Classification of Galaxies, Stars and Quasars

Akshay Sharma

September 17, 2019

1. Introduction

1.1 Background

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. Sloan Foundation, which contributed significant funding. Data collection began in 2000 the final imaging data release (DR9) covers over 35% of the sky, with photometric observations of around nearly 1 billion objects, while the survey continues to acquire spectra, having so far taken spectra of over 4 million objects. The main galaxy sample has a median redshift of $z = 0.1$; there are redshifts for luminous red galaxies as far as $z = 0.7$, and for quasars as far as $z = 5$; and the imaging survey has been involved in the detection of quasars beyond a redshift $z = 6$.

1.2 Problem

Data that might contribute to determining a Star, Galaxy or a Quasar might include previous redshift data for each class and metrics that describe what kind of class that portion in space is. This project aims to predict whether the observed portions in space is a Star, Galaxy or Quasar based on these data.

1.3 Interest

Obviously, various astronomers would be very interested in accurate prediction of a portion in space as a Star/Galaxy/Quasar, for observational, academic and research values. Others who are interested in astronomy (as a hobby or out of curiosity) may also be interested.

2. Data Acquisition and Cleaning

2.1 Data Sources

The SDSS data is available publicly but requires CasJobs (http://skyserver.sdss.org/casjobs/) to actually retrieve data and create our own database. However, the dataset doesn’t lack any data.

2.2 Data Cleaning & Feature Selection

Data downloaded from the SDSS website using CasJobs was combined into one table. There wasn't any single missing value found because of responsible record keeping. Therefore, data was used as is and data cleaning was not a necessity.

After data cleaning, there were 10,000 samples, 17 features and 1 target column in the dataset. Upon examining the meaning of each feature, it was clear which variable were unlikely to be related to the target variable ‘class’. For example, Photo object identifier (objid) and Space object identifier (specobjid) are just identifiers that are for accessing the rows back when they were stored in the original databank. Hence, they can be ignored for classification as they are not related to the outcome. Moreover, the features 'run', 'rerun', 'camcol' and 'field' are values that are used to describe parts of the camera at the moment in time when the telescope was making the observation (like, 'run' represents the corresponding scan which captured the object).

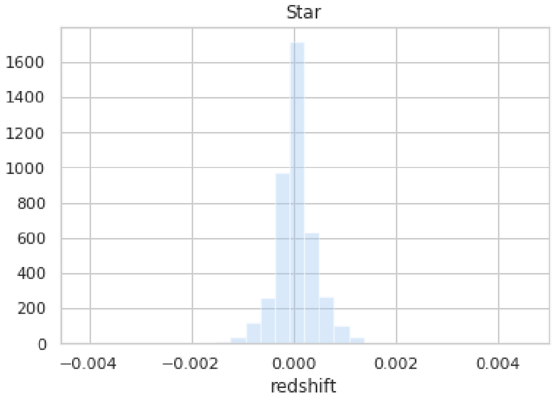
After discarding redundant features, I inspected the correlation of independent variables, and found several pairs that were highly correlated (Pearson correlation coefficient > 0.9). For example, u, g, r, i, z — filter bands (a.k.a. photometric system or astronomical magnitudes) and class were highly correlated. This makes sense, after all, different light sources would emit different filter bands. In the end, 10 features were selected (including the target column).

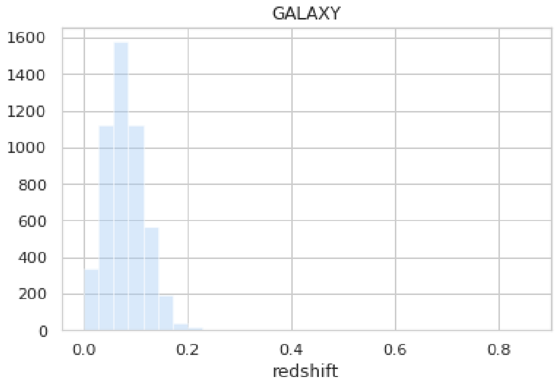
Kept Features	Dropped Features
1. Filter bands – u, g, r, i, z 2. redshift – Final Redshift 3. plate – Plate Number 4. mjd – MJD of Observation 5. fibered – Fiber ID	1. ra - J2000 Right Ascension (r-band) 2. dec – J2000 Declination (r-band) 3. objid – Photo object identifier 4. run – Run Number 5. rerun – Rerun Number 6. camcol – Camera Column 7. field – Field Number 8. specobjid – Space object identifier

