Mai Ngo

DSC 540 – Advanced ML

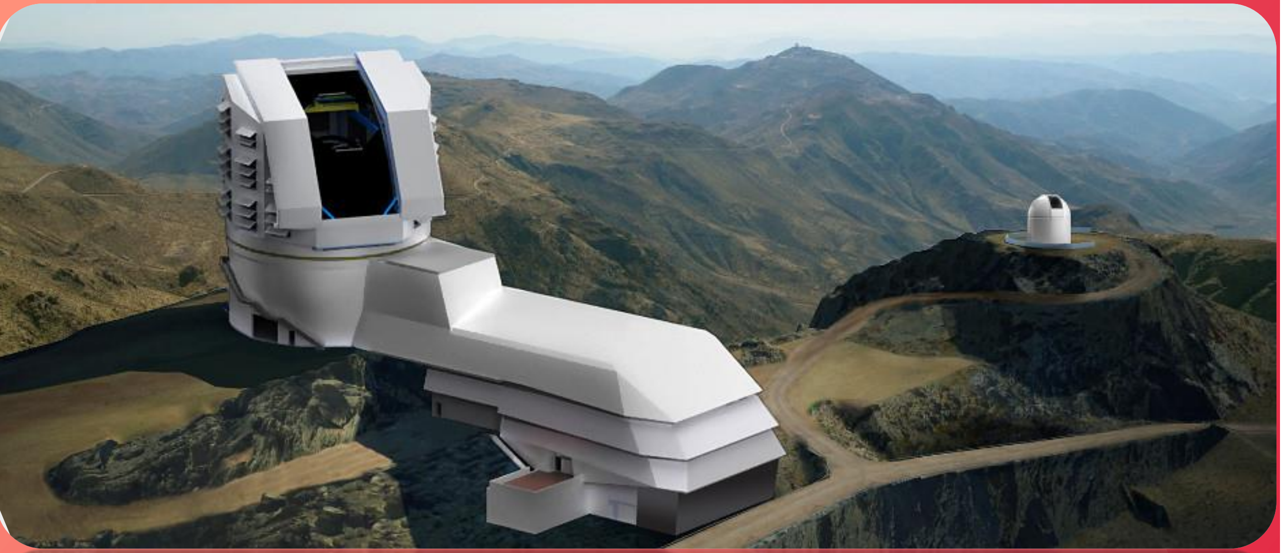
DePaul University

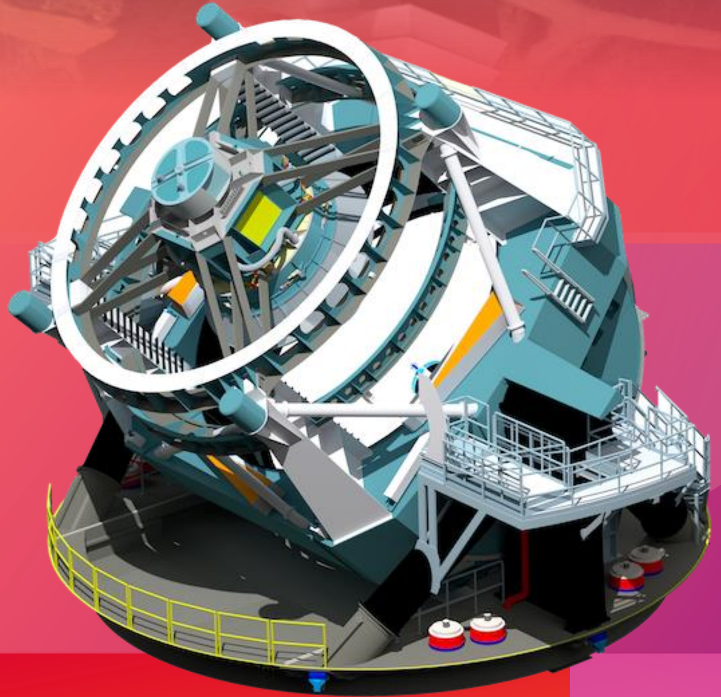
Prof. Casey Bennett

Final Project Presentation

Astrophysical Object Photometric Classification

3.18.2024





AGENDA

Large-aperture Synoptic Survey Telescope (LSST) Data

- Data Introduction
- Visualization
- Data Construction
- Class Imbalance (Down Sampling)

Classification Models

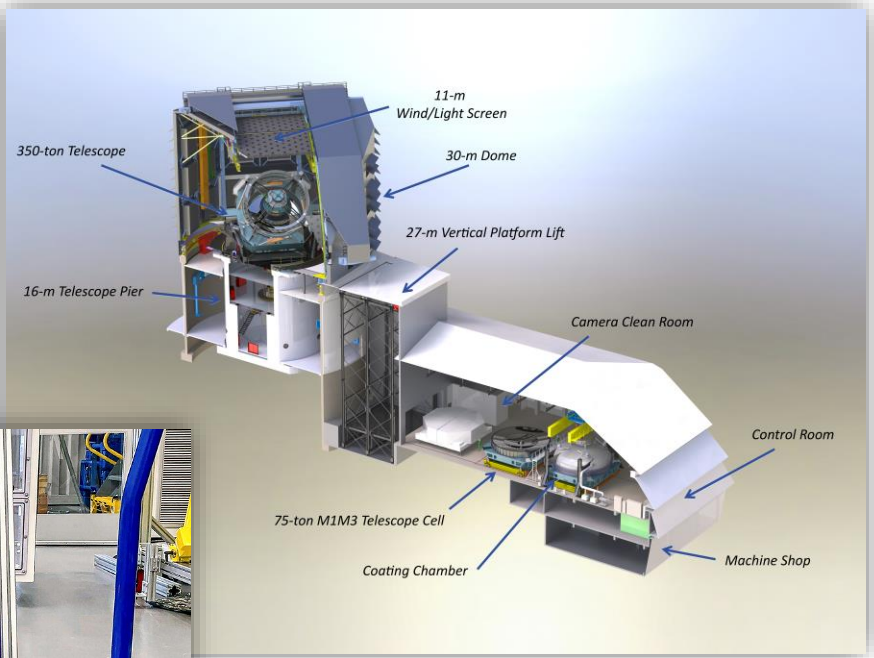
- Random Forest
- Gradient Boosting
- SVM
- Neural Network

