NASA:Asteroids Classification

Tejaswini Jadhav
AI Tech Systems,
www.ai-techsystems.com
Navi mumbai, India
tejswini.jadhav74@gmail.com

Abstract—the data is about Asteroids - NeoWs. NeoWs (Near Earth Object Web Service) is a RESTful web service for near earth Asteroid information. Main objective is, using this Asteroids data performed a comparison using original data and using principal components of the data. Bu using PCA and logistic regression classifier applied for solving the binary classification problem.by this model got 83% Accuracy

Keywords—

PCA, logistic regression, classifier, EDA, matrix.

Introduction

This Asteroids .data collected from NASA website. This asteroid data is total (4687x40) dimension.40 different feature columns of asteroid is given. From this features we have to predict how many near earth asteroids are hazardous and non-hazardous to earth. Firstly wee performed some EDA Techniques to analyze the data in terms of graph. Like histogram, pdf, bar plots, box plot, violin plot, then data reduction technique called PCA is used to reduce dataPrincipal Component Analysis (PCA) is a **linear dimensionality reduction** technique that can be utilized for extracting information from a high-dimensional space by projecting it into a lower-dimensional sub-space.in PCA you take a dataset having many features, and you simplify that dataset by selecting a few Principal Components from original features. Principal components are the key to PCA; they represent what's underneath the hood of your data. Principal components have both direction and magnitude.

As you know Classification techniques are an essential part of machine learning.so we performed Logistic Regression classifier and built a model. Logistic Regression is one of the most simple and commonly used Machine Learning algorithms for two-class classification.

1. Advantages of Logistic Regression

1. Because of its efficient and straightforward nature, doesn't require high computation power, easy to implement, easily interpretable,
2. used widely by data analyst and scientist. Also, it doesn't require scaling of features.
3. Logistic regression provides a probability score for observations.

2. Equations:

Logistic Regression:

Logistic regression is a statistical method for predicting binary classes. The outcome or target variable is dichotomous in nature. Dichotomous means there are only two possible classes. For example, it can be used for cancer detection problems. It computes the probability of an event occurrence.

It is a special case of linear regression where the target variable is categorical in nature. It uses a log of odds as the dependent variable. Logistic Regression predicts the probability of occurrence of a binary event utilizing a logic function.

Linear Regression Equation:

$$y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots\ldots + \beta nXn$$

Where, y is dependent variable and x1, x2 ... and Xn are explanatory variables.

Sigmoid Function:

$$p = 1/1 + e^{-y}$$

Apply sigmoid function on linear regression:

$$p = 1/1 + e^{-(\beta 0 + \beta 1X1 + \beta 2X2 \ldots\ldots \beta nXn)}$$

Sigmoid function:
The sigmoid function, also called logistic function gives an 'S' shaped curve that can take any real-valued number and map it into a value between 0 and 1. If the curve goes to positive infinity, y predicted will become 1, and if the curve goes to negative infinity, y predicted will become 0. If the output of the sigmoid function is more than 0.5, we can classify the outcome as 1 or YES, and if it is less than 0.5, we can classify it as 0 or NO

$$f(x) = \frac{1}{1 + e^{-(x)}}$$

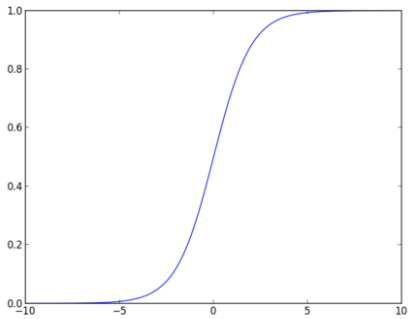

Fig.1.sigmoid function plot

Linear Regression vs. Logistic Regression*:*

Linear regression gives you a continuous output, but logistic regression provides a constant output. An example

of the continuous output is house price and stock price. Examples of the discrete output is predicting whether a patient has cancer or not, predicting whether the customer will churn. Linear regression is estimated using Ordinary Least Squares (OLS) while logistic regression is estimated using Maximum Likelihood Estimation (MLE) approach.

