Astronomical classification on

DESI

Dataset

Yifei Wang

Background

What is Dark Energy?

DESI (Dark Energy Spectroscopic Instrument)

Why is DESI Spectral Data Ideal for Classification?

Target Variables

SPECTYPE=("STAR","GALAXY","QSO")

GALAXY: 79%

STAR: 16%

QSO: 5%

IMBALANCED!

	subtype
	21696490
K	2323062
G	1572042
HIZ	1114057
LOZ	748526
M	533414
F	354030
WD	52585
A	27586
B	3785
CV	386

Objectives

1.Build a classification model for SPECTYPE

....on a random sample

2.Build a classification model for subtype on the subset of stars

Dataset

https://data.desi.lbl.gov/public/

28,425,963 Samples

– Random Sample(size = 10,000)

136 Features

– Feature Engineering

Independent Variables

Z,ZERR,CHI2

FLUX

MORPHTYPE

Technical Parameters: How to detect?

COEFFs: dim=4,5,10

Data Cleaning

NAs

Useless Columns: Indexes

Data Leaks

Scaling

```
for col in df.columns:
    print(col+":",df[col].dtype)
```

```
z: float64
zerr: float64
chi2: float64
spectype: category
morphtype: category
ebv: float64
flux_g: float64
flux_r: float64
flux_z: float64
flux_w1: float64
flux_w2: float64
flux_ivar_g: float64
flux_ivar_r: float64
flux_ivar_z: float64
flux_ivar_w1: float64
flux_ivar_w2: float64
fiberflux_g: float64
fiberflux_r: float64
fiberflux_z: float64
fibertotflux_g: float64
```

Dealing with Outliers

```
df['zwarn'].value_counts()
```

zwarn	count
0	7373
5	662
2053	540
4	468
518	202
2564	176
1	115
516	106
2	91
2049	75
646	49
2560	43
519	24
512	22
7	16
2048	10
6	10
2055	8
2566	5
3	2
2562	2
2052	1

Dealing with Outliers

```
df['deltachi2'].describe()
```

	deltachi2
count	1.000000e+04
mean	1.046260e+04
std	6.986589e+04
min	0.000000e+00
25%	1.424442e+01
50%	2.280390e+02
75%	1.193293e+03
max	2.046469e+06

dtype: float64

```
df[df['zwarn']==0]['deltachi2'].describe()
```

	deltachi2
count	7.373000e+03
mean	1.401333e+04
std	8.078843e+04
min	9.005392e+00
25%	1.578616e+02
50%	4.837253e+02
75%	2.279004e+03
max	2.046469e+06