Best Model

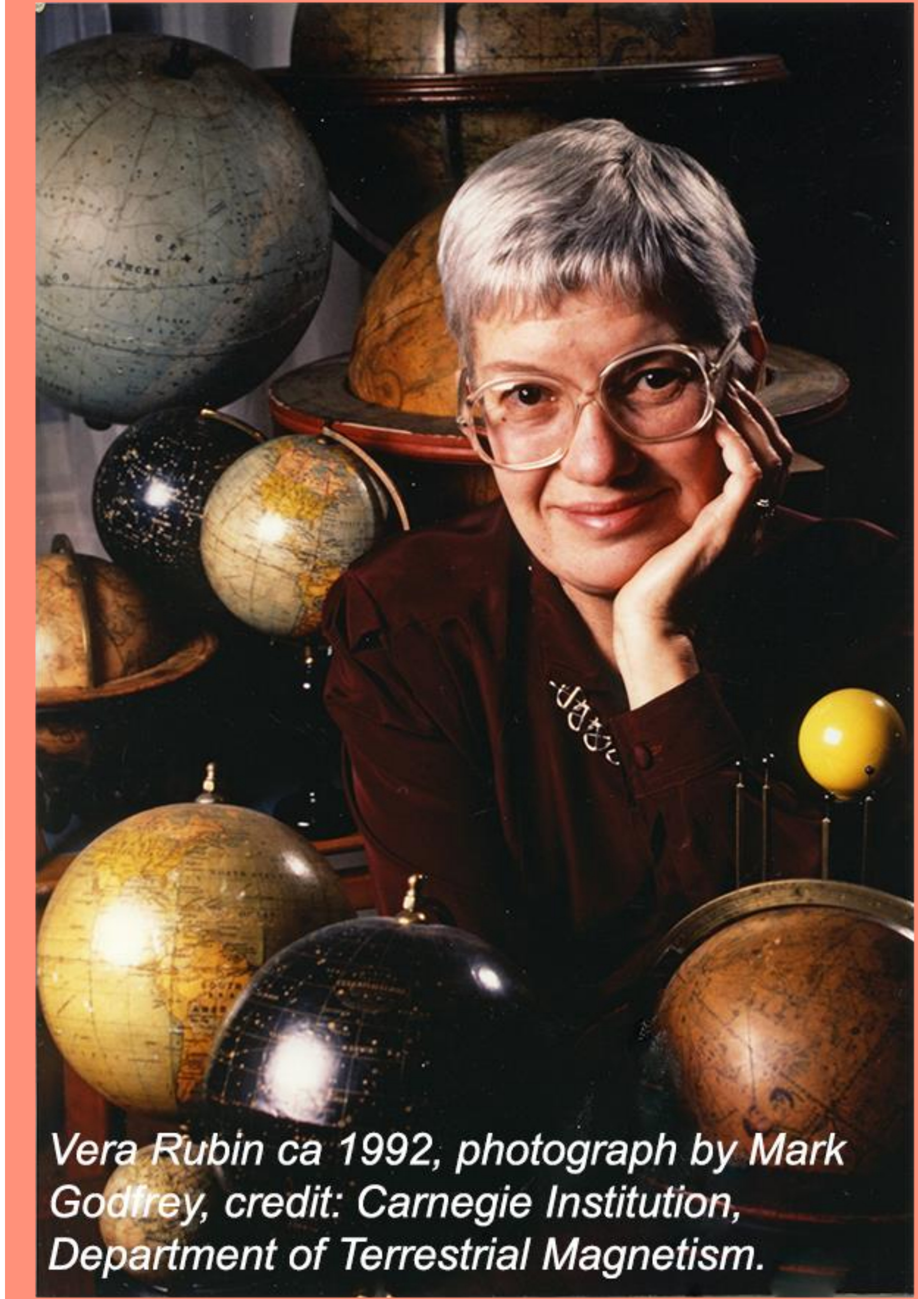
Conclusions & Takeaways

LARGE-APERTURE SYNOPTIC SURVEY TELESCOPE (LSST)

LLST : A highly efficient optical telescope observes activities of near-Earth astrophysical objects.



Location: Vera C. Rubin Observatory in Chile



Vera Rubin ca 1992, photograph by Mark Godfrey, credit: Carnegie Institution, Department of Terrestrial Magnetism.

LSST DATA INTRODUCTION

Meta Data - Obs: 7,848 | Features: 12

object_id: Object identifier.

ra: right ascension | decl: declination.

gal_l: Galactic longitude | gal_b: Galactic latitude

ddf: Flag if object in the Deep Drilling Fields survey.

hostgal_specz: Spectroscopic redshift.

hostgal_photoz: Photometric redshift.

hostgal_photoz_err: Uncertainty on hostgal_photoz.

distmod: Distance to the objects.

mwebv: Extinction of light due to Milky Way dust .

target: 14 classes.

