Stellar classification

- **Dataset**: Sloan Digital Sky Survey DR17, 100 000 observations, 17 feature columns, 3 classes
- **Business Case**: Suppose a team of astrophysicists needs a model to classify celestial objects reliably. Given the high cost associated with further research on classified objects, maximizing the **precision** of the classification model is paramount.
- **Metrics**: weighted precision, accuracy

Authors: Igor Kołodziej, Kamil Eliaszuk

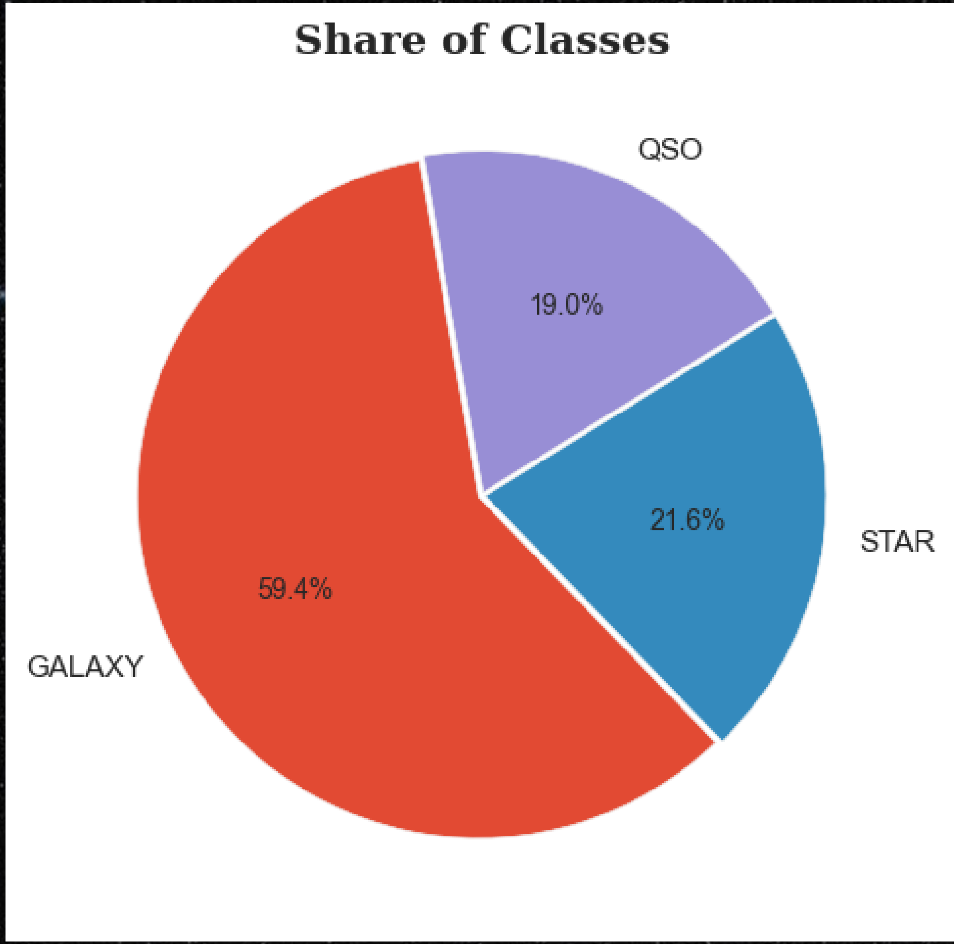

Stars
Distance from earth:
~ light-years


Galaxies
Distance from earth:
~ millions of light-years

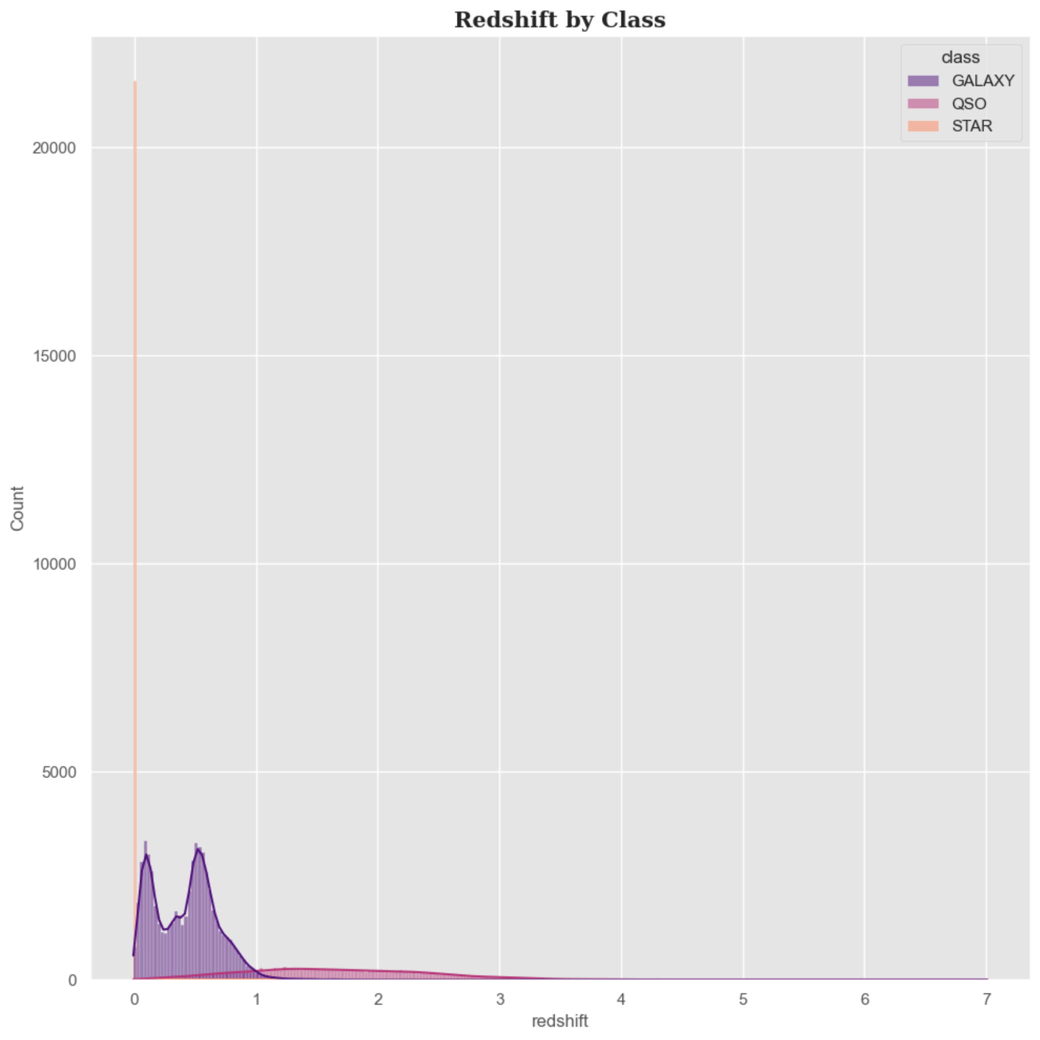

Quasars
Distance from earth:
