Module 5 Project

Astronomical Classification

Identifying Objects with Machine Learning

Matthew Onstott

Table of Contents

Problem

How can we determine if an object is a Star, Galaxy or Quasar?

A **galaxy** is a gravitationally-bound system of stars, stellar remnants, interstellar gas, dust, and dark matter.

A **star** is a luminous spheroid of plasma held together by gravity.

A **quasar**, or quasi-stellar object (**QSO**), is an active galactic nucleus that is extremely luminous and emits an extremely large amount of energy.

Dataset

- Observations = 10,000
- Features = 17
- Class Levels = 3 (*Galaxy*, *Star*, *Quasar*)

Space observations from **Data Release 14** of the **Sloan Digital Sky Survey**.

The **Sloan Digital Sky Survey** is a major multi-spectral imaging and spectroscopic redshift survey that has created the most detailed three-dimensional maps of the Universe that currently exist.

It operates from Apache Point Observatory in New Mexico, USA, and has been in service since 2000.

In its lifetime it has investigated more than one-third of the sky and taken spectra for more than three million astronomical objects.

Background

1. **objid** = object identifier (within the photometric data)
2. **ra** = J2000 right ascension (measured in the r-band)
3. **dec** = J2000 declination (measured in the r-band)
4. **u** = filter with wavelength 3,551Å and magnitude limit 22.0
5. **g** = filter with wavelength 4,686Å and magnitude limit 22.2
6. **r** = filter with wavelength 6,165Å and magnitude limit 22.2
7. **i** = filter with wavelength 7,481Å and magnitude limit 21.3
8. **z** = filter with wavelength 8,931Å and magnitude limit 20.5
9. **run** = run number
10. **rerun** = rerun number
11. **camcol** = camera column
12. **field** = field number
13. **specobjid** = object identifier (within the spectral data)
14. **class** = identifies an object as a galaxy, star, or quasar
15. **redshift** = redshift of the object
16. **plate** = plate number
17. **mjd** = modified julian date
18. **fiberid** = fiber identification number

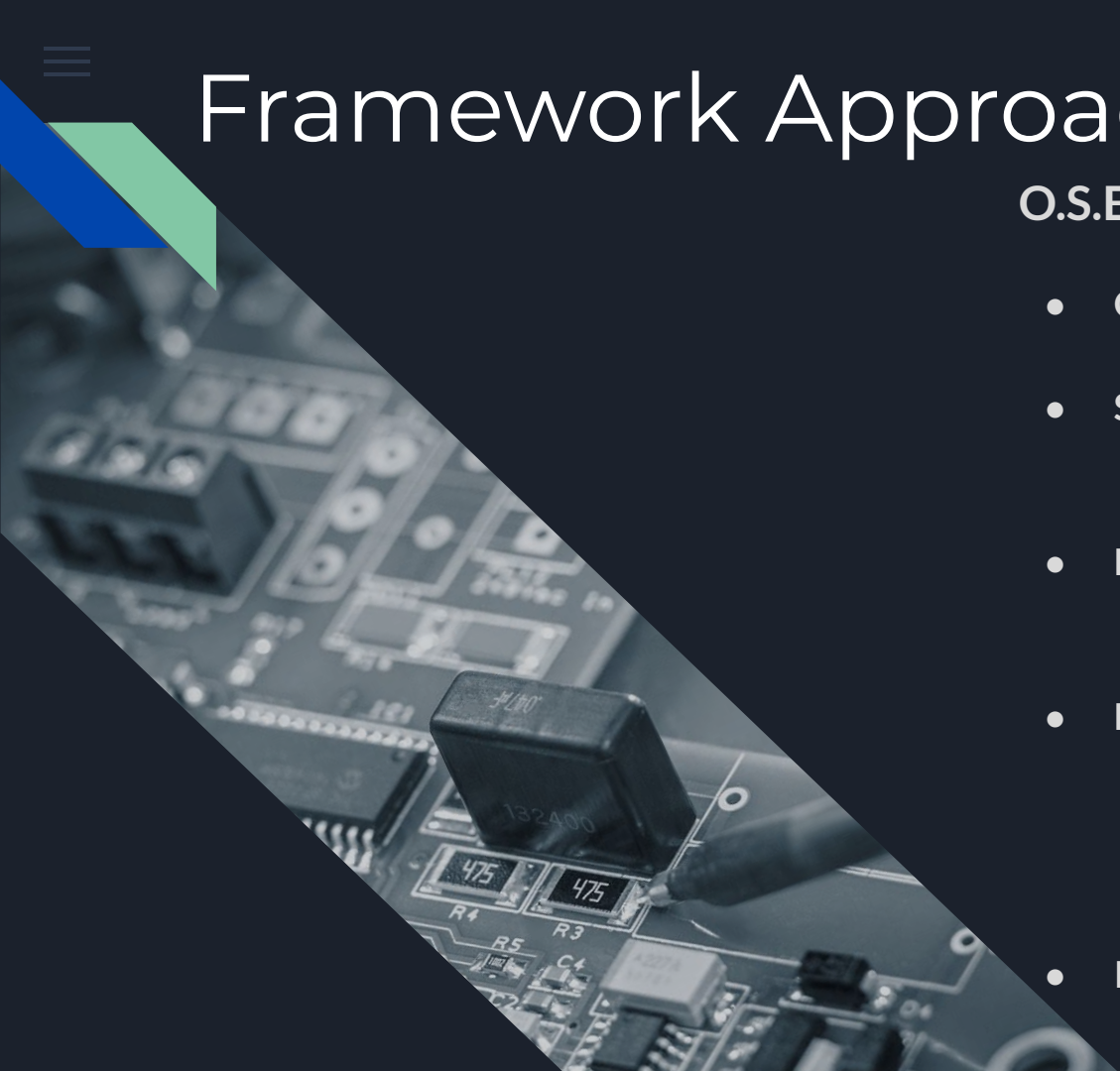
Together, **right ascension** and **declination** specify the astronomical coordinates of a point in space (on the celestial sphere in the equatorial coordinate system). **J2000** is the current standard epoch used to correct for precession of the Earth's rotation.

