Using a series of ML models for the detection of high-redshift Radio Galaxy candidates

+ **Rodrigo Carvajal**

+ Institute of Astrophysics and Space Sciences – U. of Lisbon

+ I. Matute, J. Afonso, S. Amarantidis, D. Barbosa (IA – U. Lisbon), P. Cunha, A. Humphrey (IA – U. Porto)

ia instituto de astrofísica e ciências do espaço

FCT Fundação para a Ciência e a Tecnologia

PD F FCT PhD PROGRAMMES

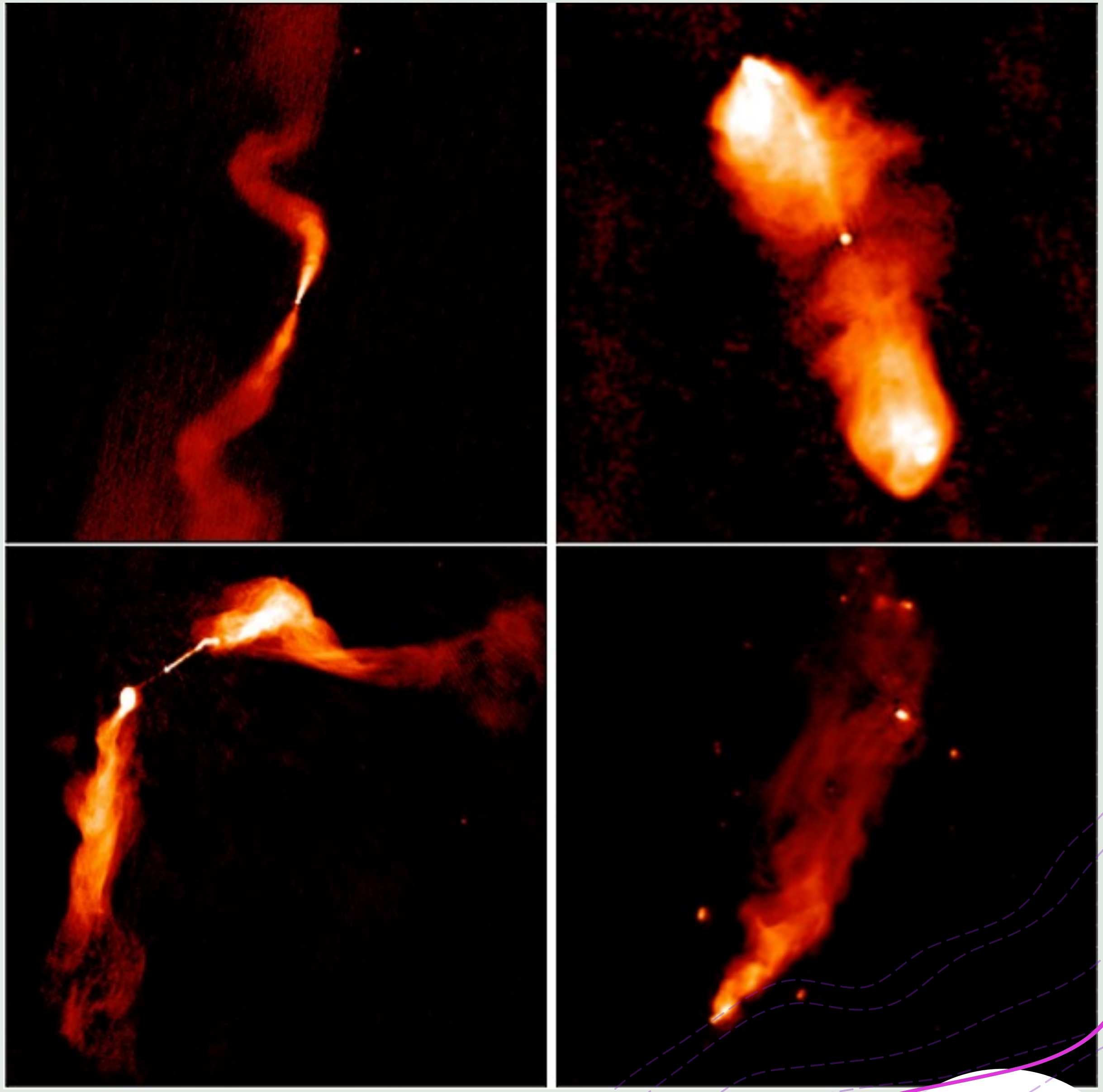
Ciências ULisboa

OCTOBER 22, 2021

DEBATING THE POTENTIAL OF ML IN ASTRONOMICAL SURVEYS

Radio Galaxies

+ Also known as Radio-Loud Active Galactic Nuclei (RLAGN).
+ AGN with radio emission strong enough to be detected.
+ In general, we focus on high-redshift RGs → EoR epoch and AGN evolution.



Hardcastle & Croston, 2020

Issues with RGs

Table 1. Big Data 3V characteristics in astronomical sky surveys.

Sky Survey	Volume	Velocity	Variety
SDSS *Sloan Digital Sky Survey*	50 TB	200 GB per day	images, catalogs, redshifts
GAIA	100 TB	40 GB per day	more then 100 parameters
Pan-STARRS *Panoramic Survey Telescope and Rapid Response System*	5 PB	5 TB per day	images, catalogs
LSST *Large Synoptic Survey Telescope*	60 PB	10 TB per day	images, catalogs
SKA *Square Kilometer Array*	3 ZB	150 TB per day	images, catalog, redshifts

Notes:
The column Volume refers to raw data produced at the end of the experiment.
Values regarding Pan-STARRS, LSST, and SKA surveys refer to expected Volume and Velocity values.

Garofalo et al., 2016

+ High-redshift AGN hard to detect.
+ Redshift determination (SED fit) takes long time.
+ Most detections in optical/NIR. We lack radio observations.
+ Future (and present) radio surveys produce large data volumes.
+ **Traditional (radio) AGN detection methods will be inefficient.**

Issues with RGs

Table 1. Big Data 3V characteristics in astronomical sky surveys.

Sky Survey
SDSS *Sloan Digital Sky Survey*
GAIA
Pan-STARRS *Panoramic Survey Telescope and Rapid Response System*
LSST *Large Synoptic Survey Telescop*
SKA *Square Kilometer Array*

Notes:
The column Volume refers to raw data
Values regarding Pan-STARRS, LSST, and SKA surveys refer to expected Volume and Velocity values.

Garofalo et al., 2016

+ High-redshift AGN hard to detect.