~ billions of light-years

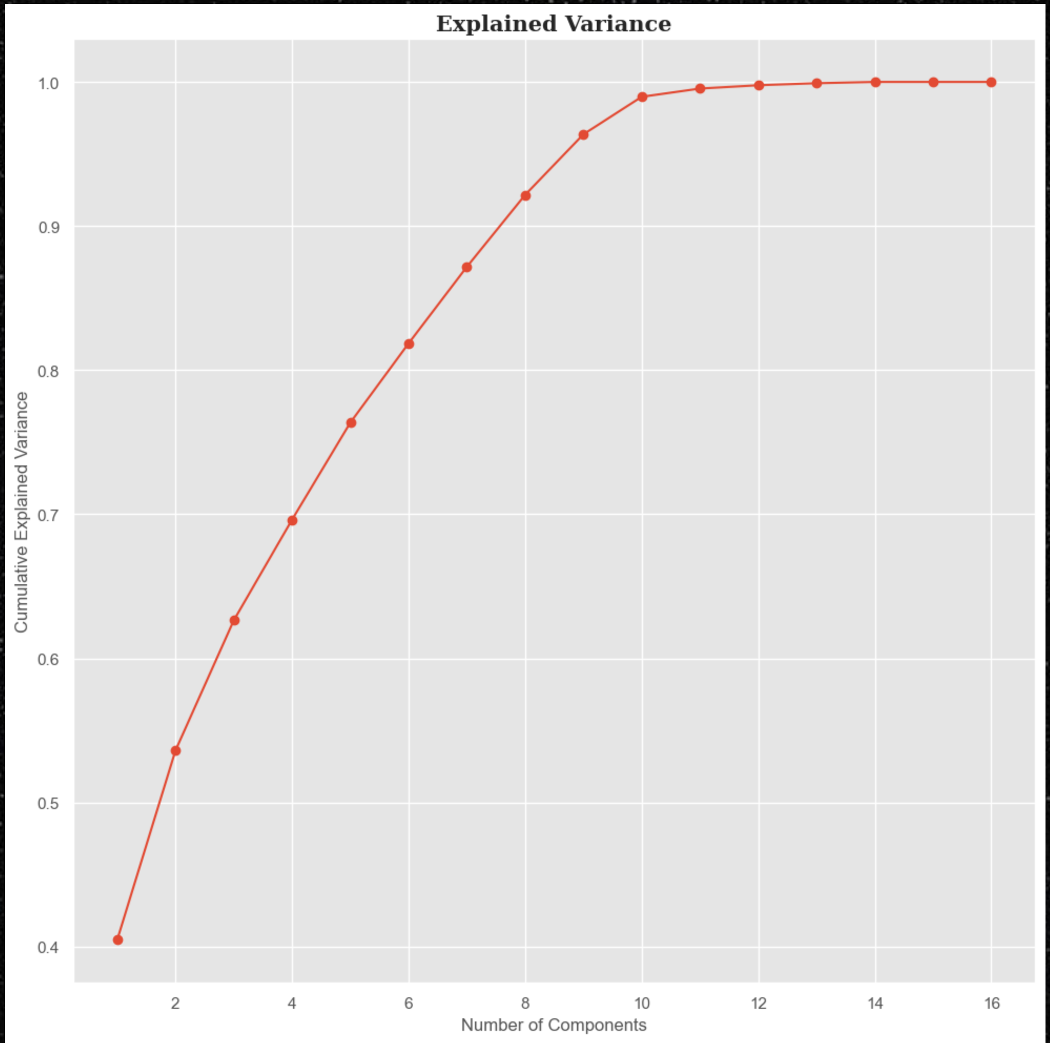
A look at the data

- **obj_ID:** Object Identifier
- **alpha:** Right Ascension angle (at J2000 epoch)
- **delta:** Declination angle (at J2000 epoch)
- **u:** Ultraviolet filter
- **g:** Green filter
- **r:** Red filter
- **i:** Near Infrared filter
- **z:** Infrared filter
- **run_ID:** Run Number
- **rereun_ID:** Rerun Number
- **cam_col:** Camera column
- **field_ID:** Field number
- **spec_obj_ID:** Unique ID for optical spectroscopic objects
- **class:** Object class (galaxy, star, or quasar)
- **redshift:** Redshift value
- **plate:** Plate ID
- **MJD:** Modified Julian Date
- **fiber_ID:** Fiber ID