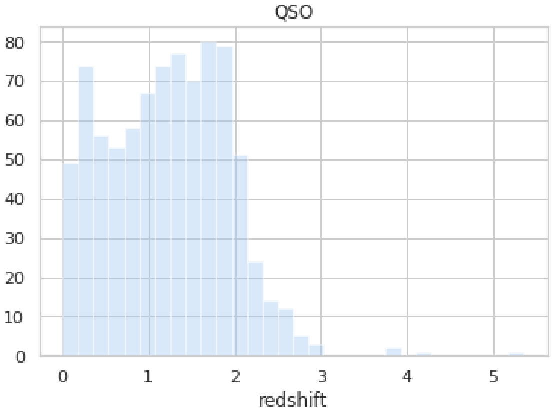
3. Exploratory Data Analysis

3.1 Relation between redshift for different classes

To understand the relation between the redshift distributions over their range, histograms for the redshift feature of each class was plotted.







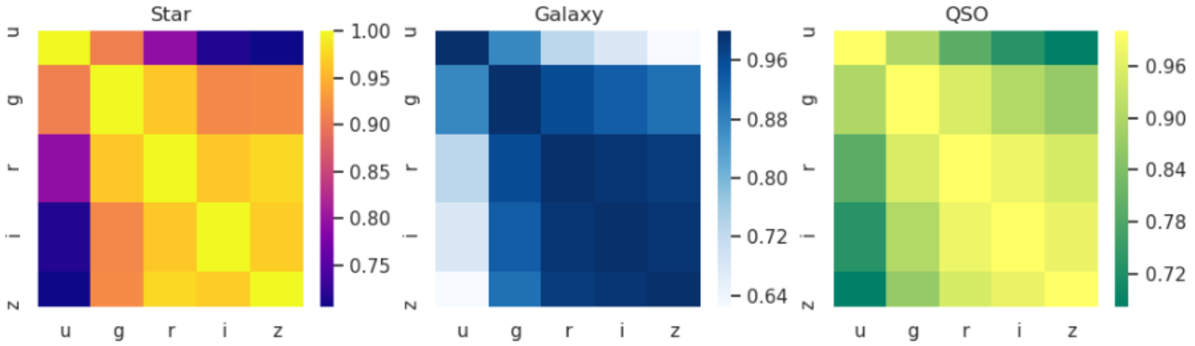
From the above plots, it can be concluded that the redshift values for different classes vary a lot. As we can see from the plots, most of the stars were observed somewhat closer to the earth than either the galaxies or quasars. Galaxies tend to be a little further away and quasars are quite distant from very close to very far.

Possibly, due to the size of the galaxies and quasars in comparison to the stars they could be observed from very far because of the galaxy and quasars emitting strong radiations.

Hence, redshift is a very valuable value that can help us distinguish apart stars, galaxies and quasars.

