Predict the age of abalone from physical measurements

1.1 Abstract— This paper tries to explore different methodologies that are available in data science, statistics and machine learning with the objective of predicting the age of an abalone shell. It is important to predict the age of an abalone because the accurate estimate helps persons involved in cultivation and supply chain of the abalone. The economic realisation of an abalone is proportional to the age of the abalone. The age of an abalone is ideally calculated by the count of the rings on the inner surface of the shell the better alternative to this methodology is leveraging data science tools and technologies to be more specific python along with libraries are used to build the machine learning model. The data to carry out the process is obtained from UCI machine learning repository, q quite old data set, ages back to 1980s. A number of classification and regression algorithms are used to obtain the best fit model and get the final model. The results are validated against the default model evaluation and error metrics

Fig : Image showing the typical abalone shell

1.2 Introduction

Abalones is a good resource for economics and recreational activities and always face threats from numerous factors including pollution, disease, predators, loss of the natural habitat, expansion of commercial harvesting, and sport fishing. Over the last 2 decades it is thought that commercially the rate of catching and the quantity of the abalone catching has decreased by around 40%. The abalones are easily harvested owing to the quasi static growth rates and variable reproductive success. The ability to quickly estimate the age of a regional abalone group of individuals is a salient capability. The information obtained from the observational study, the intent was to estimate the age of the abalone with the help of physical measurements alone thereby leading to avoidance of calculating the number of rings for aging. In an idea case scenario the growth ring is formed at each year of age. Presently the age is calculated by performing a drilling activity on the shell and counting the total number of rings on the shell with the usage of microscope which is a difficult and time taking process and the clarity on the number of rings always pose a challenge for the accuracy of the values, similar difficulties can also be identified while attempting to calculate the sex of an immature abalone. It is concluded that extra piece of information like weather patterns, geological positioning and food accessibility can shed more light into the age determination. Leveraging machine learning and data science concepts would be very reliable and self assuring route to determine the age of an abalone shell and also addresses most of the important concerns discussed above.

2. Literature review

Machine learning algorithms use mathematical reasoning, algorithms and statistical models which try to learn patterns on the data provided to it, popularly called the learning phase. From the sample inputs the machine learning algorithms learn a pattern and the algorithm performs tasks entirely based on the learned patterns and does not act on a program instructions. Machine learning has a wide array of applications into many different areas ranging from commerce, critical healthcare, insurance, banking, industrial applications, and many more sectors. Any machine learning problem can be into put into any of the two methods namely supervised machine learning and unsupervised machine learning, supervised machine learning is sub divided into classification and regression algorithms. The machine learning problem that is being addressed in this paper is a combination of a classification and regression problem. Classification machine learning approach can be leveraged when the problem is to classify certain aspects for example if a bank wants to know if a particular prospective customer for their loan products would default on payments or not, ability to identify a given email is spam or not a spam, identify a piece of information to be legit or fake, identify if a prospective customer to make a claim on the insurance policy being purchased.

Regression machine learning algorithms are to be leveraged when the requirement is to predict a quantifiable measure for example predict the height of the person, predict age of the person, predict the sales for the next month, predict the insurance claim amount, predict the price of gas in the US, predict the demand of a clothing brand, predicting the total premium customers for the next year etc. some time there is mixed based approach where both classification and regression algorithms are used in a single problem like if the requirement is to know if a customer would make a claim on his insurance, if yes what the claim amount would be when the customer makes a claim.

Another type of machine learning is unsupervised machine learning where we try to identify similar data points in the humongous amounts of data we never have a predefined output but try to find out patterns in the data some examples would be identifying different sets of customer bases for the rollout of different type of promotions based on the characteristics of purchase, anomaly detection, build recommendation systems, etc

3. Methodology

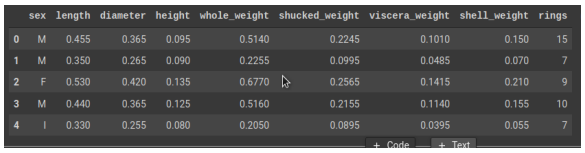
In this paper the data set is used from UCI machine learning repository that is dated back in 1995 for the first time. The data set is structured data having a total of 4177 records and 9 features. The data is provided as flat file in the form of csv. Python is the programming language used, the IDE used is jupyter notebook, the libraries and packages that are majorly leveraged in the exercise are Pandas, Numpy, matplotlib, seaborn and plotly. All the major steps that would playout in a typical machine learning solution development life cycle are implemented based on scope, requirement and data availability. These are the main steps involved in the whole process of predicting the age of the abalone shell.