dtype: float64

```
len(df)
```

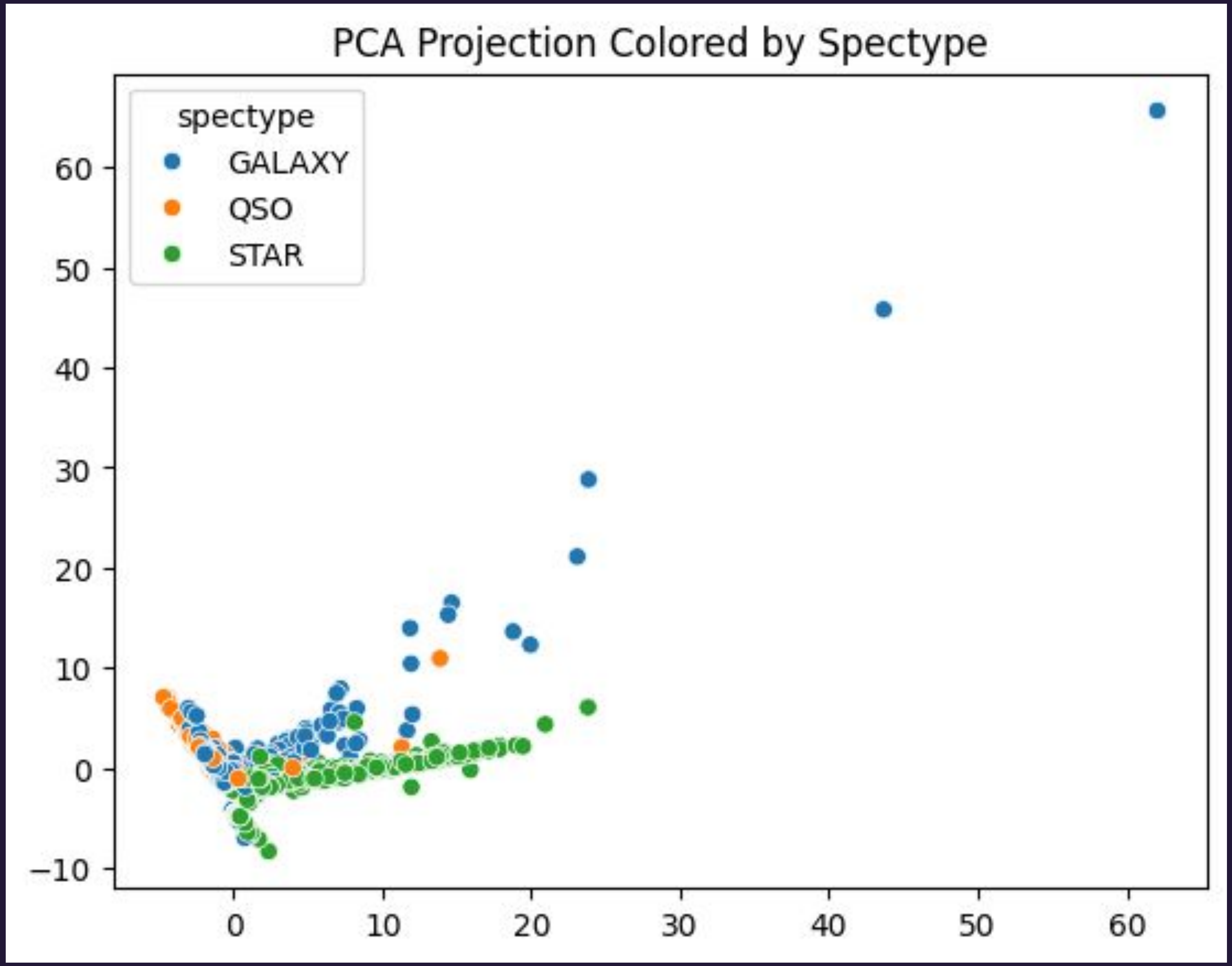
7373

Feature Engineering

```
z: float64
zerr: float64
chi2: float64
spectype: category
morphtype: category
ebv: float64
flux_g: float64
flux_r: float64
flux_z: float64
flux_w1: float64
flux_w2: float64
flux_ivar_g: float64
flux_ivar_r: float64
```

```
flux_ivar_z: float64
flux_ivar_w1: float64
flux_ivar_w2: float64
fiberflux_g: float64
fiberflux_r: float64
fiberflux_z: float64
fibertotflux_g: float64
fibertotflux_r: float64
fibertotflux_z: float64
gaia_phot_g_mean_mag: float64
gaia_phot_bp_mean_mag: float64
gaia_phot_rp_mean_mag: float64
```

Visualization



PCA

```
             z      zerr      chi2       ebv    flux_g    flux_r    flux_z  \
PC1  -0.165966 -0.082550  0.071949 -0.000090  0.224187  0.239503  0.240204
PC2   0.231722  0.141737 -0.121760 -0.047044  0.323007  0.301345  0.283124

       flux_w1   flux_w2  flux_ivar_g  ...  flux_ivar_w2  fiberflux_g  \
PC1   0.195565  0.176801    -0.146488  ...     -0.028543     0.292089
PC2   0.256805  0.237756     0.267001  ...      0.173839     0.015682

     fiberflux_r  fiberflux_z  fibertotflux_g  fibertotflux_r  fibertotflux_z  \
PC1     0.313657     0.302737        0.287254        0.311077        0.300790
PC2     0.020597     0.024465        0.014057        0.018982        0.022724

     gaia_phot_g_mean_mag  gaia_phot_bp_mean_mag  gaia_phot_rp_mean_mag
PC1              0.204483               0.206049               0.205442
PC2             -0.248240              -0.256246              -0.254486

[2 rows x 23 columns]
Explained variance ratio: [0.3549595  0.17472376]
```