SDSS employs 5 **filters** designed to let in light around specific wavelengths. The imaging camera collects photometric imaging data using an array of 30 charge-coupled device (CCD) cameras arranged in six columns of five CCDs each.

A **field** is segment of an entire SDSS image with size 1361x2048 pixels. **Run** number identifies the specific scan of an image. **Camcol** identifies the scanline within the run. **Rerun** specifies how an image was processed.

Redshift occurs when electromagnetic radiation from an object increases in wavelength. Aluminum **plates** placed in the focal plane of the telescope measure spectra for a specific patch of sky. Modified julian date (**MJD**) is an integer corresponding with the night of observation. Optical **fibers** bring an object's light from the telescopic focal plane to the pseudo-slit of the spectrographs.

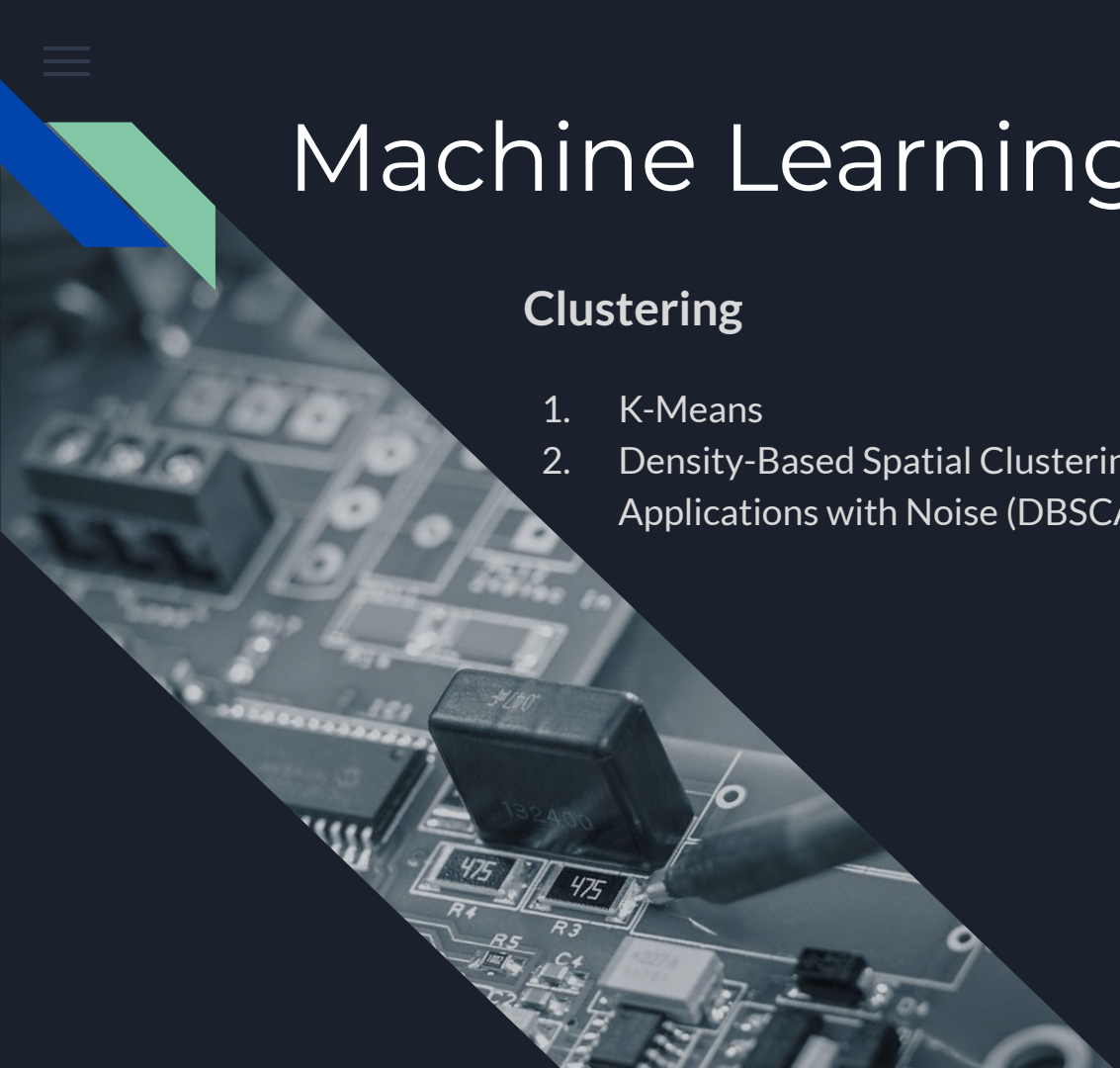
Framework Approach



O.S.E.M.N.

- **Obtain**
  - Gather data using the kaggle API
- **Scrub**
  - View metadata, check data types, handle missing values & duplicates
- **Explore**
  - Derive statistics, investigate relationships, and create visualizations
- **Model**
  - Set baseline expectations with clustering algorithms and build & tune predictions with classification algorithms
- **Interpret**
  - communicate the meaning of results