A look at the data



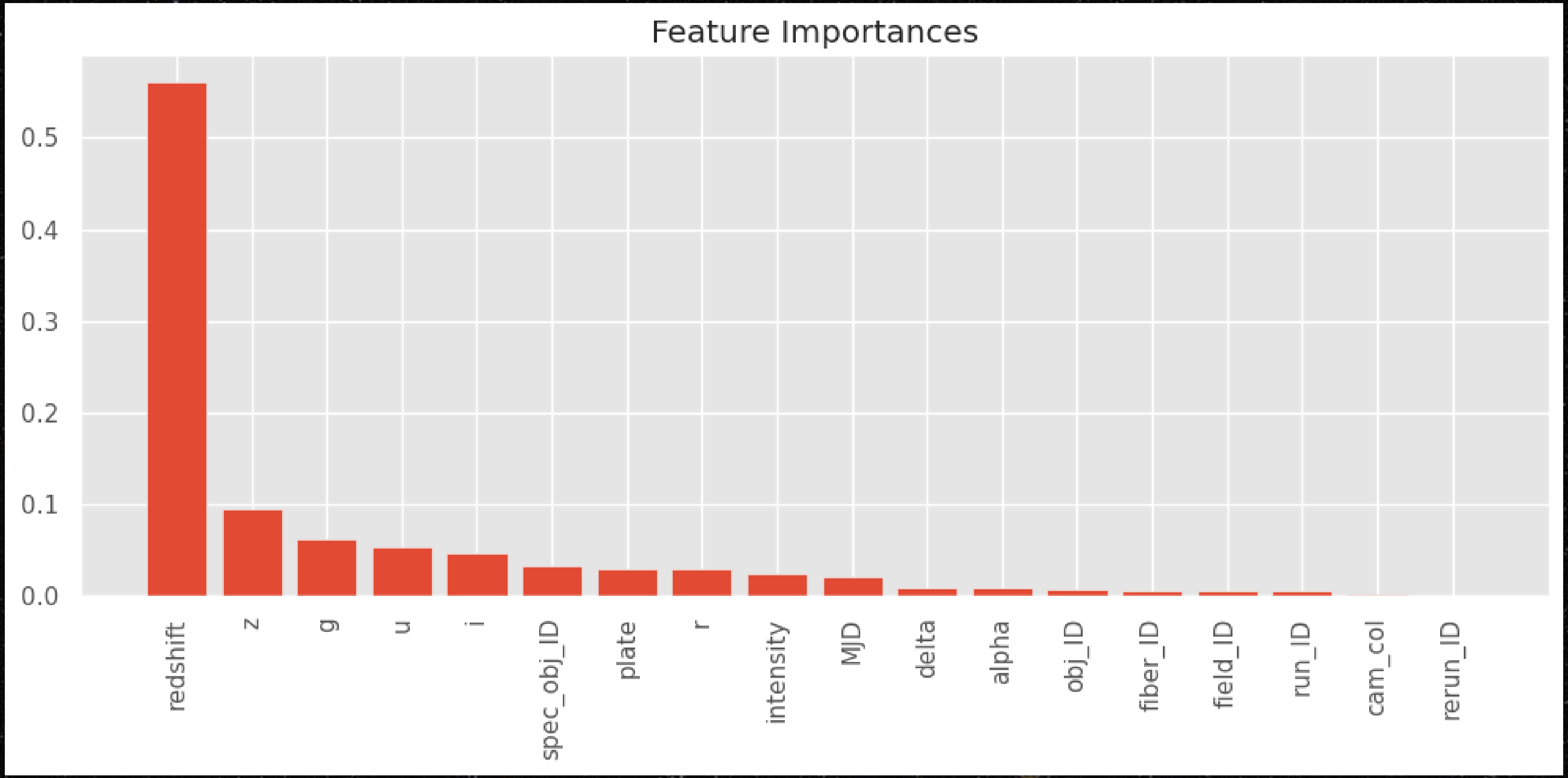


Key feature:
redshift (proportional to the distance from the earth)

PCA: high explainability with a low number of components

Data preparation

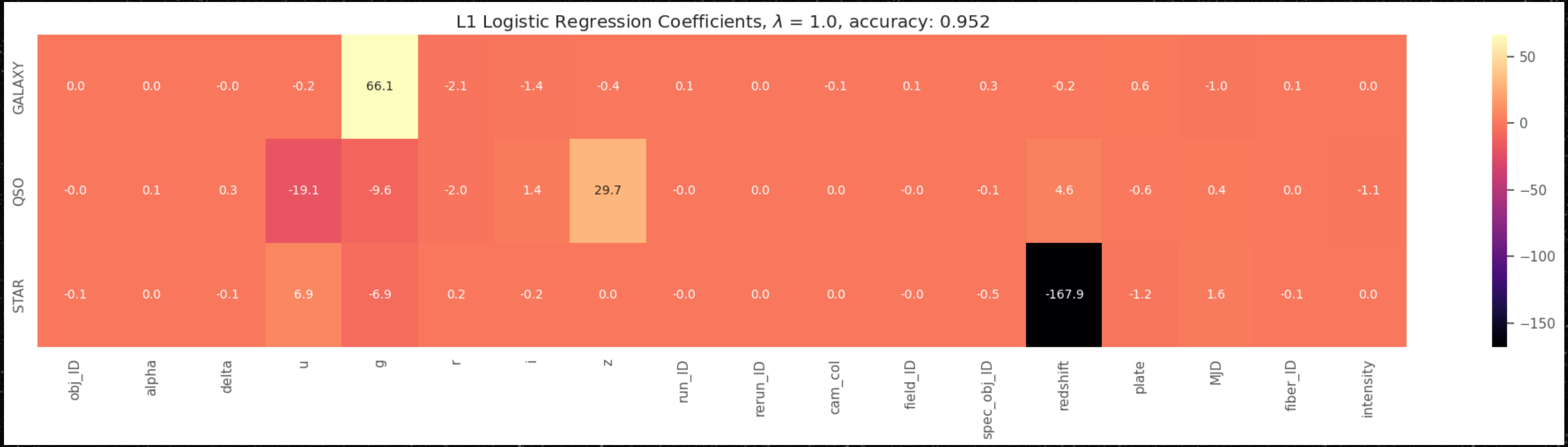
- No missing data
- Only continuous features
- Target class encoding
- Feature standarization
- Feature selection



Random forest feature importance

Data preparation

Logistic regression with regularization



We reject variables of the type ID, date, or variables that are strongly correlated with each other.

We only leave 4 features: u, g, z and redshift.
The model practically does not lose accuracy, but it is much simpler and easier to explain.

Model choice

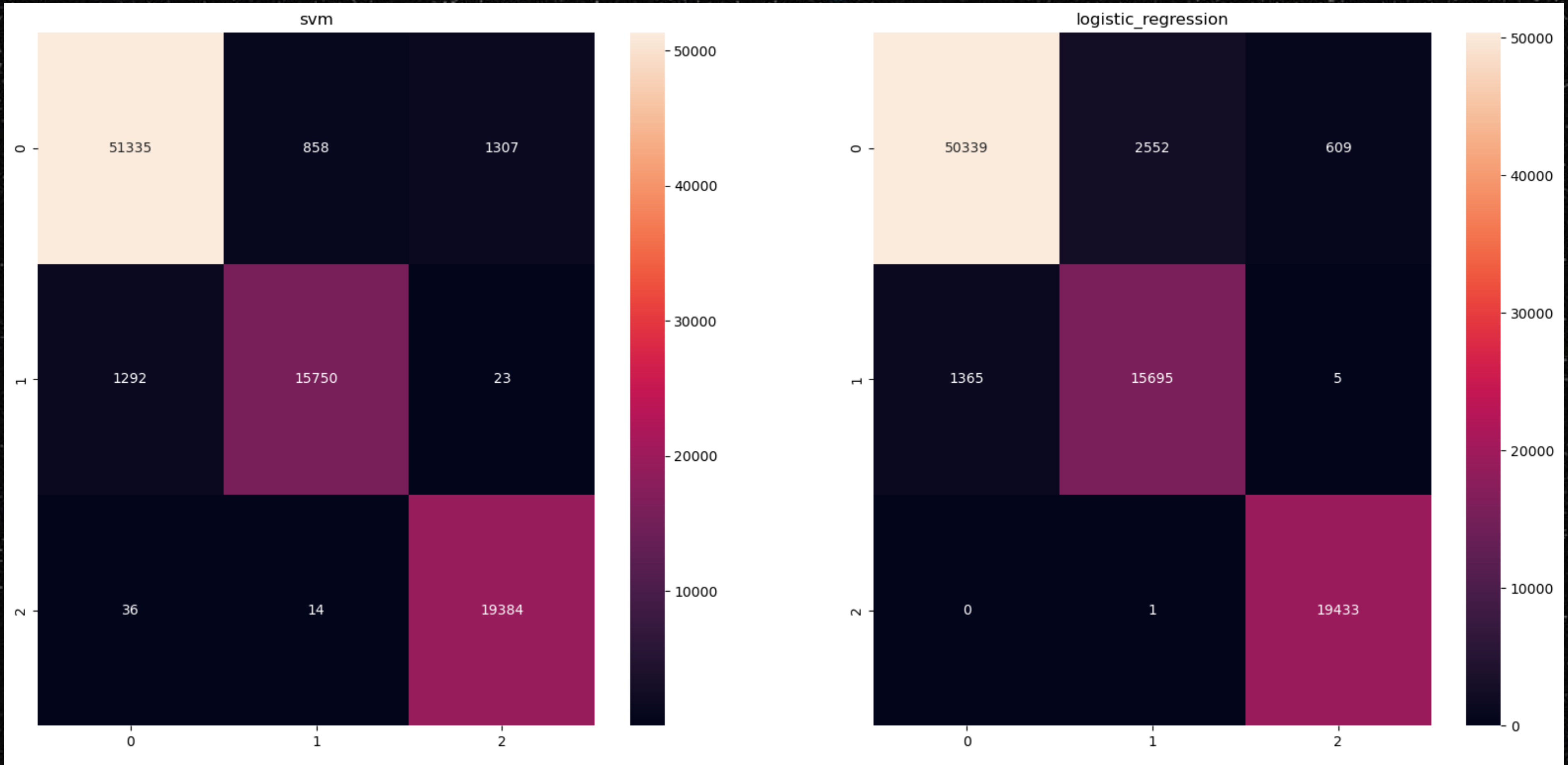
Hyperparameter tuning – Bayesian optimalization