[6, 15, 16, 42, 52, 53, 62, 64, 65, 67, 88, 90, 92, 95]

Time Series Data - Obs:1,421,705 | Features: 6

object_id: Object identifier.

mjd: Time of the observation.

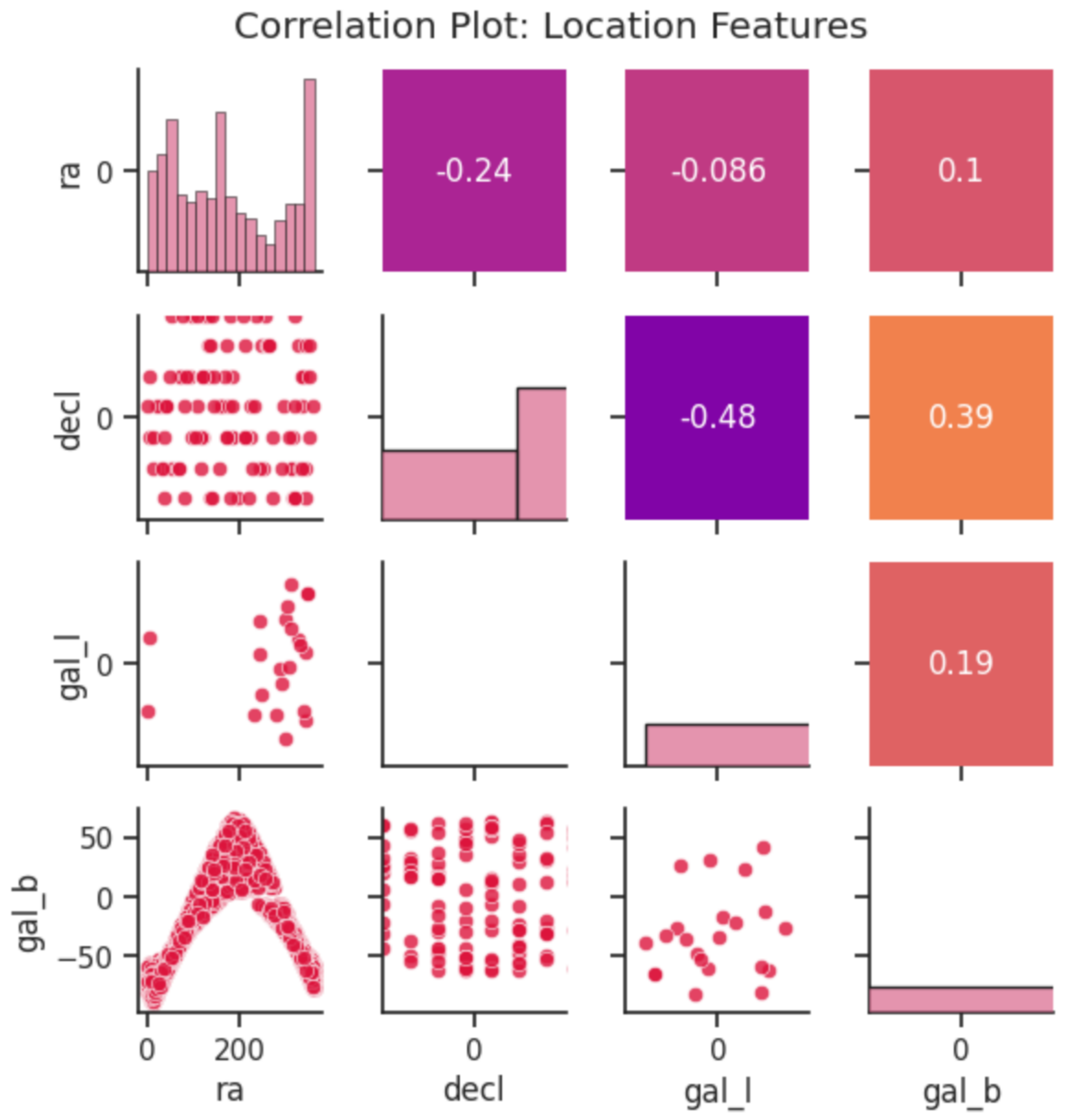
passband: The specific LSST passband.

flux: Measures brightness in the passband of observation.

flux_err: Uncertainty on flux.

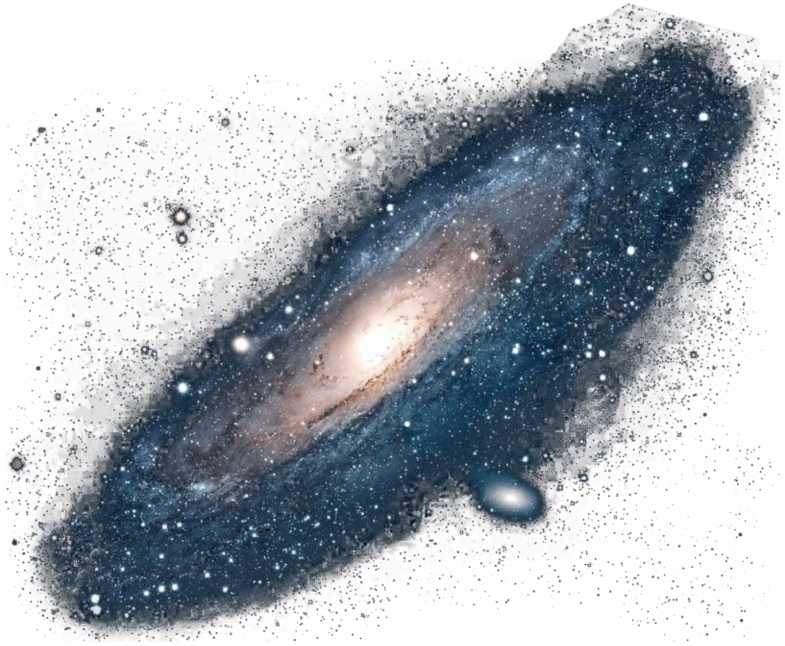
detected: Flag if object's brightness exhibits statistically significant.

