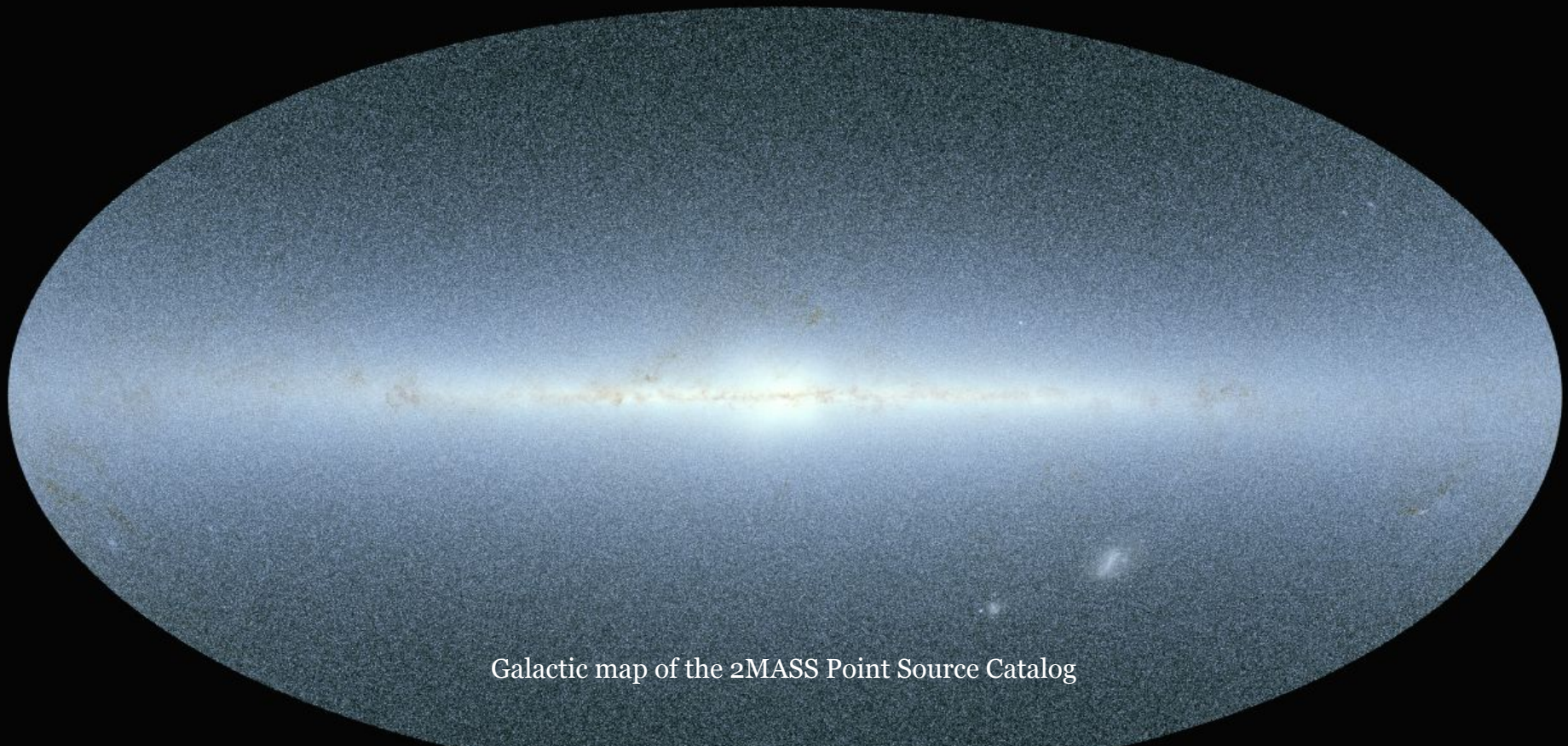
How I Wonder What You Are

Classifying Stellar Objects Using Logistic Regression

Jenica Andersen
April 20, 2022
Metis DSML, Classification Module

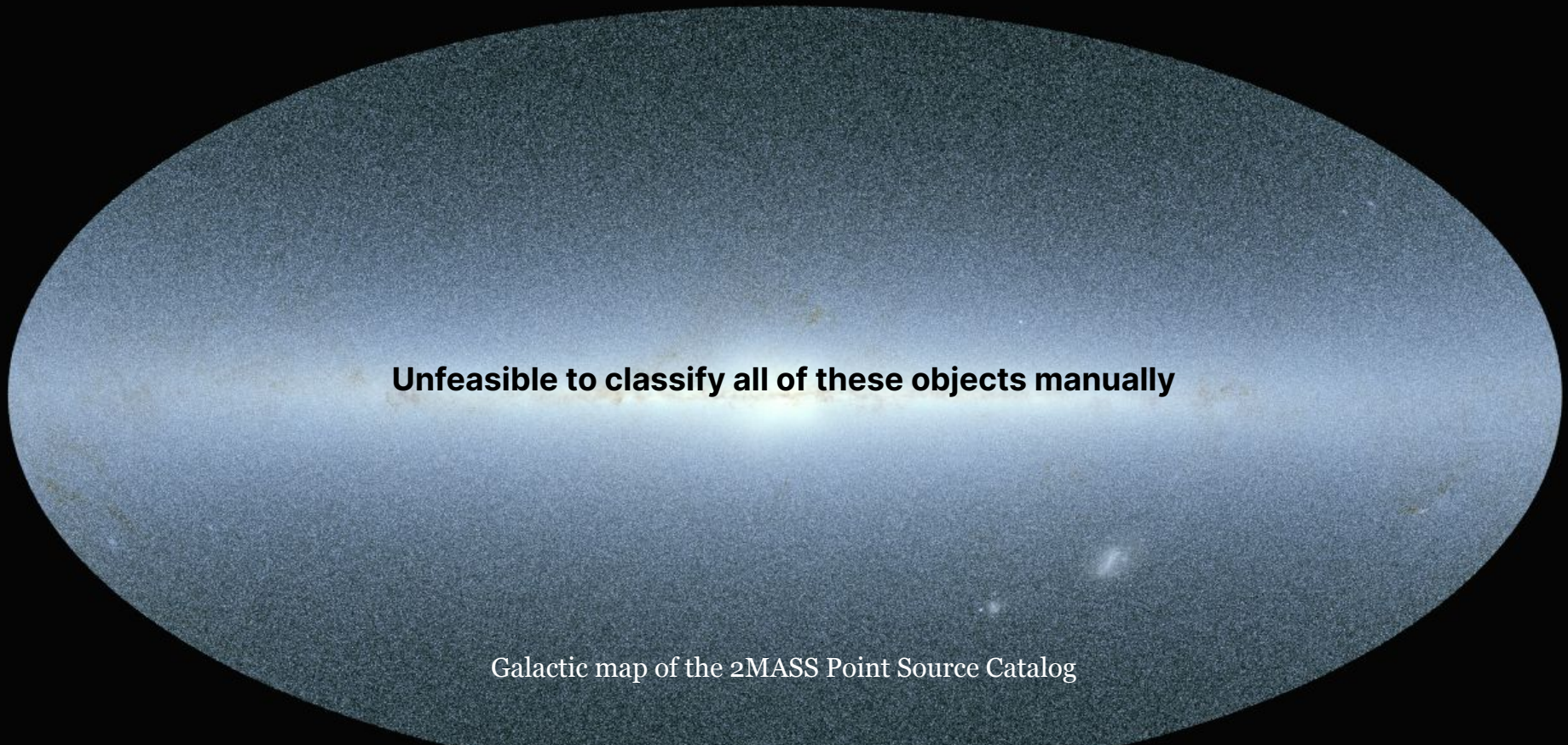
METIS®

Panoramic View of the Sky



Galactic map of the 2MASS Point Source Catalog

Panoramic View of the Sky



Galactic map of the 2MASS Point Source Catalog

Why Map the Sky?

Make Discoveries

Think cartography when Earth was uncharted

Origins of the Universe

Understand cosmic evolution and the history of where we all came from

Dark Matter

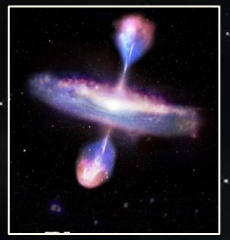
Map what’s seen, to better understand what’s unseen

Light Point-Sources

Stars

Galaxies

Quasars



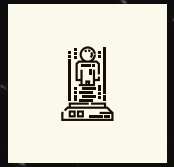
The Data

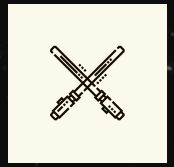
Kaggle Stellar Classification Dataset:

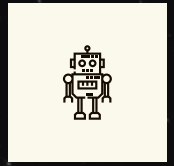
- 100,000 “sources”
- 17 features
  - (8-10 “of interest”, all continuous)
- Target: “class”
  - 18% Quasar
  - 60% Galaxy
  - 22% Stars

(No class imbalance handling)

Features

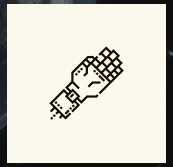


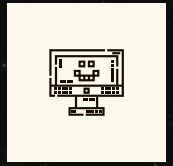




Features Used:

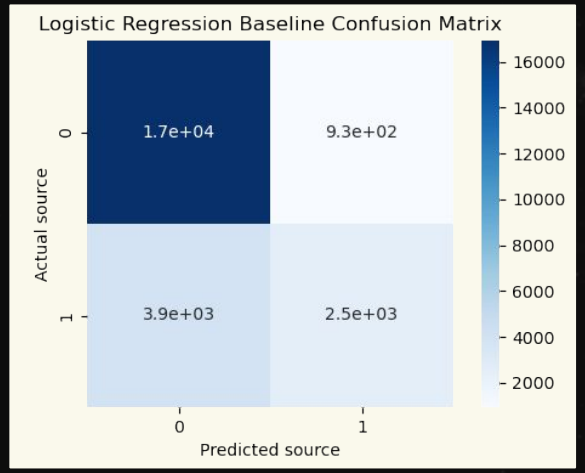
- **Location: Coordinates** on celestial sphere, **Field #**
- Photometric data (“**brightness”**): u, g, r, i, z
- **Redshift** data: increase in wavelength

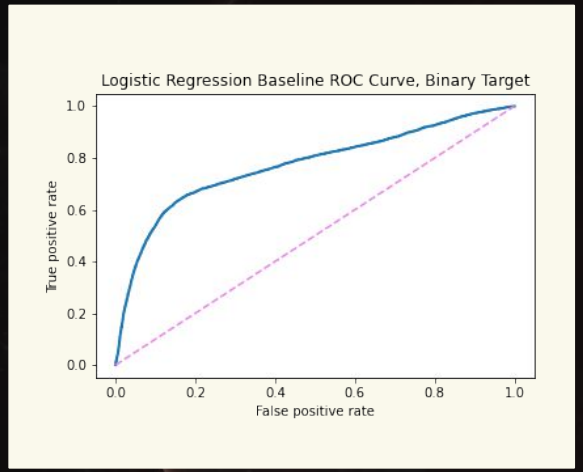




Features NOT Used:

- Object, Equipment and Run ID’s
- Date

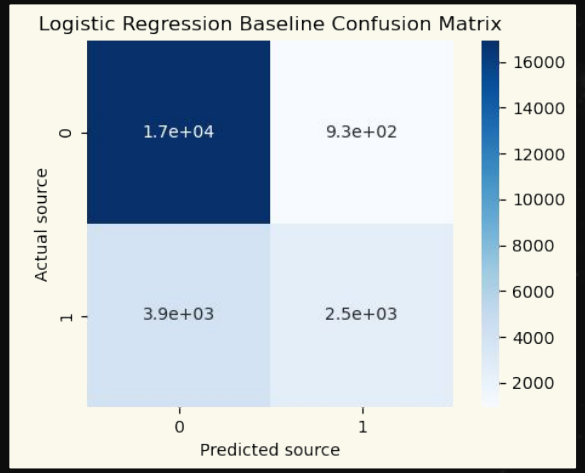


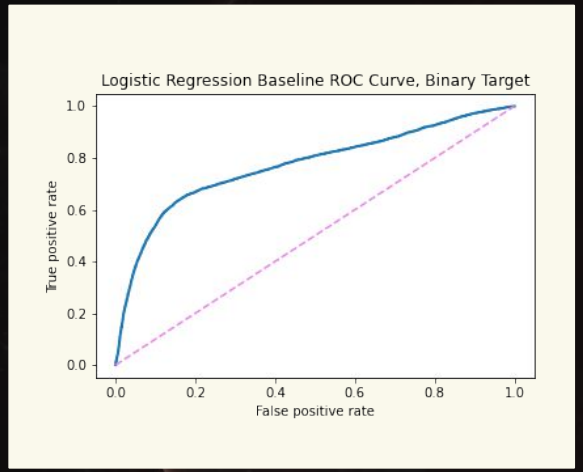


EDA: Baseline Binary Logistic Regression

Positive Class: “Star”

ROC AUC Score = 0.77





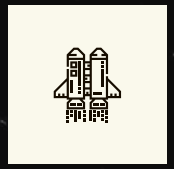
EDA: Baseline Binary Logistic Regression

Positive Class: “Star”

ROC AUC Score = 0.77

4,830 sources misclassified!