SED fit) takes long

al/NIR. We lack

io surveys produce

detection

ient.

How can we try to solve some of these problems?

OCTOBER 22, 2021

DEBATING THE POTENTIAL OF ML IN ASTRONOMICAL SURVEYS

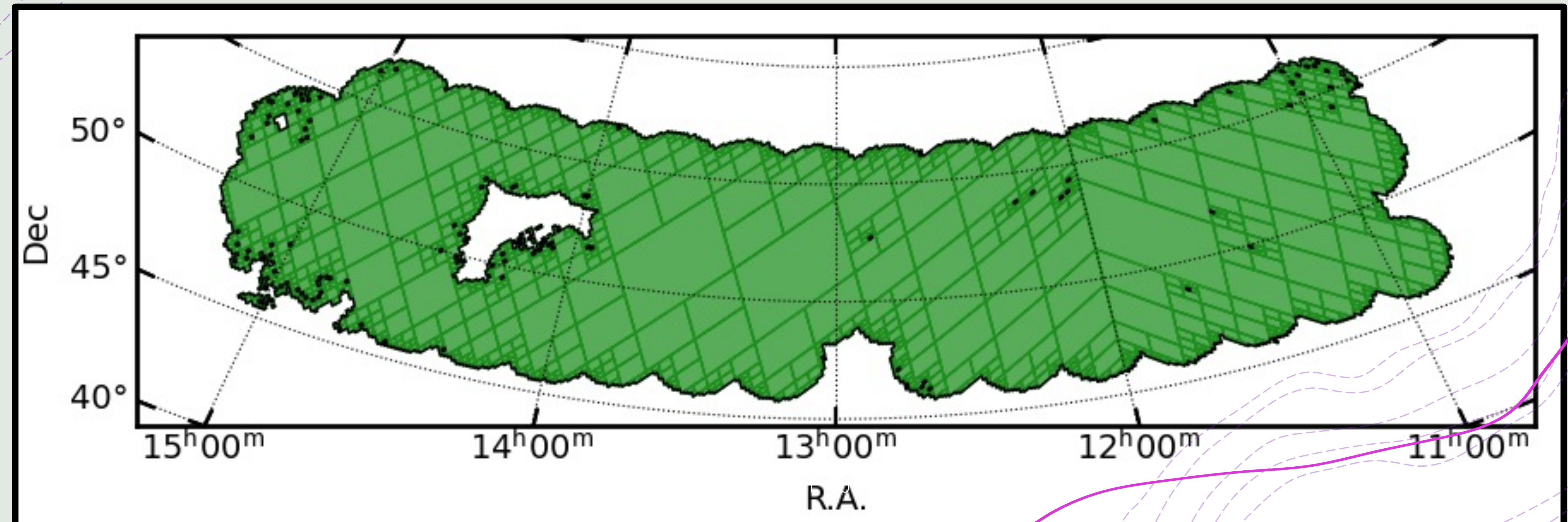
We aim to obtain…

- High-redshift RG candidates
  - AGN
  - Radio emission
  - Redshift

- Series of models
  - Control over features
  - **Interpretability**

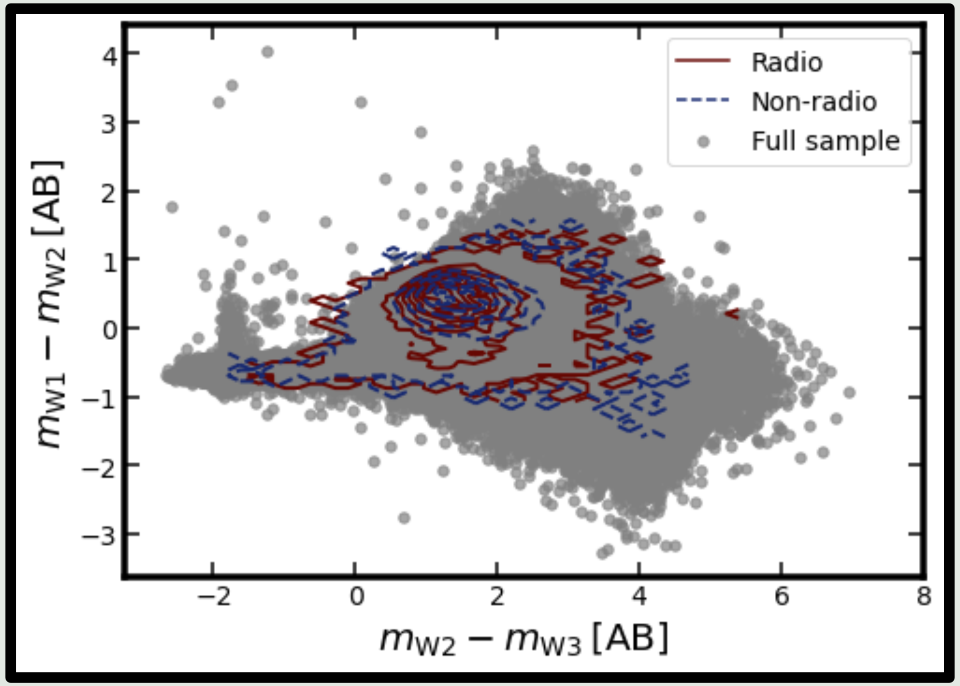
Data: HETDEX Spring Field

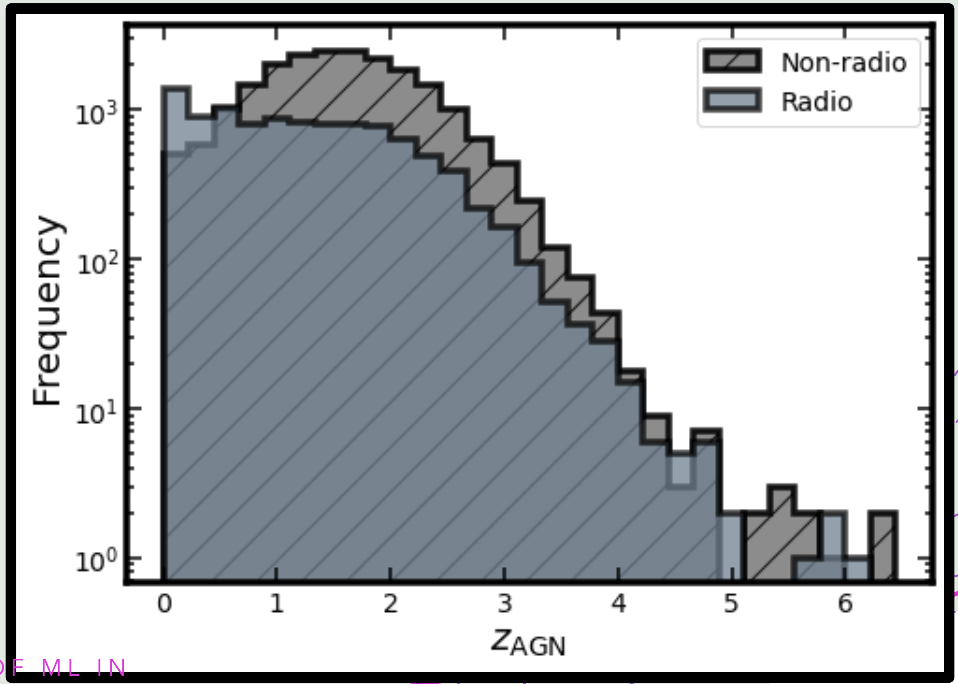
- ~400 deg^2 in the northern sky (as covered by LoTSS DR-1).
- 6,729,647 detections in NIR (CatWISE2020, Marocco+2020).
- Counterparts in:
  - Radio (LOFAR, GMRT, VLASS)
  - IR (AllWISE, 2MASS)
  - Visible+NIR (Pan-STARRS)
  - UV (GALEX)
  - X-ray (XMM-Newton)
- Cross-match with Million Quasar Catalog (v7.2, Flesch 2021)



Data Preparation