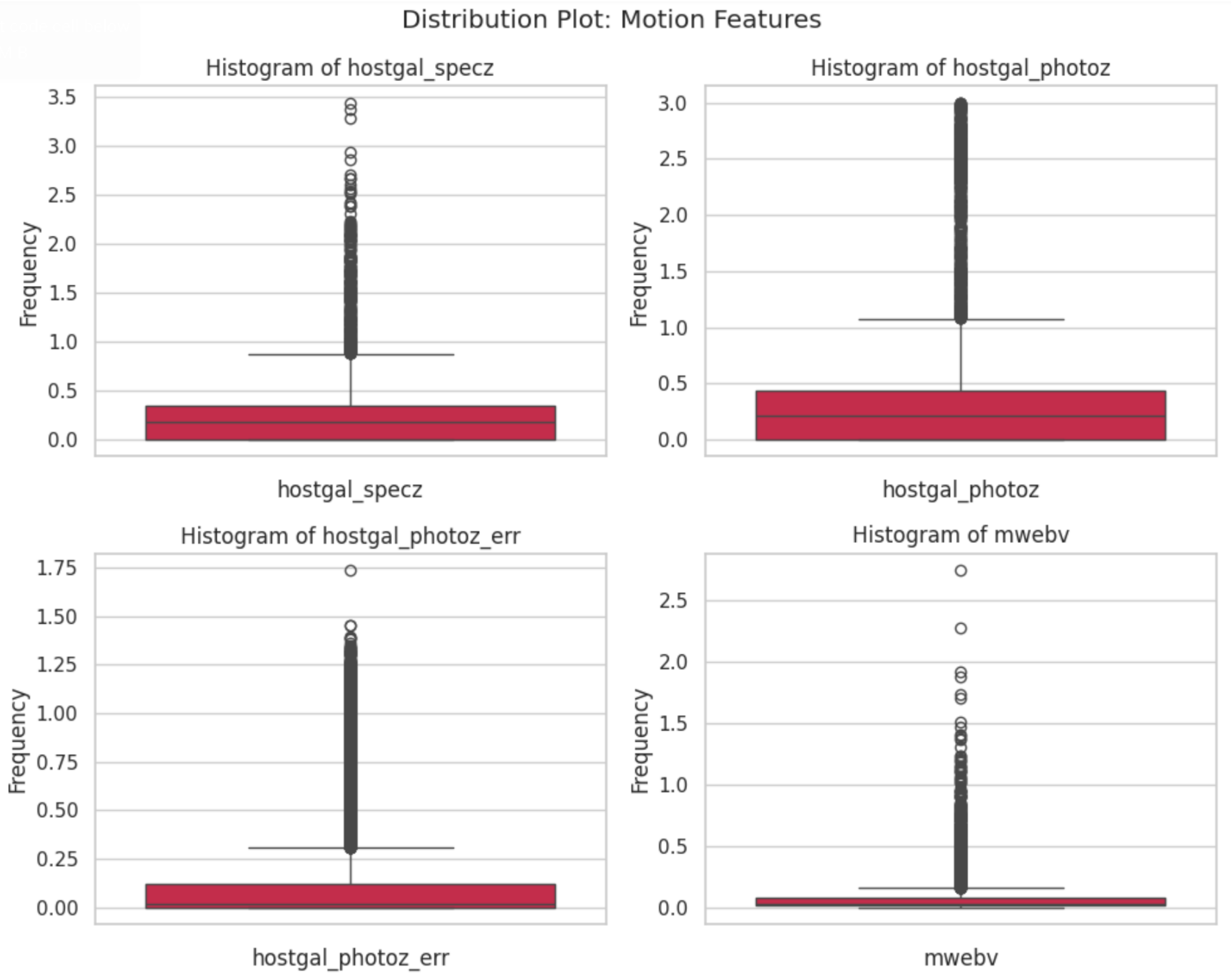
Multiple object_id under ONE target class!



