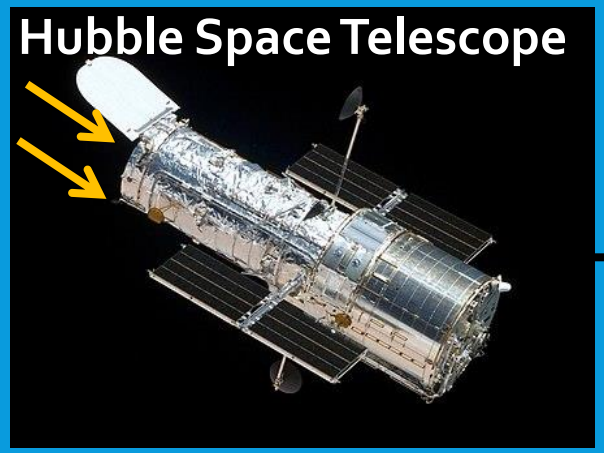
QuGaSt: Classification of Observed Astronomical Objects



?

?

Joseph P. Martin
Phase 2, Module 22: Final Capstone
May 5, 2022

Introduction

- The James Webb Space Telescope (JWST), launched in December 2021, is renewing interest in learning about remote objects in space. The JWST will make measurements in a section of the electromagnetic spectrum that extends from visible red light to non-visible infrared radiation. Compared to the Hubble Space Telescope, which makes measurements from the near-ultraviolet to near-infrared section of the band, the JWST's enhanced infrared measurements will allow it to observe older and more distant objects.

- There are many new observations of astronomical objects on the way, which will not yet have a classification. This will add to any existing observed but not yet classified objects.

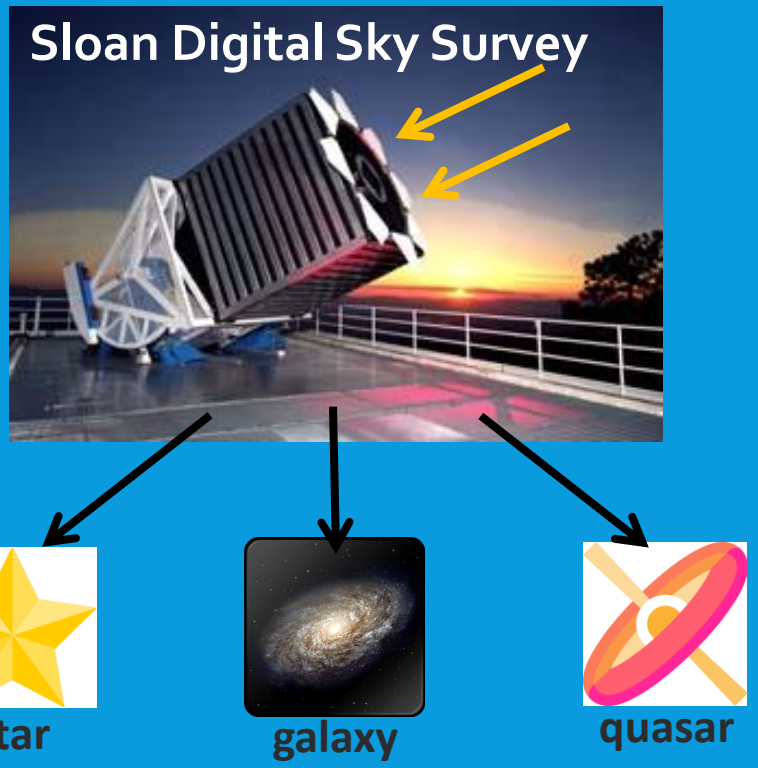
- The problem we aimed to solve in this project is: Can one quickly determine the classification of an observed object of unknown class at a high accuracy? Is it a star, galaxy (gravitationally bound system of stars), or quasar (extremely bright active galactic nucleus)?

- The Sloan Digital Sky Survey (SDSS) provides a catalog for astronomical objects with both their optical properties and classifications (star, galaxy, or quasar). We use a subset of this catalog to develop our product.

- The product of this effort is a model framework that can be customized for each telescope observing the objects that need to be classified. Different models were explored to determine the best one, which is the product model, QuGaSt (QuasarGalaxyStar).

- The end users would be the astrophysicists and other scientists who need to quickly determine an observed object's classification or who have to do this for many unclassified objects.

Dataset for Building Model Product



- The Sloan Digital Sky Survey (SDSS) was made with a telescope located on-the-ground in New Mexico.
- SDSS provides the Catalog Data set that summarizes properties obtained from camera images and spectra, which show how energy is distributed with electromagnetic frequency.
- Built model product using a 16 MB subset of the Catalog Data with 100,000 object observations, 17 feature columns, and 1 classification column (star, galaxy, or quasar). There is much more data available: image data for 0.5×10^9 objects and spectra data for 3×10^6 objects.

Feature Column	Description (all are numerical)
obj_ID	object Identifier, the unique value that identifies the object in the image catalog
alpha	right ascension angle (at J2000 epoch, a measure of east-west position)
delta	declination angle (at J2000 epoch, a measure of north-south position)
u, g, r, i, & z	magnitude measured in each bandpass filter in increasing wavelength order: u = ultraviolet, g = green, r = red, i = near infrared, z = infrared
run_ID	the run number that identifies a specific scan; 1 of 3 numbers identifying image
rerun_ID	number that specifies how the image was processed
cam_col	the camera column that identifies scanline # among 6 parallel scanlines; 1 of 3 numbers identifying image
field_ID	the field number that identifies a patch of 10 × 13 arcminutes of sky; 1 of 3 numbers identifying image

Dataset for Building Model Product


star


galaxy


quasar

- The Sloan Digital Sky Survey (SDSS) was made with a telescope located on-the-ground in New Mexico.
- SDSS provides the Catalog Data set that summarizes properties obtained from camera images and spectra, which show how energy is distributed with electromagnetic frequency.
- Built model product using a 16 MB subset of the Catalog Data with 100,000 object observations, 17 feature columns, and 1 classification column (star, galaxy, or quasar). There is much more data available: image data for 0.5×10^9 objects and spectra data for 3×10^6 objects.
- Observations from March 2000 to March 2020.

Feature Column	Description (all are numerical)
spec_obj_ID	unique ID used for optical spectroscopic objects
redshift	redshift is an increase in the wavelength, and corresponding decrease in the frequency and photon energy, of electromagnetic radiation (fractional increase in wavelength)
plate	integer indicating which SDSS plug plate was used to collect the spectrum; 1 of 3 numbers identifying spectrum
MJD	integer for modified Julian date of the night when the observation was carried out; 1 of 3 numbers identifying spectrum
fiber_ID	integer denoting the optical fiber number; 1 of 3 numbers identifying spectrum

Data Exploration and Cleaning

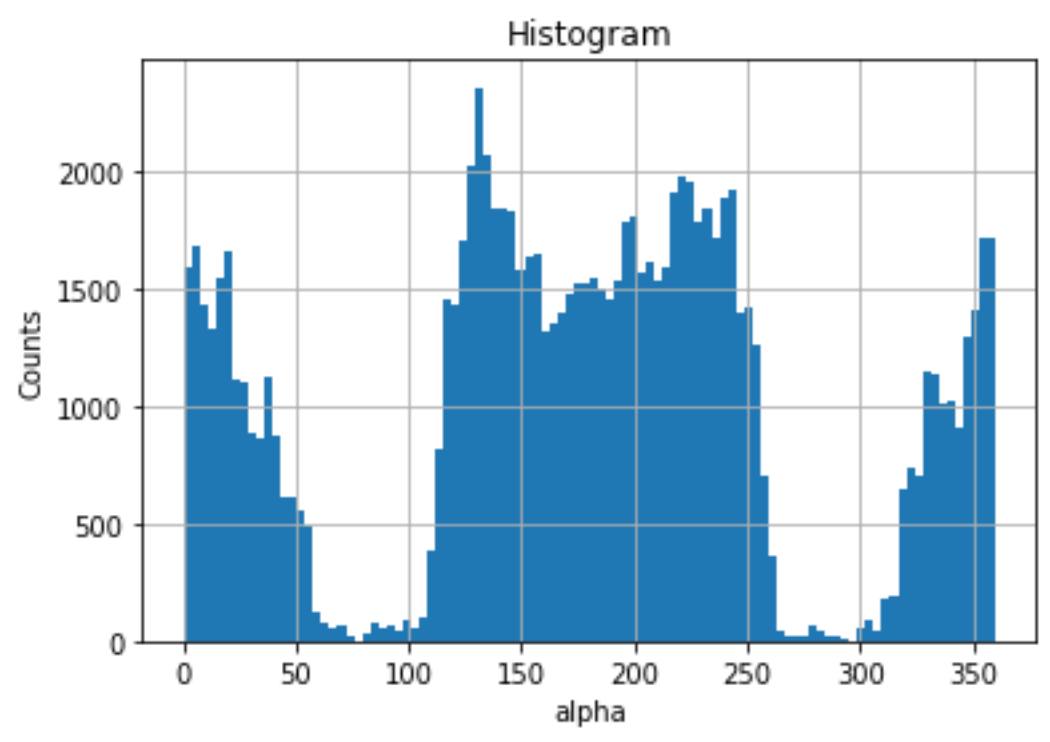
obj_ID

- Out of 100,000 objects, 78,053 are unique, so some objects appear 2 or more times
- Allow multiple independent observations of same object