+ Imputation: limiting magnitude (20 bands).
+ Colours and magnitude ratios.
+ Flags: AGN, radio, X-ray.
+ 32,365 identified AGN (0.48%)





Carvajal et al., submitted

Models Preparation

+ Train (90%) – Validation (10%)
+ Feature selection with Boruta.
+ Model stacking.
+ Fix unbalance for radio model.

$$\Delta z^N = \frac{|z_{\text{true}} - z_{\text{pred}}|}{1 + z_{\text{true}}}$$

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN})}}$$

Combining All Predictions

Combining Predictions

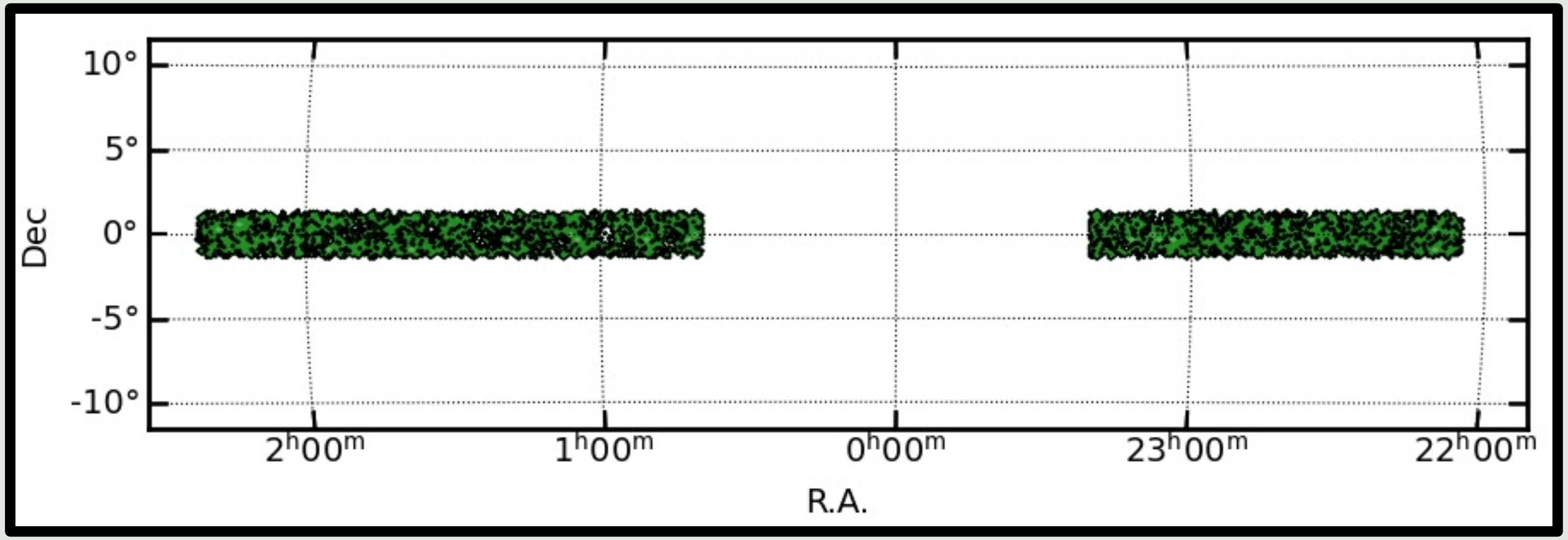
AGN Prediction

↓

Radio Prediction

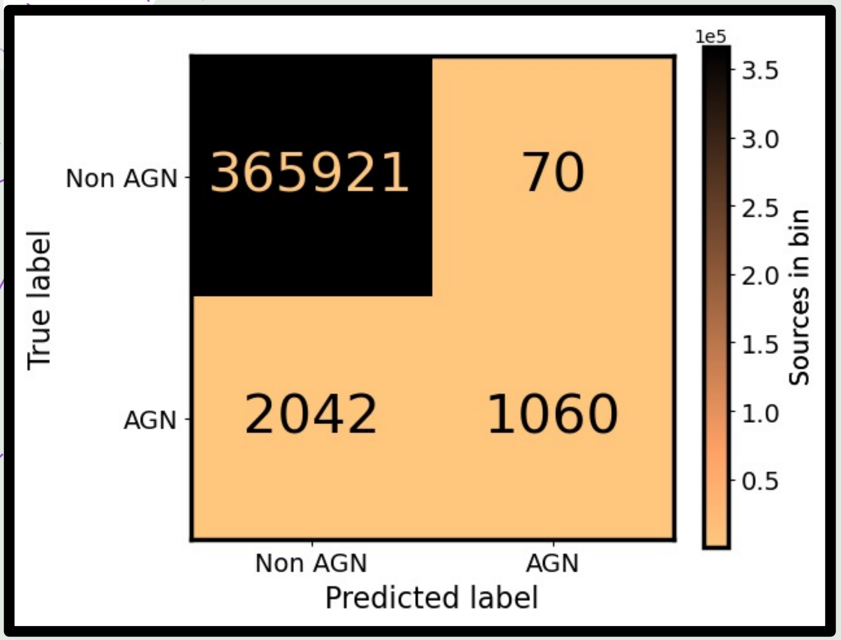
↓

Redshift Prediction



- SDSS Equatorial Strip in the Southern Galactic Cap (92 deg^2).
- Equal data collection as with HETDEX (minus LOFAR 150 MHz).
- 369,093 objects in CatWISE2020
- 2,941 objects labelled as AGN.