Machine Learning Solutions



Clustering

1. K-Means
2. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

Classification

Single Models

1. K-Nearest Neighbors (KNN)
2. Logistic Regression
3. Support Vector Machine (SVM)
4. Naive Bayes
5. Decision Tree

Ensemble Models

6. Random Forest
7. AdaBoost
8. Gradient Boosting

Obtain and Scrub Phases

Obtain

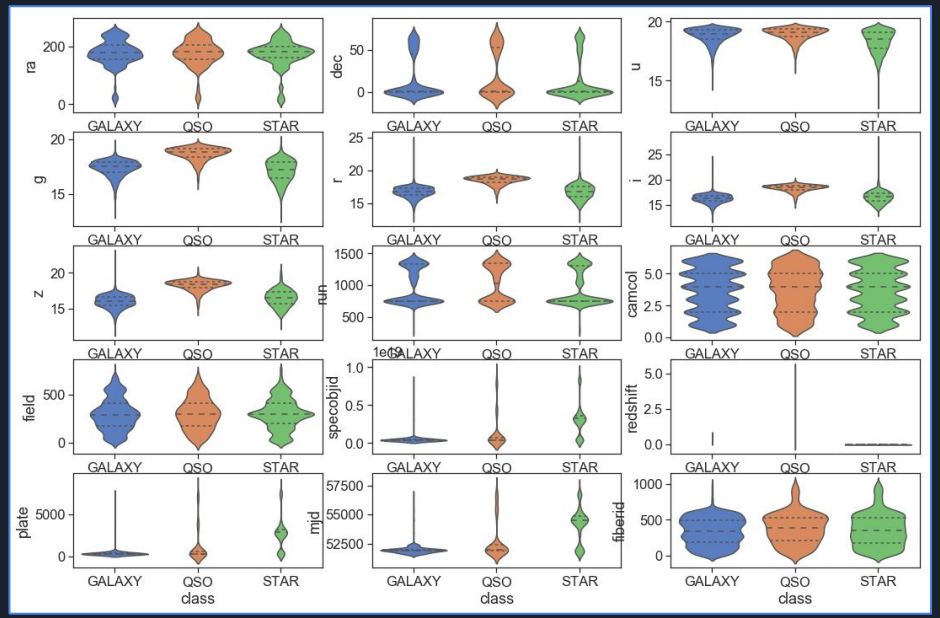
```
!kaggle datasets download -d lucidlenn/sloan-digital-sky-survey

!unzip sloan-digital-sky-survey.zip
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 18 columns):
objid         10000 non-null float64
ra            10000 non-null float64
dec           10000 non-null float64
u             10000 non-null float64
g             10000 non-null float64
r             10000 non-null float64
i             10000 non-null float64
z             10000 non-null float64
run           10000 non-null int64
rerun         10000 non-null int64
camcol        10000 non-null int64
field         10000 non-null int64
specobjid     10000 non-null float64
class         10000 non-null object
redshift      10000 non-null float64
plate         10000 non-null int64
mjd           10000 non-null int64
fiberid       10000 non-null int64
dtypes: float64(10), int64(7), object(1)
memory usage: 1.4+ MB
```

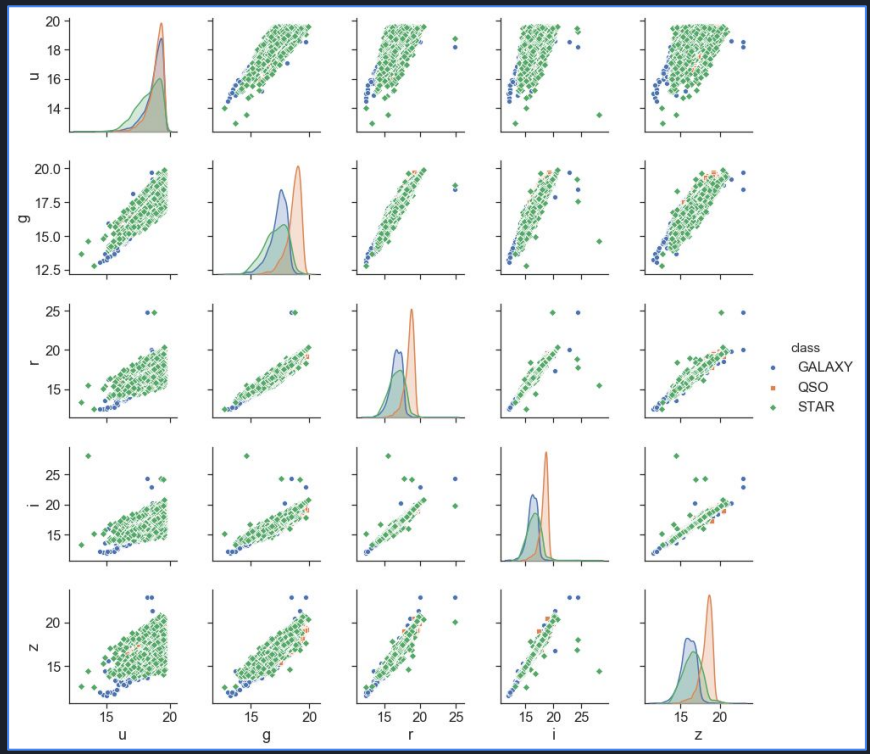
Scrub

- Observations
  - **Galaxy** = 4,998 (49.98%)
  - **Star** = 4,152 (41.52%)
  - **QSO** = 850 (8.50%)
- Data Quality
  - Incorrect Data Types = 0
  - Missing Values = 0
  - Duplicate Rows = 0
- Mean Range
  - redshift = 0.14
  - objid = 1.24e18