Data Exploration and Cleaning

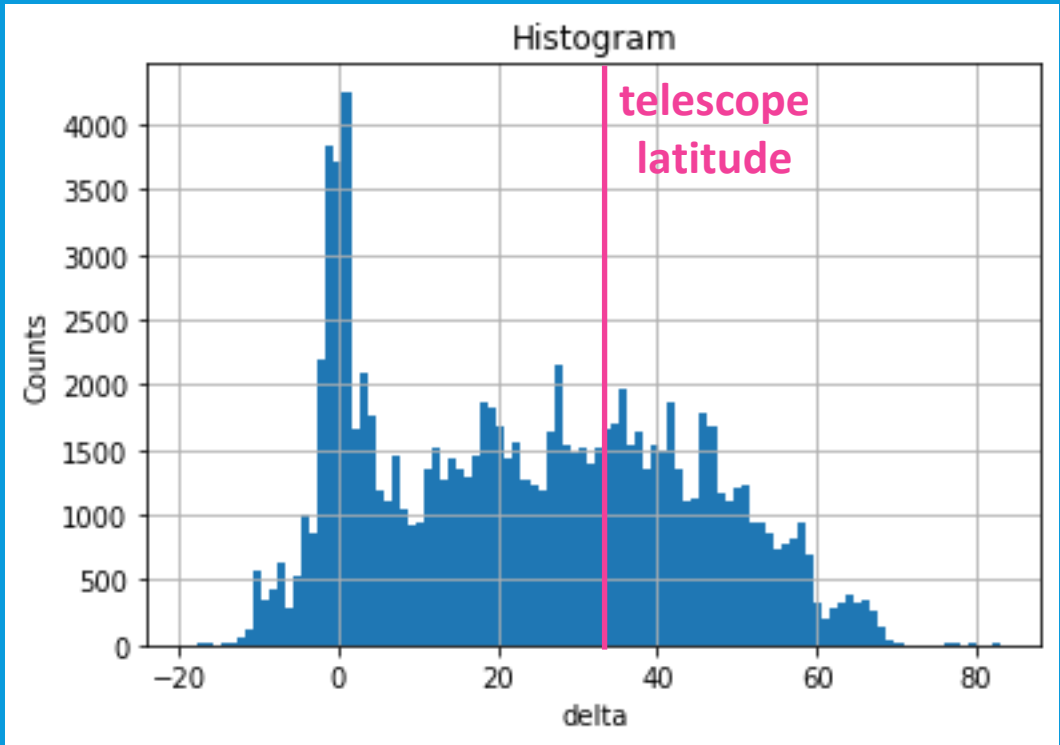
Positions: alpha & delta

alpha: right ascension (type of east-west position)



- Some positions are observed more than others, because some sections of sky sphere are visible at night during different times of year.

delta: declination (type of north-south position)

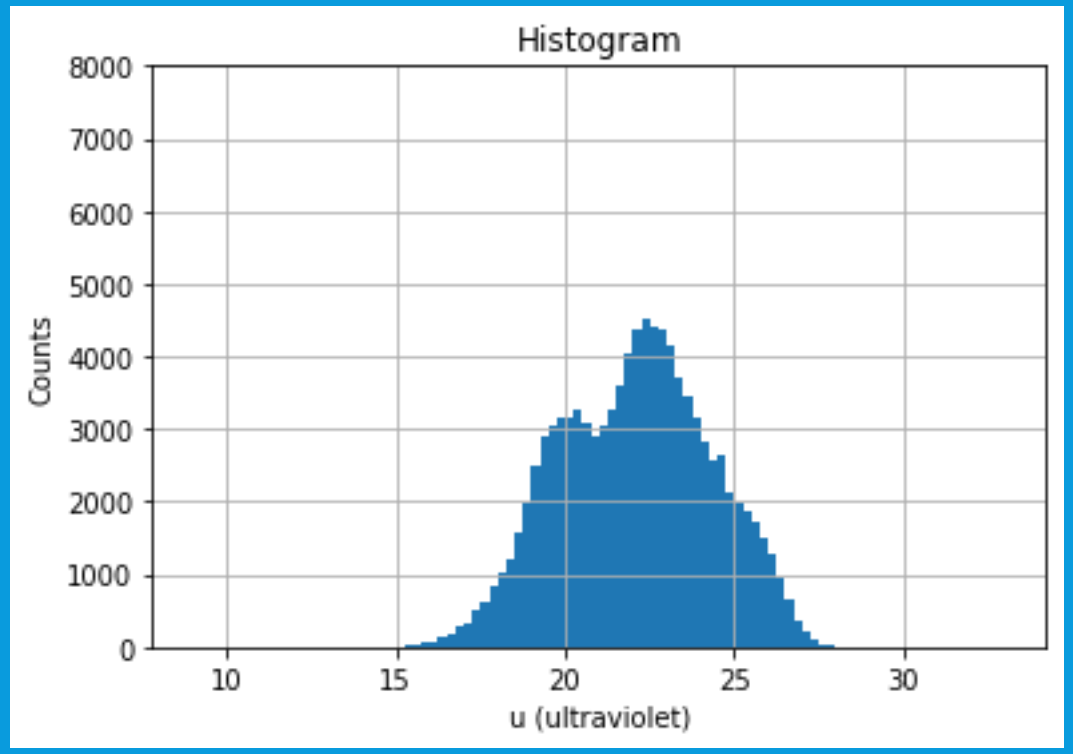


- Limited by vertical position range of telescope and centered as expected, but with a large peak at celestial equator.

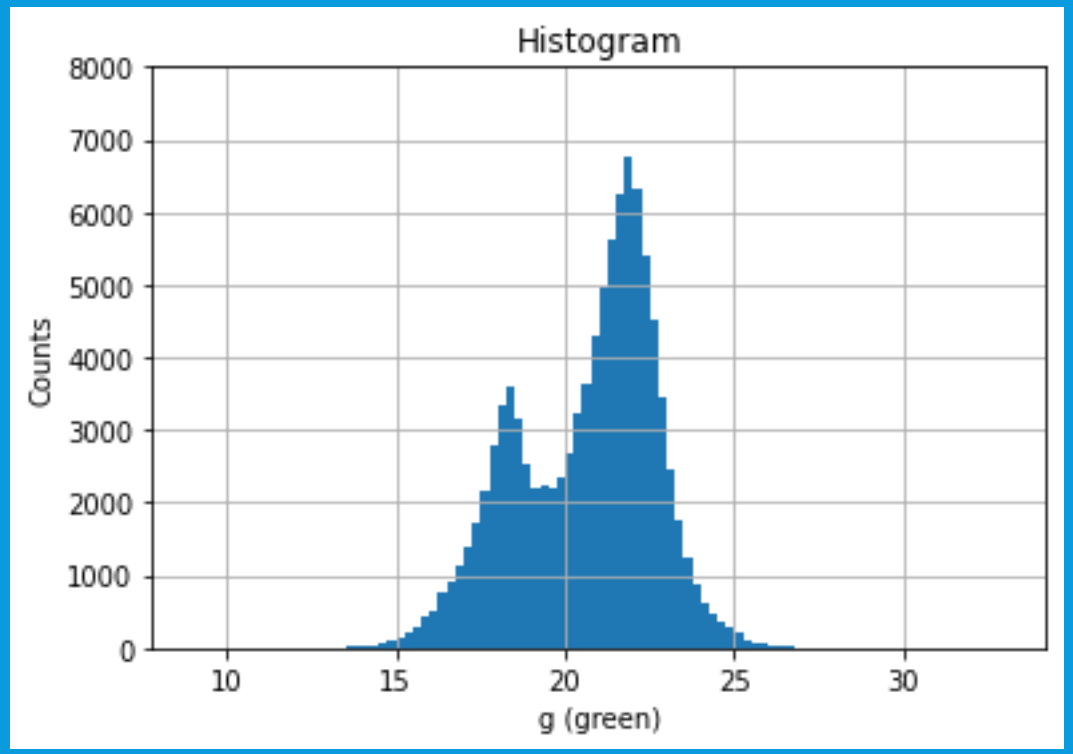
Data Exploration and Cleaning

Radiation Magnitudes in Filter Bands: u, g, r, i, z

u (ultraviolet)