AGN Detection

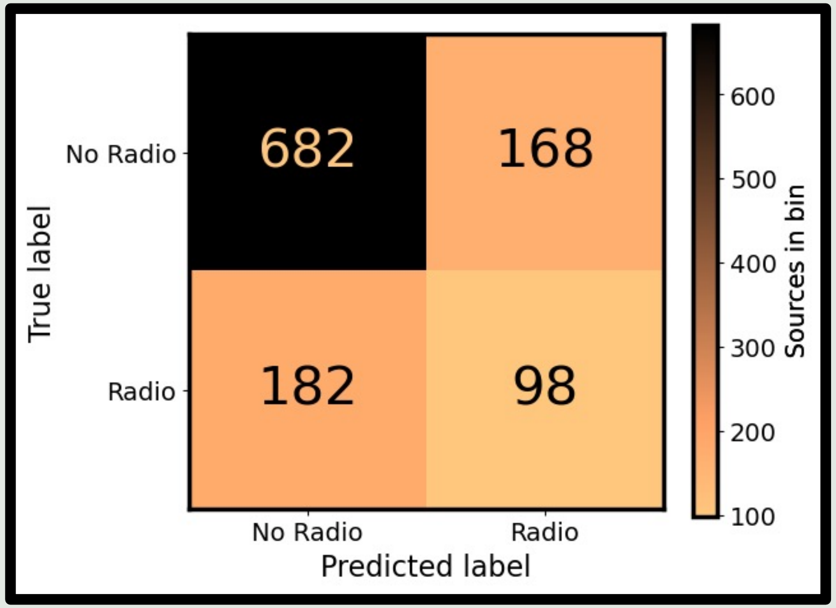


A

MCC = 0.564

1,130 AGN

Radio Detection



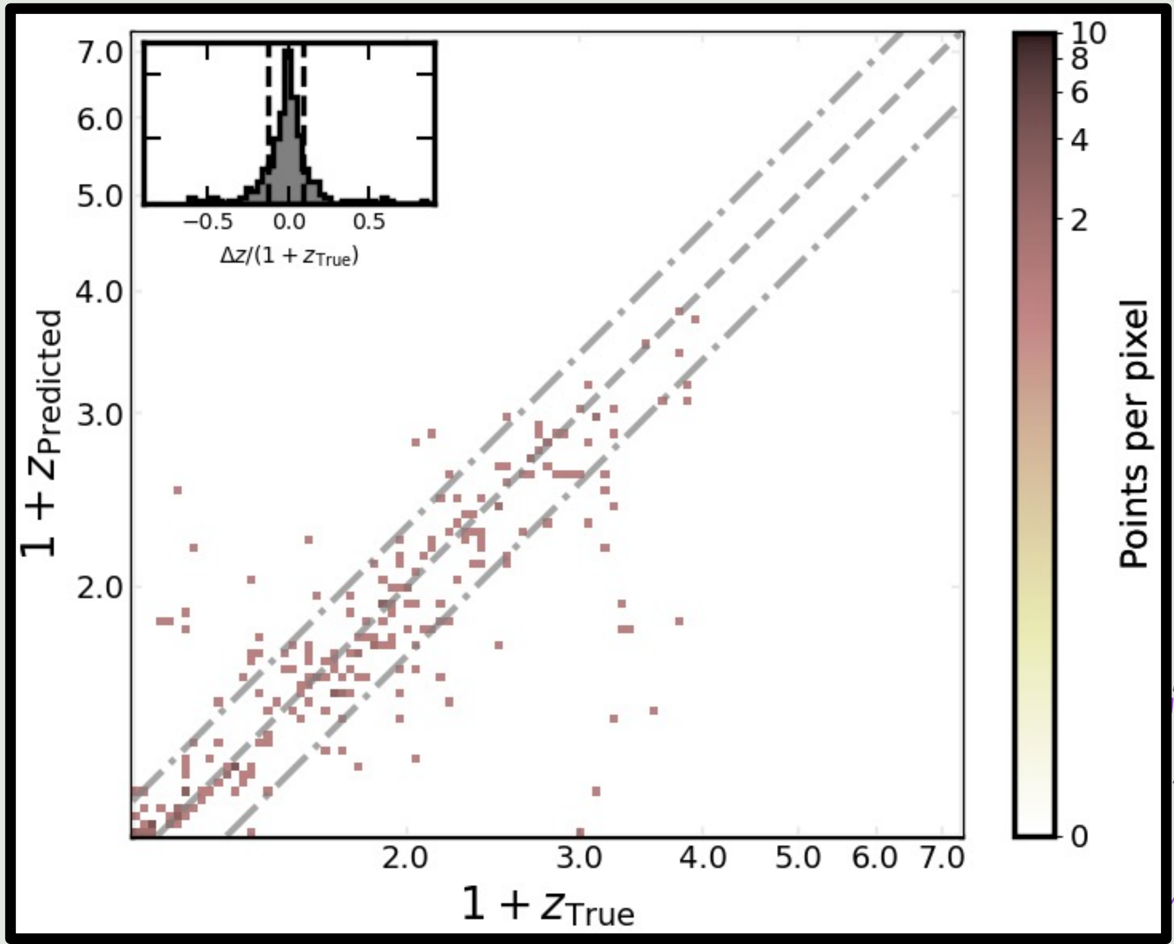
B

MCC = 0.155

Redshift

$\Delta z^{N} = 0.091$

C



All predicted radio AGN

Thus, we have 266 radio AGN candidates!

