data was created with 80:20 ratio such that each class had approximately equal representation. 10fold Cross Validation meant that the dataset was split in the 90::10 and the reported error was averaged over the ten folds

Results and Conclusions:-

the Mean Absolute Error for the various models with the corresponding hyper-parameters in brackets. The results for 10 fold CV has been averaged on the ten folds. Other penalized regression methods which were tired were Lasso Lar and Elastic Net. All of these performed similar to Ridge regression with their best performance being bit lower than the performance of OLS. It is clearly visible that RANSAC yielded the performance.

OLS:-

OLS (without SMOTHE):-1.500

OLS(with SMOTHE):-1.875

OLS (with 10 fold CV):-1.636

Ridge:-

Ridge(Without SMOTHE):-1.499(alpha=0.1)

Ridge(With SMOTHE):-1.882(alpha=0.01)

Ridge(With 10 fold CV):-1.627(alpha=0.01)

Conclusion and Future work:-

In the task of predicting age of an abalone (by predicting number of rings) through its physical characteristics, the OLS regression model works best with a MAE of 1.5.. All over, robustness regression models do a good job in dealing with outliers present in the abalone dataset. Techniques such as SMOTE and Cross Validation do not improve the performance of the models . This cements its position as the best model. Normalization of attributes seems to result in the same performance as un normalized data. The scatter plots for individual folds of cross validations show that Mean Absolute Error can be brought down below 1 for certain arrangements of data, the best being 0.936. However, it also shows that the error is still above the acceptable limits in some regions. It is the author's belief that given an adequate and balanced dataset, OLS along with SMOTE and Cross Validation can achieve the goal of Mean Absolute Error less than 0.5 across all the labels