**ABALONE PROJECT (link of the complete project is given at the end)**

**PROBLEM STATEMENT**

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 age. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem.

PURPOSE: To predict the rings of each abalone using machine learning. According to the research, the age of abalone is equal to the number of rings plus 1.5, thus we can predict the age of abalone using the number of rings. This is done by calculating the number of layers of a shell (rings). For this, the sample of the shell is collected, stained, and then a microscope is used to count the numbers of rings.

DEFINITION OF ABALONE: Abalone is a shellfish that is found in cold coastal waters. In other words, abalone is a kind of marine snail.

VARIABLES:

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Data Type**       e | **Measurement Unit**       Measurement Unit | **Description** |
| Sex | nominal | -- | M, F, and I (infant) |
| Length | continuous | mm | Longest  shell measurement |
| Diameter | continuous | mm | perpendicular to length |
| Height | continuous | mm | with meat in the shell |
| Whole weight | continuous | grams | whole abalone |
| Shucked weight | continuous | grams | weight of meat |
| Viscera weight | continuous | grams | gut-weight (after bleeding) |
| Shell weight | continuous | grams | after being dried |
| Rings | integer | -- | +1.5 gives the age in years |

**DATA ANALYSIS**

Data analysis implies cleaning, transforming, and modeling data in order to make a data-driven solution to a problem. According to the definition, the age of abalone is equal to the number of rings plus 1.5, so we will substitute the rings column with the age column and make it the target variable. The target variable is a continuous variable, thus it is a regression problem.

In the abalone dataset, there are 4178 rows and 9 columns. This means that there are 8 features and 1 target variable and each variable has 4178 data values. Among all features, Sex is categorical data and all other columns are numerical namely, Length, Diameter, Height, Whole weight, Shucked weight, Viscera weight, Shell weight, Rings.

There is no null value in the data. But, height has a minimum value equal to zero which is logically not possible. There exists skewness and outliers in the data. We need to fix these issues and then proceed further to build the models for predicting the age of abalone.

Our task is to figure out how well the features are correlated with the target variable as this will tell how important a particular feature is in estimating the age of the abalone. Also, along with this, it is important to make sure that no two features are related to each other as it will lead to the problem of multicollinearity.

**EXPLORATORY DATA ANALYSIS**

Exploratory data analysis refers to bringing out insights from the raw data by investigating and discovering patterns present in the data. This is done using visualization techniques and statistical features.

The initial step is to import necessary libraries, for example, Pandas, NumPy, matplotlib, seaborn, etc. Then, the next step is to upload the dataset to the current working directory The data is stored in a CSV file, so the read\_csv function of pandas is used to read the file. To take a glimpse of the data, the head() function is used.

To get useful insight into the data, a statistical summary feature is used:

1.    Info() function depicts that sex is object type variable, rings is an integer type variable and all other variables are float type.

2.    Describe() function indicates that the features shucked weight, viscera weight and shell weight have high standard deviation. Also, the minimum value of height is zero which is technically not possible. There is a presence of outliers in the rings column.

3.     Isna() function shows that there is no null value, thus there is no problem of missing values in the data

Now, we need to substitute the ring column with the age column using the definition of age of abalone.

Now, let’s explore data with the help of graphs. Python has a visualization library called seaborn. It can be used to perform both univariate and multivariate analyses.

1.    Histogram: It shows that there is high skewness in the features: height, Shucked weight, Viscera weight, Shell weight.

2.    Boxplot: It shows high number of outliers in the target variable: age.

3.    Heatmap: It shows that the age is related to all the variables. It also shows a correlation of height with other features.

4.    Pairplot: It depicts high correlation between various features.

5.    Regression plot: It shows how each feature is related to the column age.

**PRE-PROCESSING PIPELINE**

There are various problems in the data and we need to fix those problems:

1.    OUTLIERS: To deal with the problem of outliers, the IQR method is used and it eliminates all the values that lie either below the quartile 1 minus IQR times 1.5 or above the quartile plus IQR times 1.5. We figured out that by doing this, only 9% of the data is being lost, so we can apply the IQR method to deal with the outliers.

2.     SKEWNESS: After dealing with the problem of outliers, the skewness has been reduced drastically. This is because skewness is one of the consequences of outliers.

