Identifying galaxies, quasars, and stars with machine learning: A new catalogue of classifications for 111 million SDSS sources without spectra

A. O. Clarke, A. M. M. Scaife, R. Greenhalgh, and V. Griguta

Published in the Journal Astronomy and Astrophysics, Volume 639, A84 (2020)

Survey Data Is Growing

- Sky surveys such as the Sloan Digital Sky Survey (SDSS) and the Widefield Infrared Survey Explorer (WISE) scan the sky and collect data from billions of sources.

Survey Data Is Growing

- Sky surveys such as the Sloan Digital Sky Survey (SDSS) and the Widefield Infrared Survey Explorer (WISE) scan the sky and collect data from billions of sources.
- The Sloan Digital Sky Survey has catalogued
  - Hundreds of millions of objects in multiband photometry.
  - Millions of objects with detailed spectra (with the Baryon Oscillation Spectroscopic Survey (BOSS), part of SDSS).

Survey Data Is Growing

- Sky surveys such as the Sloan Digital Sky Survey (SDSS) and the Widefield Infrared Survey Explorer (WISE) scan the sky and collect data from billions of sources.
- The Sloan Digital Sky Survey has catalogued
  - Hundreds of millions of objects in multiband photometry.
  - Millions of objects with detailed spectra (with the Baryon Oscillation Spectroscopic Survey (BOSS), part of SDSS).
- The Legacy Survey of Space and Time (LSST) is expected to observe
  - 20 billion galaxies.
  - Produce 20 terabytes of data every night for ten years (60 petabytes)

Why We Need Astronomical Classifications

- Of the approximately 500 million unique sources catalogued in SDSS data, only 3 million have detailed (spectral) data associated with the sample.

Why We Need Astronomical Classifications

- Of the approximately 500 million unique sources catalogued in SDSS data, only 3 million have detailed (spectral) data associated with the sample.
- The vast majority of data samples are just small image cutouts and a set of numbers (photometric measurements) describing the amount of light collected at a specific source, in a specific wavelength band, contained in a specific fit function.

Why We Need Astronomical Classifications

- Of the approximately 500 million unique sources catalogued in SDSS data, only 3 million have detailed (spectral) data associated with the sample.
- The vast majority of data samples are just small image cutouts and a set of numbers (photometric measurements) describing the amount of light collected at a specific source, in a specific wavelength band, contained in a specific fit function.
- Many sources only have photometric data available.

Why We Need Astronomical Classifications

- Of the approximately 500 million unique sources catalogued in SDSS data, only 3 million have detailed (spectral) data associated with the sample.
- The vast majority of data samples are just small image cutouts and a set of numbers (photometric measurements) describing the amount of light collected at a specific source, in a specific wavelength band, contained in a specific fit function.
- Many sources only have photometric data available.
- Detailed spectroscopic observations are expensive and time-consuming.

Why We Need Astronomical Classifications

- Our goal is to learn photometric features (cheap) using a limited dataset of spectroscopic labels as ground-truth (expensive).

Why We Need Astronomical Classifications

- Our goal is to learn photometric features (cheap) using a limited dataset of spectroscopic labels as ground-truth (expensive).
    - Why? To write an encyclopedia of predictions?

Why We Need Astronomical Classifications

- Our goal is to learn photometric features (cheap) using a limited dataset of spectroscopic labels as ground-truth (expensive).
  - Why? To write an encyclopedia of predictions?
  - Not exactly. Our goal is to identify candidate objects for further study.

Why We Need Astronomical Classifications

- Our goal is to learn photometric features (cheap) using a limited dataset of spectroscopic labels as ground-truth (expensive).
  - Why? To write an encyclopedia of predictions?
  - Not exactly. Our goal is to identify candidate objects for further study.
  - Surveys use very expensive instruments to obtain spectra. Sometimes, it is necessary to take measurements from space telescopes which can cost billions of dollars.

Why We Need Astronomical Classifications

- Our goal is to learn photometric features (cheap) using a limited dataset of spectroscopic labels as ground-truth (expensive).
  - Why? To write an encyclopedia of predictions?
  - Not exactly. Our goal is to <u>identify candidate objects</u> for further study.
  - Surveys use very expensive instruments to obtain spectra. Sometimes, it is necessary to take measurements from space telescopes which can cost billions of dollars.
  - Our goal is to use machine learning to <u>prioritize</u> our limited time, money, and hardware resources.

Why We Need Astronomical Classifications

- Our goal is to learn photometric features (cheap) using a limited dataset of spectroscopic labels as ground-truth (expensive).
  - Why? To write an encyclopedia of predictions?
  - Not exactly. Our goal is to identify candidate objects for further study.
  - Surveys use very expensive instruments to obtain spectra. Sometimes, it is necessary to take measurements from space telescopes which can cost billions of dollars.
  - Our goal is to use machine learning to prioritize our limited time, money, and hardware resources.
  - Using high-probability candidates for spectroscopic follow-up means we're not wasting resources on low-probability candidates.