Best models of different types (weighted precision):

- Random forest: 0.977
- SVM: 0.970
- XGBoost: 0.974
- Regresja logistyczna: 0.959

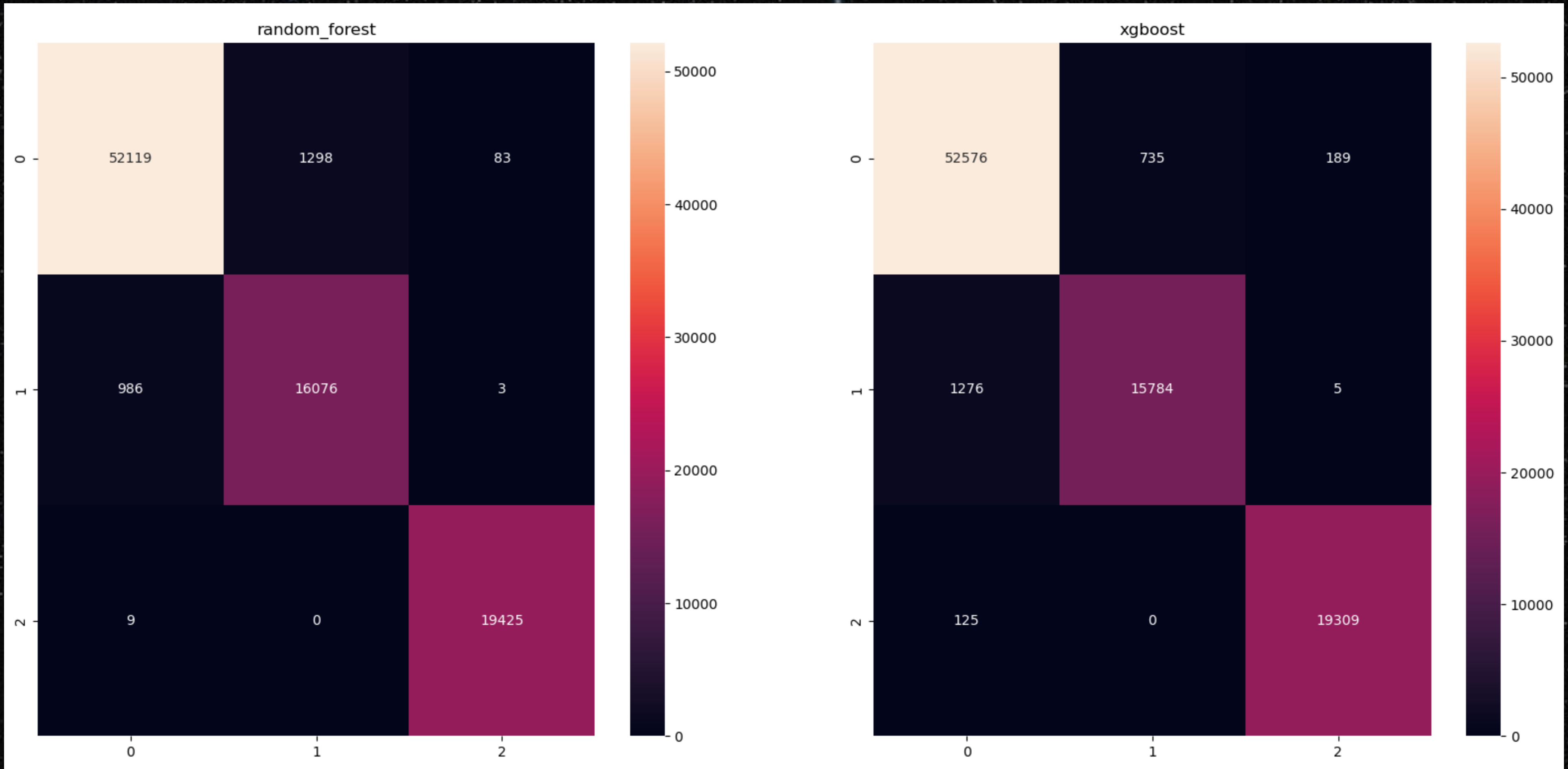
Stacking: 0.977 – high score but not higher than RF

