Comparing Performance of Machine Learning Algorithms for Galaxy Classification

F. Korhan YELKENCİ[1,3], E. Kaan ÜLGEN[1], Sinan ALİŞ[1,2], Süleyman FİŞEK[1,2]

[1]İstanbul University, Dept. of Astronomy and Space Sciences - Turkey
[2]İstanbul University Observatory Application and Research Center - Turkey
[3]İstanbul University, Dept. of Informatics - Turkey

E-mail: yelkenci@İstanbul.edu.tr
Web: cosmology.İstanbul.edu.tr

AIA2019, 22-26 July 2019, ESO HQ, Garching, Germany

Motivation

- Galaxy Data: Deep, wide, big and good quality.
- We don’t have enough number of eye to classify all galaxies.
  - Visual classification may have some biases or misclassifications.
  - At high redshift, human eye is ineffective to classify galaxies.
  - We can use parametric features of galaxies to classify (colurs, structural parameters, sersic indices, etc.).
  - Try to understand galaxy morphologies.
- Machine learning algorithms are now more efficient and reliable.
  - Comparing performances of ML algorithms on galaxy classifications
  - How accuracy changes by using different parametric features.
  - To see whether accuracy of ML algorithms change with redshift.
- Astronomical data will burst in next years.

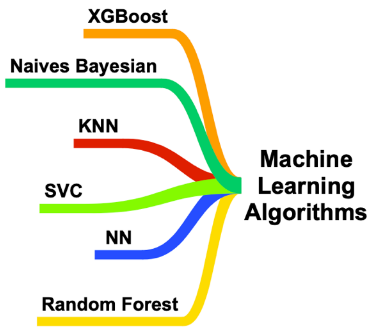
Background

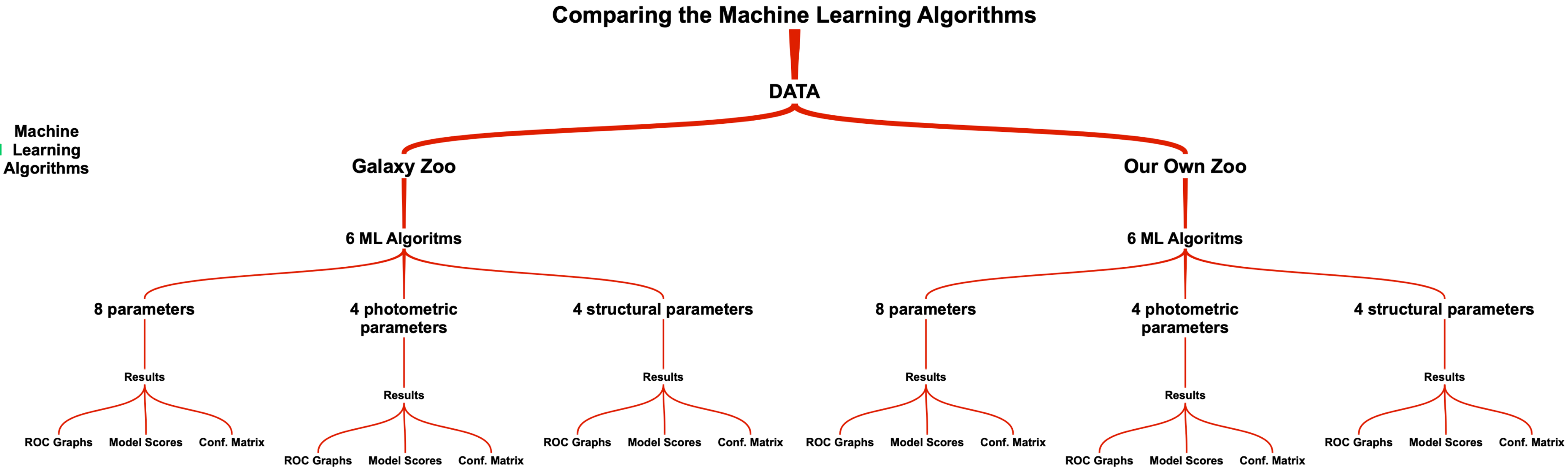
Statistical learning method	Total sample	Training set	Test set	Number of classes	Dimensions	Accuracy	Reference
SVM	7528	6022 (80 per cent)	1506 (20 per cent)	5	10	75.8 per cent	Results from our work
NN						76.0 per cent	
CT						69.0 per cent	
CTRF						76.2 per cent	
SVM	$\sim$1500	500 (33 per cent)	1000 (67 per cent)	2 (early-type, late-type)	12	80 per cent	Huertas-Company et al. (2007)
NN	$\sim$1 000 000	$\sim$75 000 (7.5 per cent)	$\sim$925 000 (92.5 per cent)	3 (early-type, spirals, point sources/artefacts)	12	90 per cent	Banerji et al. (2010)
Oblique CT	5217	$\sim$4174 (80 per cent)	$\sim$1043 (20 per cent)	5 (E, S0, Sa+Sb, Sc+Sd, Irr)	13	63 per cent	Owens et al. (1996)
Three CT algorithms including CTRF	75 000	67 500 (90 per cent)	7500 (10 per cent)	3 (ellipticals, spirals, unknown)	13	96.2 per cent	Gauci et al. (2010)
ConvNet	58 000	47 700 ($\sim$82 per cent)	5000 ($\sim$9 per cent) 5300 ($\sim$9 per cent) used for real-time evaluation during training	5 (probablities[a])	Run on images	$\sim$99 per cent	Huertas-Company et al. (2015) Dieleman et al. (2015)

Note. [a]Probabilities for each galaxy having a disc or a spheroid, being a point source, having an irregularity or being unclassifiable are the outputs.

Sreejith et al., MNRAS **474,** 5232–5258 (2018)

Schema of Application





8 Parametric features: ui, gr, gi, deVAB_r(b/a), deVRad_r(eff radius), CI (petroR50_r/petroR90_r), absMagR , sersic_n)
4 Structural features: deVAB_r(b/a), deVRad_r(eff radius), CI (petroR50_r/petroR90_r), sersic_n)
4 Photometric features: ui, gr, gi, absMagR)

Machine Learning Algorithms

	Unsupervised	Supervised
Continuous	• Clustering & Dimensionality Reduction ○ SVD ○ PCA ○ K-means	• Regression • Decision Trees • Neural Networks • Ensemble Methots ○ XGBoost ○ Random Forest
Categorical	• Association Analysis ○ Apriori ○ FP-Growth • Hidden Markov Model	• Classification ○ KNN ○ Trees ○ Logistic Regression ○ Naive-Bayes ○ SVM ○ SVC

Supervised Learning

- KNN (K-Nearest Neighbour)
- SVM (Support Vector Machine)
  - C-SVC
- Ensemble Methods
  - XGBoost
  - Random Forest
- Gaussian Naive Bayesian
- Neural Network

DATA

- SDSS - Galaxy Zoo:

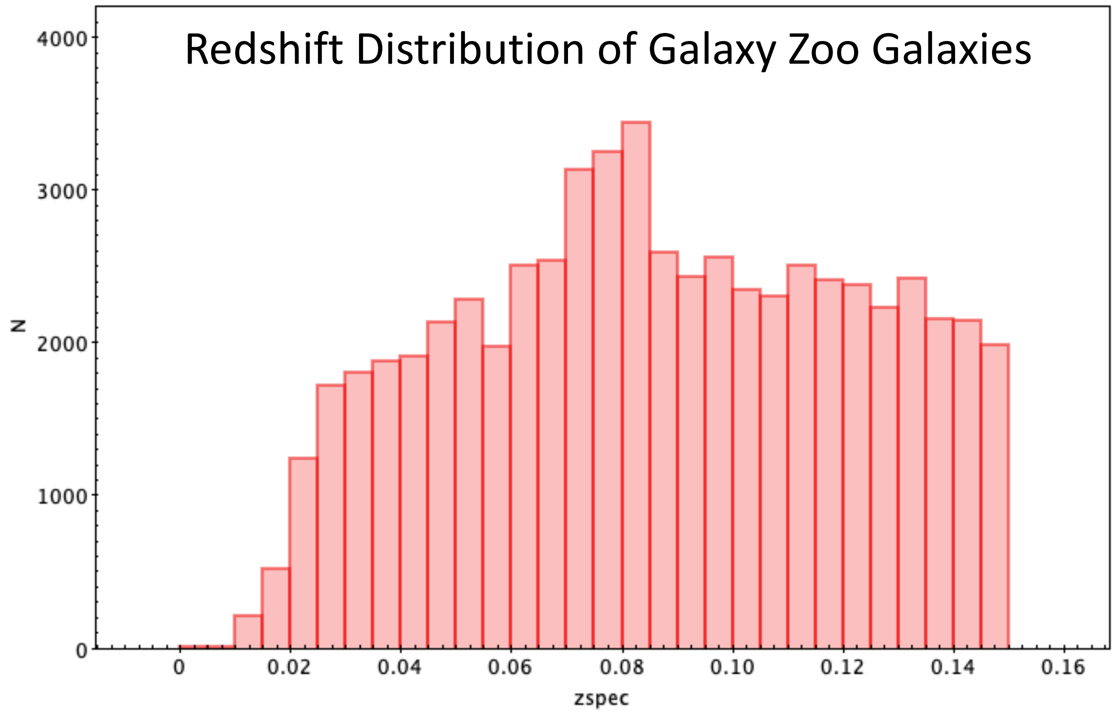
https://data.galaxyzoo.org
https://www.sdss.org
Lintott et al. 2008, MNRAS, 389, 1179
Lintott et al. 2011, 410, 166