Goal of This Work

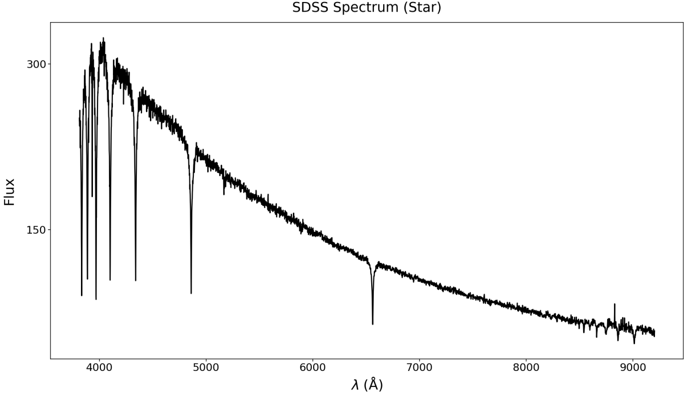
- Specifically, we want to train a machine learning model using 3.1 million spectroscopically labeled sources in SDSS DR15 to learn photometric features.

Goal of This Work

- Specifically, we want to train a machine learning model using 3.1 million spectroscopically labeled sources in SDSS DR15 to learn photometric features.
- Then use this model to classify 111 million unlabeled photometric sources as candidate quasars, galaxies, and stars.

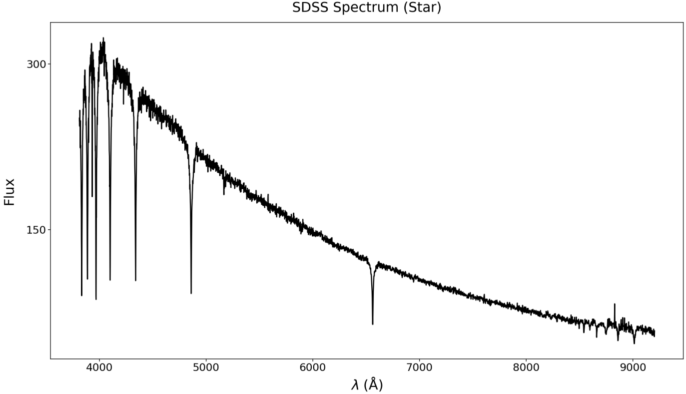
Data Representations

- First, let's establish how ground-truth labels are typically assigned in astronomy.



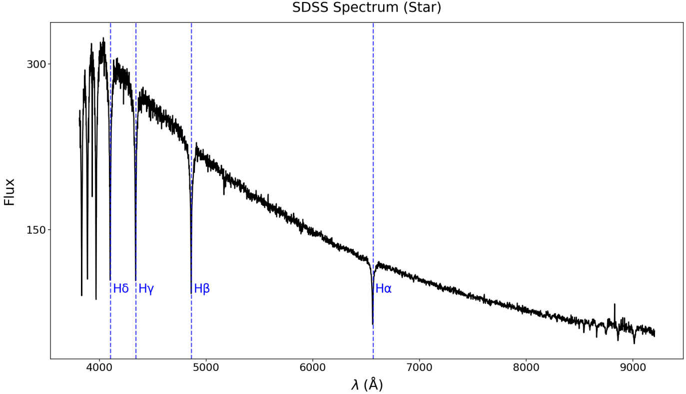
Data Representations

- First, let's establish how ground-truth labels are typically assigned in astronomy.
- Classification is typically done by analyzing spectral absorption and emission lines.



Data Representations

- First, let's establish how ground-truth labels are typically assigned in astronomy.
- Classification is typically done by analyzing spectral absorption and emission lines.



Class Labels

- Spectral fitting can distinguish between three main classes of sources:

Class Labels

- Spectral fitting can distinguish between three main classes of sources:
  - `STAR`: In the optical bands, a star's spectrum is a black body curve.

Class Labels

- Spectral fitting can distinguish between three main classes of sources:
  - `STAR`: In the optical bands, a star's spectrum is a black body curve.
  - `GALAXY`: Galaxies have spectra that are the sum of many stars, plus emission/absorption lines from gas.