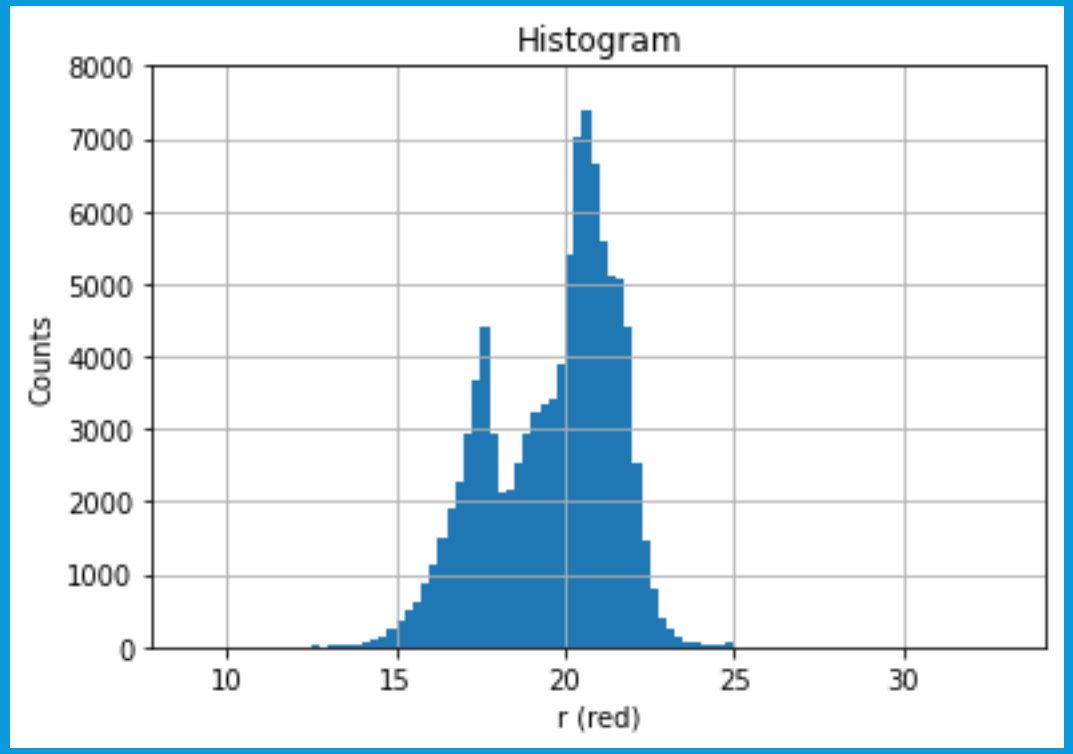
g (green)



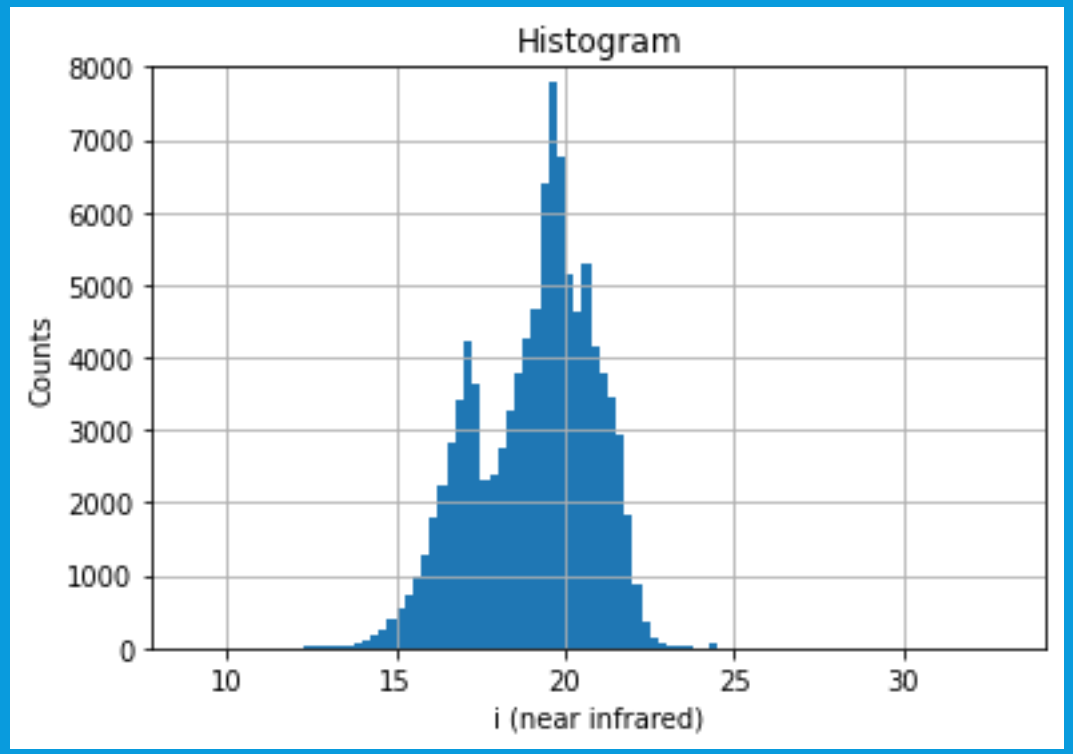
Data Exploration and Cleaning

Radiation Magnitudes in Filter Bands: u, g, r, i, z

r (red)



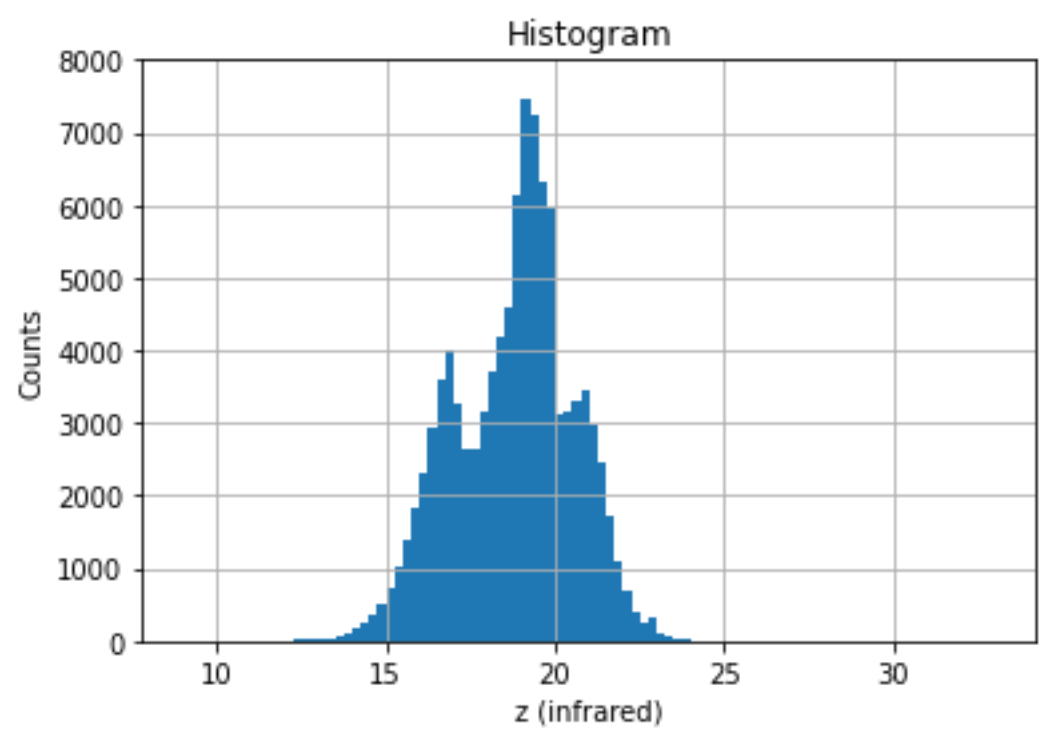
i (near infrared)