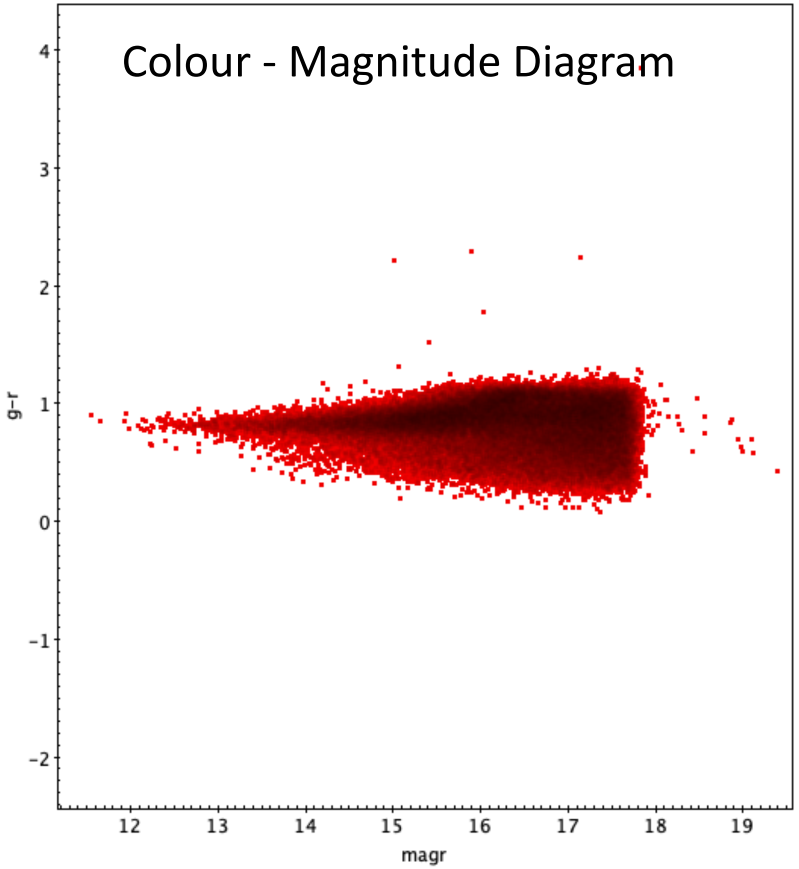
- CFHT-LS

https://www.cfht.hawaii.edu/Science/CFHTLS/

Galaxy Zoo DATA

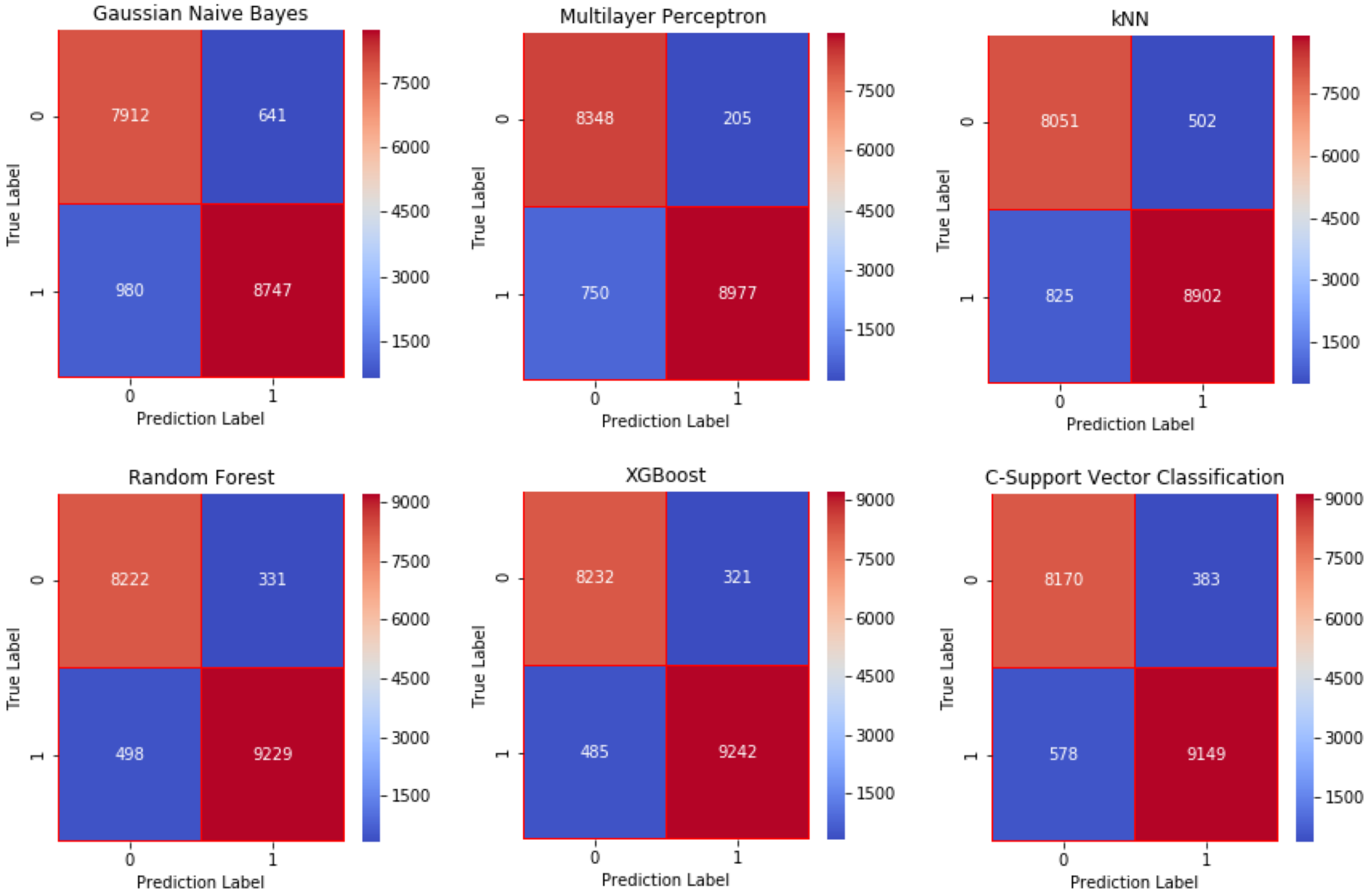
- **60932** galaxies → Elliptical: 28591, Spiral: 32341 visually classified in Galaxy Zoo.
- Redshift range: 0 < z < 0.15
- with Sersic Index
- Train and Test sets (%70/30 : 42652 / 18280)





Galaxy Zoo DATA: Results for all 8 parameters

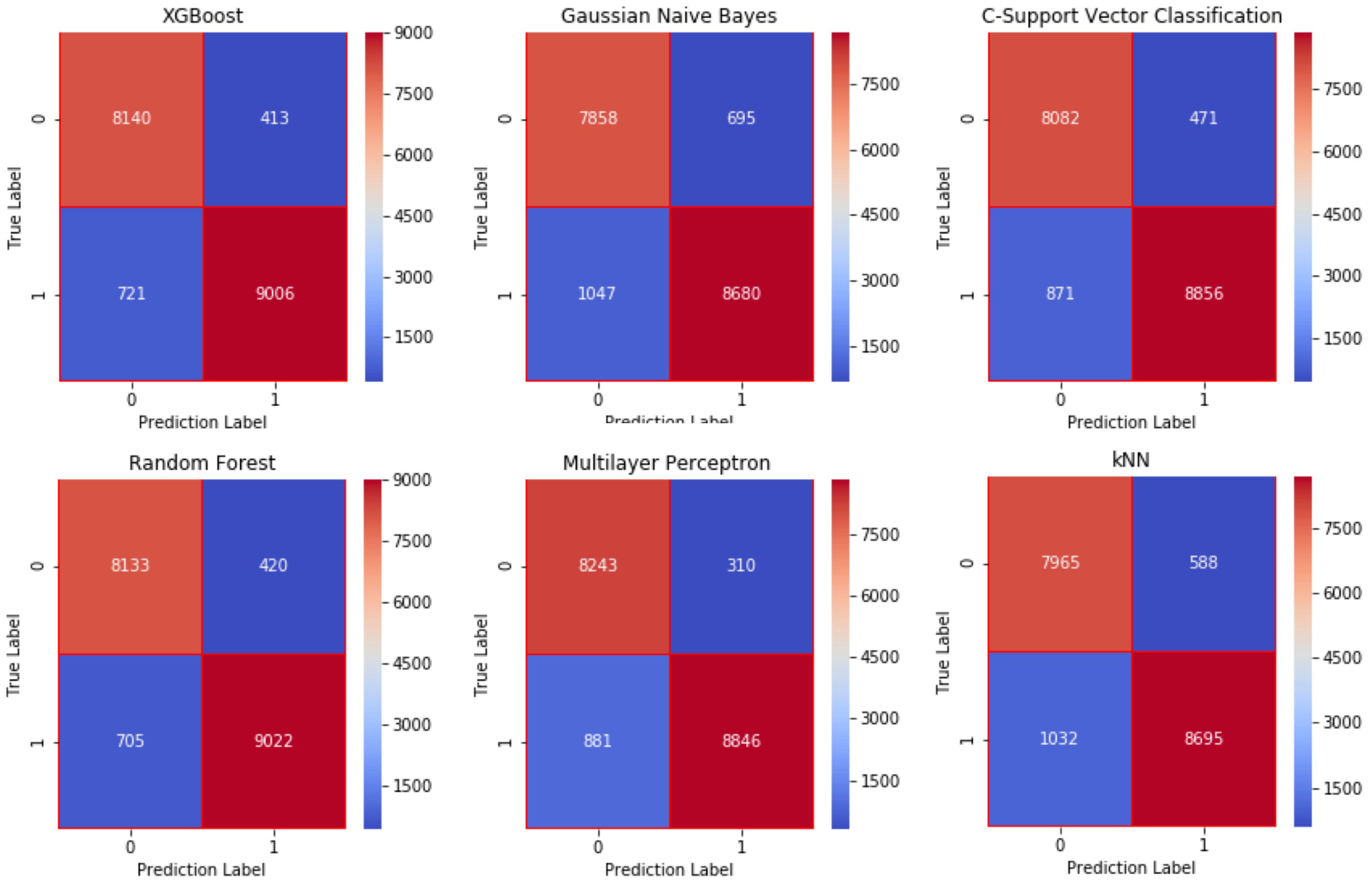
(ui, gr, gi, deVAB_r(b/a), deVRad_r(eff radius), CI (petroR50_r/petroR90_r), absMagR , sersic_n)