Class Labels

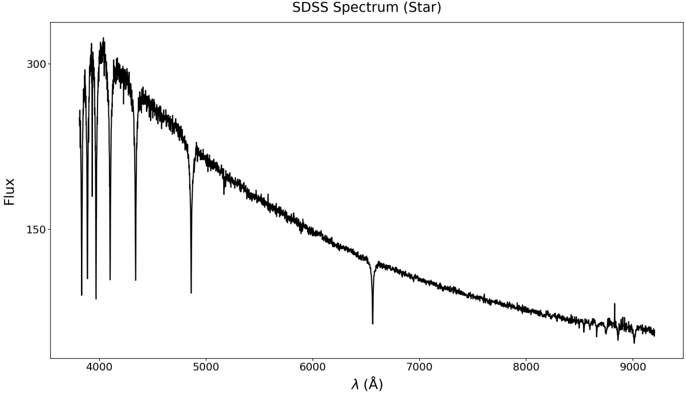
- Spectral fitting can distinguish between three main classes of sources:
  - `STAR`: In the optical bands, a star's spectrum is a black body curve.
  - `GALAXY`: Galaxies have spectra that are the sum of many stars, plus emission/absorption lines from gas.
  - `QSO`: Quasars have flat spectra with strong emission lines from AGNs around supermassive black holes.

Class Labels

- Spectral fitting can distinguish between three main classes of sources:
  - `STAR`: In the optical bands, a star's spectrum is a black body curve.
  - `GALAXY`: Galaxies have spectra that are the sum of many stars, plus emission/absorption lines from gas.
  - `QSO`: Quasars have flat spectra with strong emission lines from AGNs around supermassive black holes.
  - There are other classes, e.g. white dwarfs and unknown objects, but this work focuses on the main three.

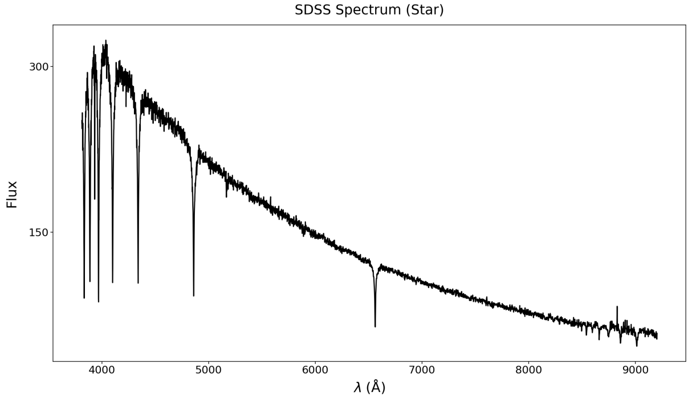
Data Representations

- Data without spectra instead captures the overall brightness in different bands.



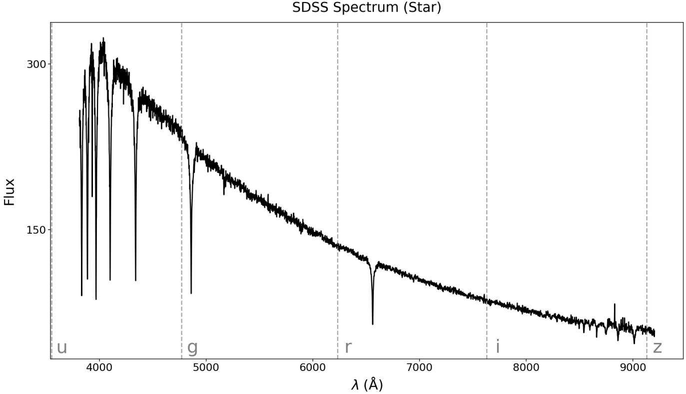
Data Representations

- Data without spectra instead captures the overall brightness in different bands.
- Classification must rely on the overall shape of the spectrum, rather than fine details.



Data Representations

- Data without spectra instead captures the overall brightness in different bands.
- Classification must rely on the overall shape of the spectrum, rather than fine details.



Classification Strategy

- Photometric data measures brightness in broad wavelength bands, in terms of the total flux magnitude. In other words, the sum of image pixels constrained by a fitting model.

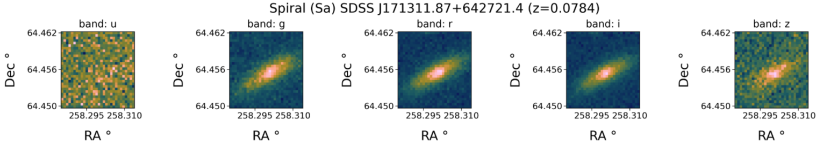


Figure: Multiband image data

Classification Strategy

- Photometric data measures brightness in broad wavelength bands, in terms of the total flux magnitude. In other words, the sum of image pixels constrained by a fitting model.
- This is much cheaper to obtain, allowing for larger datasets.

