## Space Server Dataset

Space Exploration has always been of interest to scientists and governments across the globe as it holds the key to the origin of mankind and many marvelous wonders of the universe including the possibility of alien lives. The visible universe represents the parts of space that we can see using telescopes. Yet, scientists and explorers do believe that the universe may be larger than that.

To date, scientists have explored roughly only 4% of the visible universe that is made up of planets, stars, galaxies, and other astronomical objects that astronomers and scientists can see and are aware of. Rest 96% is still left unexplored.

**Problem Definition**

I was looking for an unused and interesting dataset to improve my data science skills on when my professor mentioned the Sloan Digital Sky Survey which offers public data of space observations. As I found the data to be super insightful I want to share the data.

The data 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)

Right ascension (abbreviated RA) is the angular distance measured eastward along the celestial equator from the Sun at the March equinox to the hour circle of the point above the earth in question. When paired with declination (abbreviated dec), these astronomical coordinates specify the direction of a point on the celestial sphere (traditionally called in English the skies or the sky) in the equatorial coordinate system.

* u = better of DeV/Exp magnitude fit
* g = better of DeV/Exp magnitude fit
* r = better of DeV/Exp magnitude fit
* i = better of DeV/Exp magnitude fit
* z = better of DeV/Exp magnitude fit

The Thuan-Gunn astronomic magnitude system. u, g, r, i, z represent the response of the 5 bands of the telescope.

* run = Run Number
* rereun = Rerun Number
* camcol = Camera column
* field = Field number

Run, rerun, camcol and field are features which describe a field within an image taken by the SDSS. A field is basically a part of the entire image corresponding to 2048 by 1489 pixels. A field can be identified by:

* **run** number, which identifies the specific scan,
* the camera column, or "**camcol**," a number from 1 to 6, identifying the scanline within the run, and
* the **field** number. The field number typically starts at 11 (after an initial rampup time), and can be as large as 800 for particularly long runs.
* An additional number, **rerun**, specifies how the image was processed.
* specobjid = Object Identifier
* class = object class (galaxy, star or quasar object)

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.

* redshift = Final Redshift
* plate = plate number
* mjd = MJD of observation
* fiberid = fiber ID

In physics, **redshift** happens when light or other electromagnetic radiation from an object is increased in wavelength, or shifted to the red end of the spectrum.

Each spectroscopic exposure employs a large, thin, circular metal **plate** that positions optical fibers via holes drilled at the locations of the images in the telescope focal plane. These fibers then feed into the spectrographs. Each plate has a unique serial number, which is called plate in views such as SpecObj in the CAS.

**Modified Julian Date**, used to indicate the date that a given piece of SDSS data (image or spectrum) was taken.

The SDSS spectrograph uses optical fibers to direct the light at the focal plane from individual objects to the slithead. Each object is assigned a corresponding **fiberID**.

Please go through the project( <https://github.com/AshleshBR/Dynamic_projects-M20/blob/main/Dynamic%20project%208(space).ipynb>)

**Data Analysis**

After processing of data and exploring the various variables in the data set we can come to the conclusion that there are no null values and some of the columns have very low correlation with target so I have decided to drop those columns. Summary statistics shows that

* class column has 3 unique categories
* Most of the columns have high difference between mean and median
* There is a large difference between 75% and max values in columns like decso cfield, specobjid, redshift ,fiberid so chance of outliers in this coulmns is high

**Univariant analysis:**

lets start with univariant analysis of the variables which has high correlation

redshift: In physics, **redshift** happens when light or other electromagnetic radiation from an object is increased in wavelength, or shifted to the red end of the spectrum. From the analysis it is clear that redshift has very good correlation with target variable ,it has both positive and negative values but most of the values are positive

**Modified Julian Date**, used to indicate the date that a given piece of SDSS data (image or spectrum) was taken and it has found that good correlation with target variable. mjd value of

52000 has been recorded most number of times

# Bivariant analysis

The Thuan-Gunn astronomic magnitude system. u, g, r, i, z represent the response of the 5 bands of the telescope. In this data analysis we have compared the mean values of these bands with various classes and it is found that average values of all 5 bands are high for class 2.

Run, rerun, camcol and field are features which describe a field within an image taken by the SDSS.by the analysis of run with class we can conclude that most of the run values are recorded are class 1 ,class 2 has recorded least number of run values

# Multivariant analysis

Multivariant analysis deals with the analysis of more than 2 variable and how they affecting the target variable.We have done multivariate analysis on space object id and mjd for various classes and it is found the both space object id and mjd are linearly related to each other for each class.

**Pre-processing Pipeline**

Data pre-processing is a predominant step in machine learning to yield highly accurate and insightful results. Greater the quality of data, the greater is the reliability of the produced results. Incomplete, noisy, and inconsistent data are the inherent nature of real-world datasets. Data pre-processing helps in increasing the quality of data by filling in missing incomplete data, smoothing noise, and resolving inconsistencies.