Model	Score
XGBoost	0.955383
Random Forest	0.954773
NN	0.949405
SVC	0.947247
KNN	0.928444
Naive Bayes	0.912900

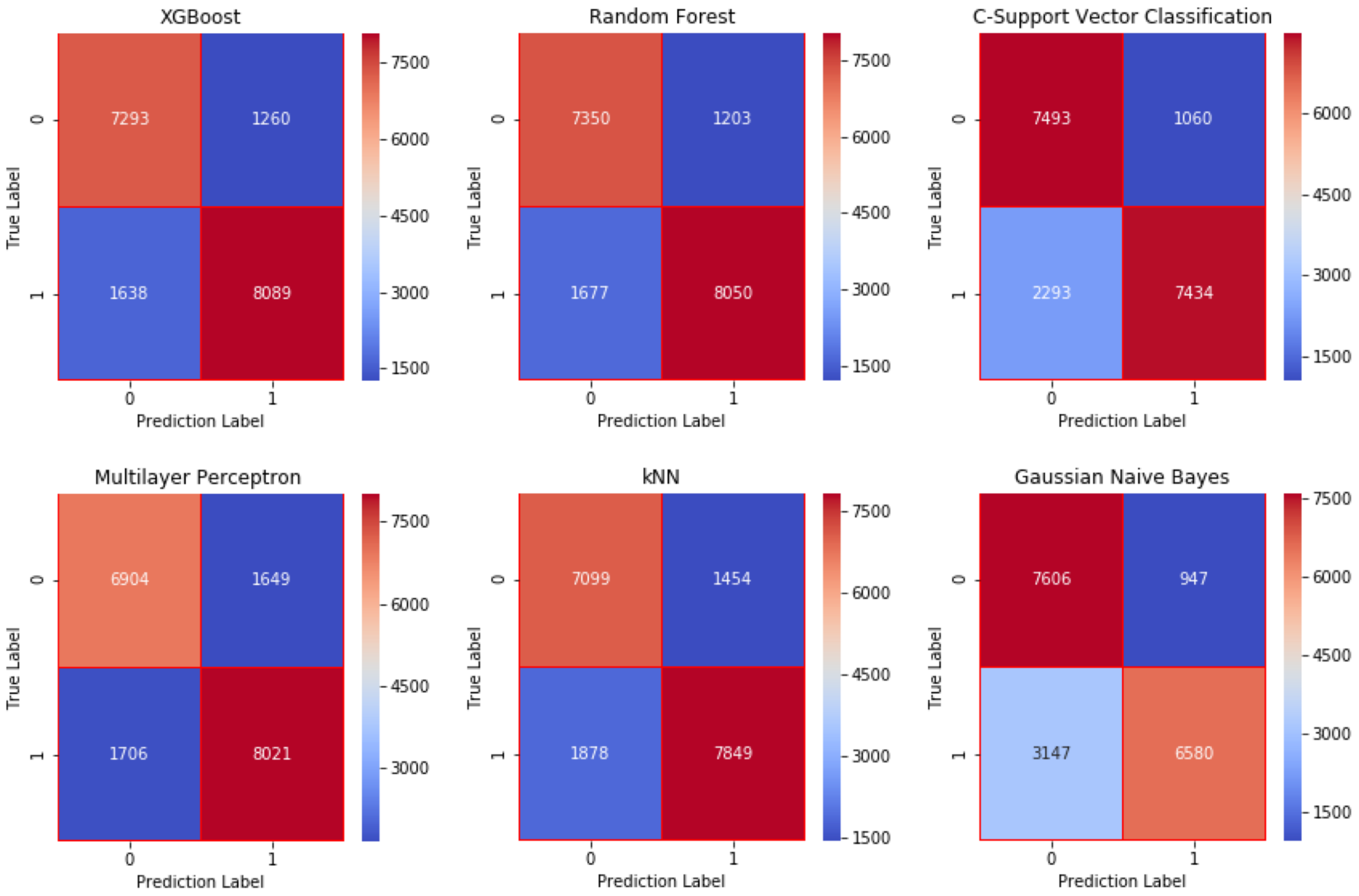
Galaxy Zoo DATA: Results for 4 structural parameters

(deVAB_r(b/a), deVRad_r(eff radius), CI (petroR50_r/petroR90_r), sersic_n)



Model	Score
Random Forest	0.938151
XGBoost	0.937682
NN	0.936697
SVC	0.925068
KNN	0.909031
Naive Bayes	0.903521

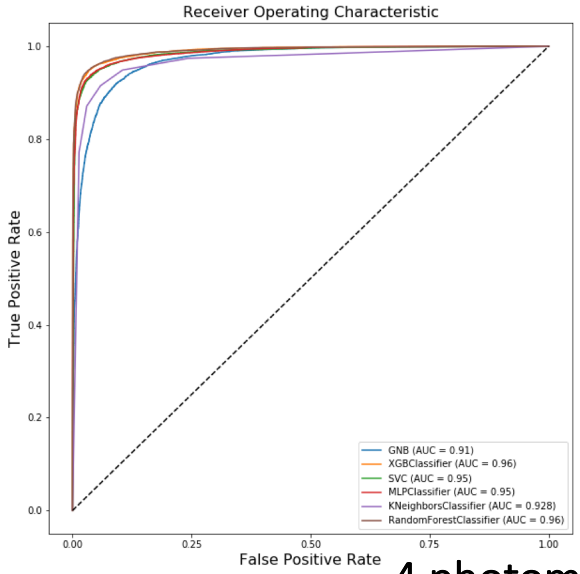
Galaxy Zoo DATA: Results for 4 photometric parameters
(ui, gr, gi, absMagR)



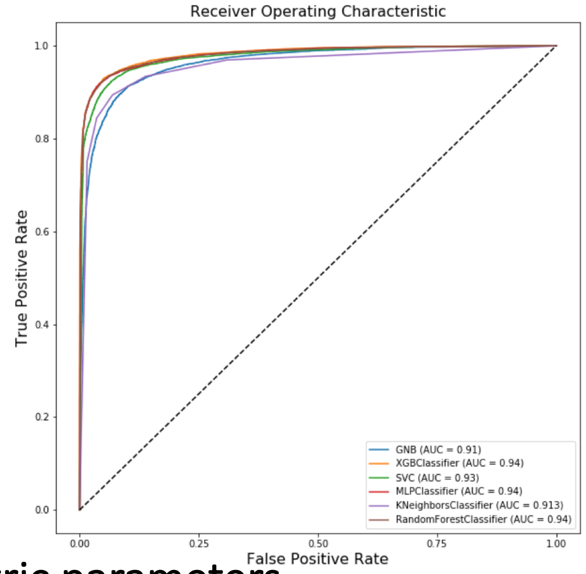
Model	**Score**
Random Forest	0.840242
XGBoost	0.838484
KNN	0.823057
SVC	0.816000
NN	0.812506
Naive Bayes	0.775908