DATA VISUALIZATION - META DATA

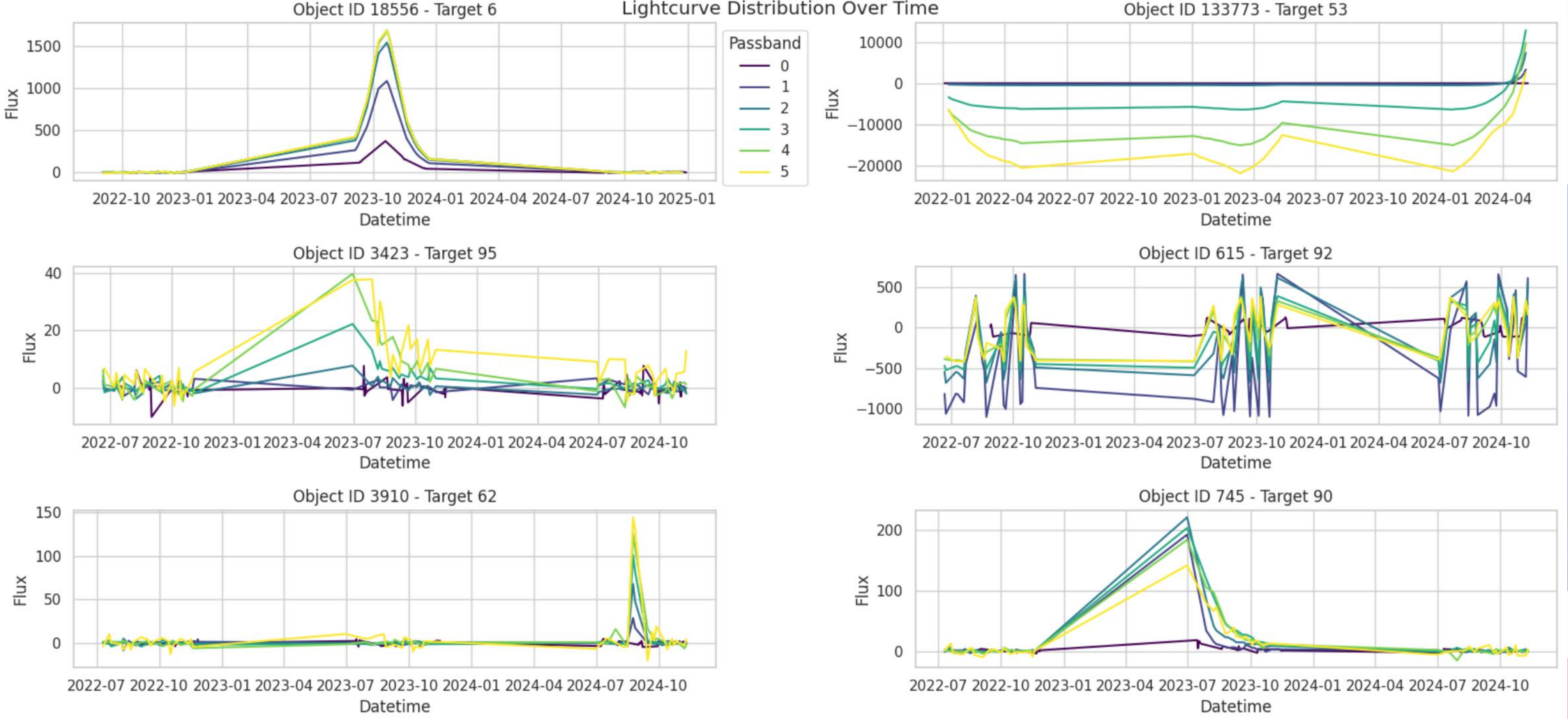
DATA VISUALIZATION – META DATA





Passband | Flux ?

DATA VISUALIZATION- TIME SERIES DATA



DATA VISUALIZATION–TIME SERIES DATA

Objective: Merged dataset contains both static numerical & time series data!

DATA CONSTRUCTION

Apply Time Rolling (1 feature | 2 time-steps)

Sequences of uniform length.

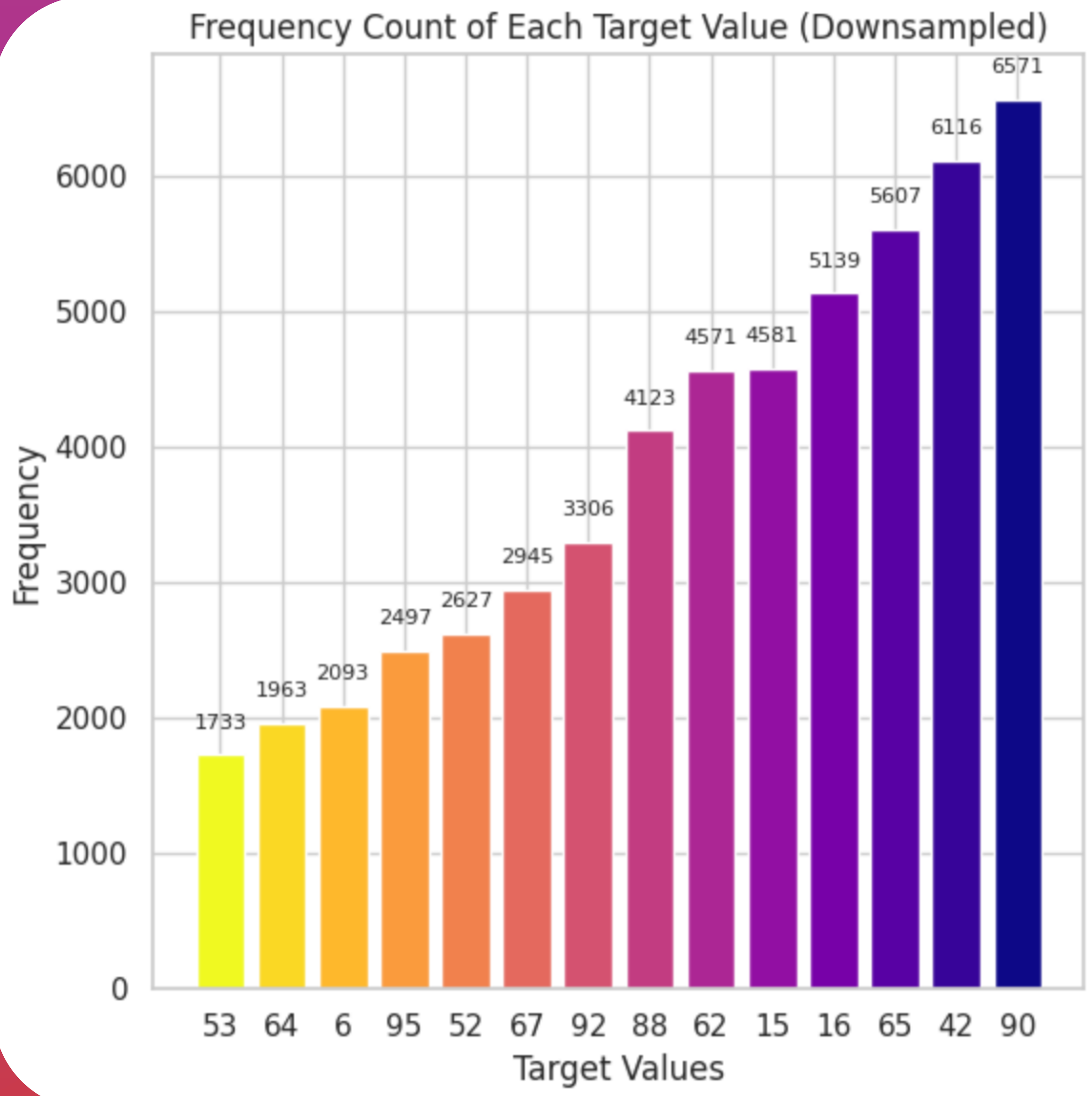
-> Each object must have the same observation time range.