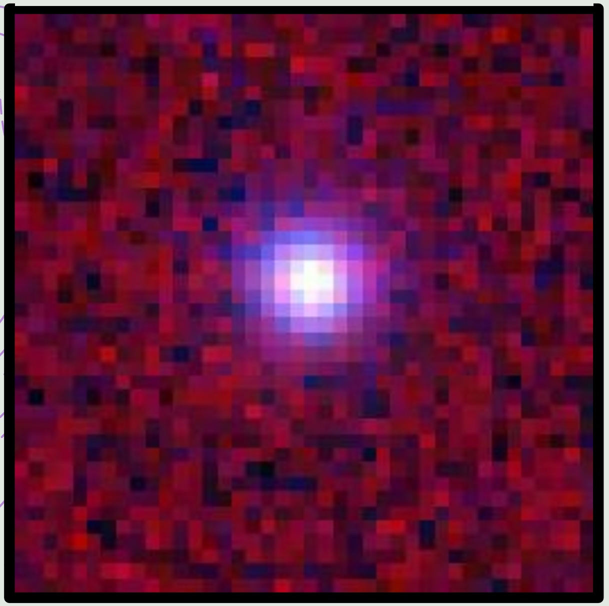
Δz^N

	RA_ICRS	DE_ICRS	is_QSO	Label_AGN	radio_detect	Label_radio	Z	Pred_Z	Z_score
180296	19.751310	-0.032450	1	1	0	1	2.750	2.786	0.010
333924	331.784688	1.023693	1	1	1	1	2.911	2.763	0.038
80636	24.262678	-0.704125	1	1	0	1	2.502	2.549	0.013
349091	14.436190	1.138420	1	1	0	1	2.762	2.449	0.083
326594	334.009676	0.974047	1	1	0	1	2.864	2.221	0.166
123079	12.557420	-0.412885	1	1	0	1	2.035	2.189	0.051
330825	349.030179	1.003573	1	1	1	1	2.638	2.109	0.145
145426	10.847652	-0.264569	1	1	1	1	2.820	2.084	0.193
261855	333.072000	0.540843	1	1	0	1	2.265	2.024	0.074
276469	17.476529	0.636962	1	1	0	1	1.975	2.014	0.013
255854	340.978137	0.501055	1	1	0	1	2.125	1.999	0.040
279768	340.109204	0.661146	1	1	0	1	2.111	1.967	0.046
178119	30.344848	-0.046815	1	1	0	1	1.514	1.956	0.176
177225	13.479837	-0.052584	1	1	0	1	1.714	1.907	0.071
76036	10.291814	-0.736649	1	1	0	1	1.823	1.883	0.021

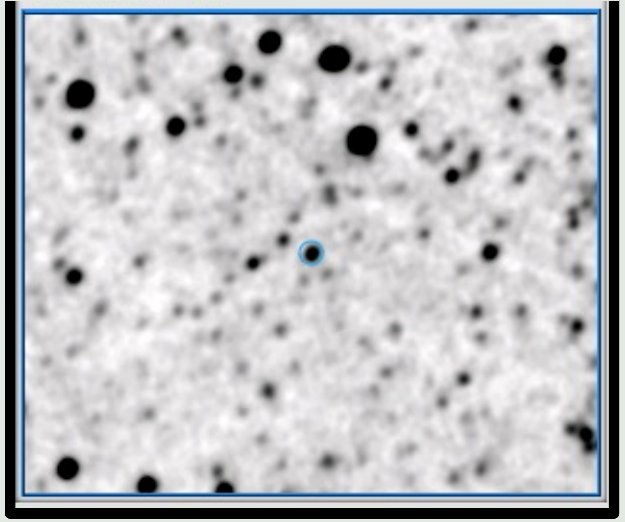
etc...

Id: 180296 Prediction: Radio-AGN z=2.786

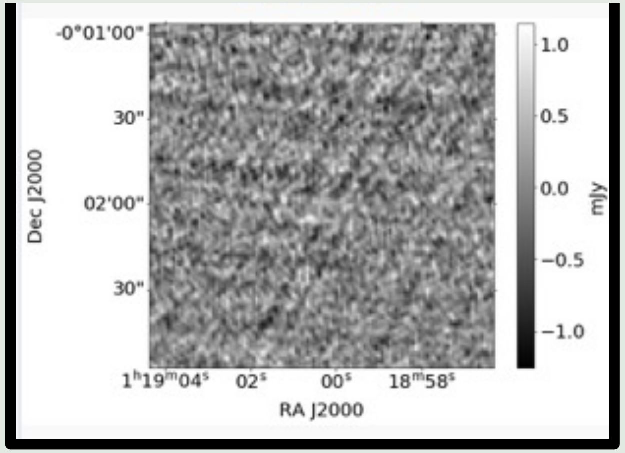
Pan-STARRS 1 (y/i/g)



WISE (W1)



VLASS (3 GHz)

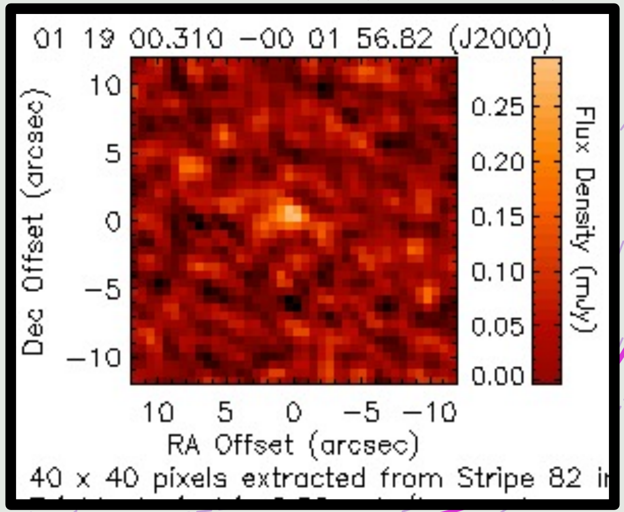


Basic data :

SDSS J011900.32-000156.7 -- Quasar

Other object types:	QSO (2012MNRAS,[MHP2012]), * (Gaia), Q? (2011AJ)
ICRS coord. *(ep=J2000) :*	01 19 00.3195546172 -00 01 56.874776256 (Optica
FK4 coord. *(ep=B1950 eq=1950) :*	01 16 26.6265000543 -00 17 42.086509446 [0.158
Gal coord. *(ep=J2000) :*	137.7938893149207 -62.1061206804555 [0.1582 0.
Proper motions *mas/yr* :	0.220 0.180 [0.373 0.220 90] A 2018yCat.1345...
Radial velocity / Redshift / cz :	V(km/s) 257538 [42] / z(spectroscopic) 2.63177 (Opt) C 2012ApJS..203...21A
Parallaxes *(mas)*:	0.0443 [0.1785] A 2018yCat.1345....0G
Fluxes (9) :	G 18.0710 [0.0038] C 2018yCat.1345....0G
	J 16.931 [0.015] D 2012MNRAS.424.2876M
	H 16.475 [0.031] D 2012MNRAS.424.2876M
	K 16.118 [0.027] D 2012MNRAS.424.2876M
	u (AB) 19.75 [0.03] C 2012ApJS..203...21A
	g (AB) 18.554 [0.007] B 2012ApJS..203...21A
	r (AB) 18.233 [0.008] B 2012ApJS..203...21A
	i (AB) 18.123 [0.008] B 2012ApJS..203...21A
	z (AB) 18.089 [0.022] C 2012ApJS..203...21A