Galaxy Zoo DATA: Comparison of ROC graphs of 6 Machine Learning Algorithms

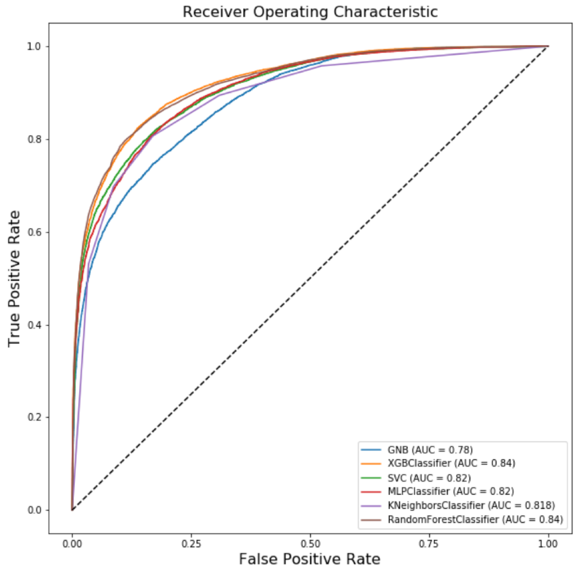
All 8 parameters

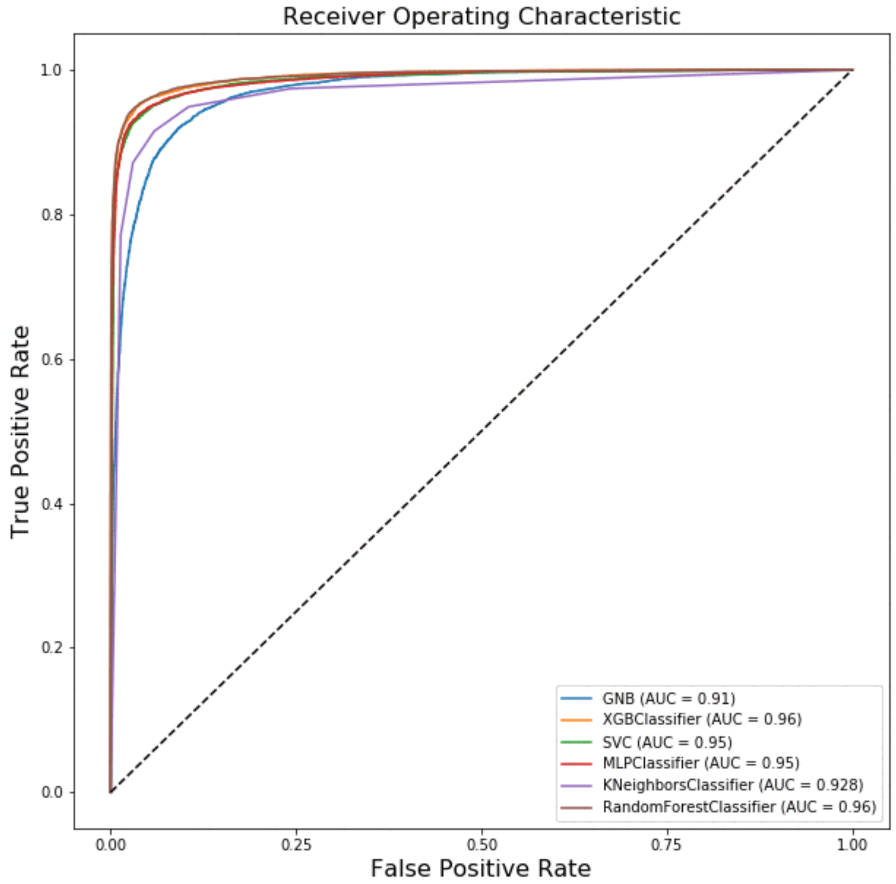


4 structural parameters



4 photometric parameters





Impact of 8 Parameters in Random Forest and XGBoost

Random Forest – Parameters

Feature	Importance
deVAB_r	0.268648
CI	0.241121
sersic_n	0.188087
gi	0.086840
ui	0.083662
gr	0.051916
absMagR	0.039950
deVRad_r	0.039776

XGBoost - Parameters

Feature	Importance
CI	0.485405
deVAB_r	0.157679
sersic_n	0.106899
ui	0.083637
gi	0.074488
gr	0.038916
deVRad_r	0.034296
absMagR	0.018678

4 structural

XGBoost - Parameters

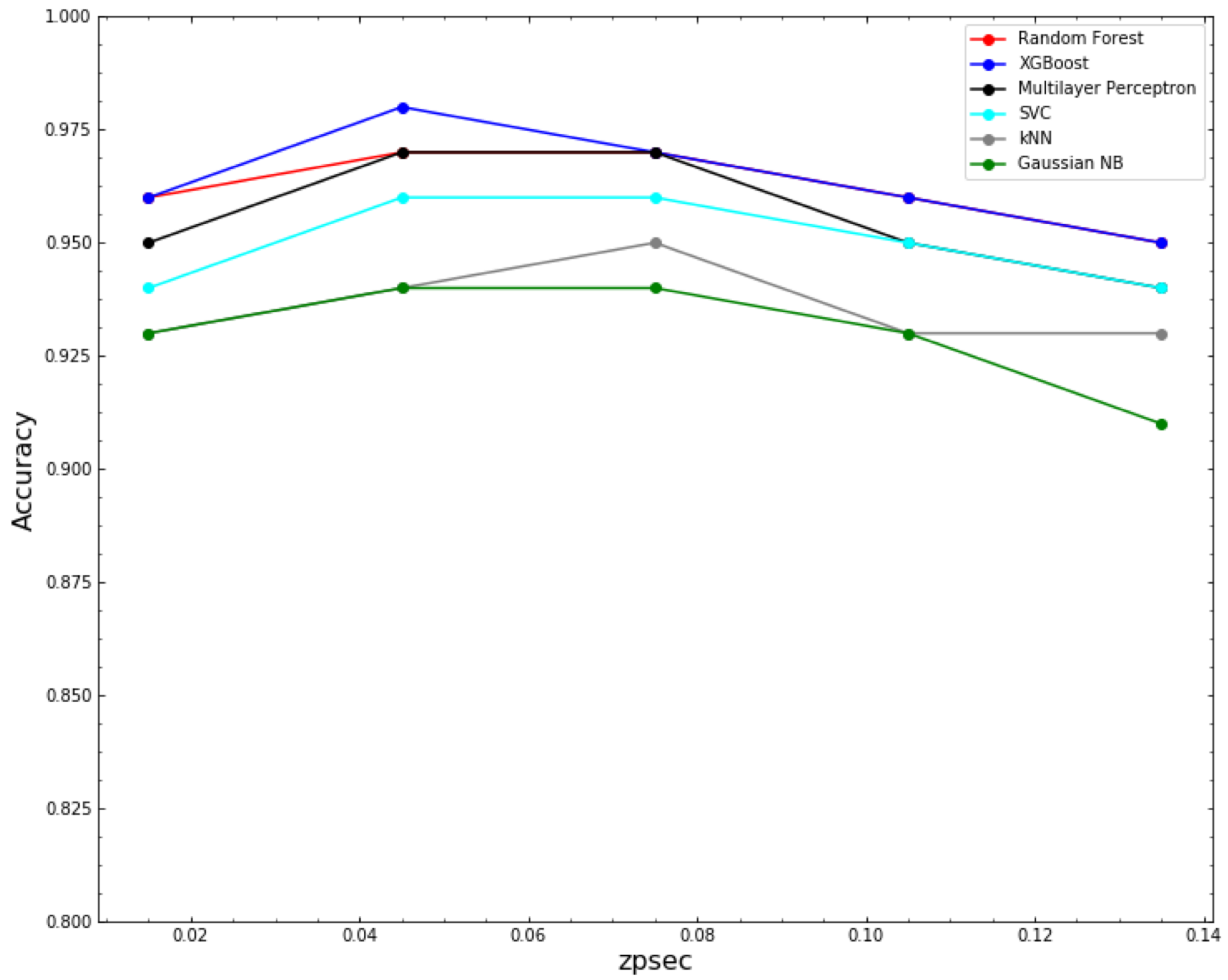
Feature	Importance
CI	0.551122
deVAB_r	0.219104
sersic_n	0.171079
deVRad_r	0.058695

4 photometric

XGBoost - Parameters

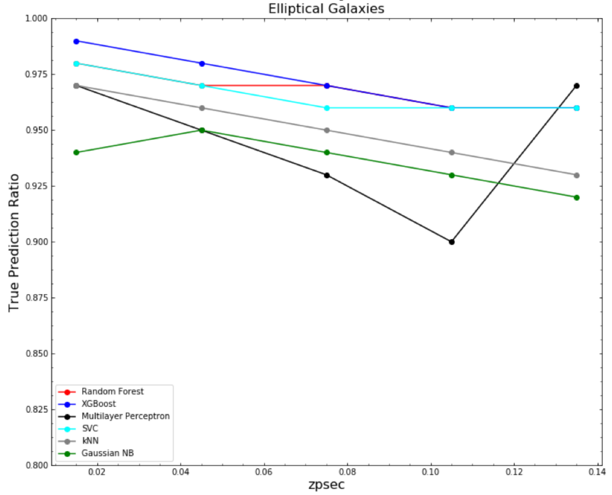
Feature	Importance
ui	0.584734
gr	0.182217
absMagR	0.171967
gi	0.061082

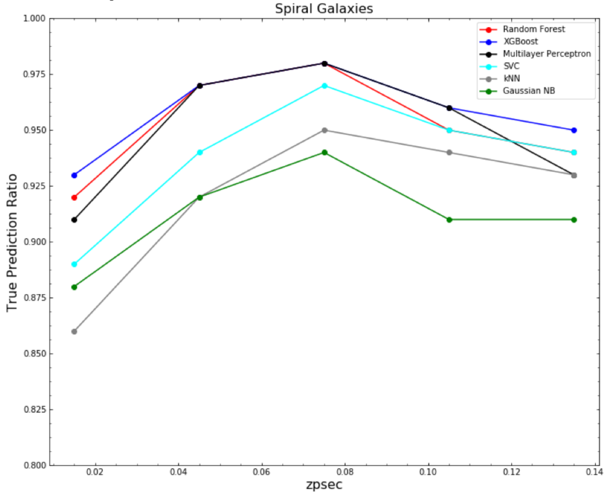
Galaxy Zoo DATA: Accuracy distribution as a function of redshift



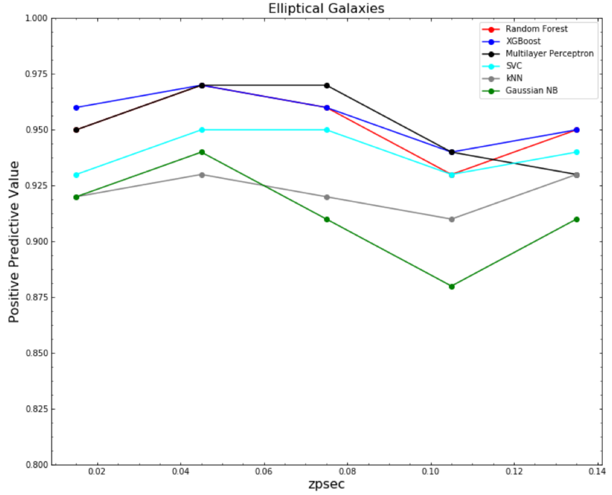
Galaxy Zoo Data: Recalls and Precisions as a function of redshift