Model choice



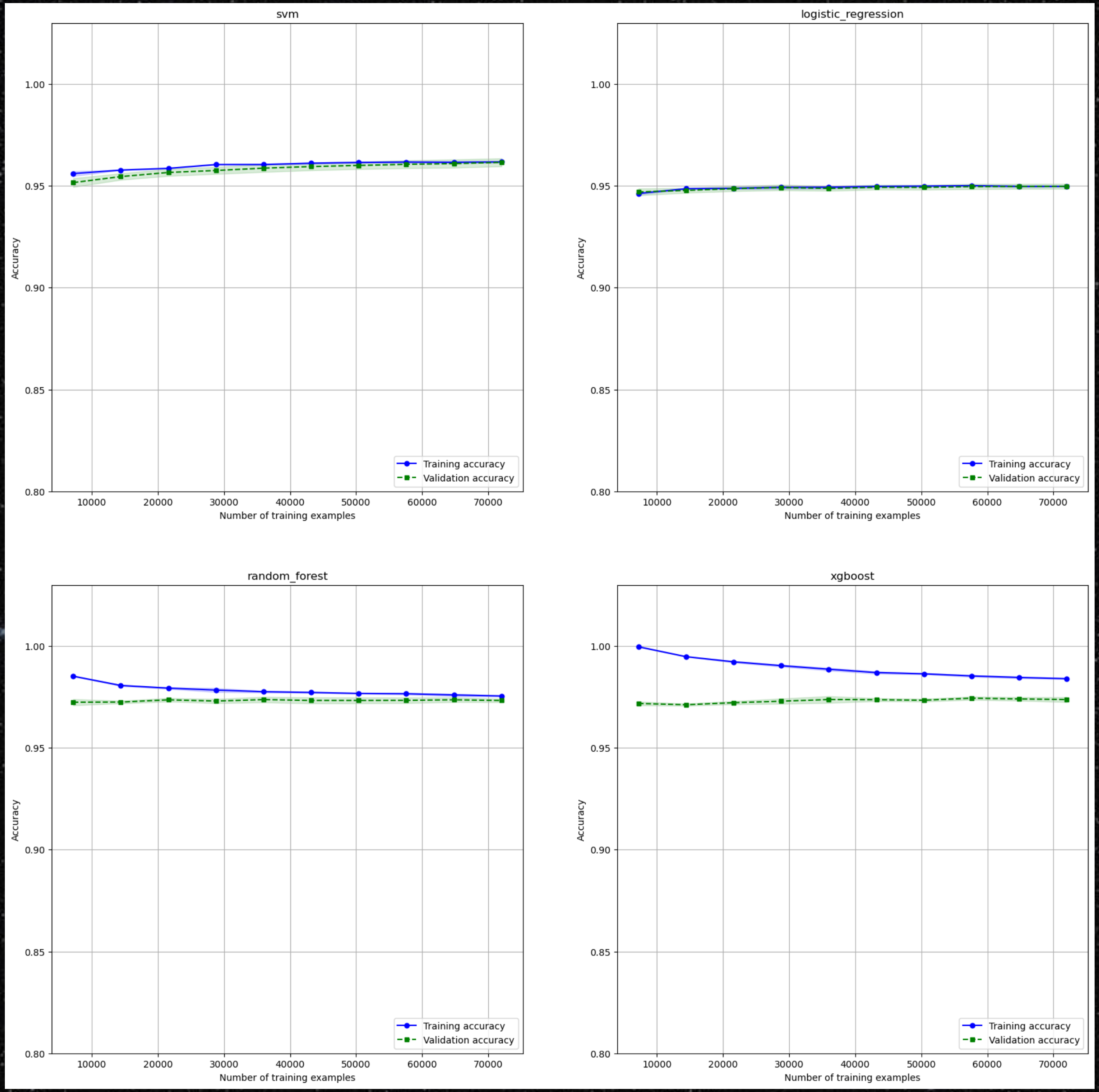
0 – galaxy, 1 – quasar, 2 - star

Model choice



0 – galaxy, 1 – quasar, 2 - star

Model choice



Model choice

Imbalanced learning techniques comparison (imbalanced-learn package) – unsatisfying results

		mean	std
random_forest	do_nothing	0.973679	0.000815
	oversample	0.973405	0.001216
	undersample	0.971543	0.001455
	smote	0.971480	0.000916
	smoteenn	0.967144	0.001128
svm	do_nothing	0.962068	0.001721
	smote	0.958986	0.001918
	oversample	0.958556	0.001907
	smoteenn	0.954595	0.002268
	undersample	0.954555	0.001966
logistic_regression	oversample	0.951581	0.001154
	smote	0.951375	0.001259
	do_nothing	0.951261	0.001098
	undersample	0.950328	0.001353
	smoteenn	0.949855	0.001419

Metric: weighted precision

```
epoch 0  | loss: 0.31288 | val_0_accuracy: 0.96056 |  0:00:03s
epoch 1  | loss: 0.13329 | val_0_accuracy: 0.90372 |  0:00:06s
epoch 2  | loss: 0.12309 | val_0_accuracy: 0.97056 |  0:00:09s
epoch 3  | loss: 0.11595 | val_0_accuracy: 0.76022 |  0:00:11s
epoch 4  | loss: 0.10824 | val_0_accuracy: 0.94956 |  0:00:14s
epoch 5  | loss: 0.10618 | val_0_accuracy: 0.96817 |  0:00:17s
epoch 6  | loss: 0.10475 | val_0_accuracy: 0.96278 |  0:00:20s
epoch 7  | loss: 0.10461 | val_0_accuracy: 0.90133 |  0:00:22s
epoch 8  | loss: 0.10476 | val_0_accuracy: 0.75961 |  0:00:25s
epoch 9  | loss: 0.1039  | val_0_accuracy: 0.76022 |  0:00:28s
epoch 10 | loss: 0.10058 | val_0_accuracy: 0.7645  |  0:00:31s
epoch 11 | loss: 0.10163 | val_0_accuracy: 0.88044 |  0:00:33s
epoch 12 | loss: 0.10296 | val_0_accuracy: 0.96994 |  0:00:36s

Early stopping occurred at epoch 12 with best_epoch = 2 and best_val_0_accuracy = 0.97056
Successfully saved model at ../models/tabnet_raw.zip
TabNet precision score: 0.970419110083432
```

Neural networks: scores comparable to the simpler models

Model validation

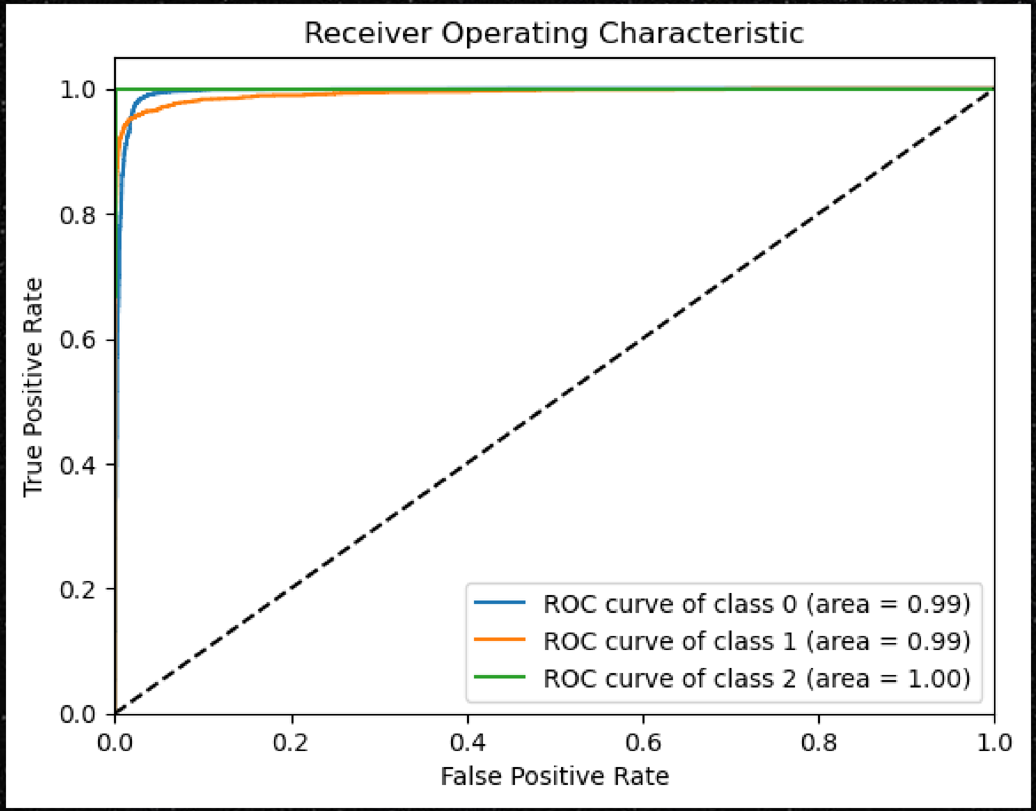
Final model choice:
Random forest with hyperparameters:
n_estimators: 259, max_depth: 13, criterion: entropy, max_features: log2, class_weight: None