Data Exploration and Cleaning

Radiation Magnitudes in Filter Bands: u, g, r, i, z

z (infrared)

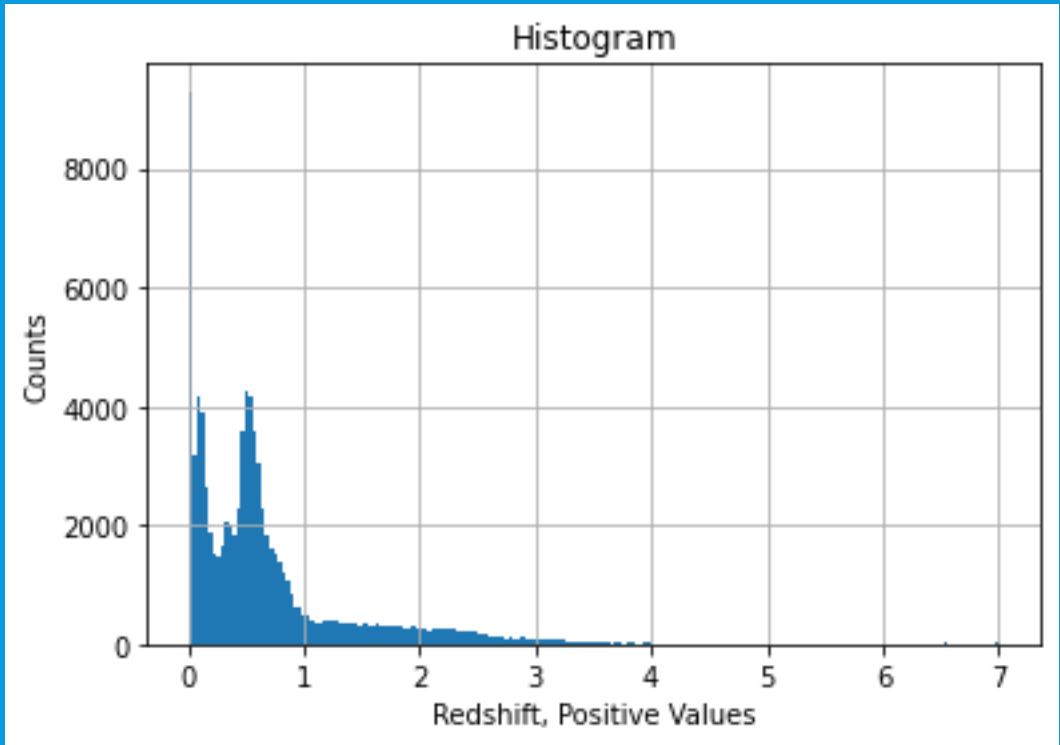


- There is interesting structure in the histograms of the magnitudes.
- There are at least 2 peaks in all of the filters.
- The peakiness generally increases with wavelength.
- The location of the main peak shifts to lower magnitudes with wavelength.
- Necessary to remove one row with missing magnitudes.

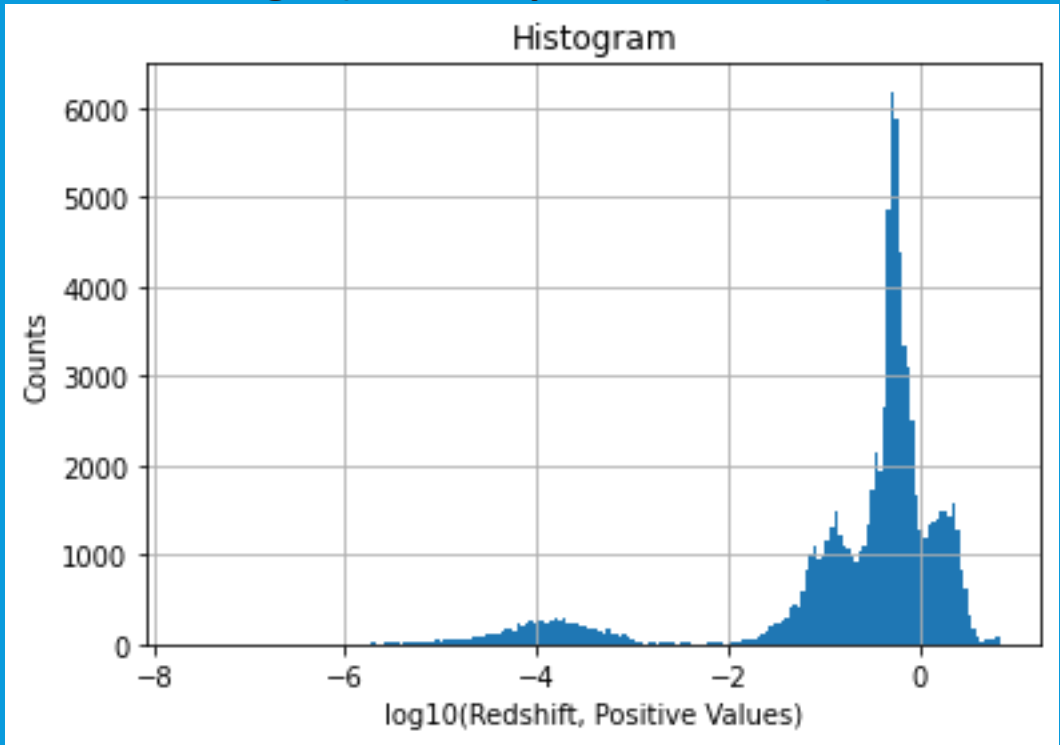
Data Exploration and Cleaning

Spectral Property: redshift

redshift, positive values



log10(redshift, positive values)



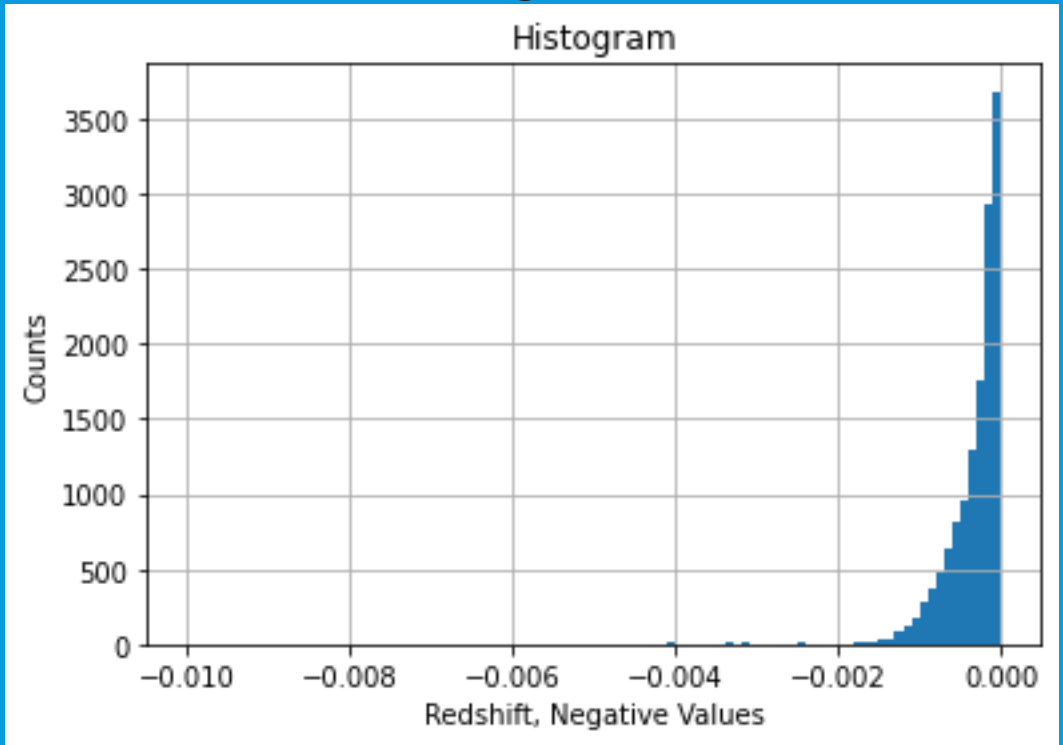
- Because of the properties of space, redshift (positive) should be more typical than blueshift (negative), with most objects moving away from each other. We find that here with 86% positive and 14% negative.

- There is a lot of interesting structure in redshift on both the positive and negative sides, especially on the positive side, and it is easier to see on the log scale.