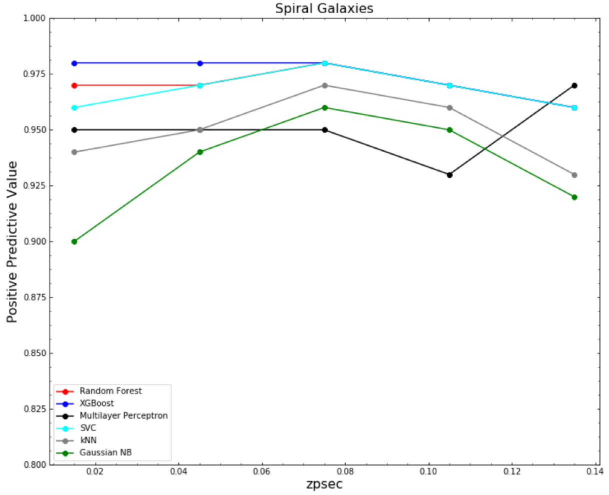
Galaxy Zoo Data: **Redshift Distribution of Recall** for 8 parameters





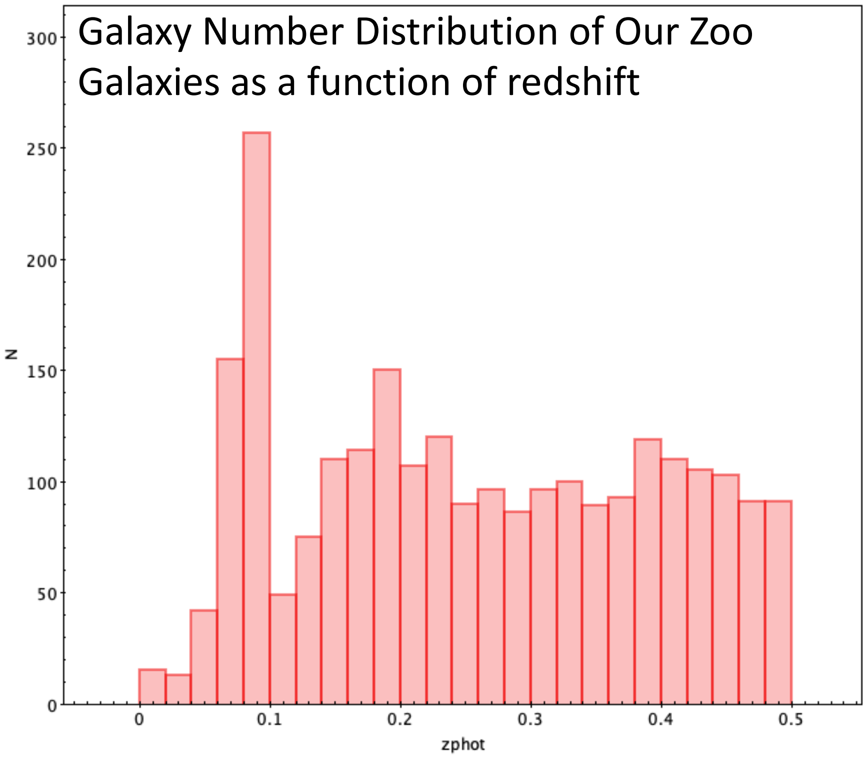
Galaxy Zoo Data: **Redshift Distribution of Precision** for 8 parameters

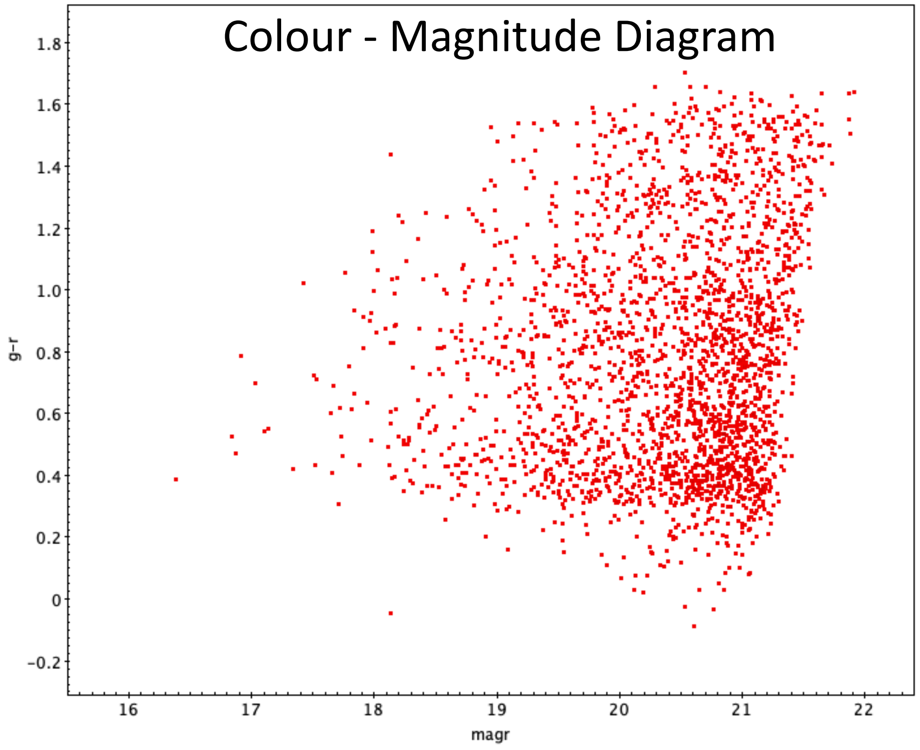




Our Own Zoo: CFHTLS - W1 catalogue

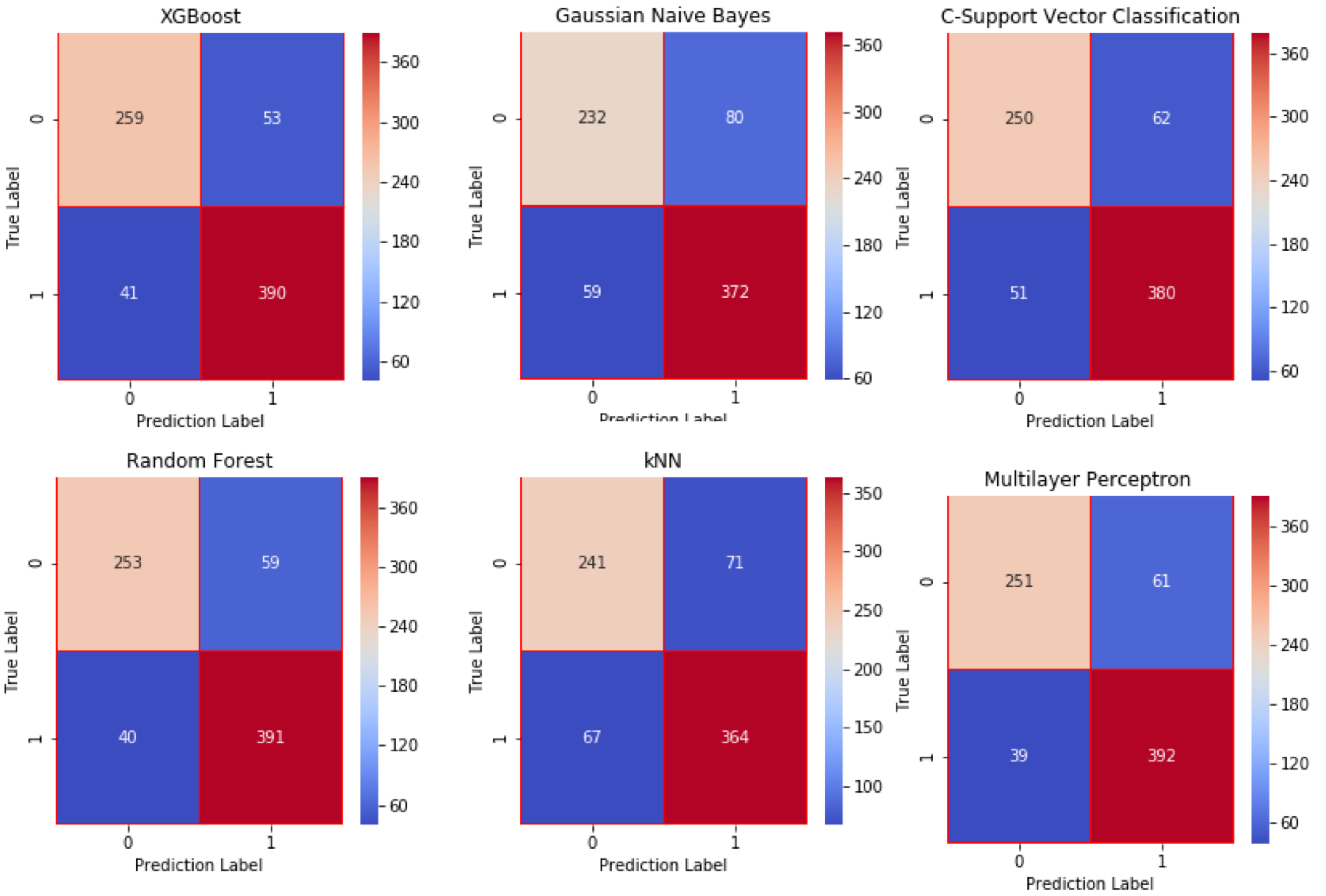
- **2500** / 180000 galaxies visually classified
- Ellipticals: 1053, Spirals: 1423 -> 2476
- Redshift range: 0 < z < 0.5
- Train and Test sets (%70/30 : 1733 / 743)





