**Space Server**

**Problem Definition**

To build a classification model to predict whether a space object is a galaxy, star or quasar based on 17 different attributes.

**Data Analysis**

The Space Server dataset is provided by the Sloan Digital Sky Survey (SDSS) and it consists of 17 different feature columns based on which the object can be classified into star, galaxy or quasar. The dataset is a combination of two different types of data which are photometric data (PhotoObj) and spectral data (SpecObj).

The 18 columns of the dataset are as follows:

1. objid = Object Identifier

2. ra = J2000 Right Ascension (r-band)

3. 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.

4. u = better of DeV/Exp magnitude fit

5. g = better of DeV/Exp magnitude fit

6. r = better of DeV/Exp magnitude fit

7. i = better of DeV/Exp magnitude fit

8. z = better of DeV/Exp magnitude fit

-In the Thuan-Gunn astronomic magnitude system, u, g, r, i, and z represents the response of the 5 bands of the telescope.

9. run = Run Number

10. rereun = Rerun Number

11. camcol = Camera column

12. field = Field number

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

- Run number identifies the specific scan,

- The camera column, or "camcol," is a number from 1 to 6 used for identifying the scanline within the run.

- The field number starts at 11 (after an initial rampup time) and can be as large as 800 for particularly long runs.

- Rerun specifies how the image was processed.

13. specobjid = Object Identifier

14. class = object class (galaxy, star or quasar object)

15. redshift = Final Redshift

16. plate = plate number

17. mjd = MJD of observation

18. fiberid = fiber ID

- Redshift happens when light or other electromagnetic radiation from an object is increased in wavelength, or shifted to the red end of the spectrum.

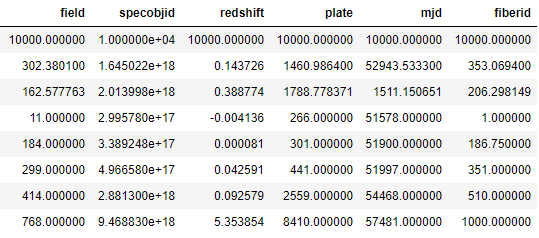
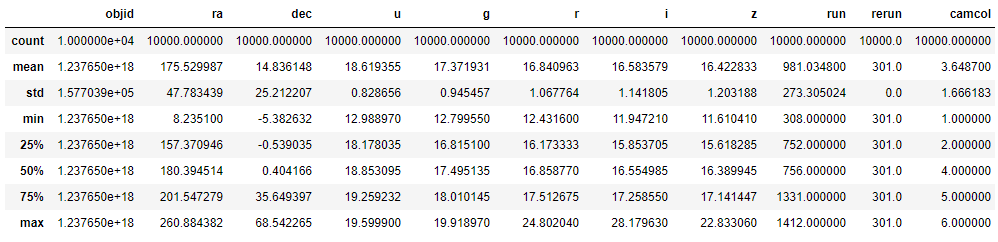
- Each spectroscopic exposure employs a large, thin, circular metal plate that positions optical fibers via holes drilled at the locations of the images in the telescope focal plane. These fibers then feed into the spectrographs. Each plate has a unique serial number.

- Modified Julian Date (MJD) is 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.

**EDA Concluding Remarks**

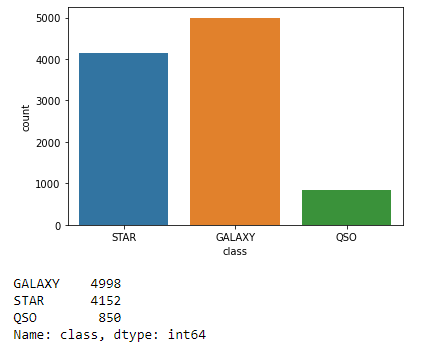
The dataset consists of total 18 columns and 10000 rows and have no null values present in it.



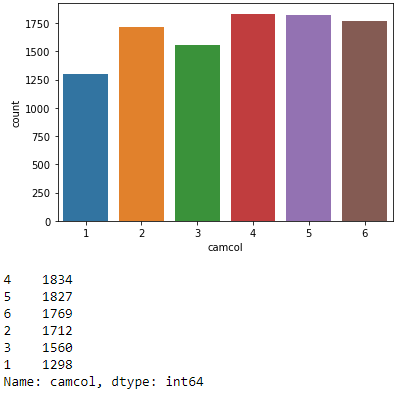
Following is the description of the dataset:

* The columns ‘objid’ and ‘rerun’ only consist of a single unique value throughout the column.
* The large difference between 75th percentile and maximum value of columns [ra, r, i, z, run, field, specobjid, reshift, plate, and fibergrid] suggests a possibility of outliers being present in them.