Explore Phase



Boxplots and Kernel Density Estimates by Object



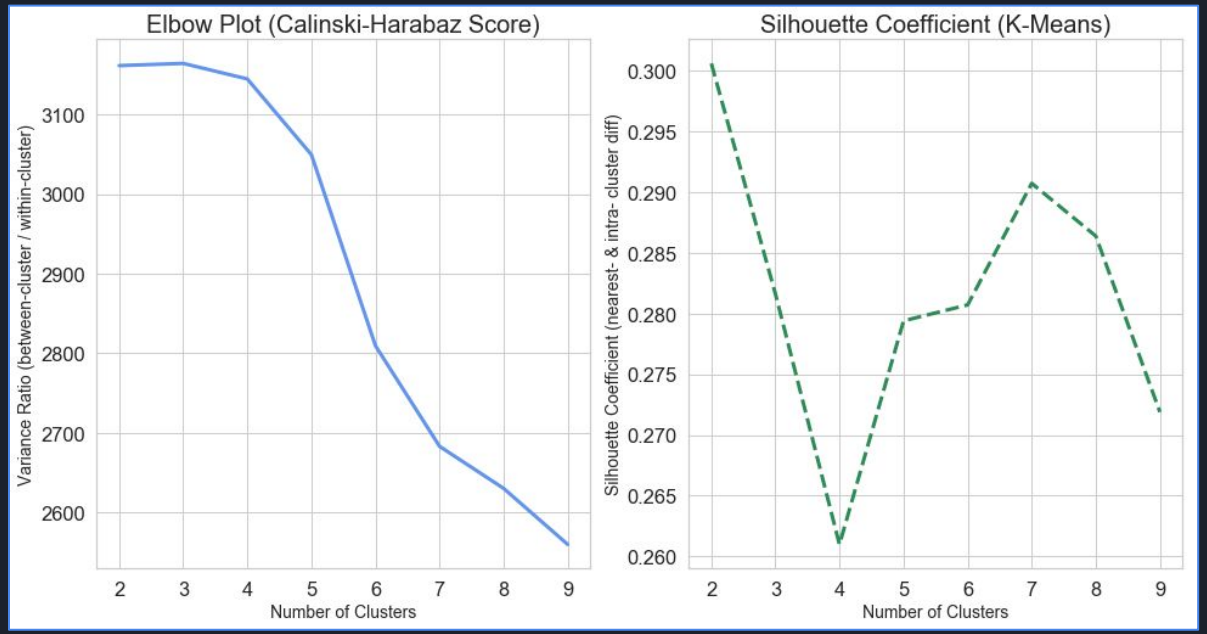
Pairwise Scatterplots and Distributions by Class

Model Phase

- **Metrics**
  - **Accuracy**: The ratio of the number of correct predictions over the total number of predictions.
  - **F1 Score:** the average of precision and recall weighted by the number of instances for each object .
    - **Precision** is the ability of a classifier not to label a negative sample as positive.
    - **Recall** is the ability of a classifier to find all positive samples.
- **Data Splits**
  - **Clustering**
    - The entire dataset is used within the algorithm.
  - **Classification**
    - The model is split into training (90%) and testing (10%) sets.
    - 10-Fold Cross-Validation is used within Grid Search on the training set to tune model hyperparameters.
    - Final model evaluation is conducted on the testing set.