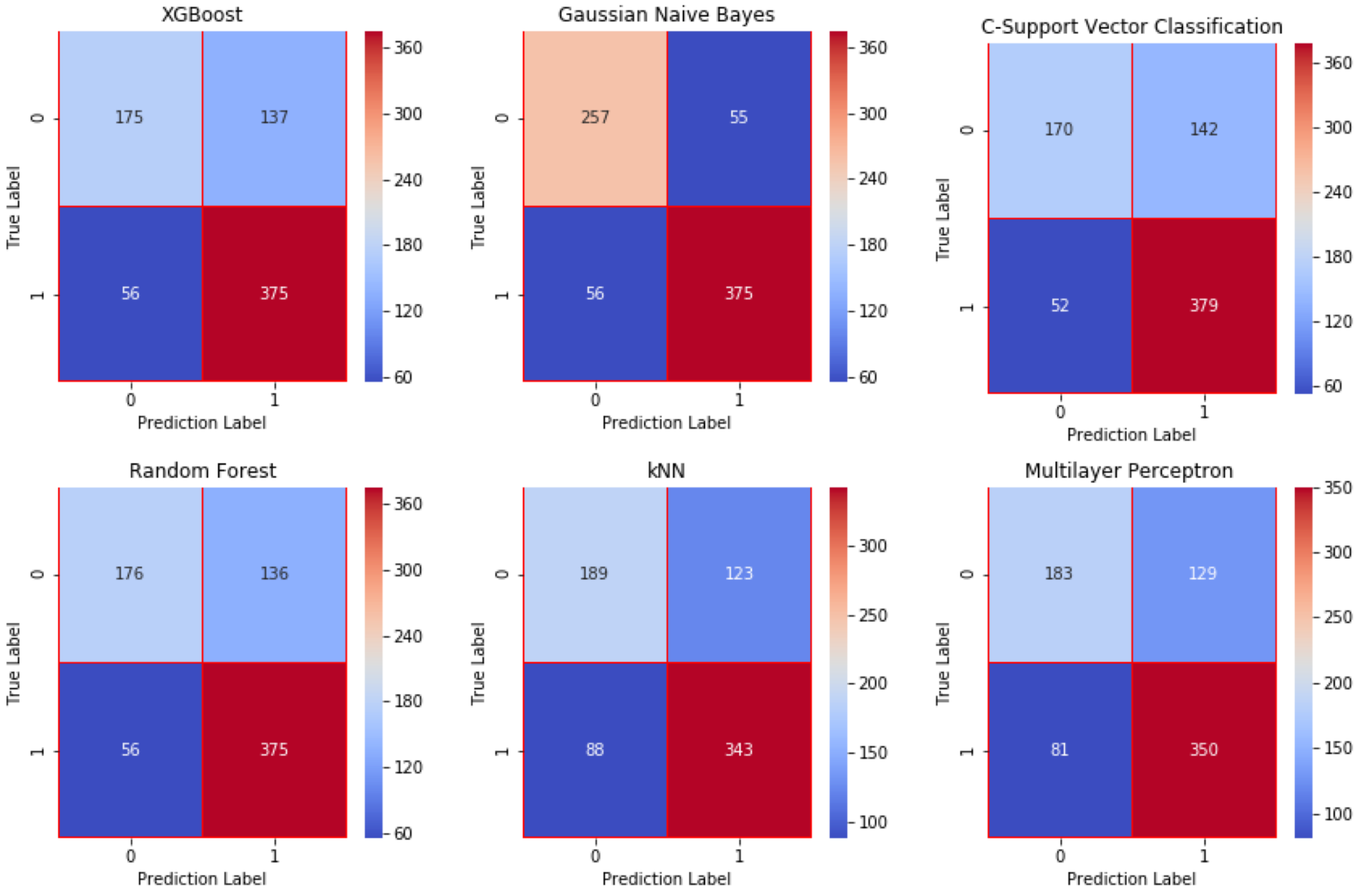
Own Zoo DATA: Results for all 8 parameters

(ui, gr, gi, deVAB_r(b/a), deVRad_r(eff radius), CI (petroR50_r/petroR90_r), absMagR , sersic_n)



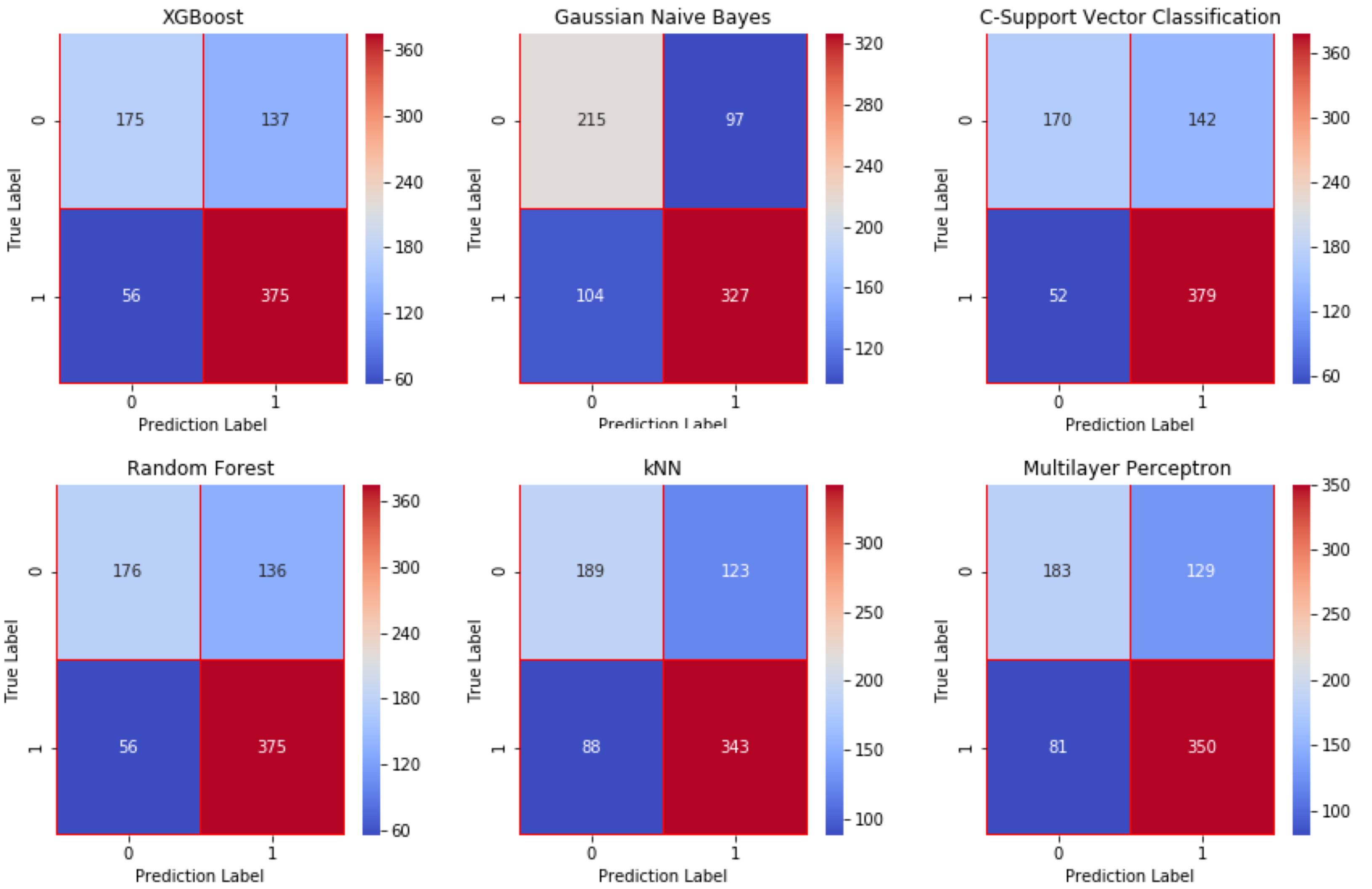
Model	Score
Random Forest	0.875356
XGBoost	0.874794
NN	0.865559
SVC	0.854586
KNN	0.834990
Naive Bayes	0.812446

Own Zoo DATA: Results for 4 structural parameters
(deVAB_r(b/a), deVRad_r(eff radius), CI (petroR50_r/petroR90_r), sersic_n)



Model	**Score**
NN	0.857467
XGBoost	0.856320
Naive Bayes	0.848819
Random Forest	0.847668
SVC	0.841901
KNN	0.830948

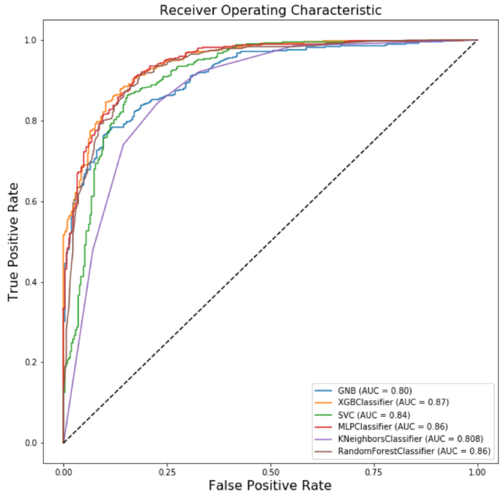
Own Zoo DATA: Results for 4 photometric parameters
(ui, gr, gi, absMagR)