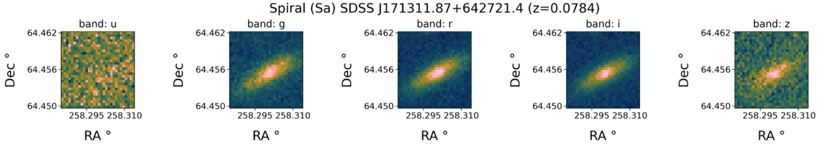


Figure: Multiband image data

Classification Strategy

- Photometric data measures brightness in broad wavelength bands, in terms of the total flux magnitude. In other words, the sum of image pixels constrained by a fitting model.
- This is much cheaper to obtain, allowing for larger datasets.
- Photometry cannot capture the fine details of a spectrum, however it can capture the overall shape of the spectrum.

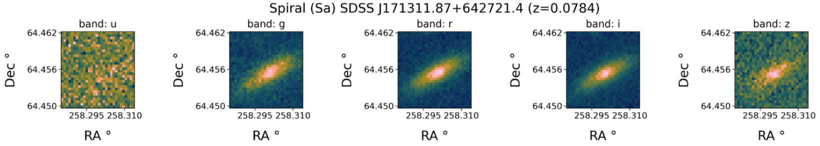


Figure: Multiband image data

Feature Engineering

- SDSS photometric data provides measurements in 5 bands:

Band	Wavelength (Å)	Color
u	3550	Ultraviolet
g	4770	Green
r	6230	Red
i	7620	Infrared
z	9130	Far Infrared

Table: Photometric Bands

Feature Engineering

- WISE photometric data provides measurements in 4 bands.:

Band	Wavelength (Å)
w1	34,000
w2	46,000
w3	120,000
w4	220,000

Table: Photometric Bands

Feature Engineering

- Feature engineering is the process of using domain knowledge to extract features from raw data.

Feature Engineering

- Feature engineering is the process of using domain knowledge to extract features from raw data.
- The Point Spread Function is used to describe how a point source of light is spread out over pixels.

Feature Engineering

- Feature engineering is the process of using domain knowledge to extract features from raw data.
- The Point Spread Function is used to describe how a point source of light is spread out over pixels.
    - PSF represents the distribution of light from a point source.

Feature Engineering

- Feature engineering is the process of using domain knowledge to extract features from raw data.
- The Point Spread Function is used to describe how a point source of light is spread out over pixels.
  - PSF represents the distribution of light from a point source.
  - PSF determines the smallest angular separation at which two point sources can be distinguished.

Feature Engineering

- Feature engineering is the process of using domain knowledge to extract features from raw data.
- The Point Spread Function is used to describe how a point source of light is spread out over pixels.
    - PSF represents the distribution of light from a point source.
    - PSF determines the smallest angular separation at which two point sources can be distinguished.
    - PSF is meaningful for Stars, Quasars, and barely resolved objects

Feature Engineering

- Feature engineering is the process of using domain knowledge to extract features from raw data.
- The Point Spread Function is used to describe how a point source of light is spread out over pixels.
  - PSF represents the distribution of light from a point source.
  - PSF determines the smallest angular separation at which two point sources can be distinguished.
  - PSF is meaningful for Stars, Quasars, and barely resolved objects
- SDSS fits three models to every object, in every band:

Feature Engineering

- Feature engineering is the process of using domain knowledge to extract features from raw data.
- The Point Spread Function is used to describe how a point source of light is spread out over pixels.
  - PSF represents the distribution of light from a point source.
  - PSF determines the smallest angular separation at which two point sources can be distinguished.
  - PSF is meaningful for Stars, Quasars, and barely resolved objects
- SDSS fits three models to every object, in every band:
  - PSF

Feature Engineering

- Feature engineering is the process of using domain knowledge to extract features from raw data.
- The Point Spread Function is used to describe how a point source of light is spread out over pixels.
  - PSF represents the distribution of light from a point source.
  - PSF determines the smallest angular separation at which two point sources can be distinguished.
  - PSF is meaningful for Stars, Quasars, and barely resolved objects
- SDSS fits three models to every object, in every band:
  - PSF
  - de Vaucouleurs profile

Feature Engineering

- Feature engineering is the process of using domain knowledge to extract features from raw data.
- The Point Spread Function is used to describe how a point source of light is spread out over pixels.
    - PSF represents the distribution of light from a point source.
    - PSF determines the smallest angular separation at which two point sources can be distinguished.
    - PSF is meaningful for Stars, Quasars, and barely resolved objects
- SDSS fits three models to every object, in every band:
    - PSF
    - de Vaucouleurs profile
    - Exponential disk

Point Spread Function (PSF)

- `psfMag` is the magnitude computed by fitting the PSF model to the source.[1]

[1] `PSF` is usually a Gaussian. See Stoughton et al. p524

Point Spread Function (PSF)

- psfMag is the magnitude computed by fitting the PSF model to the source.[1]
- $\texttt{psfFlux} = \sum_{\texttt{pixels}} \texttt{PSF}(x, y) \times \texttt{fit amplitude}$

[1] PSF is usually a Gaussian. See Stoughton et al. p524

Point Spread Function (PSF)