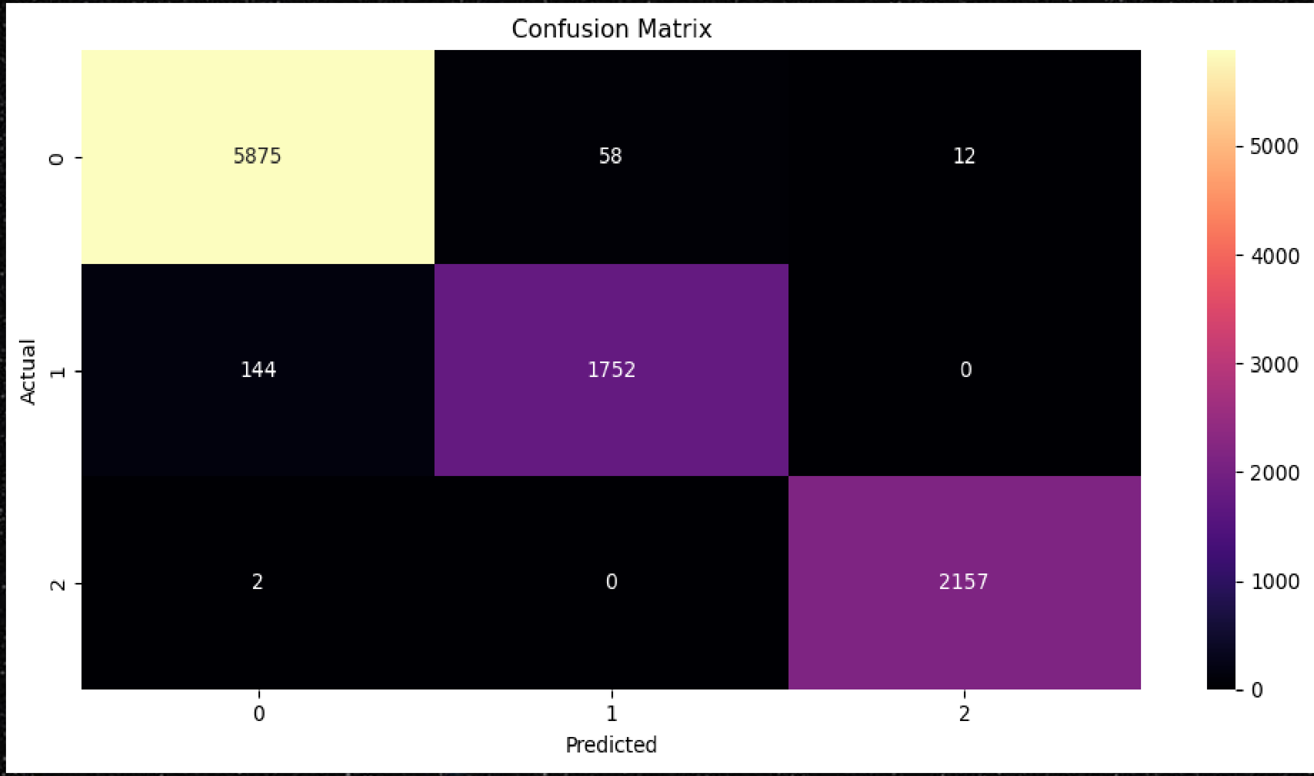
The behavior and results of the selected model have been verified by an independent validation team.

Weighted precision score: 0.9783142407171592

	precision	recall	f1-score	support
0	0.98	0.99	0.98	5945
1	0.97	0.92	0.95	1896
2	0.99	1.00	1.00	2159
accuracy			0.98	10000
macro avg	0.98	0.97	0.97	10000
weighted avg	0.98	0.98	0.98	10000