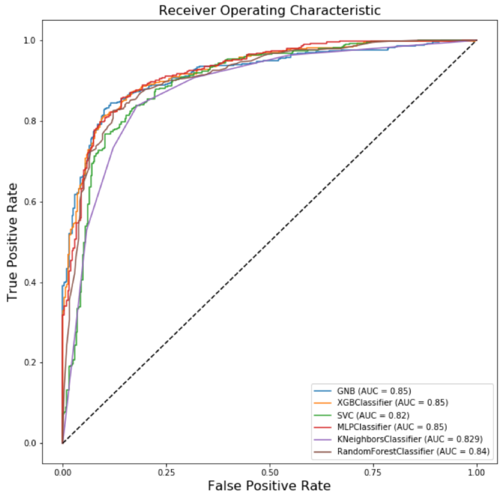
Model	Score
KNN	0.737433
SVC	0.735161
XGBoost	0.730503
Random Forest	0.721863
NN	0.721286
Naive Bayes	0.704537

Our Own Zoo DATA: Comparing of ROC graphs of 6 Machine Learning Algorithms

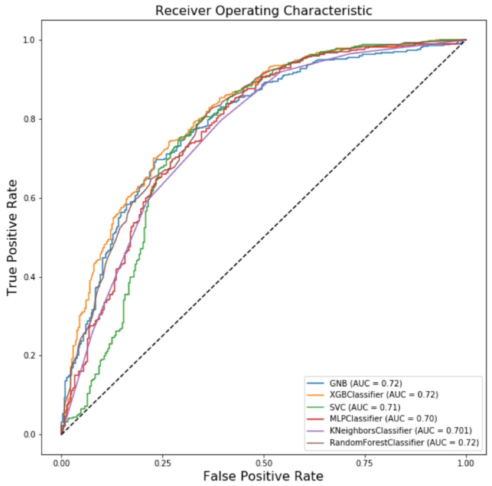
All 8 parameters

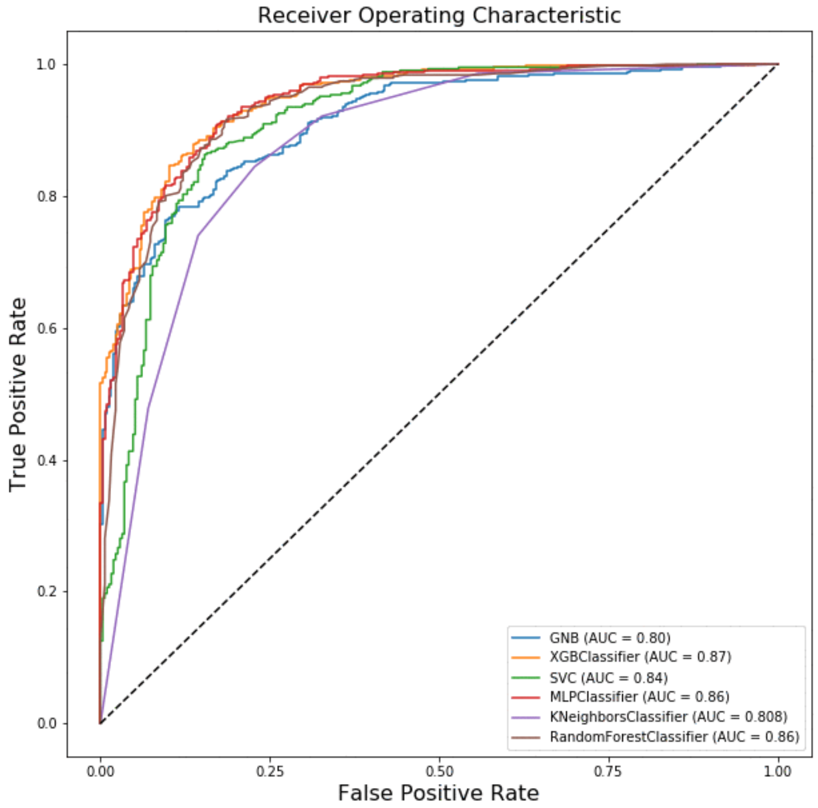


4 structural parameters



4 photometric parameters





Impact of 8 Parameters in XGBoost and Random Forest

XGBoost - Parameters		Random Forest - Parameters	
Feature	**Importance**	**Feature**	**Importance**
Re	0.310302	Re	0.245362
ui	0.168082	ba	0.213058
ba	0.144646	sersic_r	0.118683
CI	0.089578	ui	0.117339
sersic_r	0.089413	gr	0.091458
gi	0.079507	CI	0.082964
gr	0.079207	gi	0.080279
absmagR	0.039266	absmagR	0.050857

4 structural

XGBoost - Parameters

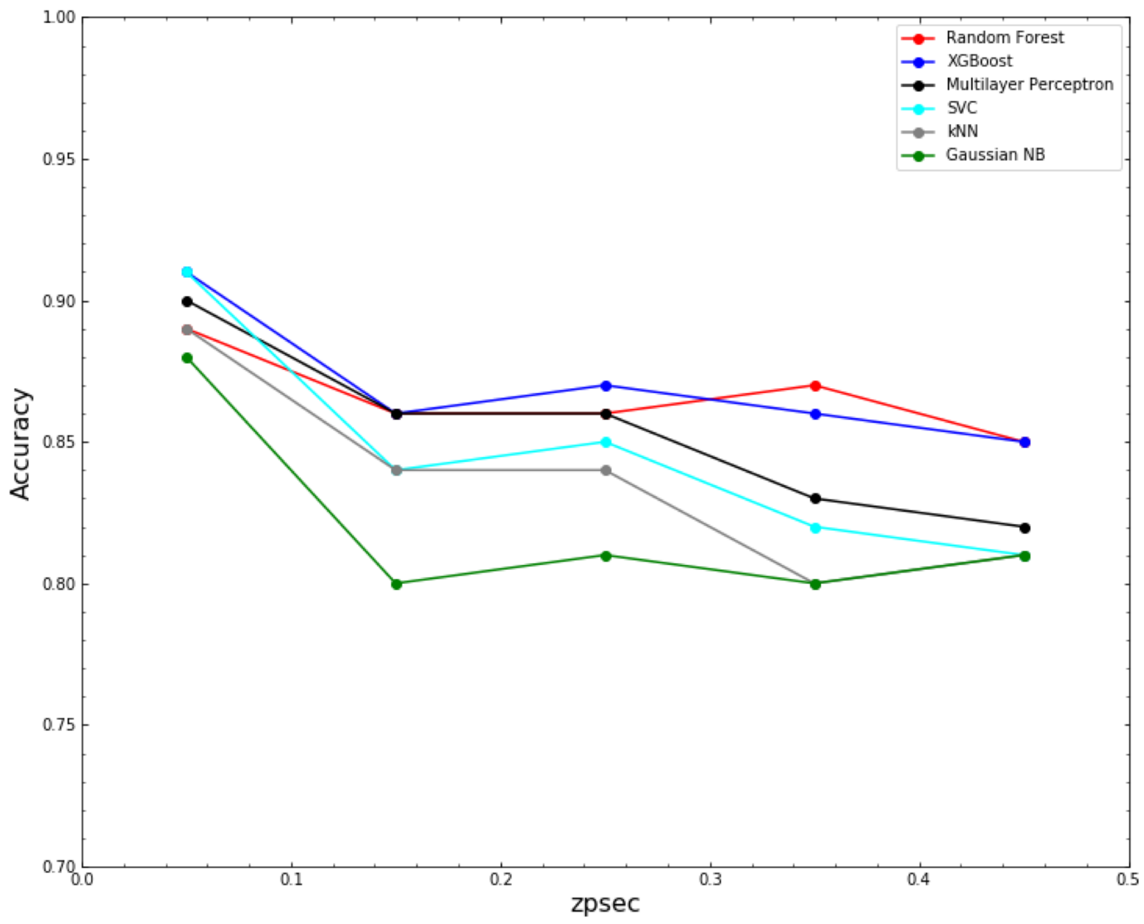
Feature	Importance
Re	0.484281
ba	0.200457
sersic_r	0.163282
CI	0.151981

4 photometric

XGBoost - Parameters

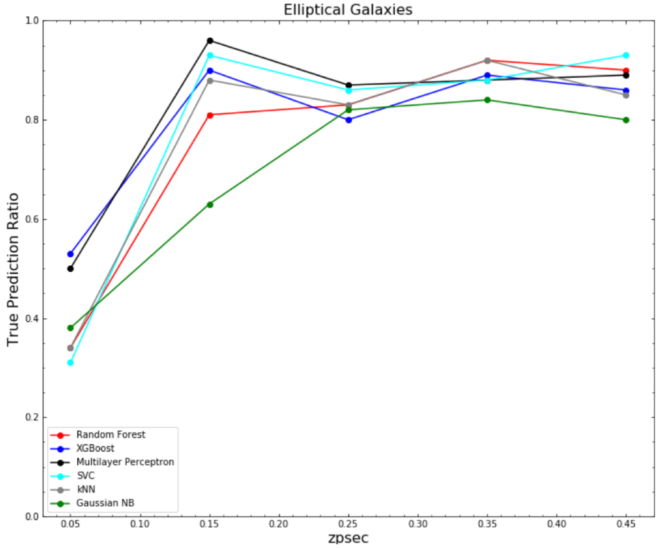
Feature	Importance
ui	0.346192
gr	0.255196
gi	0.226703
absmagR	0.171909