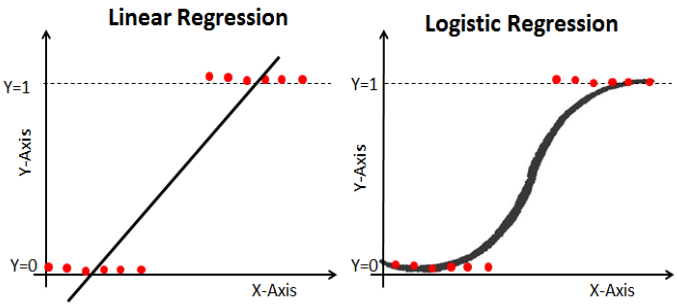

Fig.2.Linear v/s logistic Regression Graph

1. EDA

a. histogram, pdf

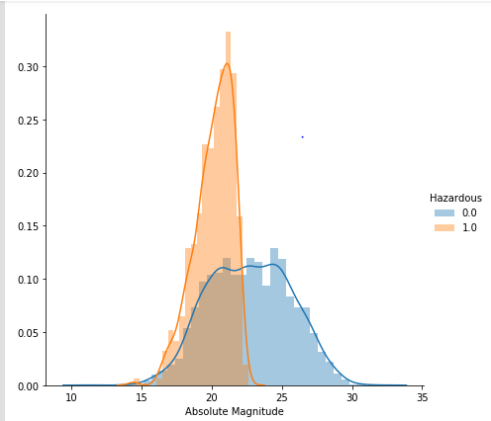

Fig.3.histogram and pdf of absolute magnitude

b.Box plot,Violin plot

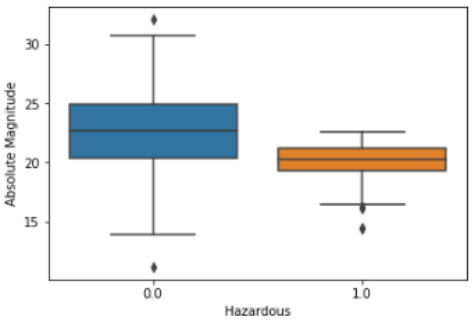

Fig.4.box plot of absolute magnitude and hazardous

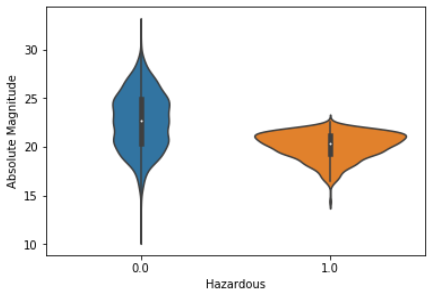

Fig.4..histogram and pdf of absolute magnitude

2.PCA

Steps:

1. Data-preprocessing: Standardizing the data
2. find the co-variance matrix which is : A^T * A
3. finding the top two Eigen-values and corresponding Eigen-vectors
4. plotting the 2d data points with seaborn

Table:

	1st component	2nd Component	Hazardous
0	0.460071	-0.693150	1.0
1	2.071261	0.304192	0.0
2	-1.446778	0.014228	1.0
3	0.718712	-2.039890	0.0
4	1.517848	-0.312831	1.0

Fig.6.reduced data table of PCA

PLOT:

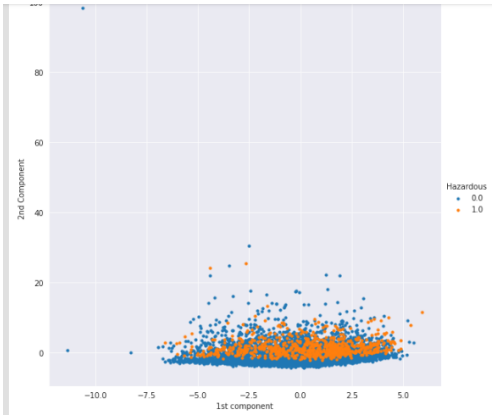

Fig.7.PCA plot

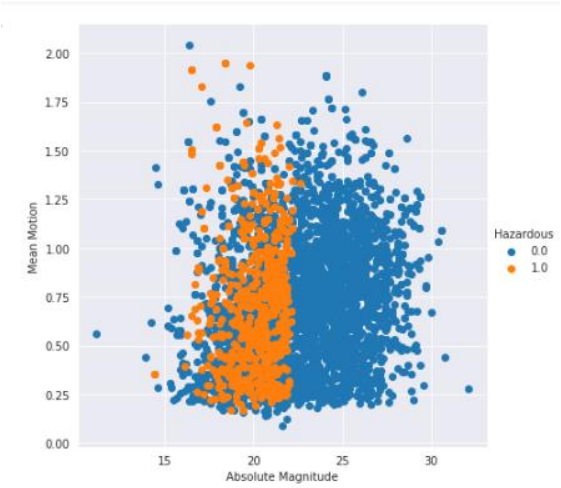

Fig.8.absolute magnitude vs mean motion plot

3. Logistic Regression

Steps:

1. selecting Features
2. Splitting data into training and test dataset: Here, the Dataset is broken into two parts in a ratio of 75:25. It means 75% data will be used for model training and 25% for model testing
3. Model Development and Prediction
4. Model Evaluation using Confusion Matrix

Confusion matrix:

A confusion matrix is a table that is used to evaluate the performance of a classification model. You can also visualize the performance of an algorithm. The fundamental of a confusion matrix is the number of correct and incorrect predictions is summed up class-wise.

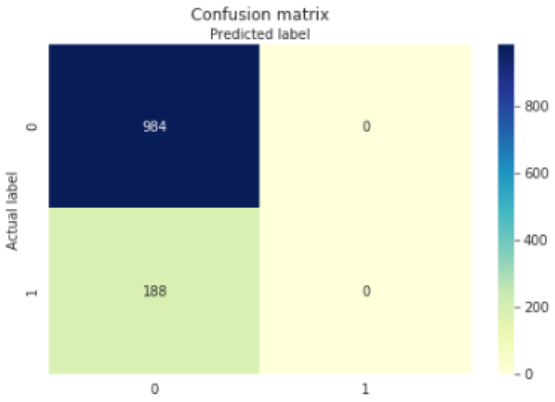

Fig.9.confusion matrix

Conclusion:

1.All the EDA plots .shows that from Absolute magnitude 22 to 30 all the asteroids are Non-hazardous to earth. And from absolute magnitude 15 to 20all asteroids are hazardous to earth
2. By using PCA dimension of the data has been reduced
3. That reduced PCA data used to run logistic regression model and got 83% accuracy

References:

[1]..https://www.kaggle.com/shrutimehta/data-preprocessing-and-correlation

[2].https://github.com/srohit0/mida/blob/master/notebooks/Dimensionality%20Reduction%20Using%20Principal%20Component%20Analysis.ipynb

[3]. https://matplotlib.org/

[4].https://github.com/mGalarnyk/Python_Tutorials/blob/master/Sklearn/PCA/PCA_to_Speed-up_Machine_Learning_Algorithms.ipynb