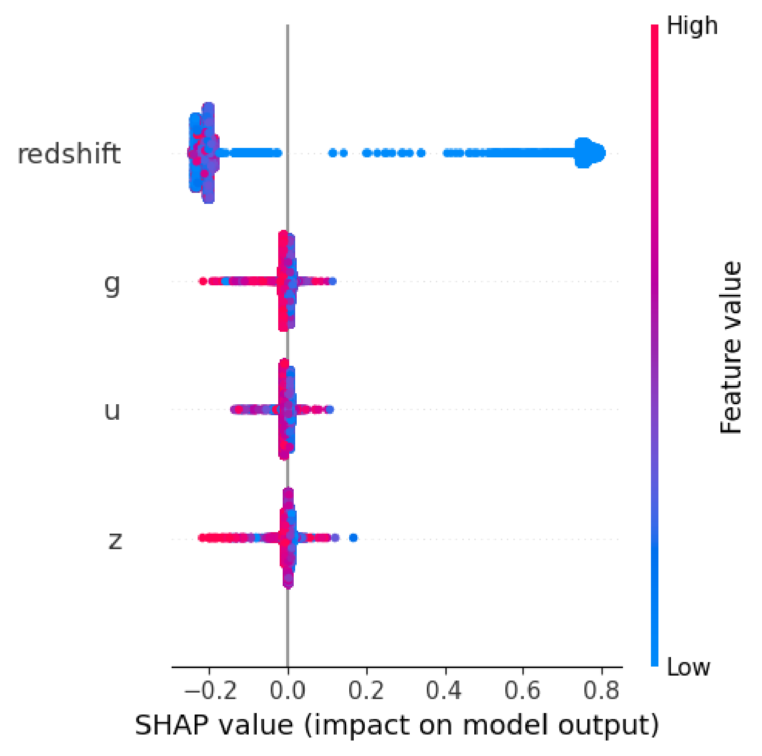
Model validation



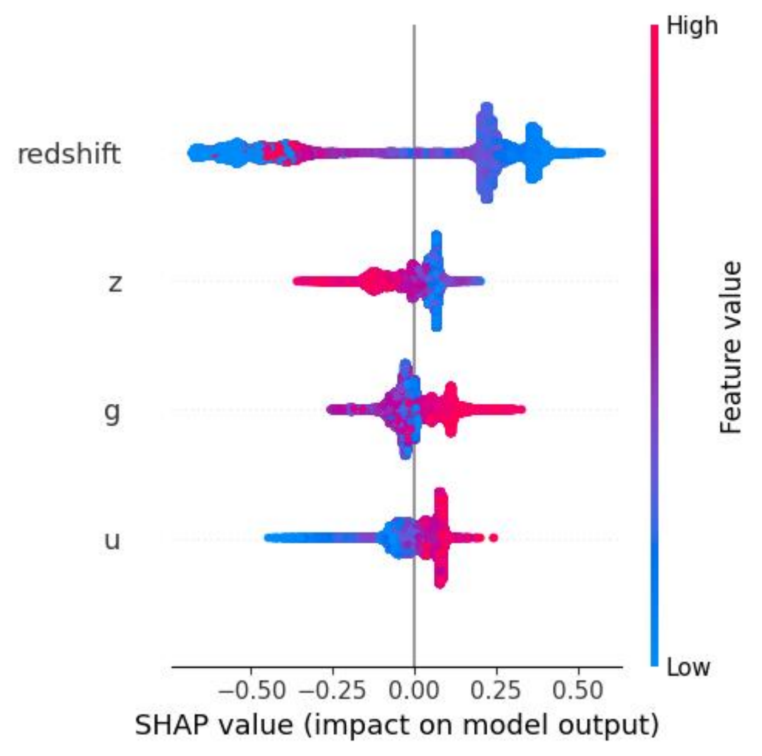


0 – galaxy, 1 – quasar, 2 - star

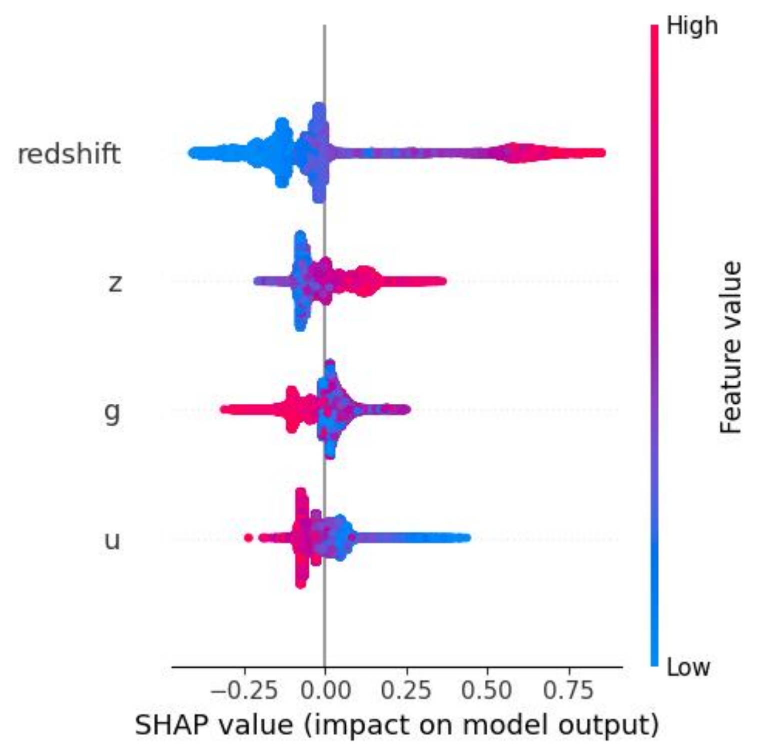
Model explainability



Star

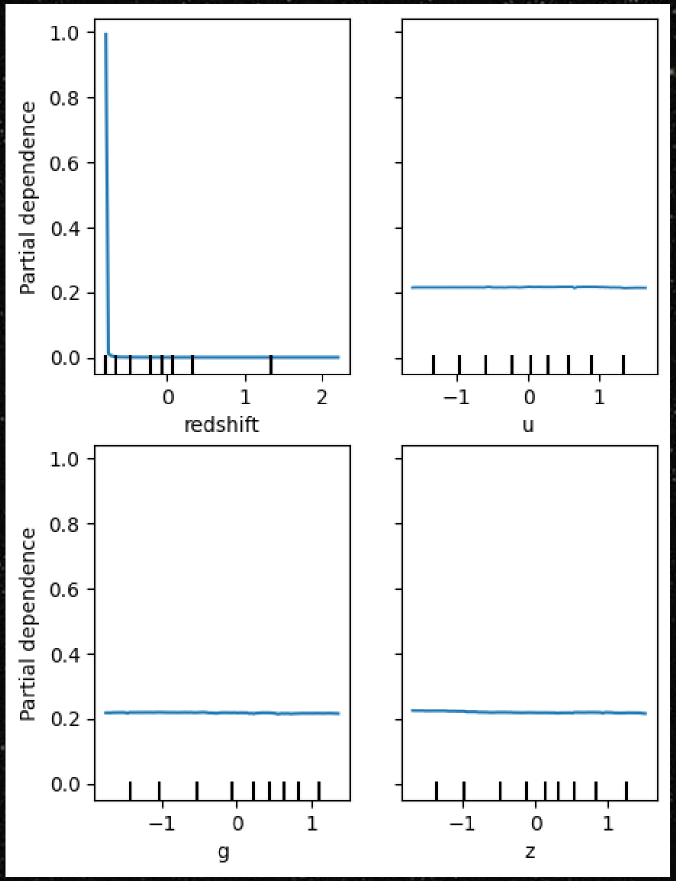


Galaxy

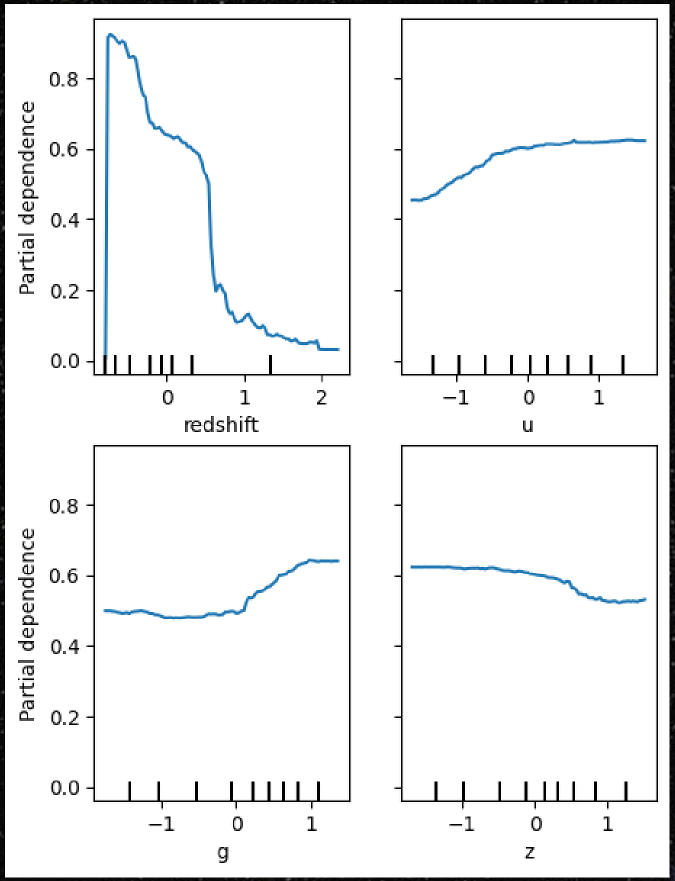


Quasar

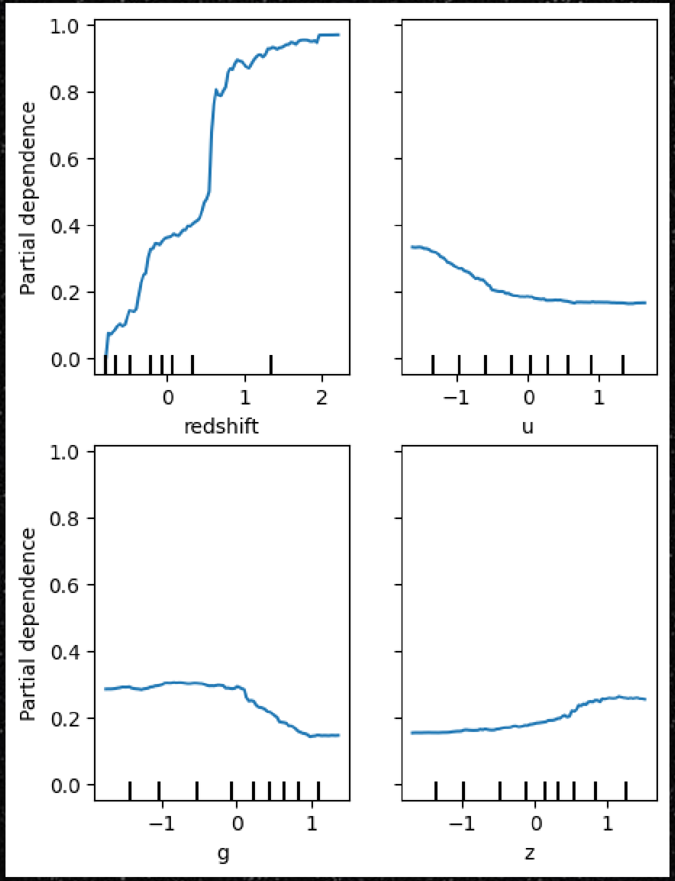
Model explainability



Star