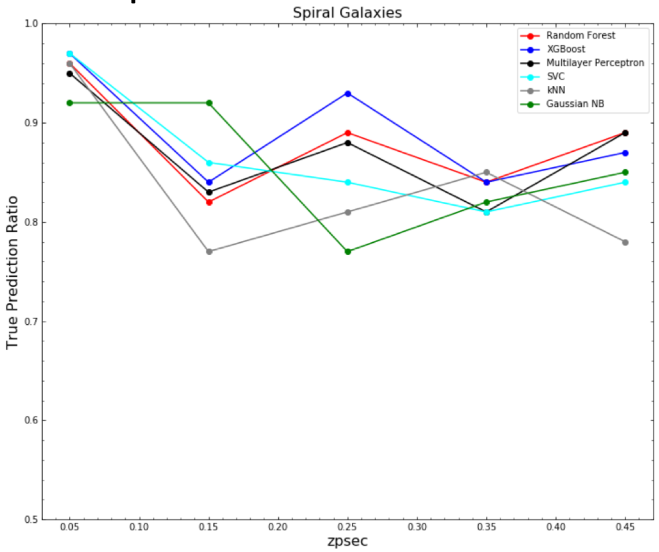
Own Zoo – Accuracy distribution as a function of redshift



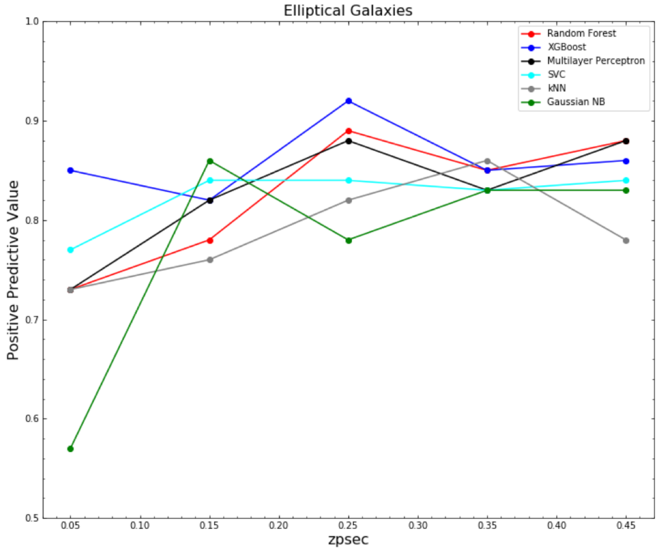
Own Zoo Data: Recalls and Precisions as a function of redshift

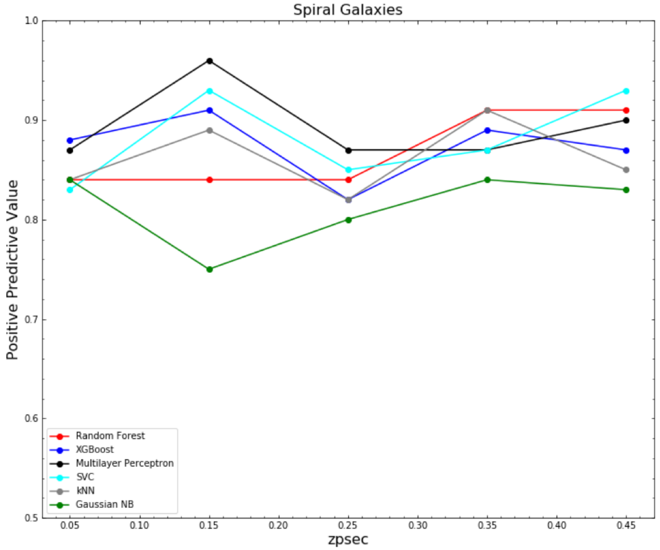
Own Zoo Data: **Redshift Distribution of Recall** for 8 parameters





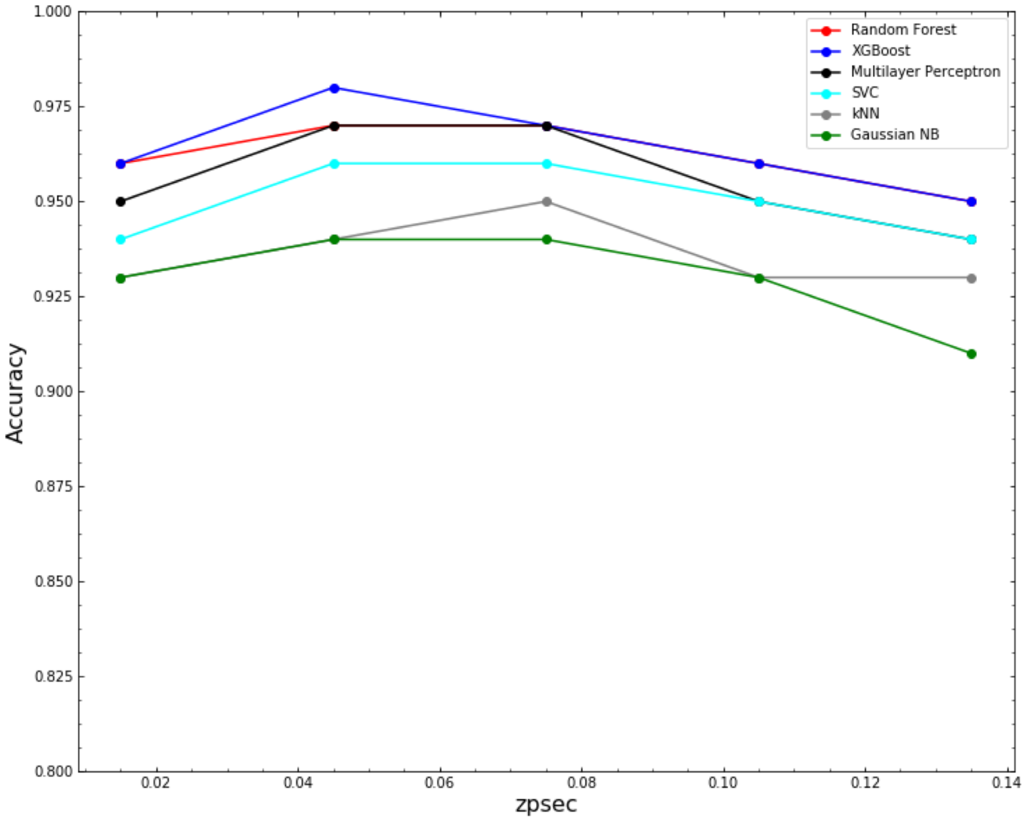
Own Zoo Data: **Redshift Distribution of Precision** for 8 parameters



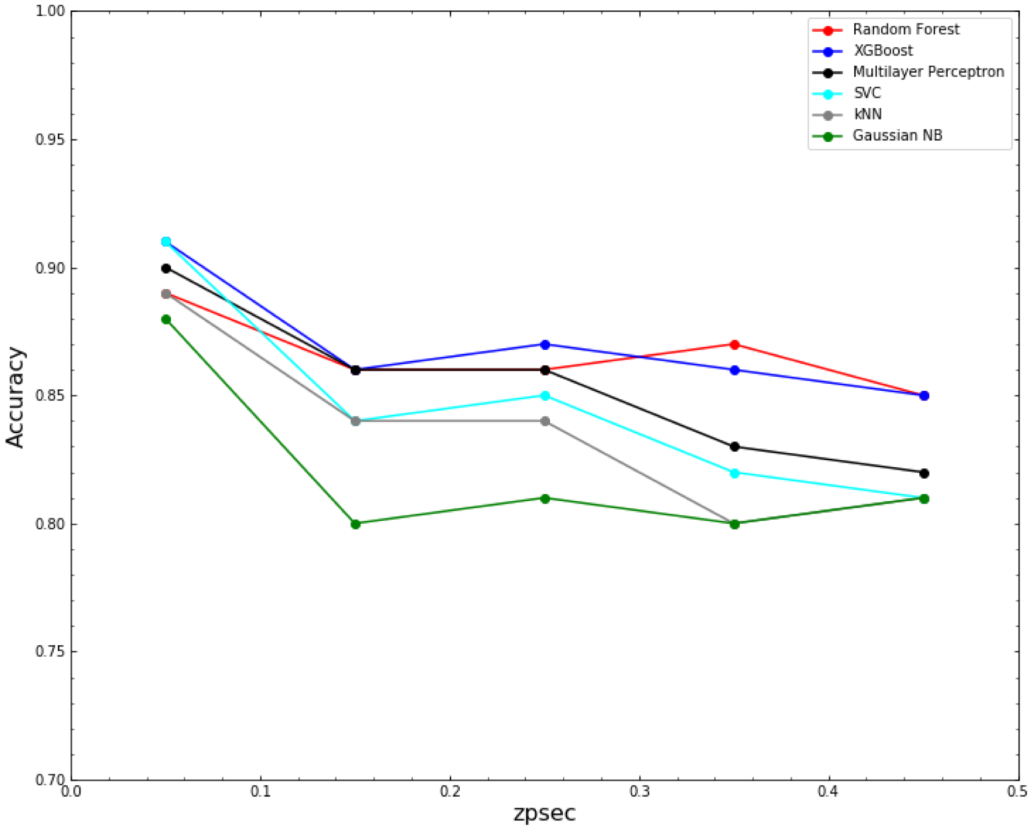


Accuracy distribution as a function of redshift

Galaxy Zoo Data 0 < z < 0.14



Own Zoo Data 0 < z < 0.5



Conclusions

- All 6 ML algorithms give almost the same results with 8 parametric features.
- The best accuray scores are obtained from RF and XGBoost algorithms in two different data.
- The photometric parameters are less effective than structural parameters.
- To select of the parametric features is crucial than the ML algorithms itself and play very important role in the scores of accuracy.
- The accuracy is slightly decreasing with higher redshifts.
- After a certain redshift human eye won't be able to distinguish galaxy classes.
- More classes of morphological types means less accuracy performances.

- Future Work:
  - To extend visually classified sample and test the algorithms.
  - To choose a robust and effective parameters by using the PCA or features selection algorithms!
  - To apply the algorithms to the higher redshifted CFHLTS-W1 field galaxy sample with sersic indices.