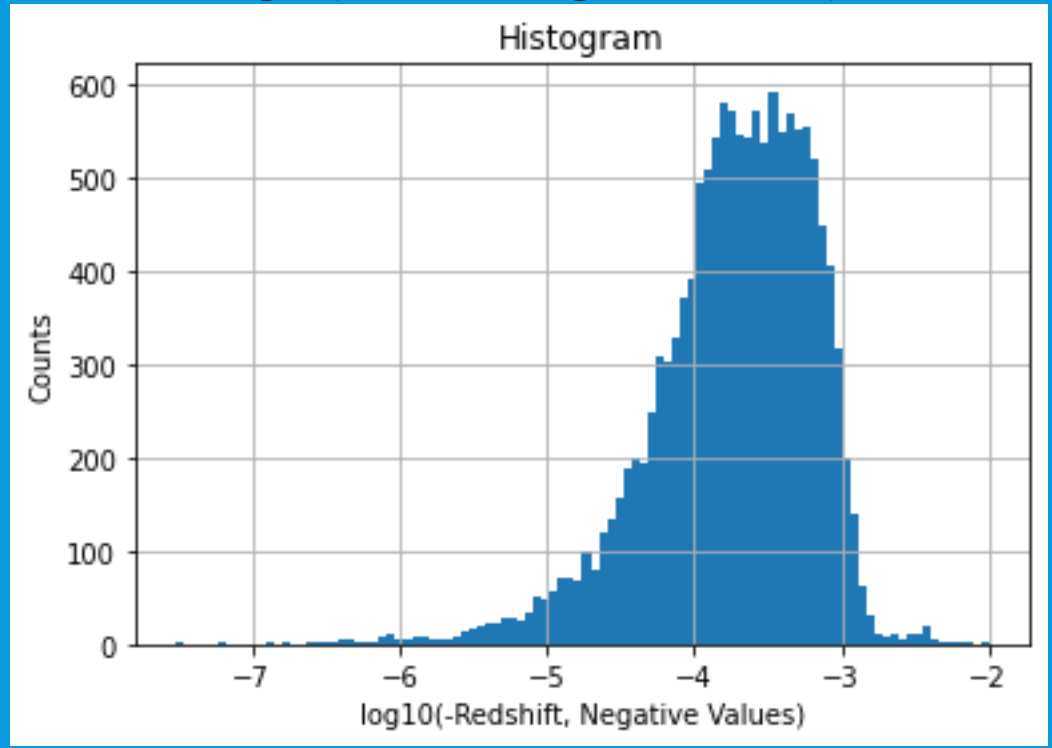
Data Exploration and Cleaning

Spectral Property: redshift

redshift, negative values



log10(redshift, negative values)



- Because of the properties of space, redshift (positive) should be more typical than blueshift (negative), with most objects moving away from each other. We find that here with 86% positive and 14% negative.

- There is a lot of interesting structure in redshift on both the positive and negative sides, especially on the positive side, and it is easier to see on the log scale.

Data Exploration and Cleaning

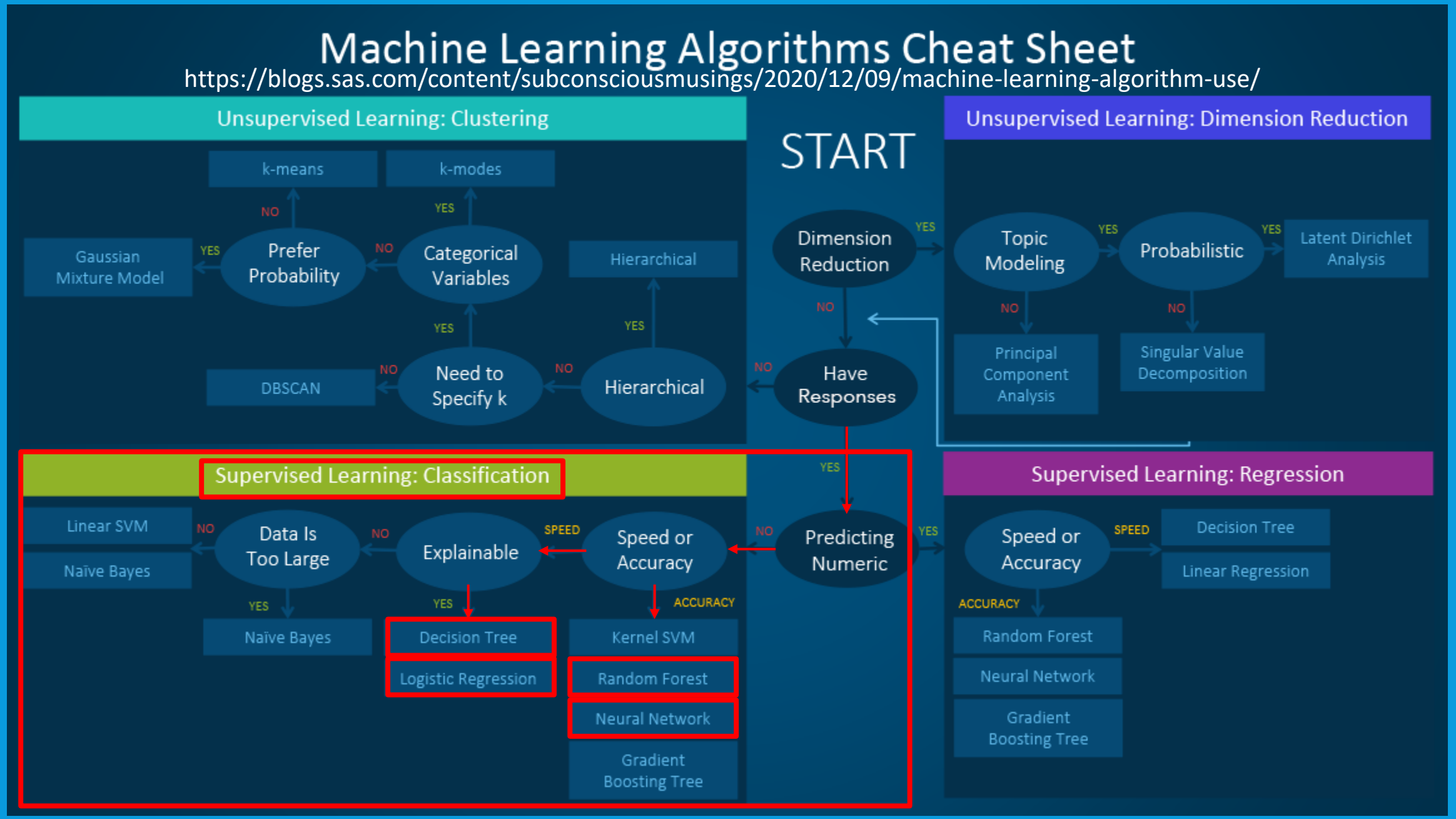
Others: run_ID, rerun_ID, cam_col, field_ID, spec_obj_ID, plate, MJD, fiber_ID

- The ranges and distributions of these features are acceptable and because we are unsure of their impact on classification, most are left in.
- rerun_ID is dropped because it only has one constant value.
- spec_obj_ID is dropped because each row has a unique ID so this would not contribute information.