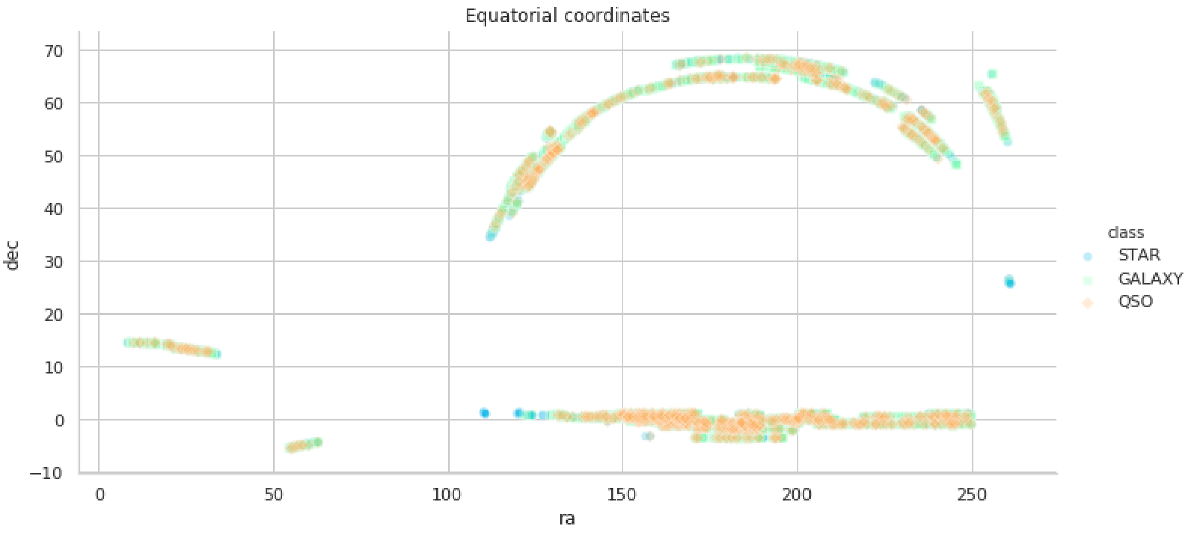
3.2 Correlation between different filter bands (u, g, r, i, z) for different Classes

At the right of the top of the correlation matrix, we observe that the correlation matrices look very similar for each class.
It can be inferred from the correlation matrices is the fact that there is a very high correlation between different filter bands (i.e. u, g, r, i, z). This was expected to some extent. Although, it can also be inferred that the filter band ‘u’ is less correlated to the other filter bands, which to be honest is quite astonishing.



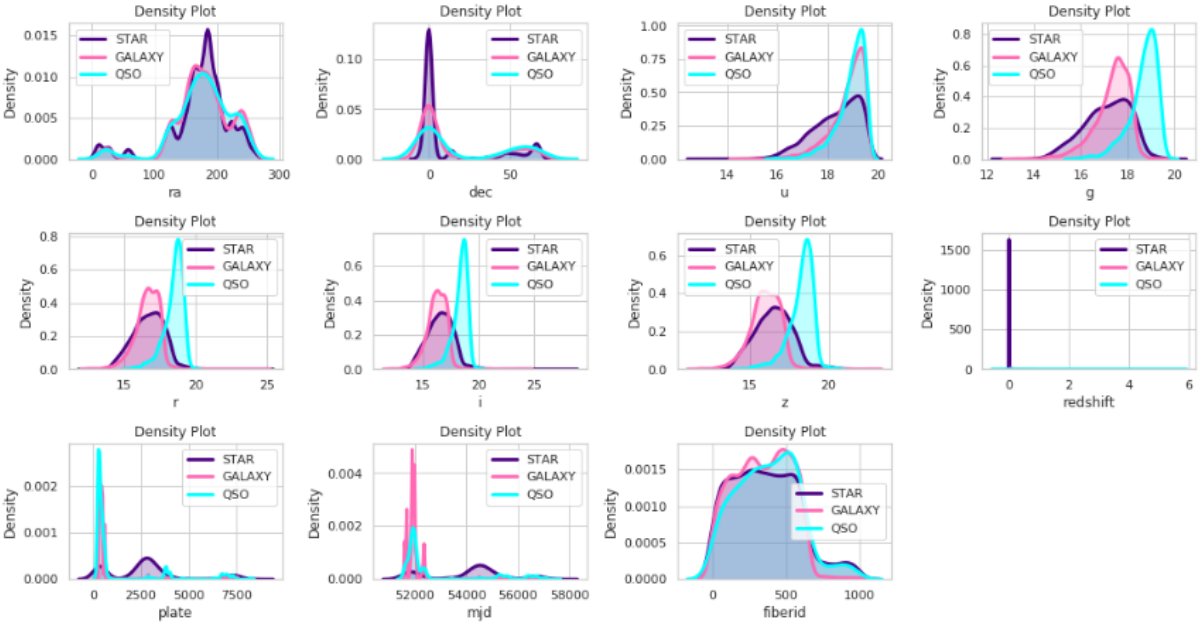
3.3 Relationship between Right Ascension (ra) and Declination (dec)