* Incomplete data can occur due to many reasons. Appropriate data may not be persisted due to a misunderstanding, or because of instrument defects and malfunctions.
* **Noisy data** can occur for a number of reasons (having incorrect feature values). The instruments used for the data collection might be faulty. Data entry may contain human or instrument errors. Data transmission errors might occur as well.

There are many stages involved in data pre-processing,

**Data cleaning** attempts to impute missing values, smooth out noise, resolve inconsistencies, removing outliers in the data.

**Data integration** integrates data from a multitude of sources into a single data warehouse.

**Data transformations**, such as normalization, may be applied. For example, normalization may improve the accuracy and efficiency of mining algorithms involving distance measurements.

**Data reduction** can reduce the data size by dropping out redundant features. Feature selection and feature extraction techniques can be used.

Lets see how we have done our data pre-processing in this data set.

* Using the the box plot we can see that there are lots of outliers in various columns and hence we removed those outliers using the z score method
* Using the hist plot we can check the skewness. usually skewness + or – 0.5 is acceptable anything above that needs to be corrected. While removing the skewness it is important to consider only independent variable and not to consider categorical variable. By applying these rules we have applied power transformation function to remove the skewness.

Data standardization is this process of making sure that your data set can be compared to other data sets. It’s a key part of research, and it’s something that everyone who uses data should consider before they even collect, clean, or analyse their first data point.in our data set we have standardised the value using StandardScaler function,science there are not categorical independent feature we can apply for whole x variable.

Now we have already separated the feature variables and target variables data is now ready for model fitting

**Building Machine Learning Models**

Building machine learning models that have the ability to generalize well on future data requires thoughtful consideration of the data at hand and of assumptions about various available training algorithms. Ultimate evaluation of a machine learning model’s quality requires an appropriate selection and interpretation of assessment criteria.

At this stage, you develop an understanding of your problem which you are trying to solve. Now your data is also in its usable shape. Now it’s time to select and train your machine model. There are many models that you can select according to your business objectives. The step of selection of models includes algorithms of prediction, classification, clustering, deep learning, linear regression, and so forth. Now you will be required to train datasets to operate smoothly. The step of training your machine model involves several algorithms and techniques. The outcome machine model can be used for evaluation to check whether it meets the operational and business requirements.

In our case the problem type is classification type and we have imported and applied all the classification algorithm and picked the best result.we got the best result in logistic regression algorithm and we have checked for any over fitting using k fold cross validation technique. It is a technique for assessing how the results of a [statistical](https://en.wikipedia.org/wiki/Statistics) analysis will [generalize](https://en.wikipedia.org/wiki/Generalization_error) to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how [accurately](https://en.wikipedia.org/wiki/Accuracy) a [predictive model](https://en.wikipedia.org/wiki/Predictive_modelling) will perform in practice. In a prediction problem, a model is usually given a dataset of known data on which training is run (training dataset), and a dataset of unknown data (or first seen data) against which the model is tested (called the [validation dataset](https://en.wikipedia.org/wiki/Validation_set) or testing set). The goal of cross-validation is to test the model's ability to predict new data that was not used in estimating it, in order to flag problems like [overfitting](https://en.wikipedia.org/wiki/Overfitting" \o "Overfitting) or [selection bias](https://en.wikipedia.org/wiki/Selection_bias) and to give an insight on how the model will generalize to an independent dataset (i.e., an unknown dataset, for instance from a real problem).after performing the cross validation I have got nearly same accuracy as of trained data set hence I have proceeded to tune the parameters of algorithm and search for best parameter combination using grid search cv function. It is also called hyper parameter tuning Different model training algorithms require different hyper parameters . A hyper parameter is a [parameter](https://en.wikipedia.org/wiki/Parameter) whose value is used to control the learning process. By contrast, the values of other parameters (typically node weights) are learned. Grid search cv and randomized search are the two famous hyper parameters used.in our case I have used grid search and extracted the best combination of parameters for model.

After the parameter tuning not it is time to predict the data and compare the results with labelled data .the model is then evaluated using different evaluation metrics. This step involves the evaluation of the machine models using a model metric approach, quality measurements, datasets, and matrix calculations. This phase is the quality assurance of a machine learning approach. This is done with the help of metrics like accuracy, confusion metrics, classification report which involves precission,recall,f1 score etc.

**Concluding Remarks**

After getting the satisfied results we can save the model for production. I use pickle function to save the model

So this is how I built a model for space object detection, which can predict whether the object is star, galaxy or quasar. This prediction and model can be really helpful for scientist to study further on space objects.

Thanks you

Ashlesh B R