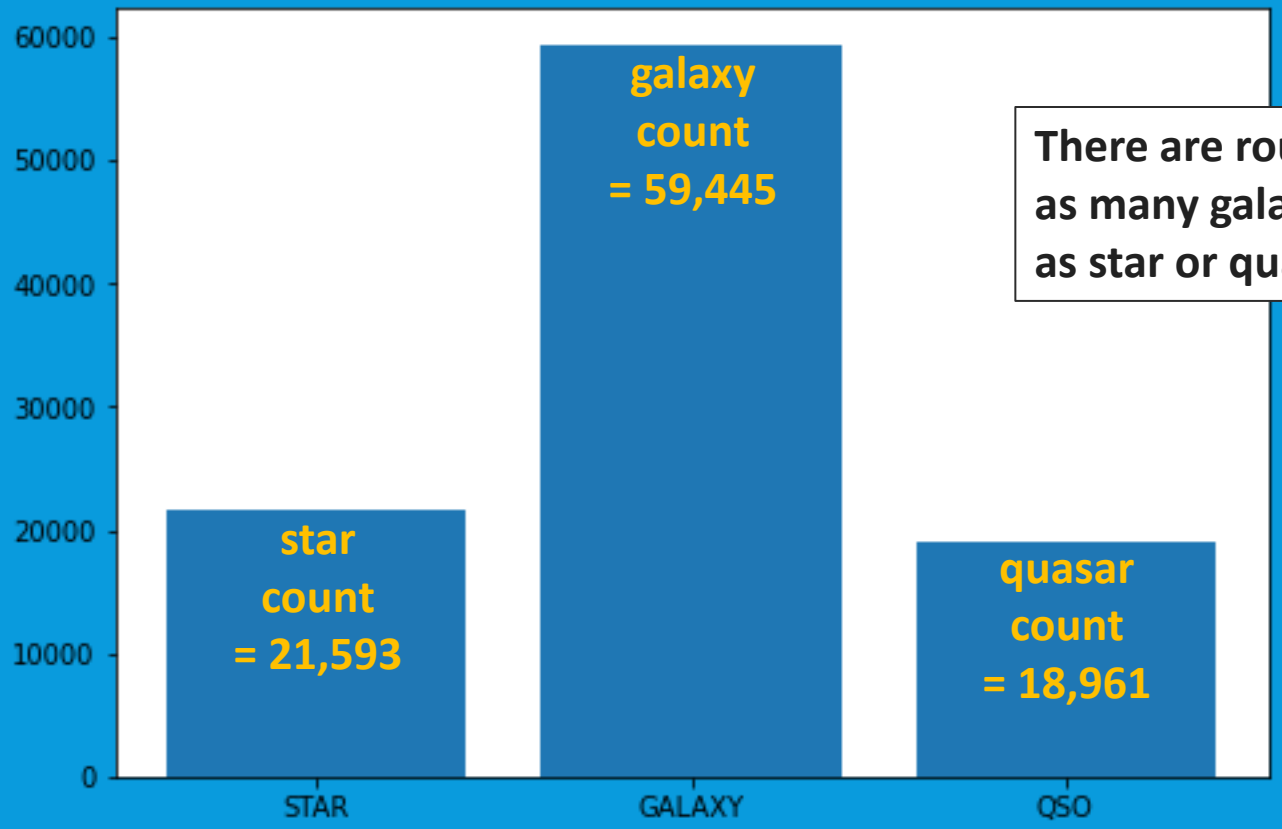
All Features

- The p-values of the normaltest (D'Agostino and Pearson's test) are well below 0.05, so none of the features can be considered normally distributed. This is as expected from what we see in the histograms.

Model Possibilities: Supervised Learning Classification



Model Preparation: Class Imbalance



There are roughly three times as many galaxy observations as star or quasar observations.

- With such an imbalance, we risk having our model(s) achieve accuracy by predicting the majority galaxy class.
- To remove this imbalance, we randomly undersample the majority galaxy class and the star class, to match the number of observations in the quasar class. There are 18,961 quasar observations, so the tradeoff is that information from 40,484 galaxy and 2,632 star observations are excluded from the model(s).
- Having a better performing model is more important than including all of the galaxy observations.

Model Preparation: Possible Transformations and Scaling

- None of the features have a skewness outside of -2 or +2, so we do not apply a logarithmic transformation to any of the features.

- The features are normalized so that they all have a mean of 0 and a standard deviation of 1. This reduces impacts from differences in ranges and variances among the features when building models.

Model Preparation: Unsupervised Learning Considerations

- Using supervised learning, but consider unsupervised learning to see if helps understand features or modeling.

Principal Component Analysis

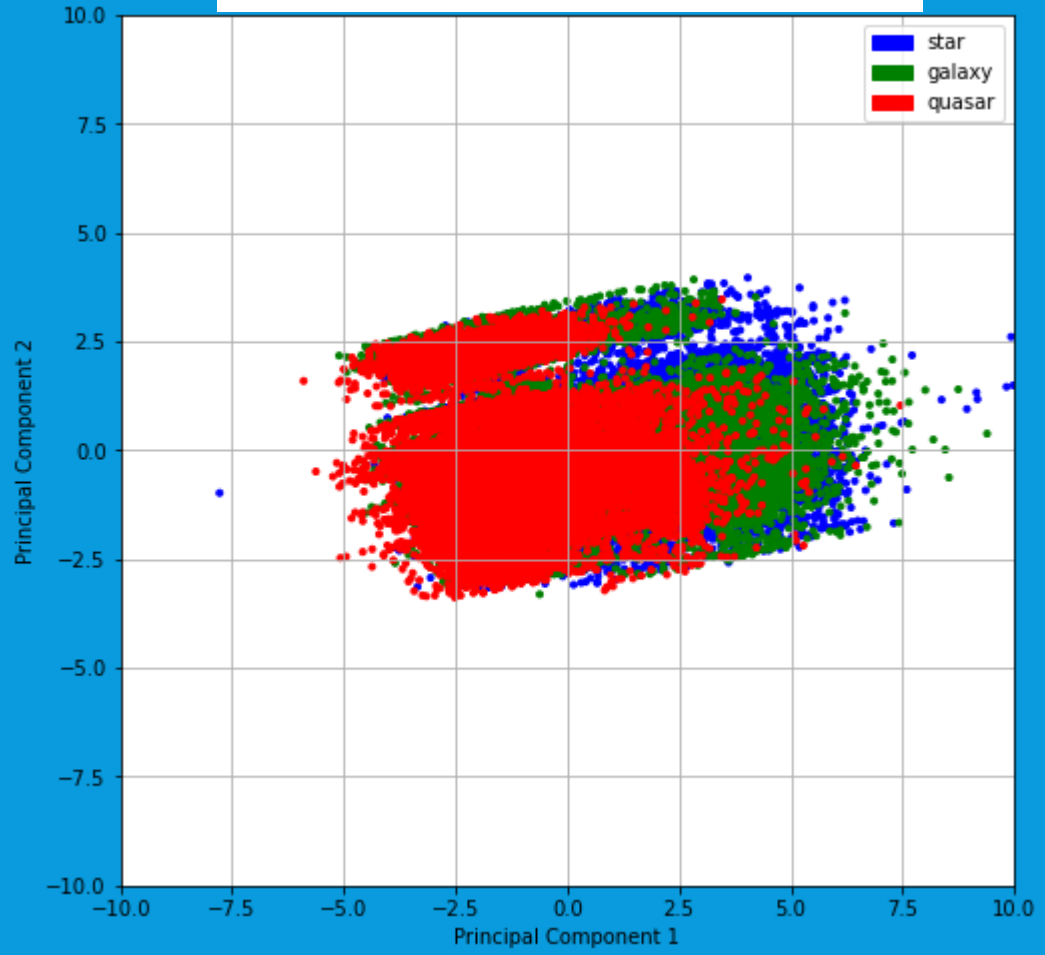
Component #	% Variance	Sum % Variance
1	37.1%	37.1%
2	14.0%	51.1%
3	9.4%	60.5%
4	7.3%	67.8%
5	7.2%	75.0%
⋮	⋮	⋮
15	7.8×10^{-21} %	100.0%

- First 2 components explain 50% of variance and first 8 more than 90%, so a lot of information in features can be shifted to smaller number of components.

- Determined later that did not need dimension reduction to have successful model fit in reasonable time.

