**Sloan Digital Sky Survey**

**INTRODUCTION**

“The **Sloan Digital Sky Survey or SDSS is** a major multi-spectral imaging and spectroscopic redshift survey using a dedicated 2.5-m wide-angle optical telescope at Apache Point Observatory in New Mexico , United States. The project was named after the Alfred P.”

The Sloan Digital Sky Survey has created the most detailed three-dimensional maps of the Universe ever made , with deep multi-color images of one third of the sky , and spectra for more than three million astronomical objects.

**PROBLEM DEFINATION FOR DATASET**

Thedata consists of 10,000 observations of space taken by the SDSS. Every observation is described by 17 feature columns and 1 class column which identifies it to be either a star , galaxy or quasar.

**Feature Description**

The table results from a query which joins two tables (actuaclly views) : "PhotoObj" which contains photometric data and "SpecObj" which contains spectral data.

* objid = Object Identifier
* ra = J2000 Right Ascension (r-band)
* dec = J2000 Declination (r-band)

The Thuan - Gunn astronomic magnitude system. U , g , r , I , z represent the response of the 5 bands of the telescope. Run , rerun , camcol and field are features which describe a field within an image taken by the SDSS.

The class identifies an object to be either a galaxy , star or quasar. This will be the response variable which we will be trying to predict. Also the class column is the perfect target for classification practices.

**DATA ANALYSIS**

I tried to understand the dataset and analysis that feature and the class of the dataset are specify for solving the problem of dataset and they also work with the band of telescope feature. I analyzed the data and did the necessary transformation , normalization , on it. Then I did some visualization and implementation of supervised algorithms to classify given data wheather it is Star, Qussaer or Galaxy using many classification algorithms such as Decisiontree , Logisticregression , Randomforestclassifier. I checked the accuracys of them. I used these tools : Pandas , Matplotlib , Sklearn. So by the use of the algorithms and tools we identifies the accuracy of dataset and their and result with comparision of our target and features of dataset , so that analysis are helpful for understand target and implementation of dataset.

**EDA**

Naturaly I started with df.head() , df.describe() and df.info() that are uses for define indicate dataset , types , null values and column , shape , heatmap. Then I started working on define the correlation between the feature of dataset and after that I used heatmap for proper understand because in figure (heatmap) are with different color and that shows the correlation between attributes. The lighter the color of the cell gets the higher correlation it exists. Then remove the object identifiers (objid , specobjid) and check shape and the dataset without that features. I then run countplot on class column we see the value of Galaxy count is 4999 (approx) , Star count is 4001 (approx) , Qso count is 850 . We see that majority of the entries are either galaxies or stars. Only 8.5% of the entries are classified as quasars. we have change the category to number in classscat column then change category of class after that the use of pairplot for 'u' , 'g' , 'r' , 'i' they represent histogram and scatterplot graphs. removing the  'run' , 'rerun' , 'camcol' , 'field' , 'class' and break the values for represent x and y so that’s reason drop the value of classcat for x and use for y.

**Data Preprocessing and transformations.**

We understand In the section we built a stages to understand the preprocessing requirement for our data. It is now time to form a preprocessing pipeline design based on our learning from the section. We will define our preprocessing pipeline in three stages: We will use a transformer to do the required transformations. It will contain 3 steps.

1. Drop the columns that are not required for model training and but for getting the target it is done for required places.
2. We are changed values in the columns using the if and else rule , also types of data.
3. Understand the target because it include the three important feature of dataset.

That are the three important steps of preprocessing and these steps are also included with the ‘run’ the dataset we understand to see working in dataset.

**MODEL BUILDING**

Now that we are done with the basic pre-processing steps , we can go ahead and build simple machine learning models over this data. We will try models here – Logistic Regressor and Random Forest Regressor , DecisionTreeClassifier to predict the result. To compare the performance of the models, we will create a validation set (or test set). Here I have randomly split the data into two parts using the train\_test\_split()  function , such that the validation set holds of the data points while the train set then let we predict on the test data and predict the x and y on dataset measure the accuracy of the model and the decisiontreeclassifier prediction of test data randomforestclassifier are we also used for model fit and score, and confusion matrix are also their for y\_test.predict(x\_test) then svc(support vector classifier) using to show train model fit and score outputs s o that are the process are using to build model and that are understandable the target value.

**CONCLUDING REMARK**

In these dataset we will do the process train , test , fit and predict the model and all are the process we will understand the target dataset and also find the their accuracy and which logistic classifier are best suited for the dataset and also understand the target class and its feature which define galaxy , star , qso so I am trying to achieve our target because I tried to cover all the dataset to given output and reduce minimum data that’s reason trying to something new from my side and trying to achieve dataset.