3.    MULTICOLLINEARITY: The feature whole weight has a very high variance inflation factor. This indicates that there exists a high correlation with any other feature. Thus, we can drop the feature whole weight. By doing this, VIF has reduced significantly but still, there is a need to drop one more feature i.e. Diameter. Now, the VIF is less than 10 for all the remaining variables. This implies that the problem of multicollinearity has been solved.

**BUILDING MACHINE LEARNING MODELS**

 The first step towards building a machine learning model is to split the columns into dependent and independent variables. The column age is the dependent variable(Y). All the remaining columns are independent variables (X).

Train test split is then used to measure the performance of the machine learning algorithm. The process is such that the entire data is split into two parts: train (80% of the data) and test(20% of the data). The models are trained using the train data and then test data is used to make predictions on data that is not used to train the model. Then we compare the values of the dependent variable with the predicted values and analyze how well models can estimate the values of the dependent variable.

In the project, four models have been used:

·         Linear regression

·         Random forest regressor

·         KNeighbors regressor

·         Gradient boosting regressor

It is a regression problem, so r^2 score and MSE are calculated for each of the models.

After this, cross validation is applied in order to check whether there is overfitting in the model or not.

Then, the difference between the r^2 score and the cross-validation score is measured as we have used r^2 score as the performance metric.

The data shows that the least difference between the r2 score and the cross-validation score is of the model gradient boosting regressor. Thus, we will apply hyperparameter tuning on gradient boosting regressor. We have received the best parameters for the model using the gridsearchCV.

Save the model after substituting the best parameters in the model.

**CONCLUDING REMARKS**

The machine learning model has made the complicated task of cutting the abalone and then counting the rings easier. Thus, the project shows that it is possible to determine the age of an abalone without using a microscope. The process of cleaning the data helped in making a better model as the chances of getting the right prediction increases when we feed the right data to the model.

**REFERENCE**:

<https://github.com/Jasmine-kaur8/Adalone_project/blob/main/Adalone_(project_4).ipynb>

[**HR Analytics Project**](https://github.com/Jasmine-kaur8/HR_Analytics_Project)**(link of the complete project is given at the end)**

**PROBLEM STATEMENT**

Employee retention is a very important factor that determines the success of the company. To understand how well a company is retaining its employees, the attrition rate is used as a metric. The attrition rate is used to measure and analyze how many employees left the company within a certain period. This helps the team to understand what are the factors determining the attrition rate.  A high attrition rate indicates that the employees are leaving the company frequently. A major problem in high employee attrition is its cost to an organization. Job postings, hiring processes, paperwork, and new hire training are some of the common expenses of losing employees and replacing them.  So, the goal of any company is to maintain a low attrition rate in order to retain the talent in the company. To make this goal possible, it is important to satisfy the needs of the employees.

PURPOSE OF THE PROJECT:  To predict and analyze the attrition rate using HR Analytics in order to maintain a low attrition rate. Machine learning helps in reducing the risk of high attrition by using predictive models. The purpose of these models is to provide an effective way to determine whether an employee will stay in the company or not on the basis of the details provided in the dataset.

**DATA ANALYSIS**

Data analysis is the science of examining a set of raw data and then converting it into useful and informative data which helps in making calculated decisions by the user. It consists of data collection, cleaning, transforming, analyzing, and modeling that helps in better decision-making.

The HR analytics data has 1471 rows and 35 columns. The target variable is Attrition takes two possible values i.e. Yes or No. Thus, this is a classification problem. There are 34 features in the model that are used as the independent variable to predict attrition.

**EXPLORATORY DATA ANALYSIS**

Insights from the data using the **statistical summary** available in python:

1.    Data.describe():

·         There are outliers in the features: Monthly Income, Total Working Hours, Years At Company, Years in Current Role, Years since Last Promotion, Years with Current Manager.

·         There is high standard deviation in the features: Monthly Income, Stock Options Level, Years At The company, Years In Current Role, Years Since Last Promotion, Years With Current Manager.

2.    Data.info():

·         There are some features that are object type and some are numerical type.

3.    Data.isna().sum():

·         There is no null value in the dataset.