To assess the relation between the Right Ascension (ra) and Declination (dec), a 2D scatter plot was plotted.



The main point to note here is that the equatorial coordinates do not differ significantly between stars, galaxies and quasars. This can be due to the fact that the SDSS telescope covers the same area of the space and the stars, galaxies and quasars are observed equally at all coordinates in this area.

3.4 Density Distribution Plots of the features for different classes



The above density distribution plots of different features for different classes provides a great overview of the relationship of a particular feature for the different classes.

4. Data Transformations and Preparation

The quantity of dimensions in the dataset were reduced by replacing the different filter band values – u, g, r, i, z by only 3 dimensions (PCA_1, PCA_2, PCA_3) using Principal Component Analysis.

Principal Component Analysis: Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called principal components.

PCA's key advantages are its low noise sensitivity, the decreased requirements for capacity and memory, and increased efficiency given the processes taking place in a smaller dimensions.

5. Predictive Modelling

There are two types of models, regression and classification, that can be used to predict the class of an object. Regression models can provide additional information on the class of the object, while classification models focus on the probabilities of the class of the object. The underlying algorithms are similar between regression and classification models, but different audience might prefer one over the other. For example, a profession astronomer might be more interested in the amount of additional information of the object (regression models), but a general hobby astronomer might find the results of classification models more interpretable. Therefore, in this study, I carried out both regression and classification modelling.

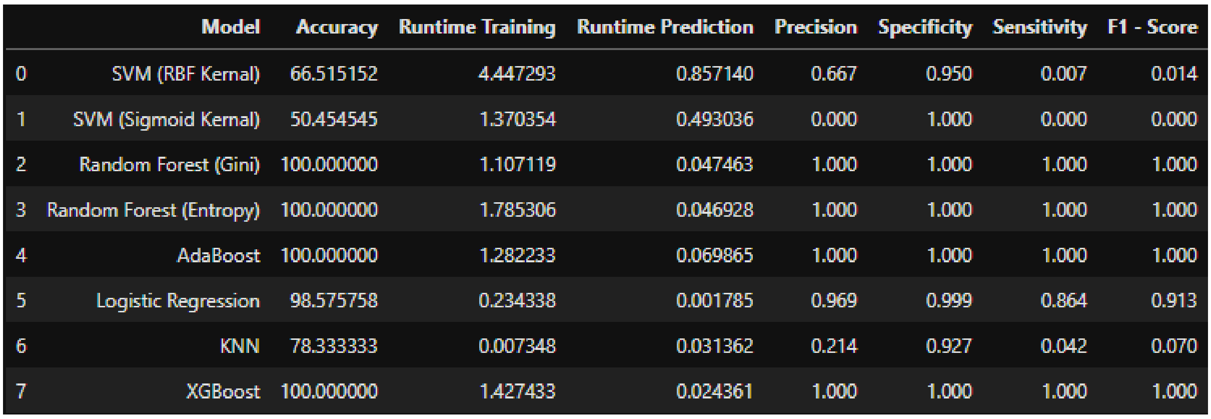
5.1 Regression Models

5.1.1 Applying standard algorithms and their problems (with solution)

I applied regression models (Logistic Regression), Support Vector Machines (SVM), Random Forest Classifier, K Nearest Neighbours, and gradient boost models to the dataset. Some models gave 100% accuracy, while it can be true that the model was trained very well as one might say but it is highly unlikely. This 100% accuracy might have been obtained due to the simple case of over-fitting data. Regardless of that, my primary focus was on the model evaluation of XgBoost and AdaBoost.