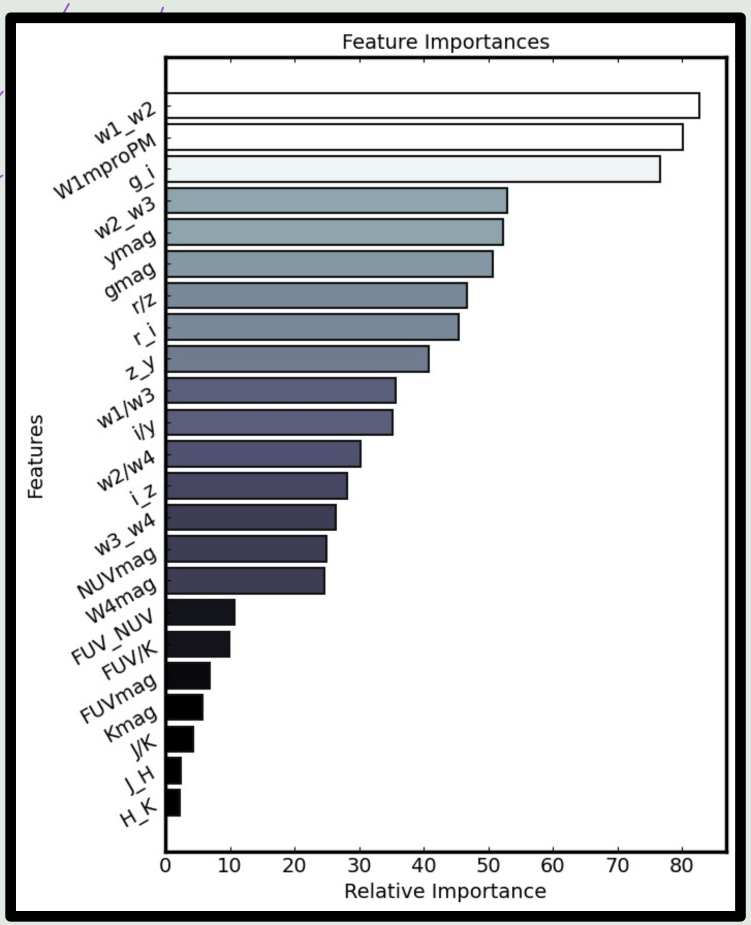
VLA SDSS 82 (1.4 GHz)



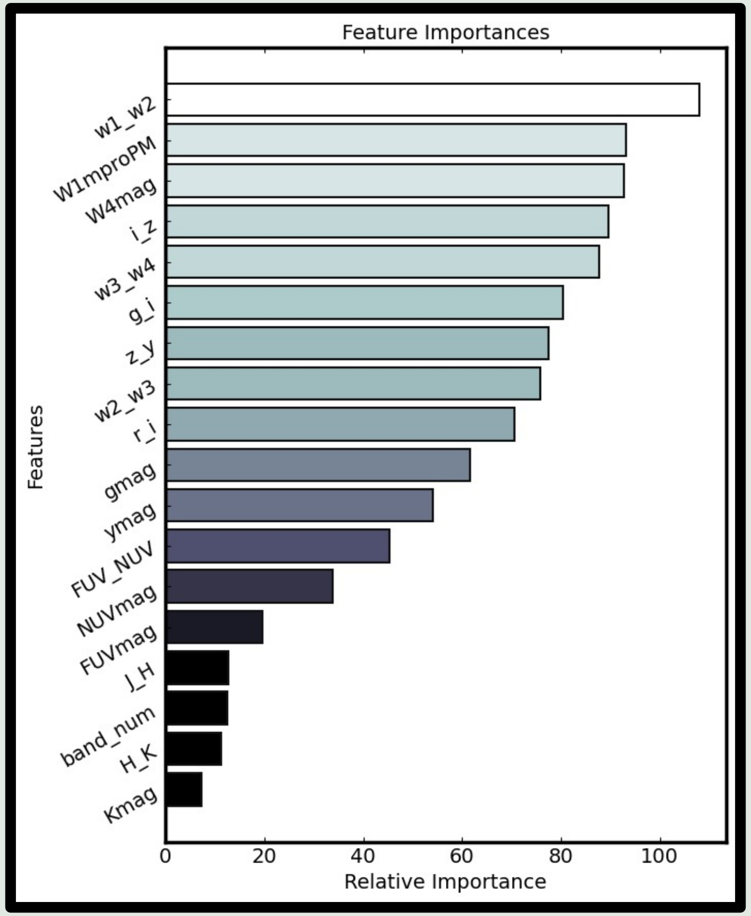
We can also extract information from the models themselves!