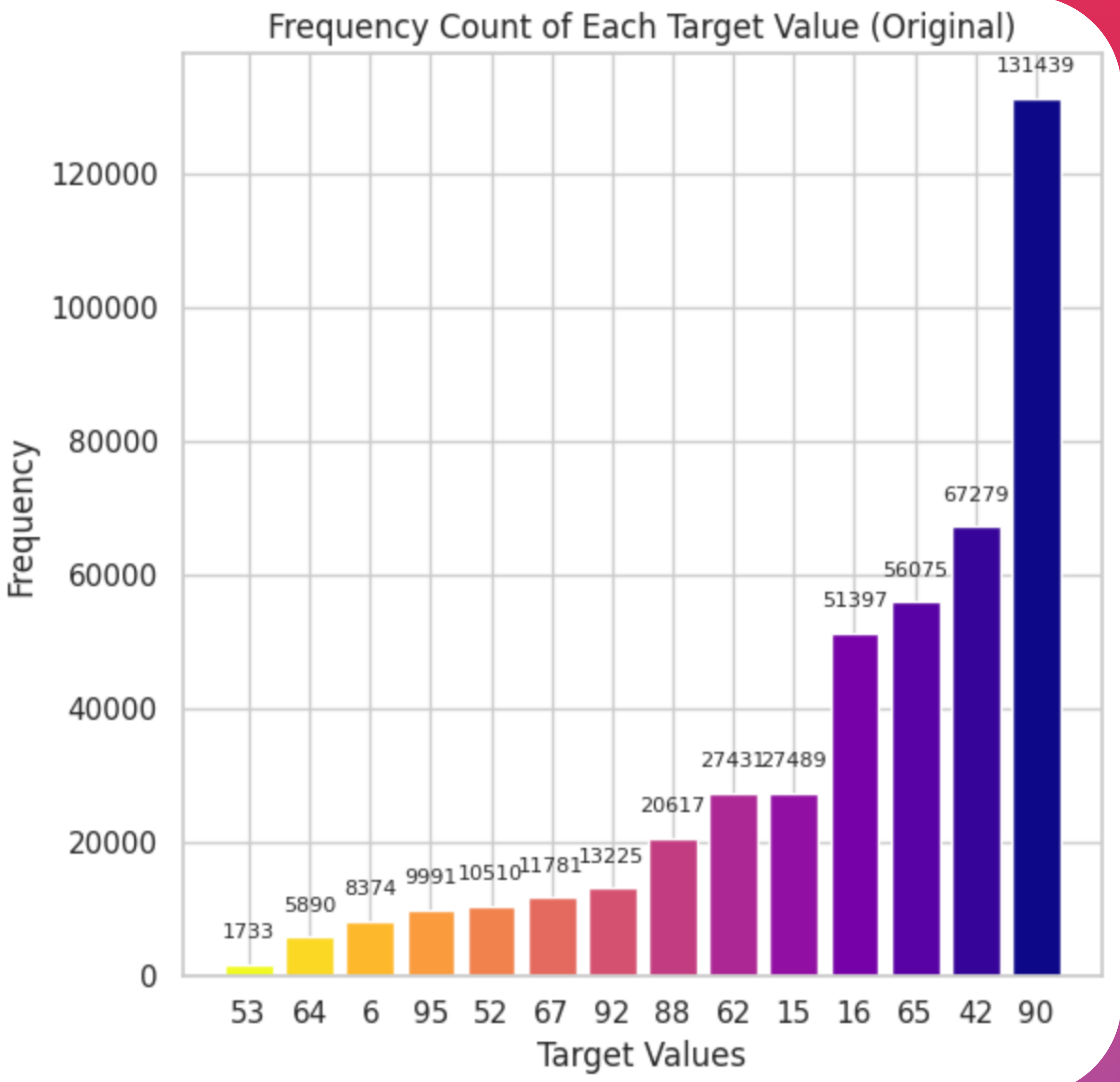
Date	Temp (° C)
01/01/24	35
01/02/24	36
01/03/24	37
01/04/24	38

Date	Temp_t1	Temp_t2
01/01/24	35	36
01/02/24	36	37
01/03/24	37	38
01/04/24	38	NaN

*Features drop as well.