Model Overview

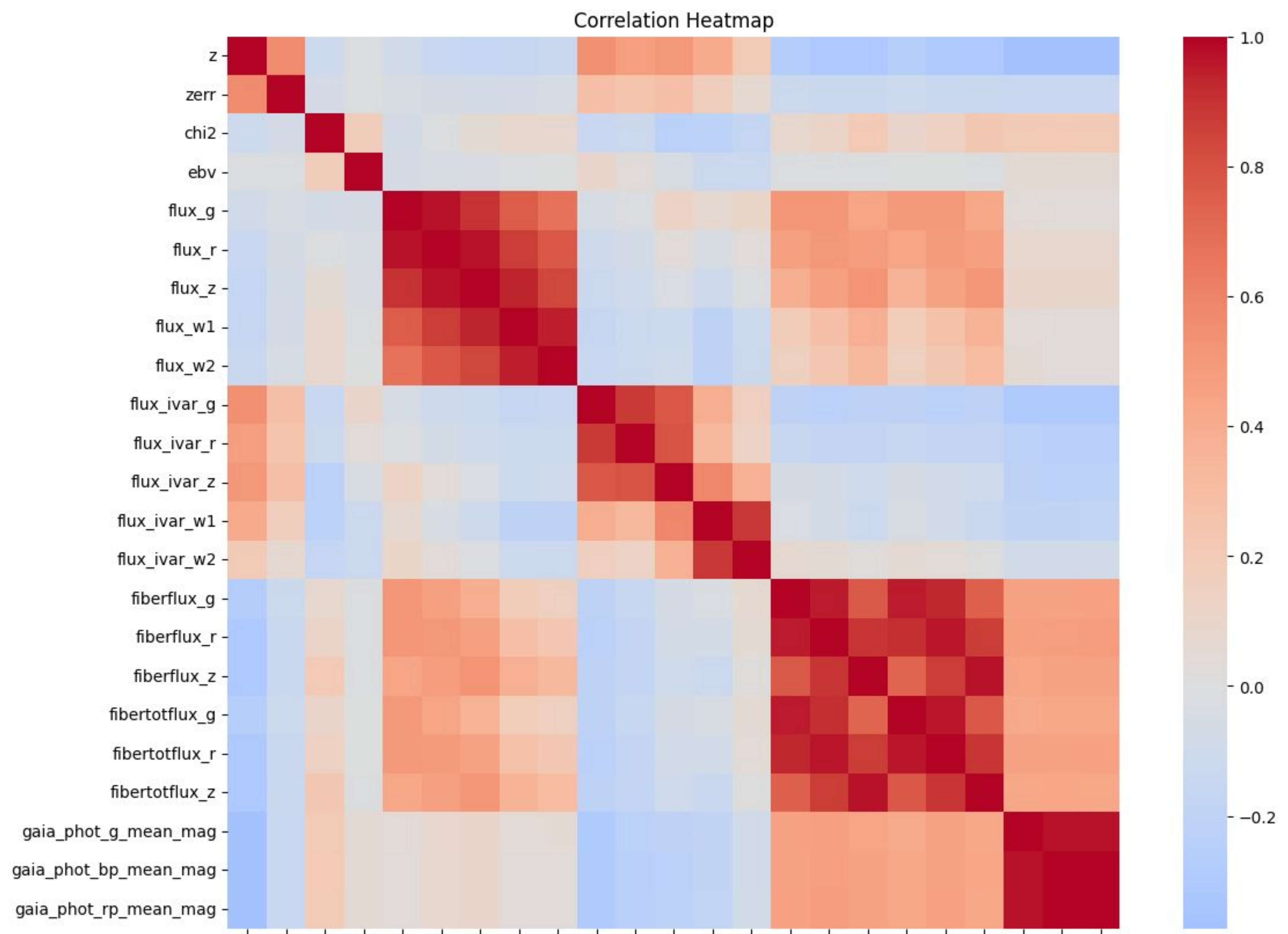
Continuous&Categorical IVs

Multi-class

Random Forest

CatBoost/ LightGBM?