Feature Importances

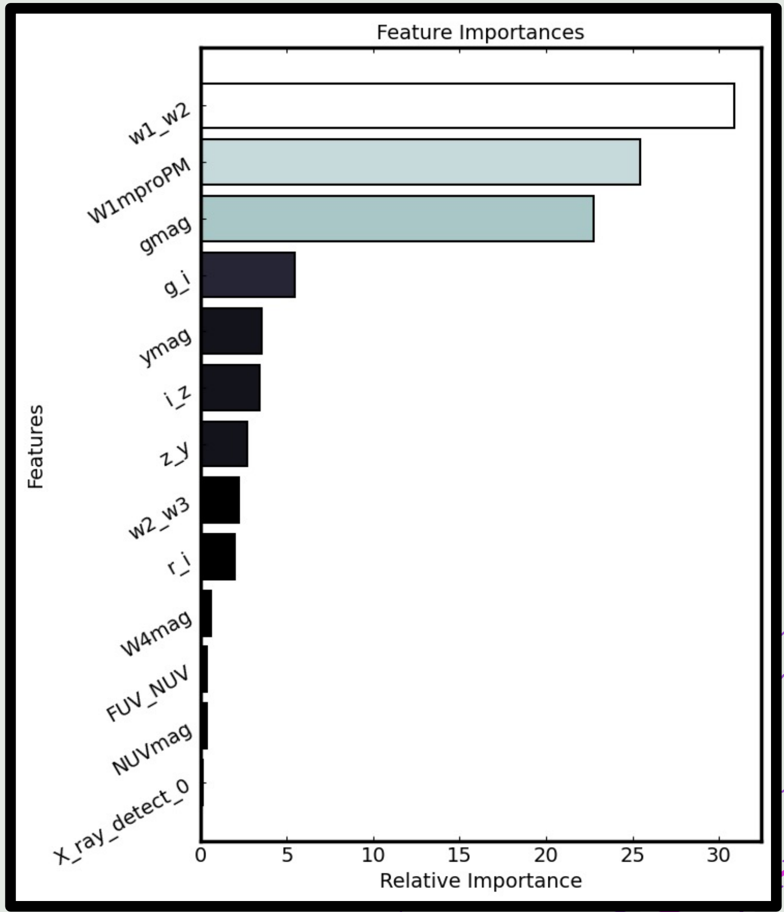
AGN detection



Radio detection

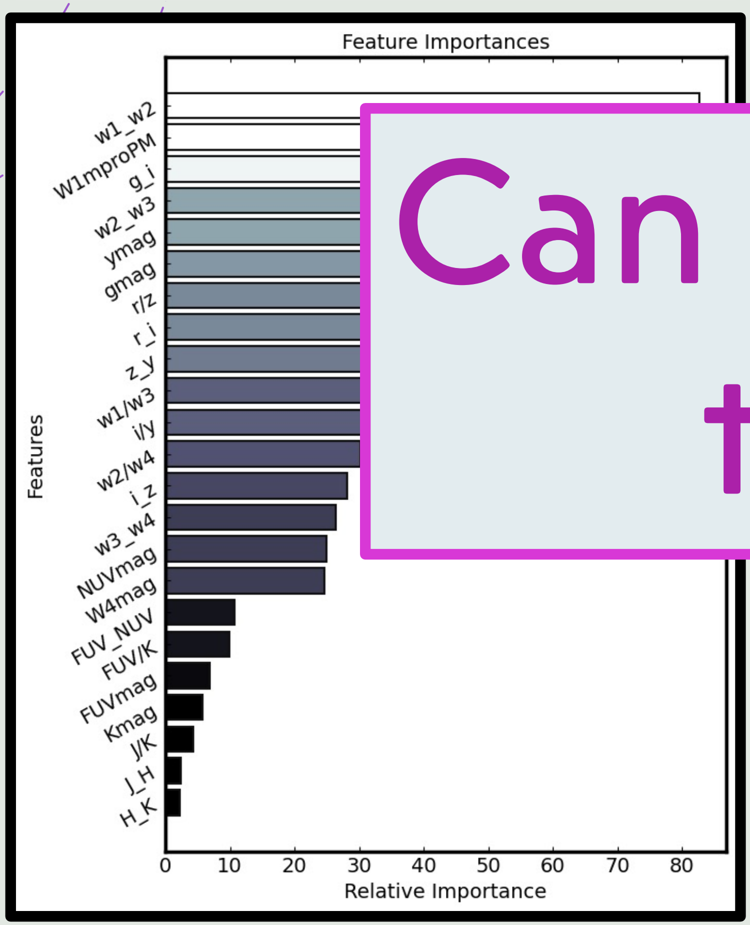


Redshift value



Feature Importances

AGN detection



Radio detection

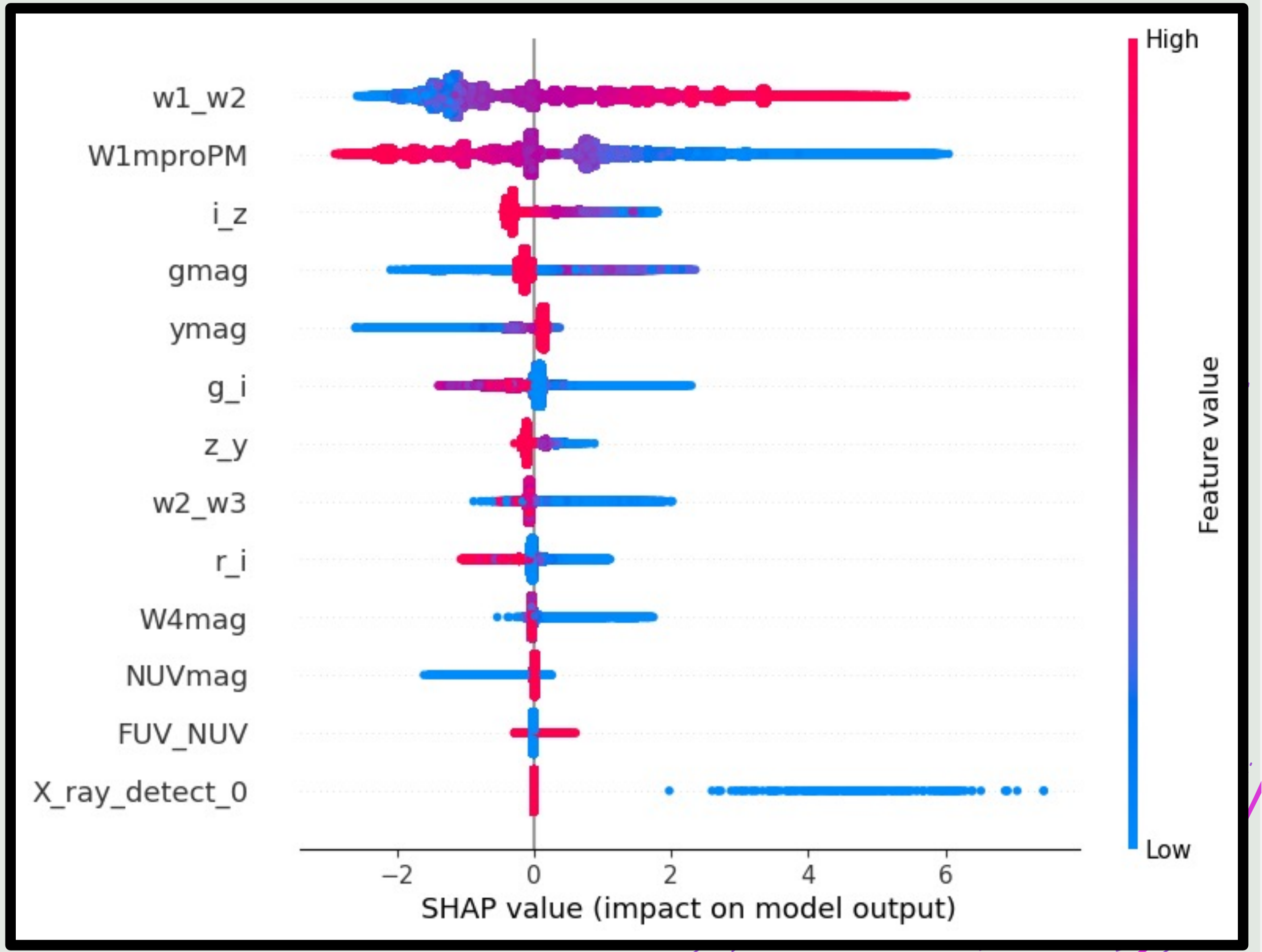
Redshift value

Can we do better than this?