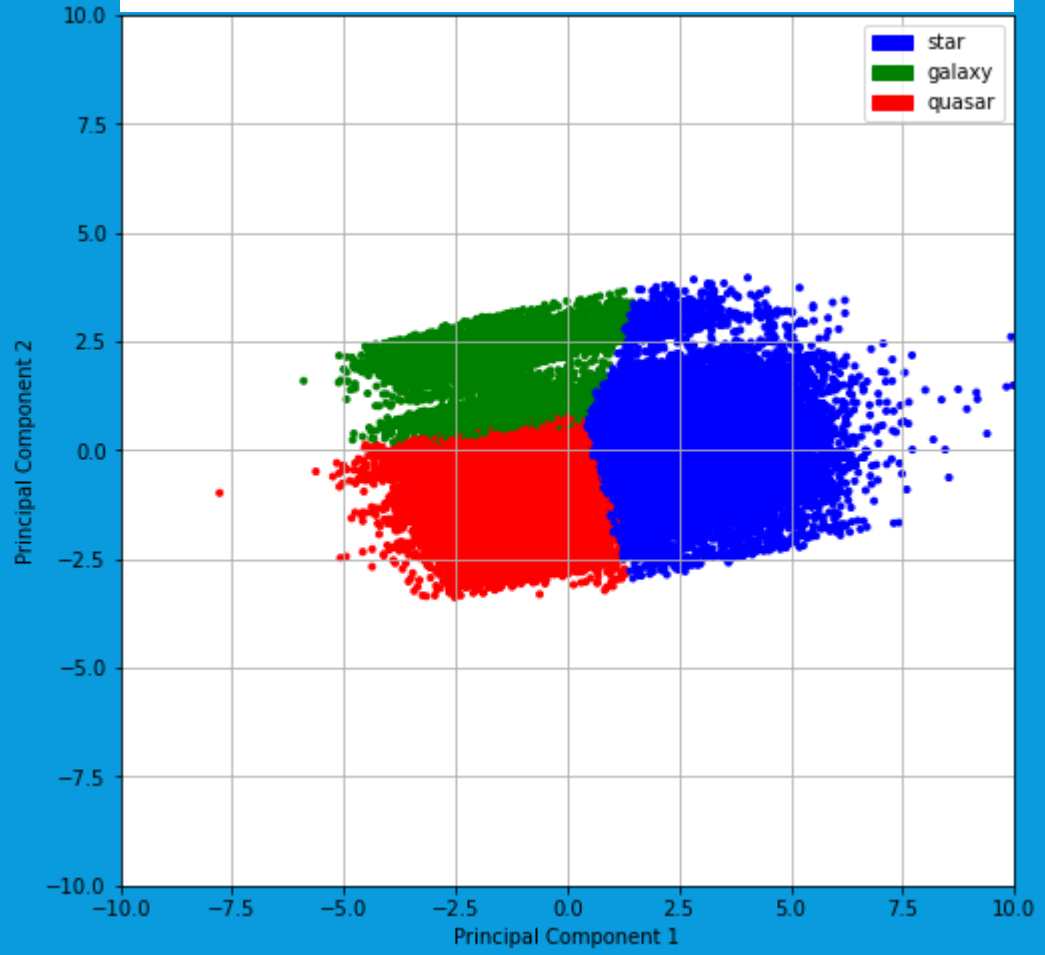
Model Preparation: Unsupervised Learning Considerations

Known Classes vs. First 2 PCs



- Classes largely occupy same space, but some separation in PC 1.

K-Means Cluster Classes vs. First 2 PCs

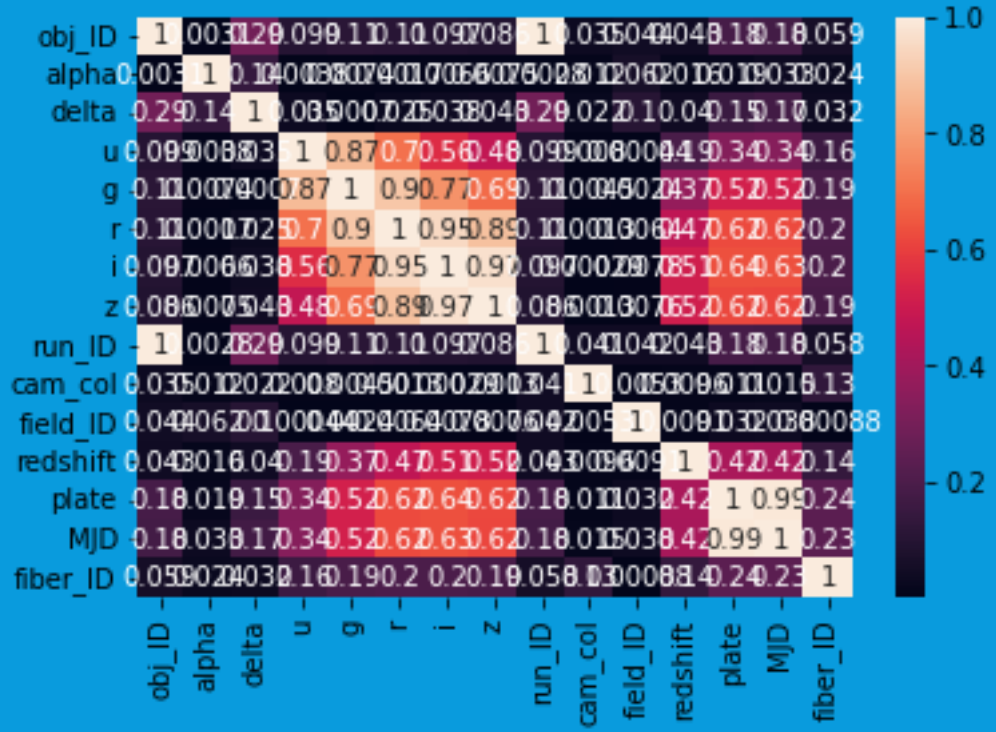


- K-Means cluster classes deviate significantly from known classes:
  - only 48% predicted correctly
  - ARI = 0.090

Model Preparation: Training/Test Split and Feature Selection

- Split the observations into random training (80%) and test (20%) sets. The training set is used to build the models and the test set is used to evaluate the performance of the models.
- Feature Selection: Spearman rank-order correlation and p-value for non-normal data
- 81 of 105 off-diagonal correlations are significant at 1% level.

Correlation Heatmap



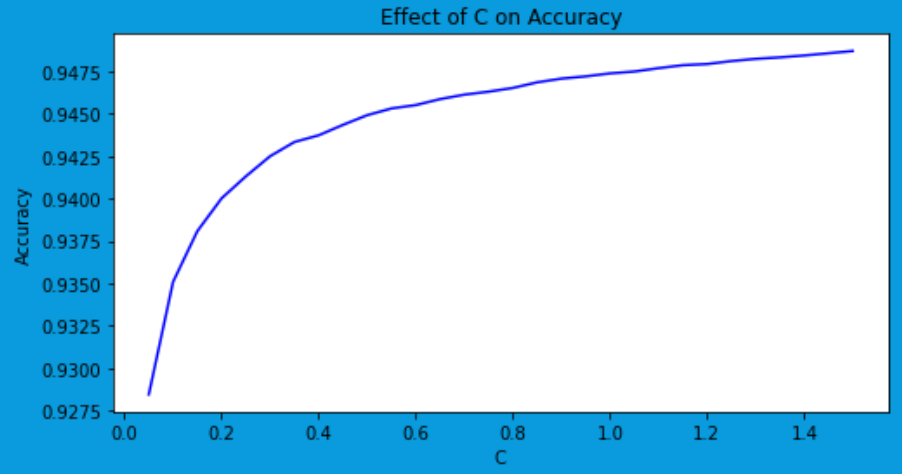
Top Absolute Correlations (>0.9)

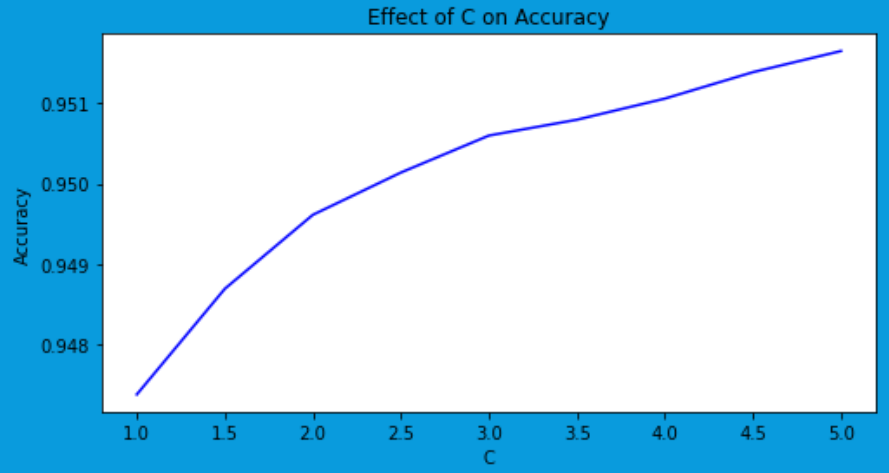
Feature 1	Feature 2	|Correlation|
obj_ID	run_ID	0.999943
MJD	plate	0.991851
i (near infrared)	z (infrared)	0.971249
i (near infrared)	r (red)	0.949284
g (green)	r (red)	0.904705

- Drop obj_ID and MJD because of their extremely high correlations with other features.
- Reluctant to remove filter magnitudes, so retain them. |Correlation| cutoff is 0.975.