~~Linear? SVM? Bayesian?~~



Cross-Validation

Automatic

Stratified

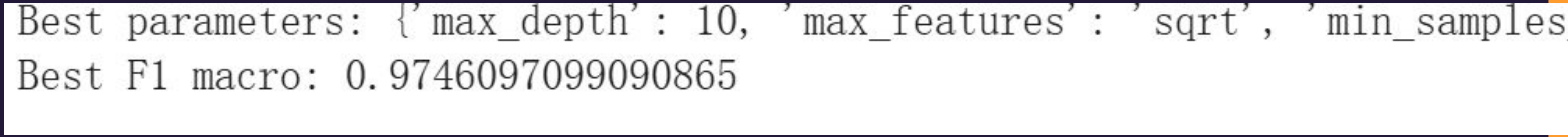
K-fold

```
cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
```

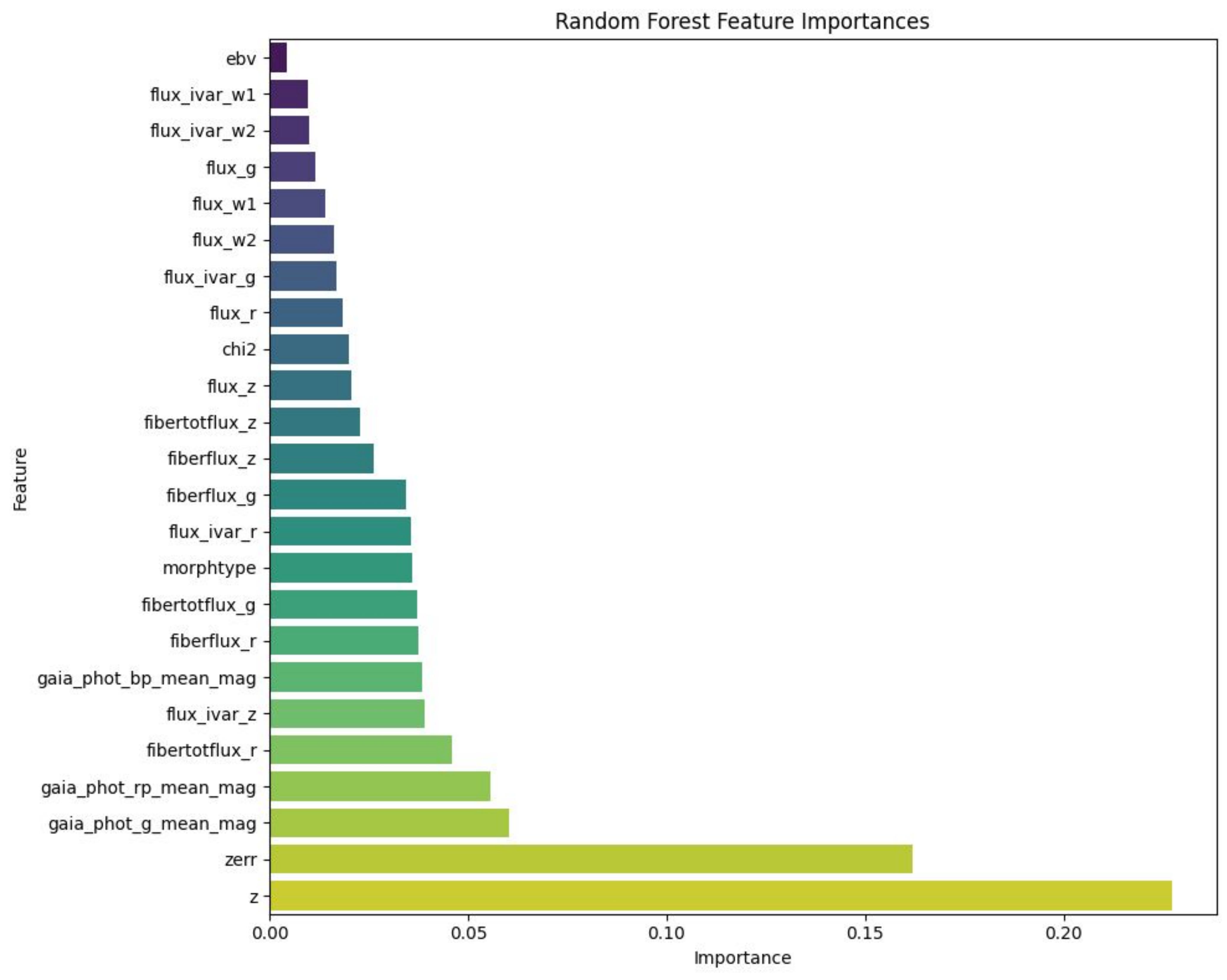
```
Cross-validation scores (F1 macro):
Fold 1: 0.9736
Fold 2: 0.9810
Fold 3: 0.9639
Fold 4: 0.9674
Fold 5: 0.9815
```