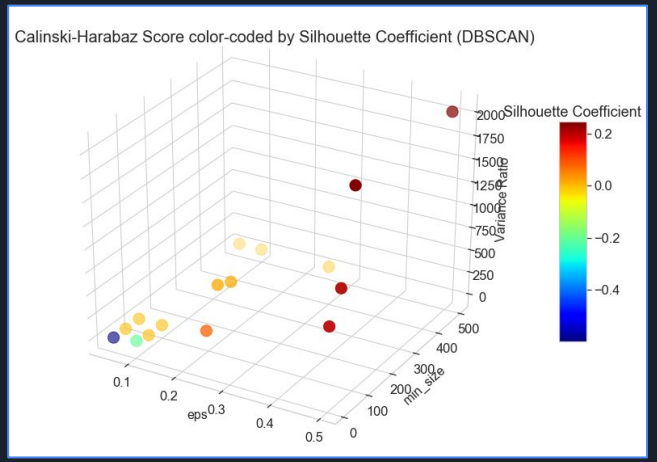
Clustering Algorithms

K-Means



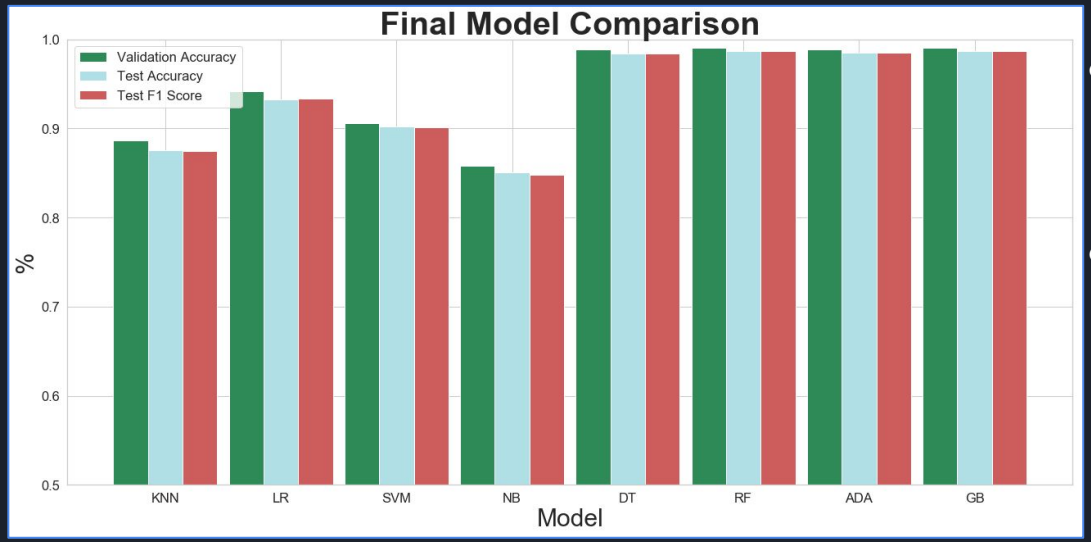
The maximum Variance Ratio in K-Means Clustering is for 3 clusters which is consistent with the problem context and expectations.

DBSCAN



Observations within the feature space are sparsely separated according to DBSCAN Clustering. Performance is shown to improve with an increase in between-points distance.

Classification Algorithms

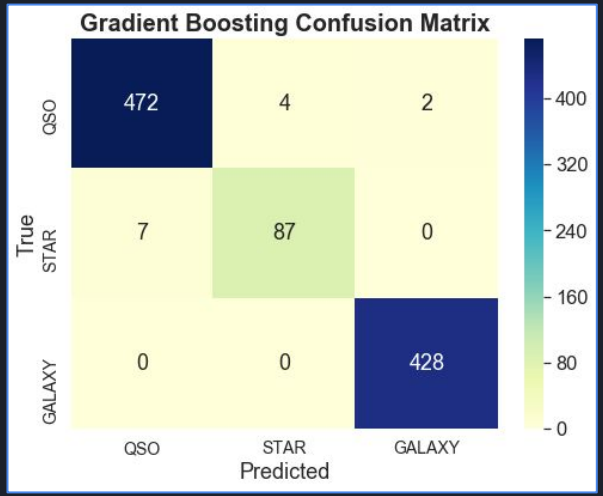


Comparison of results for 5 single and 4 ensemble machine learning models for each performance metric.