Shapley Values

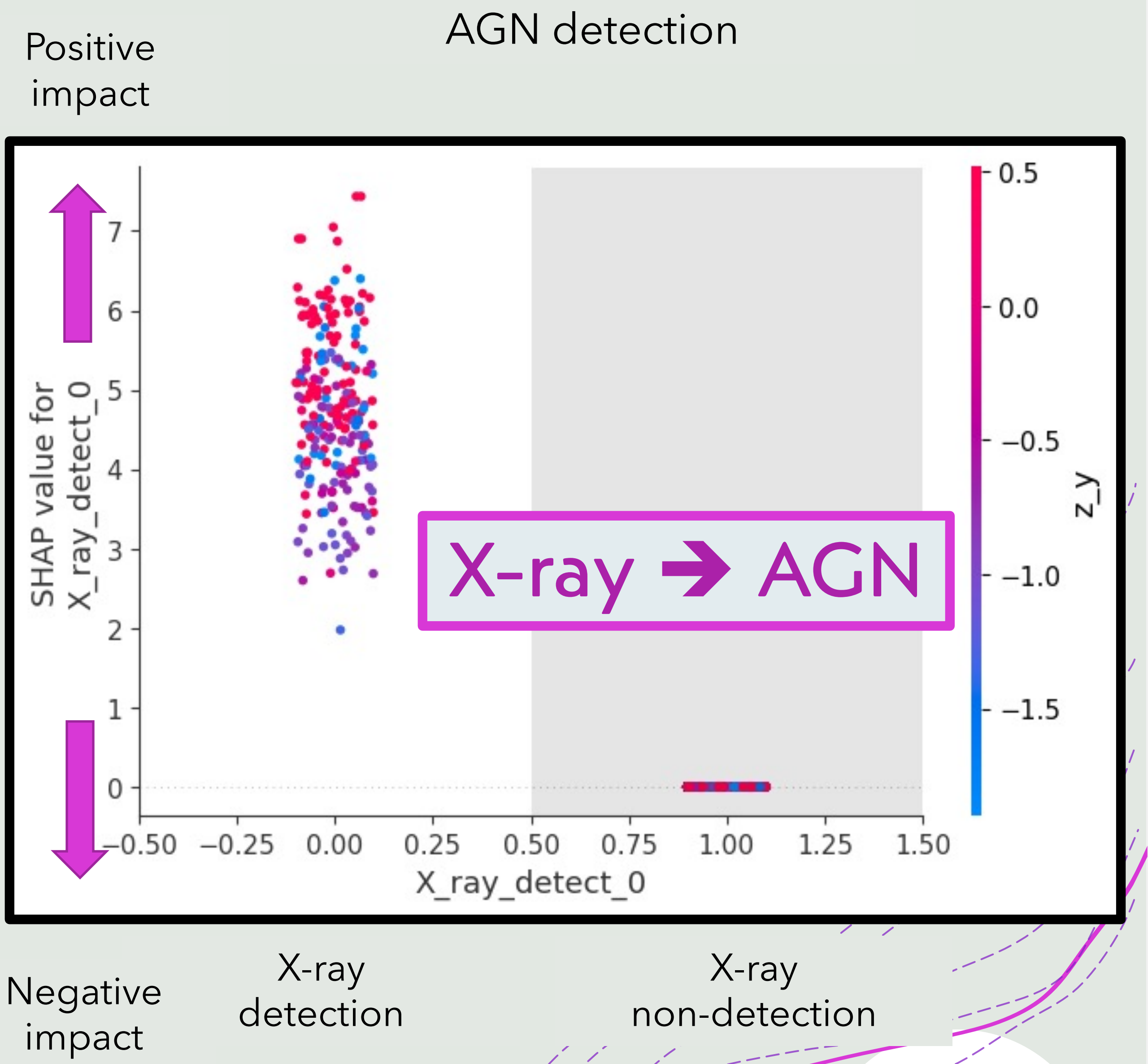
- From Game Theory (Shapley, 1953).
- They show how each feature impacts the final prediction (per source).
- High Shapley value increase probability of detection or high redshift.
- Allows analysis of interplay between features.

AGN detection



Shapley Values

- From Game Theory (Shapley, 1953).
- They show how each feature impacts the final prediction (per source).
- High Shapley value increase probability of detection or high redshift.
- Allows analysis of interplay between features.



Final Thoughts

Final Toughts

+ No need for fully clean data to obtain meaningful results.
+ Some degree of transferability with minor changes in dataset.
+ Using series of models useful to understand each step.
+ ML models can give insight over probably hidden correlations among features (new discoveries?).

Future Steps

+ Include uncertainties
+ Tackle imbalance (AGN, radio, z)
+ Include morphological properties.

Thank you for your attention!

Questions? Comments?

rcarvajal@oal.ul.pt

@r_carvajalp

racarvajal.github.io

@racarvajalp

This work has been supported by the Fundação para a Ciência e a Tecnologia (FCT) through the Fellowship PD/BD/150455/2019 (PhD::SPACE Doctoral Network PD/00040/2012) and POCH/FSE (EC), and through research grants PTDC/FIS-AST/29245/2017, UID/FIS/04434/2019, UIDB/04434/2020 and UIDP/04434/2020.

Using a series of ML models for the detection of high-redshift Radio Galaxy candidates

+ **Rodrigo Carvajal**

+ Institute of Astrophysics and Space Sciences – U. of Lisbon

+ I. Matute, J. Afonso, S. Amarantidis, D. Barbosa (IA – U. Lisbon), P. Cunha, A. Humphrey (IA – U. Porto)

instituto de astrofísica e ciências do espaço

FCT Fundação para a Ciência e a Tecnologia

PD F FCT PhD PROGRAMMES

Ciências ULisboa

OCTOBER 22, 2021

DEBATING THE POTENTIAL OF ML IN ASTRONOMICAL SURVEYS