5.1.2 Performances of different models

The table below shows the different parameters of the different machine learning algorithms used, thus, giving us a fair idea of the accuracy, precision, sensitivity, specificity and F1 score of a particular model which is important in determining the best model for the classification of the objects observed in the space into stars, galaxies or quasars.

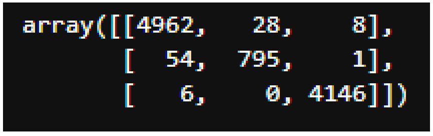


	Model	Accuracy	Runtime Training	Runtime Prediction	Precision	Specificity	Sensitivity	F1 - Score
0	SVM (RBF Kernal)	66.515152	4.447293	0.857140	0.667	0.950	0.007	0.014
1	SVM (Sigmoid Kernal)	50.454545	1.370354	0.493036	0.000	1.000	0.000	0.000
2	Random Forest (Gini)	100.000000	1.107119	0.047463	1.000	1.000	1.000	1.000
3	Random Forest (Entropy)	100.000000	1.785306	0.046928	1.000	1.000	1.000	1.000
4	AdaBoost	100.000000	1.282233	0.069865	1.000	1.000	1.000	1.000
5	Logistic Regression	98.575758	0.234338	0.001785	0.969	0.999	0.864	0.913
6	KNN	78.333333	0.007348	0.031362	0.214	0.927	0.042	0.070
7	XGBoost	100.000000	1.427433	0.024361	1.000	1.000	1.000	1.000

Accuracy: It is the fraction of predictions that our model got right.
Precision: It expresses the portion of the data points our model says was relevant was actually relevant.
Specificity: It is the measure of the portion of actual negatives that were correctly identified as such (eg. An object not being classified as, say a star, did not actually belong to that class).
Sensitivity: It is a measure of the proportion of actual positive cases that got predicted as positive (eg. An object being classified as a, say star, actually being a star).

XgBoost Classifier and AdaBoost classifier were further compared using K Fold Cross validation and by using cross value preciction model to determine the best model for the classification of objects in space as stars, galaxies or quasars.

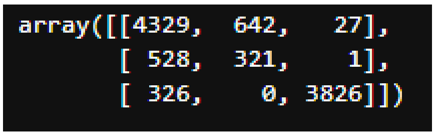
XgBoost Model Evaluation Results:



```
array([[4962,   28,    8],
       [  54,  795,    1],
       [   6,    0, 4146]])
```

From above it can be inferred that:
--> 4962 correct STAR predictions
--> 795 correct Galaxy predictions
--> 4146 correct Quasar predictions,
were done by the XGboost Model.
There were 96 incorrect predictions done by the XGBoost Model too.
Therefore, it can be said that almost all of the predictions made by the XGBoost Model were correct.

AdaBosst Model Evaluation Results:



```
array([[4329,  642,   27],
       [ 528,  321,    1],
       [ 326,    0, 3826]])
```

From above it can be inferred that:
--> 4328 correct STAR predictions
--> 321 correct Galaxy predictions
--> 3826 correct Quasar predictions,
were done by the AdaBoost Model.

There were 1525 incorrect predictions done by the AdaBoost Model too.

Therefore, it can be said that there are substantial amount of wrong predictions done by the AdaBoost Classifier in comparision to the XgBoost Classifier.

Hence, it can be concluded that XGBoost is the best suited classifier for this task.

6. Conclusion

I choose XGboost model as my preferred my approach. The XGBoost model not only has the highest accuracy, but also has the highest correct prediction of all the models developed as a part of this analysis. The advantages of using XGBoost Classifiers over other models is that they are very simple to implement. Since they are made up of weak individual learners, they are less susceptible to overfitting. However, XGBoost is sensitive to noisy data and outliers.