4.    Data[‘Attrition’].value\_counts():

·         It shows that the value ‘no’ is 1233 and the value ‘yes’ is 237. Thus, there are 1233 people who didn't leave and there are 237 people who left the company. Thus, there is a problem of class imbalance.

5.    Data.groupby(‘Attrition’):

·         There is a single value for Employee Count, Standard Hours, Over 18.

·         Those with an average age of 37 are less likely to leave the company as compared to those who are around age 33.

·         Those with an average monthly income of 6832 are less likely to leave the company as compared to those with 4787 monthly income.

·         Those who have work experience of around 11 years are less likely to leave the company as compared to those who have that of around 8 years.

·         Those who have worked in the company for around 7.5 years are less likely to leave than those who have worked in the company for around 5 years.

Insights from the data using the **graphical representation** available in python:

Ø  Boxplot:

·         There is large a number of outliers in the Monthly Income feature.

Ø  Histogram:

·         There is skewed data in Hourly Rate, Monthly Income, Job Involvement, Percent Salary Hike, Total Working Hours, Performance Rating, Training Times Last Year, Work Life Balance, Years At Company, Years In Current Role, Years Since Last Promotion, Years With Current Manager.

Ø  Correlation matrix:

There seems to have a correlation between:

* Total Working Years and Age
* Total Working Years and Job Level
* Total Working Years and Monthly Income
* Total Working Years and Years At Company
* Years At Company and Years In Current Role
* Years In Current Role and Years With Current Manager

Ø  Distplot( after applying label encoder):

·         There is high skewness in the features: Distance from home, Monthly Income, Number of companies worked, Percent Salary hike, Years at the company, years since last promotion, years with the current manager.

Ø  Scatterplot:

·         Those with greater than 11 years with the current manager are less likely to leave than those with less than 11 years.

·         Those who got promotions since last year is less likely to leave as compared to those who have not.

·         Those with greater than 10 years in the current role are less likely to leave than those with less than 10 years.

·         Those with greater than 25 years with the company are less likely to leave than those with less than 25 years.

·         Those with greater than 25 years total working years are less likely to leave than those with less than 25 years.

·         Those with greater than 1 lakh of monthly income are less likely to leave than those with less than 1 lakh of monthly income.

·         Those with age greater than 50 are less likely to leave than those with age less than 50.

**PRE-PROCESSING PIPELINE**

Ø  As there is a single value for Employee Count, Standard Hours, Over 18, it is better to drop these features as this will not give any contribution in analyzing the attrition.

Ø  There is a problem of class imbalance, thus oversampling the minority class can help in solving this problem.

Ø  The z score was applied for dealing with the problem of outliers, it leads to a 10% loss of data (which is acceptable).

Ø  VIF is calculated after applying standard scalar. The VIF is less than 10 for all the features, thus there is no problem of multicollinearity.

**BUILDING MACHINE LEARNING MODELS**

* Separated the independent variable i.e Attrition with all other features called the independent variables.
* Applied train test split and trained the classification models i.e. KNearestNeighbors, Random Forest Classifier, Decision Tree, Gradient Boosting Classifier. The accuracy score of each of the models is calculated as 0.76, 0.91, 0.81, 0.89 respectively. Applied cross-validation and got the CV score as 0.72,0.82,0.68, 0.64 respectively. The difference between the accuracy score and the CV score is minimum for the KNearestNeighbors.
* Received classification report for each of the models.
* Applied hyperparameter tuning for KNearestNeighbors using gridsearchCV.
* Received the best parameter and substituted them in the model.
* Finally, saved the model for future reference.

**CONCLUDING REMARKS**

In, this project, a machine learning model is developed to predict whether there would be attrition or not. Our model shows 81% accuracy. This model can help in predicting the attrition which will, in turn, helps the company to save time and money in hiring the employees who will work for a longer span of time. This will also help the applicant in landing a better job as a machine learning model can predict which job will satisfy which type of employees more. Therefore, this model can help both employers and employees and as a return, it benefits the company as well.

**REFERENCE**

[HR\_Analytics\_Project/HR\_Analytics\_Project.ipynb at main · Jasmine-kaur8/HR\_Analytics\_Project · GitHub](https://github.com/Jasmine-kaur8/HR_Analytics_Project/blob/main/HR_Analytics_Project.ipynb)