- psfMag is the magnitude computed by fitting the PSF model to the source.[1]
- $\texttt{psfFlux} = \sum_{\texttt{pixels}} \texttt{PSF}(x, y) \times \texttt{fit amplitude}$
- $\texttt{psfMag} = -2.5 \log_{10}(\texttt{psfFlux}) + \texttt{zero point}$

[1] PSF is usually a Gaussian. See Stoughton et al. p524

Point Spread Function (PSF)

- psfMag is the magnitude computed by fitting the PSF model to the source.[1]
- $\texttt{psfFlux} = \sum_{\texttt{pixels}} \texttt{PSF}(x, y) \times \texttt{fit amplitude}$
- $\texttt{psfMag} = -2.5 \log_{10}(\texttt{psfFlux}) + \texttt{zero point}$
- When used as a feature for point objects, this represents the magnitude (the log of the flux)

[1] PSF is usually a Gaussian. See Stoughton et al. p524

Extended Objects - de Vaucouleurs profile fit

- `devMag` is used to fit elliptical galaxies and bulge-dominated objects.

Extended Objects - de Vaucouleurs profile fit

- `devMag` is used to fit elliptical galaxies and bulge-dominated objects.
- $I(r) = I_0 \exp(-7.67(\frac{r}{r_{eff}})^{\frac{1}{4}} - 1))$

Extended Objects - de Vaucouleurs profile fit

- `devMag` is used to fit elliptical galaxies and bulge-dominated objects.
- $I(r) = I_0 \exp(-7.67(\frac{r}{r_{eff}})^{\frac{1}{4}} - 1))$
- where r_{eff} is the effective radius containing half the light.

Extended Objects - Exponential profile fit

- `expMag` fits a 2D exponential disk, used for disk galaxies, spiral arms, late-type galaxies

Extended Objects - Exponential profile fit

- `expMag` fits a 2D exponential disk, used for disk galaxies, spiral arms, late-type galaxies
- $I(r) = I_0 \exp(-1.68(\frac{r}{r_{eff}}))$

modelMag - Matched apertures across bands

- In the r-band, the better fit of devMag and expMag is chosen.

`modelMag` - Matched apertures across bands

- In the r-band, the better fit of `devMag` and `expMag` is chosen.
- That same model shape (radius, profile) is applied to u,g,r,i,z, but only the amplitude (flux) is refit.

`modelMag` - Matched apertures across bands

- In the r-band, the better fit of `devMag` and `expMag` is chosen.
- That same model shape (radius, profile) is applied to u,g,r,i,z, but only the amplitude (flux) is refit.
- Used for galaxy colors, photometric redshift, ML classification

cmodelMag - Best total light estimate

- Takes the best fit from devMag and expMag, and combines them linearly to get the best total flux estimate.

`cmodelMag` - Best total light estimate

- Takes the best fit from `devMag` and `expMag`, and combines them linearly to get the best total flux estimate.
- $\texttt{flux}_{cmodel} = \texttt{frac}_{dev} \times \texttt{flux}_{dev} + (1 - \texttt{frac}_{dev}) \times \texttt{flux}_{exp}$ where $\texttt{frac}_{dev}$ is the fraction of the `devMag` component in the fit.

Resolved or point-like

- Measure how resolved the source is by calculating the difference between the `cmodelMag` and the PSF magnitude in the r-band:

Resolved or point-like

- Measure how resolved the source is by calculating the difference between the `cmodelMag` and the PSF magnitude in the r-band:
- $\texttt{resolved}_r = |\texttt{psf}_r - \texttt{cmodMag}_r|$

Dataset - Final Feature Set

- The final feature set consists of 9 photometric plus 1 feature representing how resolved the source is.

Dataset - Final Feature Set

- The final feature set consists of 9 photometric plus 1 feature representing how resolved the source is.

$\texttt{psf}_u$, $\texttt{psf}_g$, $\texttt{psf}_r$, $\texttt{psf}_i$, $\texttt{psf}_z$, $\texttt{w1}$, $\texttt{w2}$, $\texttt{w3}$, $\texttt{w4}$, $\texttt{resolved}_r$

Three Samples of Different Classes

Galaxy	psf_u	psf_g	psf_r	psf_i	psf_z	$cmod_r$	$\lvert psf_r - cmod_r \rvert$
	23.49	22.09	20.10	19.29	19.07	19.51	0.59
	$w1$	$w2$	$w3$	$w4$	–	–	–
	14.69	14.44	12.36	8.76	–	–	–
Quasar	psf_u	psf_g	psf_r	psf_i	psf_z	$cmod_r$	$\lvert psf_r - cmod_r \rvert$
	17.99	17.82	17.81	17.78	17.68	17.72	0.08
	$w1$	$w2$	$w3$	$w4$	–	–	–
	14.12	13.03	10.21	7.46	–	–	–
Star	psf_u	psf_g	psf_r	psf_i	psf_z	$cmod_r$	$\lvert psf_r - cmod_r \rvert$
	16.88	15.77	15.81	15.87	15.88	15.81	0.00
	$w1$	$w2$	$w3$	$w4$	–	–	–
	15.17	15.17	12.66	9.18	–	–	–