CLASS IMBALANCE (DOWN SAMPLING)





FINAL DATA

Train Data:

- Observations: 667, 725
- Features: 18

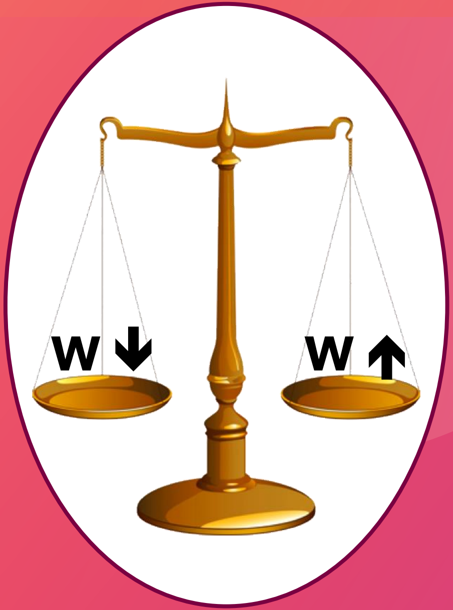
Down Sampled Data:

- Observations: 81,155
- Features: 18

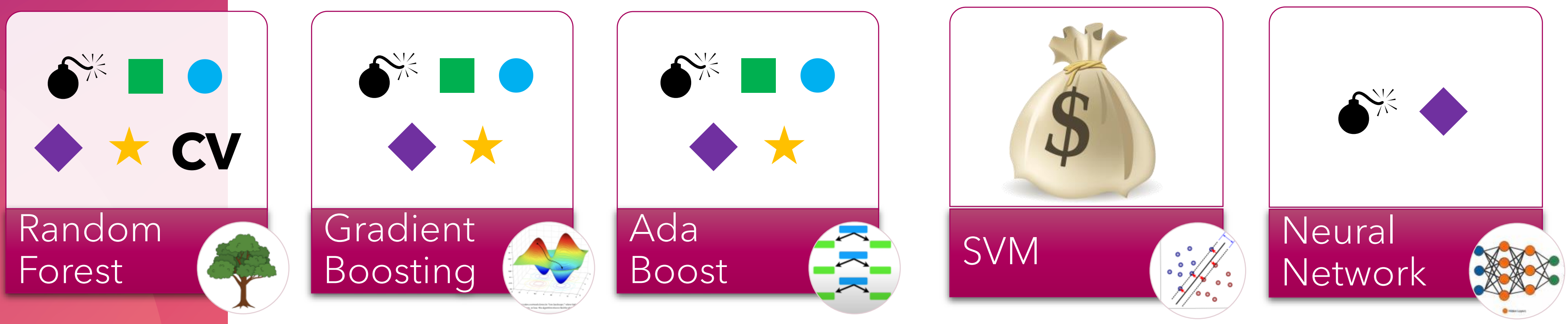
Train-Test Split | 0.65 – 0.35

-> Ensure similar class distribution in both sets.

Feature Selection ONLY on static data.



CLASSIFICATION MODELS



💣 🟩 🔵 ◆ ⭐ CV

Random Forest

💣 🟩 🔵 ◆ ⭐

Gradient Boosting

💣 🟩 🔵 ◆ ⭐

Ada Boost

SVM

💣 ◆

Neural Network

Top-down order!

💣: Parameter Tuning → Base 'best' Model.

🟩 : Feature Selection (FS) on Train Data.

🔵 : Model FS with Train Data.

◆ : Base 'best' Model on Down-sampled Data.

⭐ : Model FS on Down-sampled Data.

Top-down order!

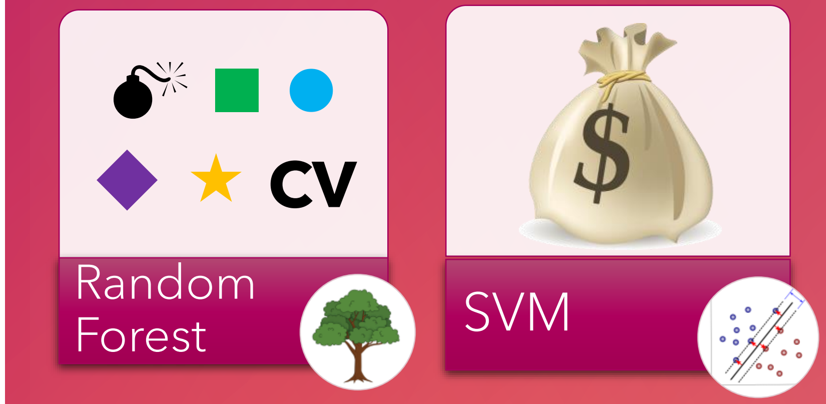
💣: Hyperparameter Tuning → Base 'best' Model.

🟩 : Feature Selection (FS) on Train Data.

🔵 : Model FS with Train Data.

◆ : Base 'best' Model on Down-sampled Data.

★ : Model FS on Down-sampled Data.