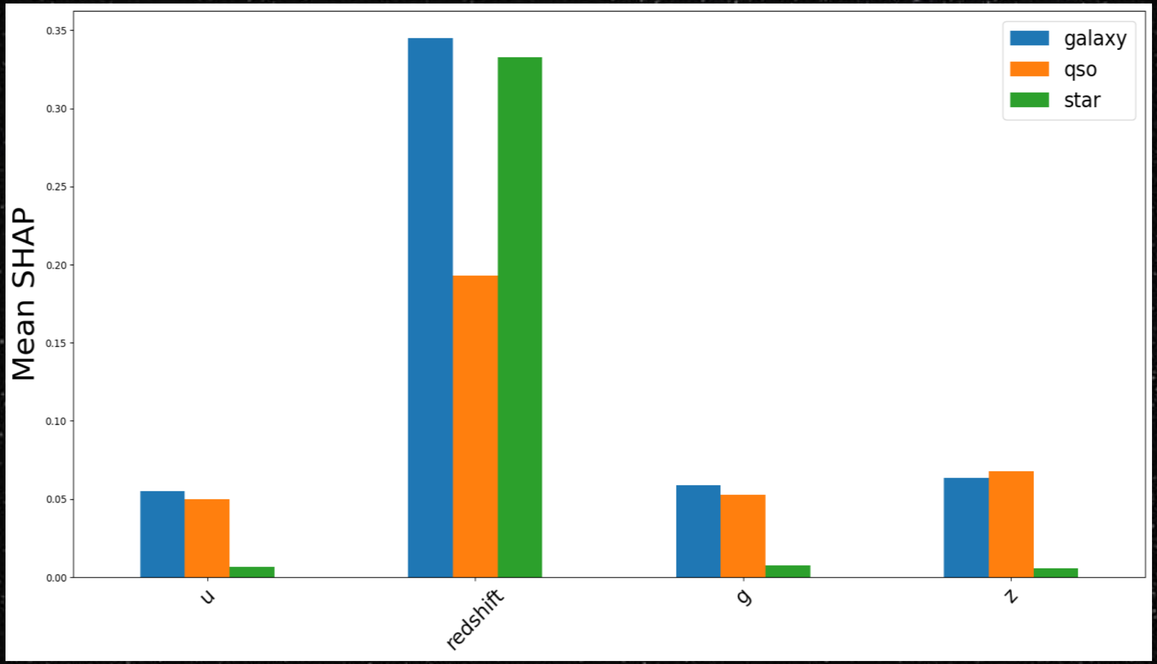


Galaxy

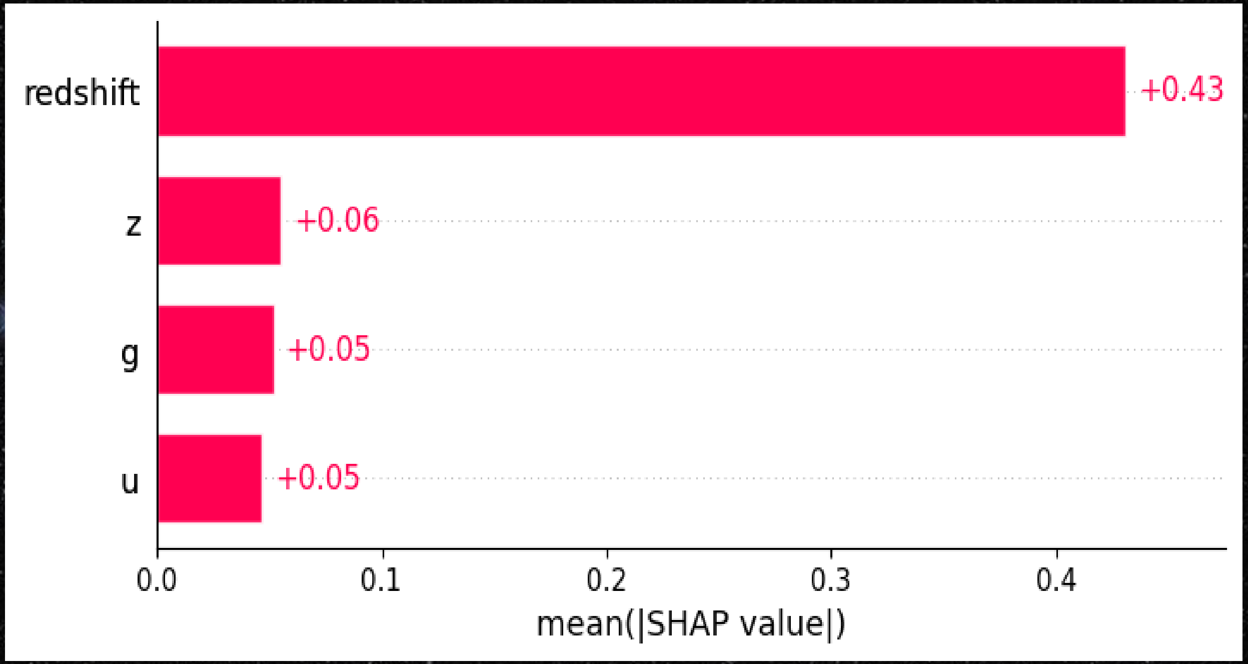


Quasar

Model explainability



Mean abs shap



Mean abs shap for the predicted class

Final result

We managed to create a model that is:

- simple
- explainable
- highly precise

This model can efficiently and reliably classify celestial bodies, thereby supporting the work of astrophysicists.