Dataset - Training Data

- The dataset includes 3,238,003 unique sources which have been spectroscopically observed and assigned a class of `STAR`, `GALAXY`, or `QSO` (quasar), based on spectroscopic models described earlier. [2]

[2] See the Baryon Oscillation Spectroscopic Survey (BOSS)

Dataset - Training Data

- The dataset includes 3,238,003 unique sources which have been spectroscopically observed and assigned a class of `STAR`, `GALAXY`, or `QSO` (quasar), based on spectroscopic models described earlier. [2]

[2] See the Baryon Oscillation Spectroscopic Survey (BOSS)

Dataset - Training Data

- The dataset includes 3,238,003 unique sources which have been spectroscopically observed and assigned a class of `STAR`, `GALAXY`, or `QSO` (quasar), based on spectroscopic models described earlier. [2]
- The dataset was constructed using SDSS Data Release 15 (DR15) data.

[2] See the Baryon Oscillation Spectroscopic Survey (BOSS)

Dataset - Preprocessing

- Removed duplicate WISE matches, keeping only the closest match.

Dataset - Preprocessing

- Removed duplicate WISE matches, keeping only the closest match.
- Removed SDSS sources with -9999 values for cMag, and WISE sources with 9999 as magnitude.

Dataset - Preprocessing

- Removed duplicate WISE matches, keeping only the closest match.
- Removed SDSS sources with -9999 values for cMag, and WISE sources with 9999 as magnitude.
- Removed sources with `zWarning` $\neq 0$. 0 indicates no problems.

Dataset - Preprocessing

- Removed duplicate WISE matches, keeping only the closest match.
- Removed SDSS sources with -9999 values for cMag, and WISE sources with 9999 as magnitude.
- Removed sources with `zWarning` $\neq 0$. 0 indicates no problems.
- In total, 138,546 sources were removed.

Dataset - Features Optimization

- Different combinations of SDSS, WISE, and $\texttt{resolved}_r$ features were tested.
- The best performance was achieved using all features.

Zoom Out - Big Picture

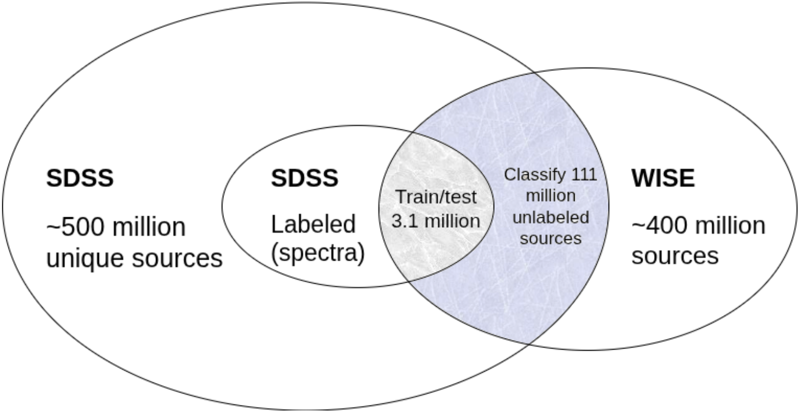


Figure: Venn Diagram of Data Overlap

The Model

- The model used was a Random Forest.[3]

[3] Specifically, `RandomForestClassifier` in the `sklearn.ensemble` package. See Louppe 2014

The Model

- The model used was a Random Forest.[3]
- It is an ensemble of independent decision trees.

[3] Specifically, `RandomForestClassifier` in the `sklearn.ensemble` package. See Louppe 2014

The Model

- The model used was a Random Forest.[3]
- It is an ensemble of independent decision trees.
  - Each tree is trained on a random subset of both features and samples.

[3] Specifically, `RandomForestClassifier` in the `sklearn.ensemble` package. See Louppe 2014

The Model

- The model used was a Random Forest.[3]
- It is an ensemble of independent decision trees.
    - Each tree is trained on a random subset of both features and samples.
    - The predicted classification comes from a majority consensus classification across the full set of models from all trees in the forest.

[3] Specifically, `RandomForestClassifier` in the `sklearn.ensemble` package. See Louppe 2014

The Model

- Reasons for choosing this algorithm:

The Model

- Reasons for choosing this algorithm:
  - Ensemble learning is robust to overfitting, and minimizes variance and bias.

The Model

- Reasons for choosing this algorithm:
  - Ensemble learning is robust to overfitting, and minimizes variance and bias.
  - Excels at numerical and categorical features over different scales.