Univariate Analysis:

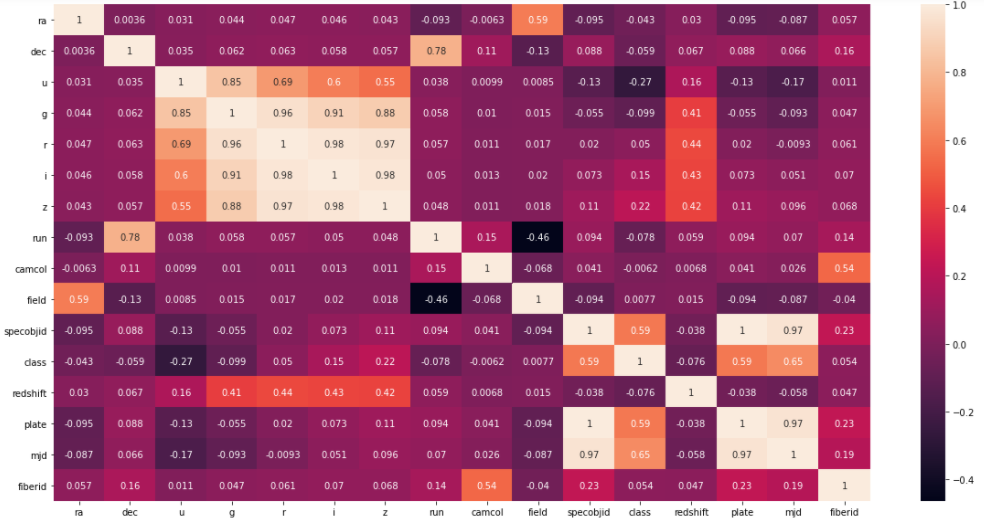


* Most of the space objects are a galaxy or a star.



* The camera columns from 1-6 or almost evenly distributed.

Checking Correlation:

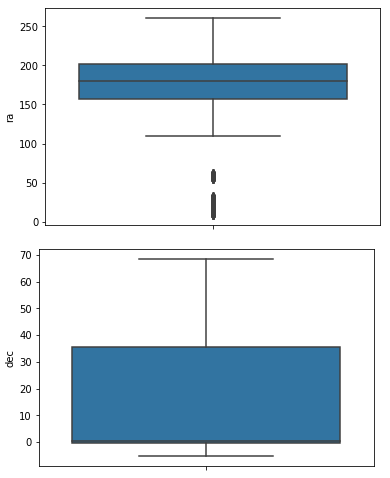
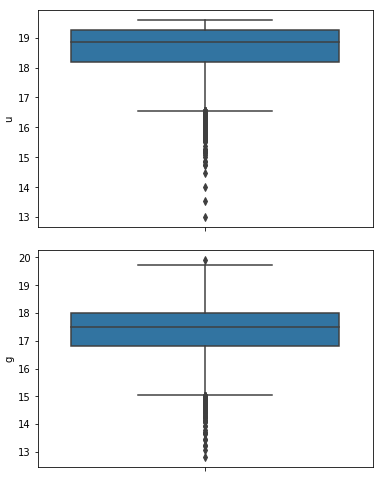
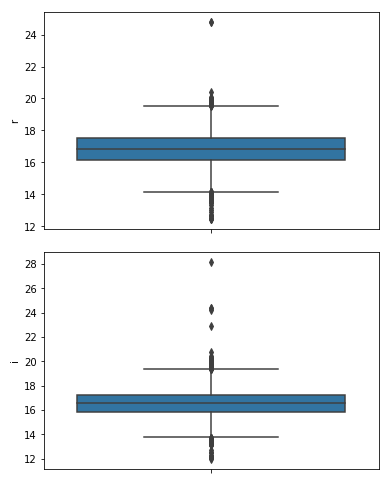


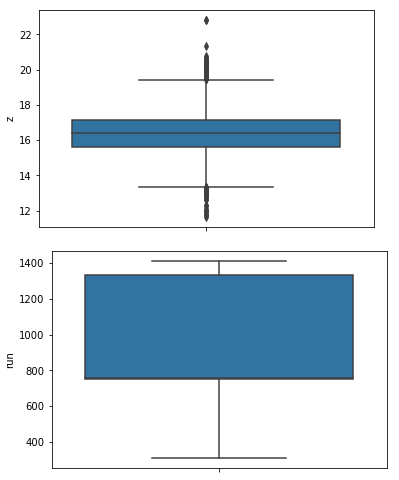
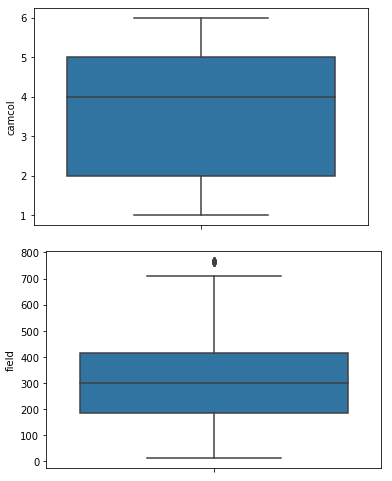
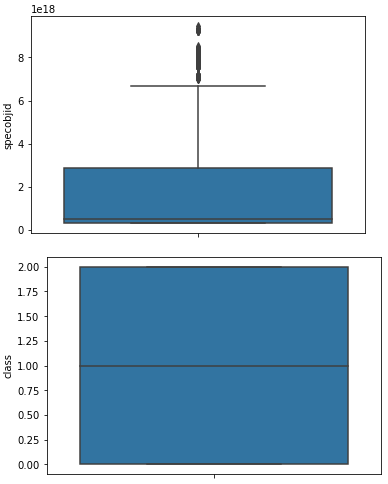
- 'specobjid', 'plate', and 'mjd' have high positive correlation with the target variable.