Hyperparameters Tuning

sklearn.model_selection.RandomizedSearchCV



Outcome



Random Forest Feature Importances
ebv
flux_ivar_w1
flux_ivar_w2
flux_g
flux_w1
flux_w2
flux_ivar_g
flux_r
chi2
flux_z
fibertotflux_z
fiberflux_z
fiberflux_g
flux_ivar_r
morphtype
fibertotflux_g
fiberflux_r
gaia_phot_bp_mean_mag
flux_ivar_z
fibertotflux_r
gaia_phot_rp_mean_mag
gaia_phot_g_mean_mag
zerr
z
Feature
0.00
0.05
0.10
0.15
0.20
Importance

Star Classification

Large Dataset

Multiple Imbalanced Class

LightGBM

```
print(df.shape)
df.head()
```

(4750477, 30)

	count
subtype	
K	2274717
G	1542632
M	508679
F	345419
WD	50880
A	25320
B	2468
CV	362

Star Baseline: RF

Subset n_samples=10000

RF is slow so no cross validation

Accuracy is high, Marco is low

Tend to recognize majority classes

Some subtypes are poorly recognized

	precision	recall	f1-score	support
A	0.92	0.74	0.82	115
B	0.00	0.00	0.00	9
CV	0.00	0.00	0.00	1
F	0.98	0.89	0.93	1453
G	0.97	0.99	0.98	6496
K	0.99	1.00	0.99	9566
M	1.00	1.00	1.00	2149
WD	0.96	0.97	0.96	211
accuracy			0.99	20000
macro avg	0.73	0.70	0.71	20000
weighted avg	0.98	0.99	0.98	20000