The Model

- Reasons for choosing this algorithm:
  - Ensemble learning is robust to overfitting, and minimizes variance and bias.
  - Excels at numerical and categorical features over different scales.
  - Effective at multi-class classification

Hyper-parameters

- Hyper-parameters are model parameters that control the learning process and affect the model's performance.

Hyper-parameters

- Hyper-parameters are model parameters that control the learning process and affect the model's performance.
- Hyper-parameters in Random Forests:
  - Number of trees: *n_estimators*
  - Maximum number of features per decision tree: *max_features*
  - Minimum samples per leaf: *min_samples_leaf*

Hyper-parameters Optimization

- Used a 5-fold cross validation scheme to optimize hyper-parameters.

Hyper-parameter	Optimal Value	Comments
n_estimators	200	Best performance/running time tradeoff
max_features	3	Increasing this will improve F_1 score, but will lead to overfitting. Used the default value of `Int`($\sqrt{n\_features}$)
min_samples_leaf	1	Higher values result in drop in F_1 score

Table: Hyper-parameter Optimization Results

Hyper-parameters Optimization

- Used a 5-fold cross validation scheme to optimize hyper-parameters.
- Meaning, the training data was randomly split into 5 parts. Each part was used to validate the model trained on the other 4 parts combined.

Hyper-parameter	Optimal Value	Comments
n_estimators	200	Best performance/running time tradeoff
max_features	3	Increasing this will improve F_1 score, but will lead to overfitting. Used the default value of $\mathtt{Int}(\sqrt{n\_features})$
min_samples_leaf	1	Higher values result in drop in F_1 score

Table: Hyper-parameter Optimization Results

Hyper-parameters Optimization

- Used a 5-fold cross validation scheme to optimize hyper-parameters.
- Meaning, the training data was randomly split into 5 parts. Each part was used to validate the model trained on the other 4 parts combined.
- The hyper-parameters were tuned to maximize the average F_1 score across all 5 folds.

Hyper-parameter	Optimal Value	Comments
n_estimators	200	Best performance/running time tradeoff
max_features	3	Increasing this will improve F_1 score, but will lead to overfitting. Used the default value of `Int`($\sqrt{n\_features}$)
min_samples_leaf	1	Higher values result in drop in F_1 score

Table: Hyper-parameter Optimization Results

Precision, Recall, F1 Score

- Precision: TP / (TP + FP).

Precision, Recall, F1 Score

- Precision: TP / (TP + FP).
- Recall: TP / (TP + FN).

Precision, Recall, F1 Score

- Precision: TP / (TP + FP).
- Recall: TP / (TP + FN).
- F1 Score: 2 * (Precision * Recall) / (Precision + Recall). The F1 Score is a harmonic mean of precision and recall. This metric balances precision and recall.
  - Galaxies: 0.991
  - Quasars: 0.952
  - Stars: 0.978

Test Results on Unlabeled Data

- 50,417,547 galaxies classified, 70% with *pred_prob* > 0.9

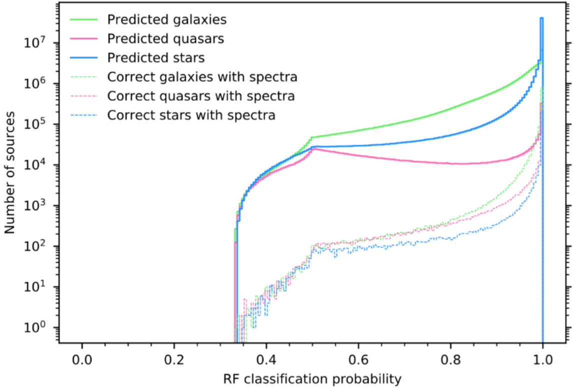


Figure: Histogram of classification probabilities using a bin size of 0.005.

Test Results on Unlabeled Data

- 50,417,547 galaxies classified, 70% with $pred\_prob > 0.9$
- 2,137,839 quasars classified, 34% with $pred\_prob > 0.9$

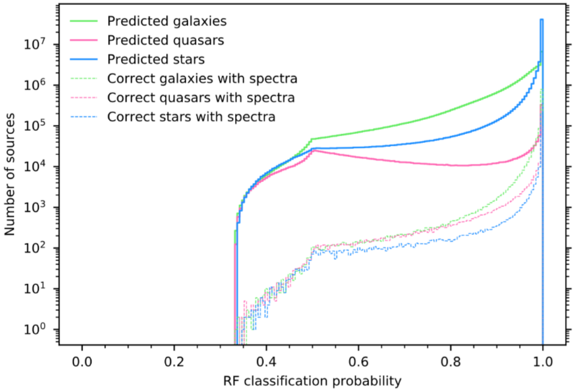


Figure: Histogram of classification probabilities using a bin size of 0.005.