1. Data Ingestion and Data Cleaning
2. EDA - Exploratory Data Analysis
3. Feature engineering and Feature selection
4. Model development

The above mentioned are the 4 phases in the process which are executed in chronological order to achieve the end objective of the exercise i.e. to predict the age of an abalone shell

3.1 Data Ingestion and Data Cleansing

This is the first section in the entire solution development process, the data file is ingested into the environment, the number of rows and columns are checked in the data read, a preliminary check is made to ensure the reading process has happened without any errors. Next step is to normalise the feature names in the data by removing spaces between the column names by underscore and converting all the feature names to lower case. Checking for the meta data of the features, Check for the null values is done it is observed that no null values are found in the data. Creation of "age" leveraging the rings feature in the data by applying the suggested rule age = number of rings + 1.5



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

+ Code + Text

3.2 EDA - Exploratory Data analysis

In this step, the data is explored end to end to better understand the characteristics of the data and know the nuances that are lying inside of the data. The first step is to generate all the descriptive statistics metrics for all the numerical features and examining the result obtained from the analysis. It is observed that the minimum values for the feature height is showing as zero there are 2 records in the data where the height is recorded as zero, these values are replaced with the mean of the height column as the height column shows close to normal distribution.

Then univariate analysis is done to understand the data for each of the features in the data using pie charts, histograms and violin plots. These are the following observations made on the univariate analysis

- The qualitative feature or the categorical feature "sex" has three values or categories and the proportion of these categories are amost same close to 33 % for each category
- The "height" feature has close to normal distribution and has very little or no outliers
- The response feature "age" has right skewed distribution and has many outliers
- All the weight features show very similar distribution and all of them are heavily right skewed also suggests that these are highly correlated features would be further investigated later
- The "length" and "diameter" features show very similar similar characteristics, left skewed distribution

Bi Variate Analysis

Analysis for gender and age comparison

- All the gender categories show few outliers
- Infant data is very heavily right skewed shows that most of the data has less than 11 years of age
- For male and female data, the characteristics are very similar characteristics have data points starting from age 6 and steadily increases

Analysis of height vs gender

- Males an females tend to have similar heights but infants show significantly lesser heights

Analysis of length vs gender

- Males tend to have more length when compared to females
- Infants have much less length when compared to males and females

3.3 Feature Engineering

The feature “sex” has to be converted into numerical feature, using label encoder.

A correlation matrix is plotted to analyse the collinearity between the features since regression is to be analysed. It is observed most of the features have a high collinearity to investigate further. We have created an inspection for VIF(Variance Inflation Factor). It is observed that most of the features show high VIF value, therefore few features to be engineered or removed from the data. For this requirement a new feature “volume” is created using length, diameter and height. Among the weights only one feature “shucked weight” is selected. The features “volume”,”sex” and “shucked weight” show a very less VIF values therefore multicolinearity problem is addressed.

3.4 Modelling

The features engineered are used to map to a high level function to predict the age of the abalone. Various models from machine learning literature are suitable for this use case. This problem comes under supervised as the predictor variable ‘Age’ is present in the dataset. In supervised machine learning, there are two types of models, classification and regression. This problem can be modeled either in the form of classification or in the form of a regression problem. Converting this problem to classification involves considering each age value as a class, then classifying each abalone into one of the classes which are ages. Another way is to convert into a regression problem, which makes the age data a continuous variable. The problem with considering it as a continuous variable is that the age will always be a floating point number but the age is almost an integer variable. This causes an unavoidable error in calculating model performance in regression. Model performance measurement methods are different for classification and regression. For classification, Accuracy can be a great simplified metric to measure performance. Other metrics like confusion matrix, precision etc. are not fairly interpretable when many classes are involved. In case of regression, root mean squared error (RMSE) and mean absolute percentage error (MAPE) can be used. RMSE helps us identify how far the predicted age is, to the true age and MAPE is used to understand the percentage error made by the model.

For classification formulation, age has to be formulated into multiple classes. However, after observing the data, the age can be split into two classes with a threshold of 10. This provides three advantages. First, the data is clearly different for ages above and below 10. Second, this partition created a balanced class labels for classification and finally, good performance on this bifurcated data indicates good separability in features and can be extended to even further splits of data.

For regression data, the age can be considered as a continuous variable as discussed above and can be directly used for modelling. However, when considered in the end use case, it is better to provide a rounded value of age or a range of ages for better user experience. The classification output can be used as an additional input to the regression model to produce a combined model utilizing both classification as well as regression model capabilities. The combination of classification and regression can be used only if the classification performance is accurate.

3.5 Results

In the case of classification, then the bifurcation of data is done, the performance of the dataset on support vector machines and random forests were underwhelming. The accuracy of the model is utmost 40% after hyperparameter tuning with grid search. The accuracy of 42% is not sufficient to combine it into regression model data. In the case of regression, the metrics MAPE 0.79 for random forest regressor

References

1. https://archive.ics.uci.edu/ml/datasets/abalone
2. https://medium.com/@ryotennis0503/analysis-of-abalone-age-9851efa5e105
3. https://rstudio-pubs-static.s3.amazonaws.com/378381_1221e0d1034b4020a38a862a76890bb6.html
4. https://www.kaggle.com/c/predict-abalone-age