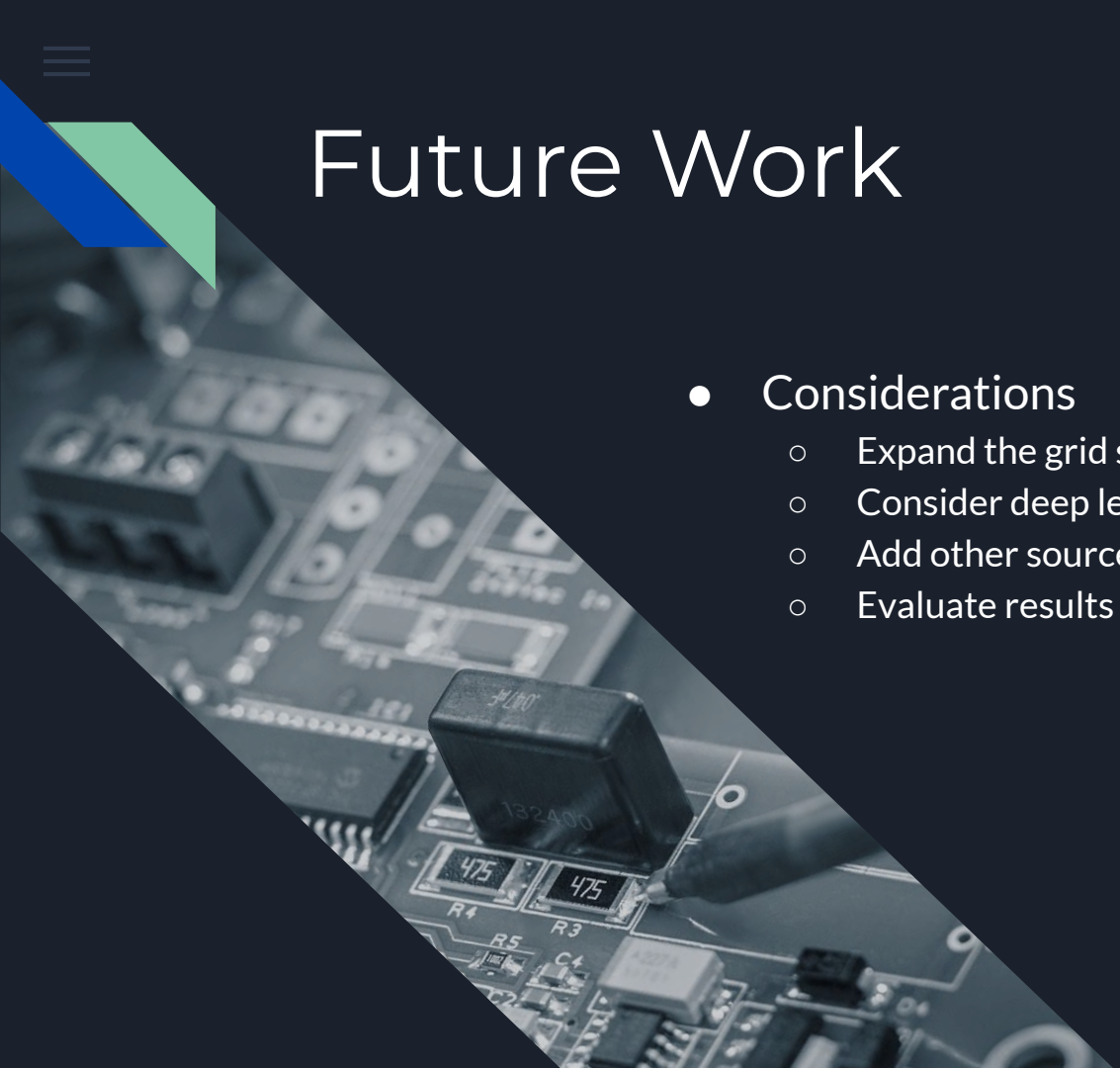
- **Highest Metrics**
  - **Validation Accuracy**: Decision Tree, Random Forest, AdaBoost, Gradient Boosting (99%)
  - **Test Accuracy**: Random Forest, Gradient Boosting (98%)
  - **Test F1 Score**: Random Forest, Gradient Boosting (98%)
- **Lowest Metrics**
  - **Validation Accuracy**: Naive Bayes (86%)
  - **Test Accuracy**: Naive Bayes (85%)
  - **Test F1 Score**: Naive Bayes (85%)
- **Effects of Class Imbalance**
  - **KNN**: **QSO** level has low recall (74%)
  - **Naive Bayes**: **QSO** and **Star** levels have low recall (72% and 74%, respectively)
- **Misclassification**
  - Highest for **Star** samples with **Galaxy** predictions in Naive Bayes (112), SVM (71), KNN (64), and Logistic Regression (32)
  - Highest for **QSO** samples with **Galaxy** predictions in Decision Tree (10), AdaBoost (8), Random Forest (8), and Gradient Boosting (7)

Recommendation

- **Model**
  - Gradient Boosting
- **Rationale**
  - Superior performance on new data
  - Able to correctly identify observations in each class level
  - Adjusts to hard-to-classify observations
  - Tuning flexibility



Future Work



- Considerations
  - Expand the grid search for hyperparameters
  - Consider deep learning algorithms
  - Add other sources and transform predictors
  - Evaluate results on data from other time periods