Classification Modeling: Logistic Regression

- multinomial logistic regression: linear in feature coefficients and classifies based on probability of class
- For ‘solver’ (algorithm used in fitting) tried: ‘lbfgs’ (large-scale bound-constrained optimization), ‘sag’ (stochastic average gradient descent), and ‘saga’ (variant of ‘sag’); ‘lbfgs’ performed best.
- For ‘penalty’ used ‘l2’ for regularization to combat overfitting. Penalty depends on square of coefficients.
- Increased number of iterations as necessary until convergence obtained.
- Used 10-fold cross validation to optimize model hyperparameters.
- Tried various values for ‘C’, which is the inverse of regularization strength:





- Accuracy just keeps increasing with C suggesting larger C values may be allowing overfitting. Just use C = 1 (default).

Classification Modeling: Logistic Regression

Results

Model	Data	Accuracy
Optimal	Training	0.947743
Optimal	Test	0.946735

The optimal logistic regression model appears to perform similarly for data within and outside the training data.

Positive Coefficients

features	coefficients
redshift	15.135622
g	0.732543
r	0.357890
u	0.179362
field_ID	0.034489
fiber_ID	0.010702
obj_ID	0.007516

Negative Coefficients

features	coefficients
i	-1.175383
plate	-0.074042
delta	-0.073576
z	-0.071234
cam_col	-0.040328
alpha	-0.030784

- The most important feature is redshift, followed by
- i (near infrared), g (green), r (red), and u (ultraviolet).
- The coefficients for other features including z (infrared) and properties of where and how the observations were made are smaller (in an absolute sense).

Predicted Class →

True Class ↓

	Star	Galaxy	Quasar
Star	3774	0	1
Galaxy	164	3514	154
Quasar	7	280	3483

Star prediction nearly totally accurate.

Some galaxies labeled stars and quasars.

Some quasars labeled galaxies and several as stars.

Classification Modeling: Decision Tree

- Used 10-fold cross validation on training data to optimize the tree.
- The trees produced with default parameters are too complex, with too much depth and too many leaves. They are likely overfitting the test set in each fold.
- Looked for the optimal max_depth, the maximum depth of the trees. Accuracy score is best at max_depth = 9.
- Looked for the optimal min_samples_leaf, which is the minimum number of samples required to be at a leaf node. Increasing it may have the effect of smoothing the model. Overall, the mean score does not vary much with a wide range of min_samples_leaf values. The optimal value is min_samples_leaf = 7.
- For the criterion parameter, 'entropy' results in a slightly larger accuracy than 'gini'.

Results

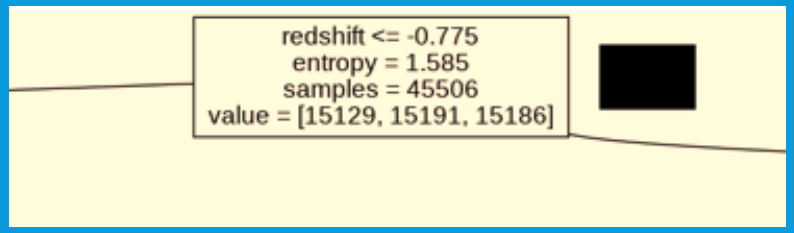
Model	Data	Accuracy
Optimal	Training	0.969982
Optimal	Test	0.967654

The optimal decision tree model appears to perform similarly for data within and outside the training data.

Classification Modeling: Decision Tree

- This is the optimal tree:

- This is the top of the optimal tree:



- The upper node rules are dominated by redshift.
- The top node rule is redshift <= -0.775.
- On the true (left) side of the top node, the next node is redshift <= -0.779, which in turn has redshift <= -0.784 on its true side and redshift <= -0.779 (to within rounding) on its false side.
- On the false (right) side of the top node, the next node is redshift <= 0.283, which in turn has redshift <= -0.045 on its true side and g <= 0.852 on its false side.

Classification Modeling: Random Forest

- The random forest model was setup with the parameters obtained in decision tree: criterion = ‘entropy’, max_depth = 9, and min_samples_leaf = 7.
- Used the default value for the number of trees in the forest, n_estimators = 100.
- Used 10-fold cross validation on training data.
- A parameter test is tried in turning off bootstrap, but that just lowers the accuracy slightly.

Results

Model	Data	Accuracy
Optimal	Training	0.969301
Optimal	Test	0.967039

The optimal random forest model appears to perform similarly for data within and outside the training data.

Classification Modeling: Artificial Neural Network (ANN)

- Investigated use of ANN from deep learning for classification by implementing tools from TensorFlow and Keras.
- Converted known classes to numbers and one-hot encoded.
- Several settings and hyperparameters to consider when setting up and fitting ANN. The number of combinations of settings and hyperparameters can quickly multiply even with very few choices.
  - Limit some choices immediately:

Name	Description	Value	Reason
activation	activation function in hidden layers	'relu'	Rectified linear unit best performer in course and "most common."
activation	activation function in output layer	'softmax'	Obtain probability of each class in the last layer.
optimizer	objective function to train model	'sgd'	Stochastic gradient descent is "most common."
loss	function for model performance	'categorical_crossentropy'	"Cross-entropy is most common loss function for classification problems."

Classification Modeling: Artificial Neural Network (ANN)

- Investigated use of ANN from deep learning for classification by implementing tools from TensorFlow and Keras.
- Converted known classes to numbers and one-hot encoded.
- Several settings and hyperparameters to consider when setting up and fitting ANN. The number of combinations of settings and hyperparameters can quickly multiply even with very few choices.
  - Two choices for others:

Name	Description	Values
nl	number of total layers	(3, 6)
units	number of neurons in hidden layers	(64, 128)
learning_rate	step rate for gradient descent	(0.01, 0.05)
batch_size	size of mini batch for gradient descent	(50, 100)

  - There are 16 combinations of hyperparameters. Used 20 epochs and found accuracy increases: always with nl, usually with units and learning_rate. Accuracy always decreases with batch_size.
  - Sought greater accuracy by switching to nl = 12, learning_rate = 0.10, and batch_size = 25, and 40 epochs. Kept units = 128.

Classification Modeling: Artificial Neural Network (ANN)

Results

Model	Data	Accuracy
Final	Training	0.975000
Final	Test	0.963523

The artificial neural network appears to perform similarly for data within and outside the training data.

Summary and Conclusions

- How well does each optimized model perform on the test data?

Model Type	Mean Accuracy
Logistic Regression	0.946735
Decision Tree	0.967654
Random Forest	0.967039
Artificial Neural Network	0.963523

- All of the nonlinear modeling approaches perform better than the only linear approach: logistic regression.
- Decision tree (DT) performs best, slightly better but nearly the same as random forest (RF) (relative difference of 0.064%).
- The ANN accuracy is slightly lower than that for DT / RF (relative difference with DT of 0.428%). There are a lot of optimization possibilities for ANN, so may be possible to get a little more out of it to bring it up to level of DT / RF.
- The most important features in the optimal logistic regression are redshift, near infrared, green, red, and ultraviolet. The optimal decision tree is dominated by redshift, but the color filter magnitudes and date are also important.

Model Product

- The product is an optimized decision tree model that can be used to determine, at high accuracy, whether an unclassified astronomical object is a star, galaxy, or quasar by using observed optical and spectral properties for the object.

- The model works largely because redshift and other properties are so definitive in predicting an object's class.

- When applying the model to observations from another telescope, the model built above would have to be rebuilt and customized by comparing the features we had available from the Sloan Digital Sky Survey (SDSS) and those available from the other telescope. It would be customized, possibly with fewer features, for each telescope (e.g., Webb and Hubble measure in different EM ranges).

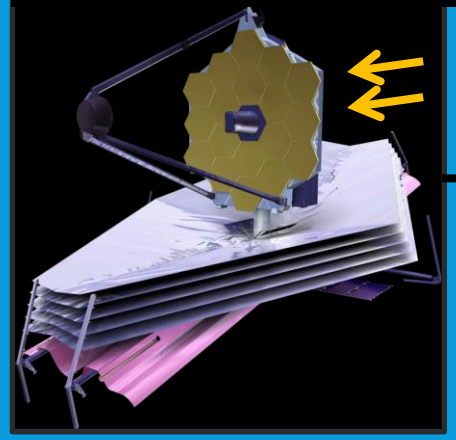
- End users (astrophysicists and other scientists) who want to determine the classification for one or more observed but not yet classified objects could do so quickly and without needing to know the exact details of the model.

- Possible additional steps:
  - Train model with more observations.
  - Work to rebuild model so it’s applicable to another telescope such as Hubble or Webb or another ground-based telescope. Any additional performance testing would require a different source for determinations of classes.

Hubble Space Telescope

QuGaSt

James Webb Space Telescope



QuGaSt