CLASSIFICATION MODELS

RANDOM FOREST

- Tune parameters: 'criterion': [**'gini'**, 'entropy’]
- Both with and without Cross-Validation gives the same outputs. Yet significantly difference in run time.

Model	Accuracy	Log Loss
Train Data	**0.97**	**1.05**
Train Data (FS - 2)	**0.81**	**6.78**
Down-sampled Data	**0.91**	**3.17**
Down-sampled Data (FS -2)	**0.64**	**12.89**

Model	Accuracy	Log Loss
Train Data	**0.68**	**11.64**
Train Data (FS - 2)	**0.65**	**12.55**
Down-sampled Data	**0.65**	**12.53**
Down-sampled Data (FS - 2)	**0.64**	**13.12**

GRADIENT BOOSTING

Tune parameters: 'criterion': ['friedman_mse', **'squared_error’**]

ADA BOOST

Tune parameters: 'n_estimators': [20, 50, **100**, 200],
'learning_rate': [**0.01**, 0.3, 0.5, 1.0]

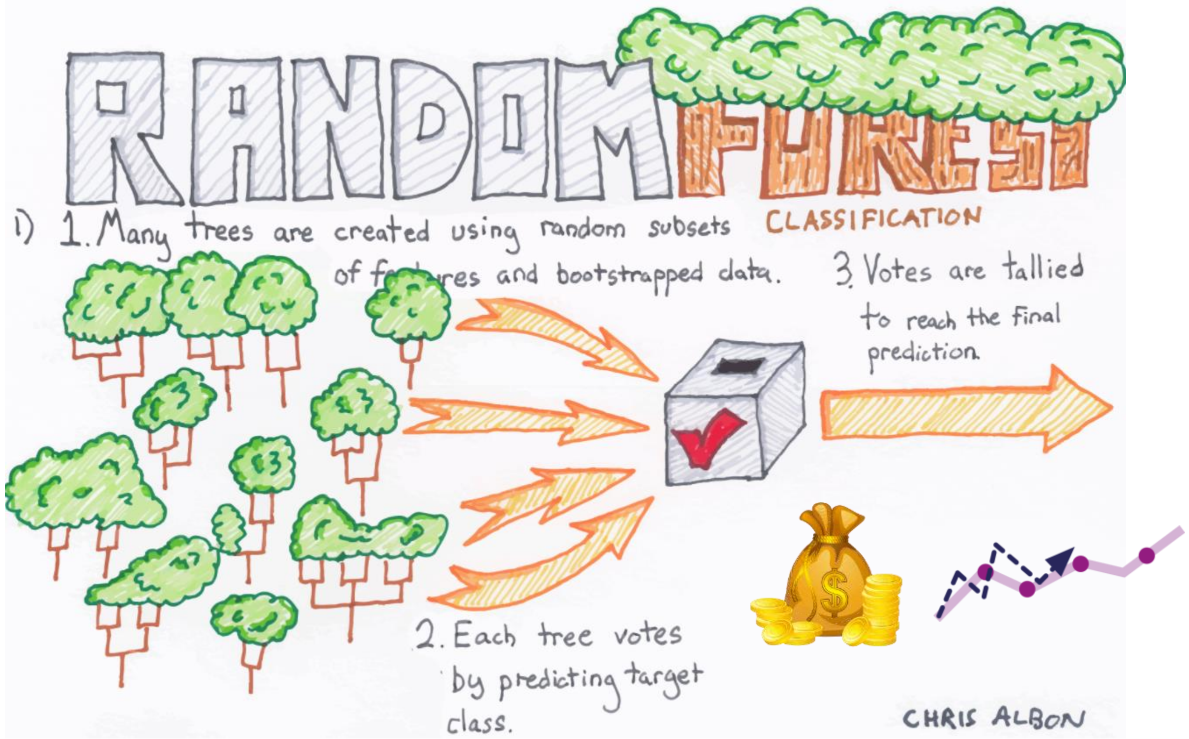
Model	Accuracy	Log Loss
Train Data	**0.46**	**19.54**
Train Data (FS - 3)	**0.31**	**24.74**
Down-sampled Data	**0.23**	**7.78**
Down-sampled Data (FS - 3)	**0.23**	**27.78**

NEURAL NETWORK

Model	Accuracy	Accuracy (weight)	Cross-entropy Loss
Down-sampled Data	**0.46**	**0.59**	**9.6**

- 1 LSTM layer for time series.
- 2 Dense layer for static.
- Drop out rate = 0.2
- Dense output layer | combine.

BEST MODEL?


RANDOM FOREST
CLASSIFICATION
1) 1. Many trees are created using random subsets of features and bootstrapped data.
2. Each tree votes by predicting target class.
3 Votes are tallied to reach the final prediction.
$
CHRIS ALBON

	Accuracy	Log Loss
Train Data	0.97	1.05
Train Data (FS - 2)	0.81	6.78
Down-sampled Data	0.91	3.17
Down-sampled Data (FS - 2)	0.64	12.89

IMPROVEMENT SUGGESTIONS

- ✓ Much more room for data pre-processing.
- ✓ Time-series | Static Data.
- ✓ Feature Engineering.
- ✓ Class Imbalance.
- ✓ Machine Learning Models | Neural Network.
- ✓ Parameters Tuning.

CONCLUSION & TAKEAWAYS

- ✓ Enjoy working on this project.
- ✓ Apply my knowledge from this course.
- ✓ Learn about astrophysical objects and how telescope works!

THANK YOU!