A1: SWISS CHEESE IN SPACE | AI 539 ML CHALLENGES

SUBMITTED BY : PUSHPAK VIJAY KATKHEDE

1. Data Set Profile :

Text

Description automatically generated

Text

Description automatically generated

Feature Name : PSF\_e1

Chart, histogram

Description automatically generated

No Missing Values

---------------------------------------------------------------------------

Feature Name : PSF\_e2

Chart, histogram

Description automatically generated

No Missing Values

----------------------------------------------------

Feature Name : scalelength

Chart, histogram

Description automatically generated

No Missing Values

----------------------------------------------------------------------------

Feature Name : model\_flux

Chart, histogram

Description automatically generated

No Missing Values

----------------------------------------------------------------------------

Feature Name : MAG\_u

Chart, histogram

Description automatically generated

Total missing values in MAG\_u feature are : 693

Feature Name : MAG\_g

Chart

Description automatically generated

Total missing values in MAG\_g feature are : 144

----------------------------------------------------------------------------

Feature Name : MAG\_r

Chart

Description automatically generated

Total missing values in MAG\_r feature are : 24

Feature Name : MAG\_i

Chart

Description automatically generated

Total missing values in MAG\_i feature are : 2

----------------------------------------------------------------------------

Feature Name : MAG\_z

Chart, histogram

Description automatically generated

Total missing values in MAG\_z feature are : 157

All Missing values in whole dataset : 1020

**Patterns in Missing Values**

Upon observing the dataset for many times I was though unable to identify any kind of pattern that might exist. I have tried relating the values with the other columns as well but there no concrete similarity that I was able to observe. However, I have observed that the missing values as mostly populated around the MAG\_g value more that ~24.5. Hence, I would classify this as in the Missing at RANDOM (MAR) category.

1. **Classifier Used and Its parameter values:**

I have used Support Vector Classifier to perform classification over the Target Label CLASS\_STAR. It is based on the technique of Support vector machines and relies on hyper planes also called as decision boundaries for the purpose of classification. There are multiple parameters that are available with SVC that can help tune the accuracy results of the classifier like kernel, gamma, coefficient, class\_size,etc etc. We have currently preferred to use the defaults provided by the sci-kit learn library. They are as follows:

SVC( C=1.0, kernel='rbf', degree=3, gamma='scale', coef0=0.0, shrinking=True, probability=False, tol=0.001, cache\_size=200, class\_weight=None, verbose=False, max\_iter=-1, decision\_function\_shape='ovr', break\_ties=False, random\_state=None )

* This is taken from official scikit learn documentation [link](C=1.0,%20kernel='rbf',%20degree=3,%20gamma='scale',%20coef0=0.0,%20shrinking=True,%20probability=False,%20tol=0.001,%20cache_size=200,%20class_weight=None,%20verbose=False,%20max_iter=-1,%20decision_function_shape='ovr',%20break_ties=False,%20random_state=None)

1. **Description of new method:**

In the new method I have tried to use median instead of mean to impute the missing values as I observed they are more concentrated towards the median rather than the mean. Although It just gave me a very little improvement in the accuracy of the prediction.

1. **ACCURACY TABLE**

This table contains the observation results from the various experiments performed in this assignment.

|  |  |  |
| --- | --- | --- |
| METHOD | Accuracy with Full Set | Accuracy with only Missing Values |
| A | 0.4425 | 0 |
| B | 0.6865 | 0.5922330097087378 |
| C | 0.71 | 0.41048 |
| D | 0.6115 | 0.40898058252427183 |
| E | 0.6213 | 0.41019417475728154 |

1. **Recommendation:**

For this dataset the most accuracy was achieved using the method to omit the dataset features and train the classifier using the remaining features only. In my observation, the density of values was high for the columns with missing values around the median i.e., more specifically between the 25th and 75th percentile which is evident from the data profile generated in the beginning of this document. Also, the histograms also show that the dataset does have outliers eventually due to (99, -99) which are the missing values and making the distribution stretch over the outliers. As well as the mean of these features are substantially differentiated from the median and that too beyond +- 1 SD. So, here the best strategy I believe would be to omit these features from the training dataset.

1. **OWN SKY OBJECT:**

This is the object that I got on following the point 6 in the assignment from the link.

|  |  |
| --- | --- |
| id | W1m4m3\_105355 30.538847270000016 |
| pos | 30.538847270000016 -9.700278699000002 |
| CLASS\_STAR | 0.57 |
| PSF\_e1 | 0.0191 |
| PSF\_e2 | 0 |
| scalelength | 0.3947 |
| model\_flux | 13.2 |
| MAG\_u | 23.919 |
| MAG\_g | 23.773 |
| MAG\_r | 23.407 |
| MAG\_i | 22.87 |
| MAG\_z | 22.495 |

The prediction I tried was using method C. The Prediction made by classifier was False which is correct for my object.

1. **Reflection:**

During this assignment, I have brushed up my skills with pandas and NumPy. I have also learned new function like take()\_ that can prove to be very handy in the upcoming parts as it will be vital to perform data pre processing in achieving the accuracy for upcoming assignments. The hardest part in the assignment was not with exemplary methods but understanding the requirements of the methods. I henceforth will try to get the doubts cleared about the exact expectation before hand in the classrooms or offc hours only just to avoid the confusions. I had to spend around 9 – 10 hours to complete the assignment.

1. **Extra Credit:**

Even when we know. The category of the sky object we might in the future come across more deviated values of the params that haven’t been yet classified as any of the category. At that time the model that we are training can help us identify and predict if that is star object or a galaxy. Such model are helpful when the data gets more complicated and on that level i=they can predict the upcoming stream of data and do the classification well.