Star Model: LightGBM

Train on 80% of the data

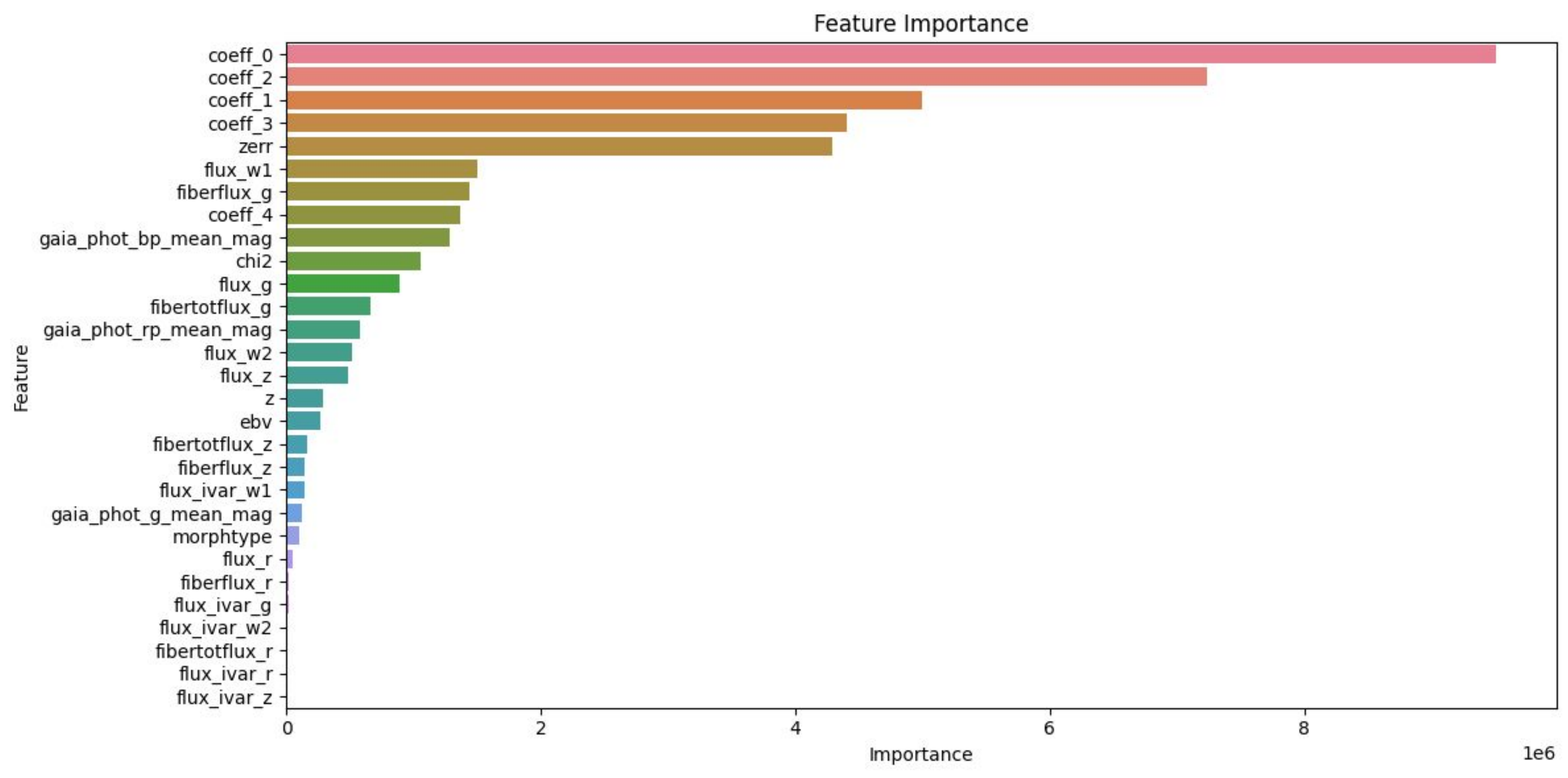
Still slow, No Cross-validation

Marco increased

Recongnized every class

	precision	recall	f1-score	support
A	0.97	0.96	0.97	5064
B	0.95	0.84	0.89	494
CV	1.00	1.00	1.00	72
F	0.99	0.99	0.99	69084
G	1.00	1.00	1.00	308526
K	1.00	1.00	1.00	454944
M	1.00	1.00	1.00	101736
WD	1.00	1.00	1.00	10176
accuracy			1.00	950096
macro avg	0.99	0.97	0.98	950096
weighted avg	1.00	1.00	1.00	950096

Feature Importance



Conclusion

RandomForest is good for classifying the Type

Most important feature: z

LightGBM is good for classifying the Subtype

Most important feature: coefficients

Future work

Train a Model that better recognizes B

Model Compression

Reference

DESI Collaboration et al. (2016). The DESI Experiment Part I: Science,Targeting, and Survey Design. arXiv:1611.00036. https://arxiv.org/abs/1611.00036

Thank you