Test Results on Unlabeled Data

- 50,417,547 galaxies classified, 70% with *pred_prob* > 0.9
- 2,137,839 quasars classified, 34% with *pred_prob* > 0.9
- 58,840,082 stars classified, 93% with *pred_prob* > 0.9

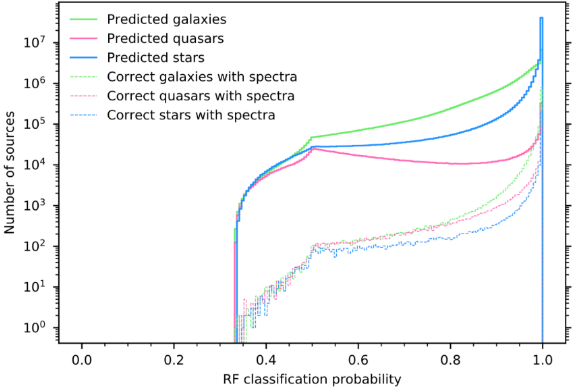


Figure: Histogram of classification probabilities using a bin size of 0.005.

Test Results on Unlabeled Data

- A non-linear dimension reduction technique (UMAP) was used to reduce the number of features from 10 to 2. This provided a 2-D visualization of the feature space, and was used for clustering visualizations.

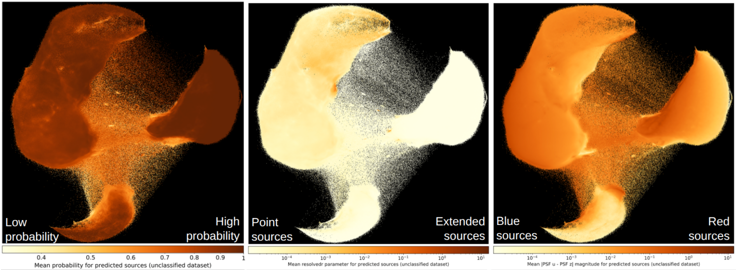


Figure: UMAP projections of photometric features.

Key Takeaways

- The number of catalogued quasars, according to this model's predictions, increased by a factor of 4.

So What?

- A spectroscopic follow-up survey targeting quasars with high classification probabilities using the model and data is the natural next step.[4]

[4] Code repository: `https://github.com/informationcake/SDSS-ML`
Dataset repository: `https://zenodo.org/records/3768398`

So What?

- A spectroscopic follow-up survey targeting quasars with high classification probabilities using the model and data is the natural next step.[4]
- When designing new surveys, machine learning models like this can help optimize target selection.

[4] Code repository: `https://github.com/informationcake/SDSS-ML`
Dataset repository: `https://zenodo.org/records/3768398`

So What?

- A spectroscopic follow-up survey targeting quasars with high classification probabilities using the model and data is the natural next step.[4]
- When designing new surveys, machine learning models like this can help optimize target selection.
- Proposed scientific workflow:

[4] Code repository: `https://github.com/informationcake/SDSS-ML`
Dataset repository: `https://zenodo.org/records/3768398`

So What?

- A spectroscopic follow-up survey targeting quasars with high classification probabilities using the model and data is the natural next step.[4]
- When designing new surveys, machine learning models like this can help optimize target selection.
- Proposed scientific workflow:
  - Select quasar candidates based on classification probabilities.

[4] Code repository: `https://github.com/informationcake/SDSS-ML`
Dataset repository: `https://zenodo.org/records/3768398`

So What?

- A spectroscopic follow-up survey targeting quasars with high classification probabilities using the model and data is the natural next step.[4]
- When designing new surveys, machine learning models like this can help optimize target selection.
- Proposed scientific workflow:
  - Select quasar candidates based on classification probabilities.
  - Perform spectroscopic follow-up observations to confirm quasar nature.

[4] Code repository: `https://github.com/informationcake/SDSS-ML`
Dataset repository: `https://zenodo.org/records/3768398`

So What?

- A spectroscopic follow-up survey targeting quasars with high classification probabilities using the model and data is the natural next step.[4]
- When designing new surveys, machine learning models like this can help optimize target selection.
- Proposed scientific workflow:
  - Select quasar candidates based on classification probabilities.
  - Perform spectroscopic follow-up observations to confirm quasar nature.
  - Use confirmed quasars for cosmological studies and surveys, such as mapping large-scale structure or studying quasar evolution.

[4] Code repository: `https://github.com/informationcake/SDSS-ML`
Dataset repository: `https://zenodo.org/records/3768398`

References

[1] Clarke et al. (2020), "Identifying galaxies, quasars, and stars with machine learning: A new catalogue of classifications for 111 million SDSS sources without spectra", *Astronomy and Astrophysics*, Volume 639, A84.