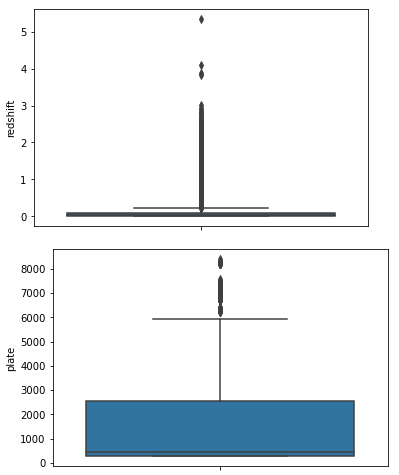
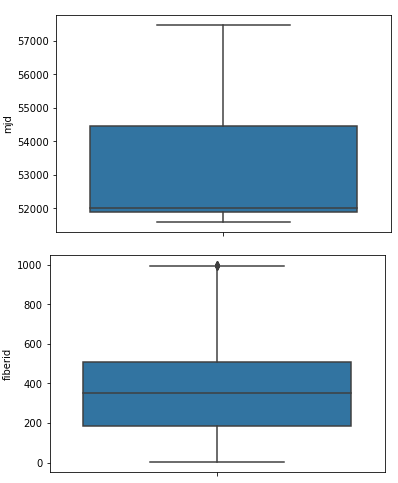
- 'u' has a moderate negative correlation with ‘class’.

- 'camcol' and 'field' have a very low correlation with target variable.

Checking for outliers using Boxplots:

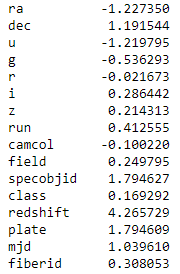
  

* The boxplots show that ‘u’, ‘i’, ‘z’, and ‘redshift’ have large number of outliers present in them.

Skewness:



* Columns ‘ra’, ‘dec’, ‘u’, ‘specobjid’, ‘redshift’, and ‘mjd’ have high skewness in the data.

**Pre-processing Pipeline**

1. First the least relevant columns are dropped from the dataset. This includes column ‘rerun’ and ‘objid’ as they contain only single unique value throughout and that cannot be used for the classification.

2. The target column ‘class’ is encoded using LabelEncoder into 0, 1, and 2 so that the correlation of the target column with the feature columns can be analyzed.

3. After checking the correlation, outliers, and skewness, it was determined that the column ‘redshift’ has large number of outliers, high skewness, and low correlation with the target column due to which, it was dropped from the dataset.

4. 4. In the next step, the target column i.e. ‘class’ is assigned to variable ‘y’ and the rest of the dataset is assigned to variable ‘x’.

5. The skewness in the dataset is removed using the method of ‘yeo-johnson’ of power transformation.

6. The data is standardized using StandardScaler as there was large variation in the ranges among the columns.

**Building Machine Learning Models**

As our target variable is of categorical type consisting of three different categories, the classifier algorithms are used to build the model. The accuracy of the model is determined using accuracy score, confusion matrix, and classification report. To check the over-fitting and under-fitting, the cross-validation score is used.

1. First a loop function is used to determine the best random state for the train-test split using the SVC algorithm. The function yielded the highest accuracy score of 93.59 at the random state of 259 for a test size of 20%.

2. For building the best model, several classifier algorithms were used which gave the following

|  |  |  |
| --- | --- | --- |
| Algorithm | Accuracy score | Cross-validation Score |
| SVC(kernel=’rbf’) | 0.935 | 0.904 |
| SVC(kernel=’linear’) | 0.956 | 0.95 |
| SVC(kernel=’poly’) | 0.911 | 0.885 |
| DecisionTreeClassifier | 0.903 | 0.861 |
| KNeighborsClassifier | 0.88 | 0.846 |
| RandomForestClassifier | 0.935 | 0.92 |
| AdaBoostClassifier | 0.894 | 0.863 |
| GradientBoostingClassifier | 0.932 | 0.907 |
| BaggingClassifier | 0.933 | 0.897 |
| ExtraTreesClassifier | 0.929 | 0.908 |

3. The SVC(kernel=’linear’) model showed the highest accuracy and cross-validation score due to which, it was further hypertuned using GridSearchCV for several parameters and it resulted in the final accuracy score of 95.97.

4. The hypertuned SVC(kernel=’linear’) model is saved using pickle method.

**Concluding Remarks**

1. The space objects can be classified into star, galaxy or quasar on the basis of the photometric and spectral data.
2. The most important features in the classification of the space objects are 'specobjid', 'plate', and 'mjd'.
3. The dataset is best fitted into a model using the Support vector classifier algorithm and hypertuning.