Models & Algorithms



kNN

Logistic Regression

Random Forest

Pipeline

Organized models, preprocessing & parameters

Randomized- & GridSearch CV

Random- was faster, honed by **Grid-.** 30% test data

Scaled Data

MinMaxScaler and Standard Scaler

Metrics

ROC AUC

tells how much the model is capable of distinguishing between classes

ROC

R

Recall

indicates how good the classifier is at minimizing false negatives

Precision

indicating how good the classifier is at identifying true positives

P

F1

F1 Score

an overall performance metric

Metrics

Used for GridSearchCV

ROC AUC

tells how much the model is capable of distinguishing between classes

ROC

R

Recall

indicates how good the classifier is at minimising false negatives.

Precision

indicating how good the classifier is at identifying true positives

P

F1

F1 Score

an overall performance metric

Metrics

Used for GridSearchCV

ROC AUC

tells how much the model is capable of distinguishing between classes

ROC

Precision

indicating how good the classifier is at identifying true positives

P

Also key

R

Recall

indicates how good the classifier is at minimising false negatives.

F1

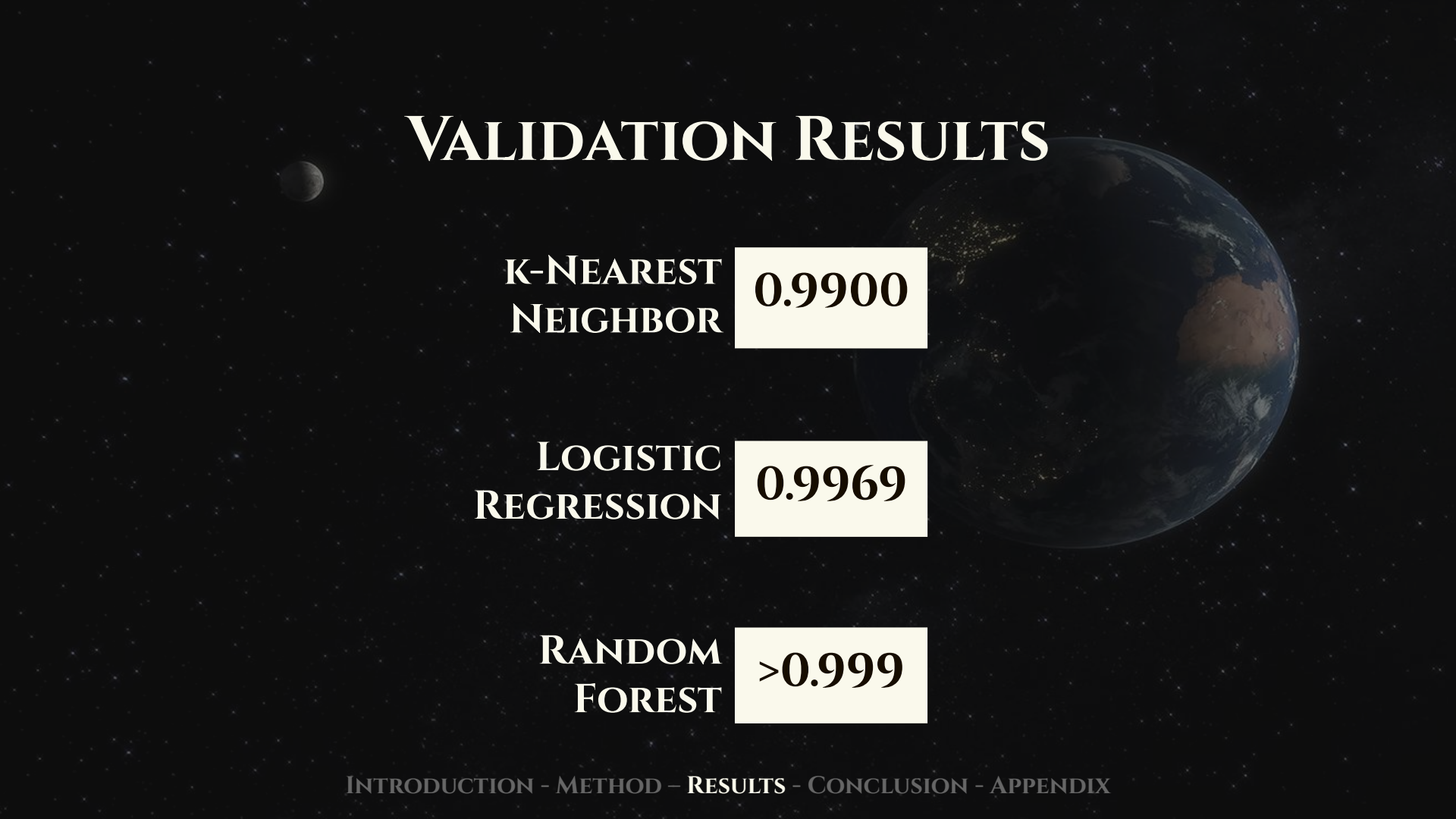
F1 Score

an overall performance metric

Recall: Cost of False Negatives > False Positives (in Binary Model)

- **Nearby stars** are brighter → prone to misclassification
- Strong interest in nearby **extra solar planets** (exploring for life, richer scientific results)
- **Fewer stars** than galaxies (single misclassification represents a larger fraction)
- Potential for **compounded mistakes**, far-reaching and unknown

Validation Results



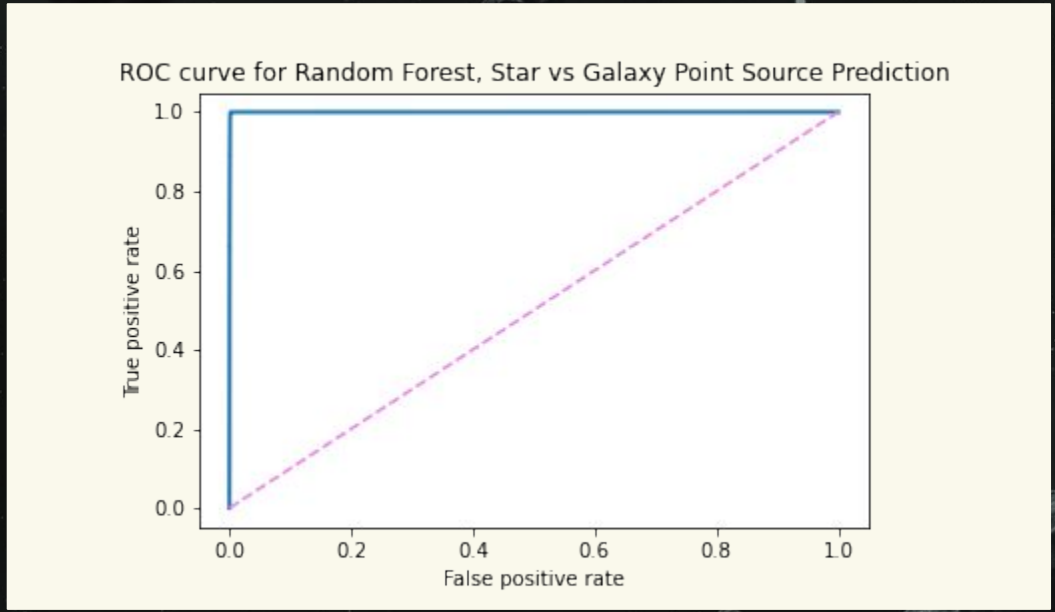
k-Nearest Neighbor	0.9900
Logistic Regression	0.9969
Random Forest	>0.999

Results: Validation Model Scores

k-Nearest Neighbor	0.9900
Logistic Regression	0.9969
Random Forest	>0.999

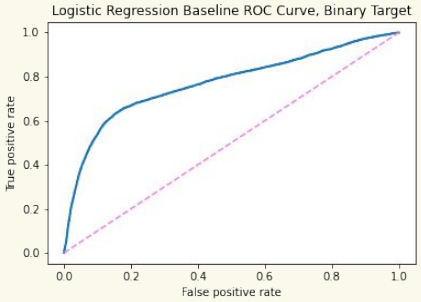
Results: Final ROC AUC Scores
(Binary Class Only)

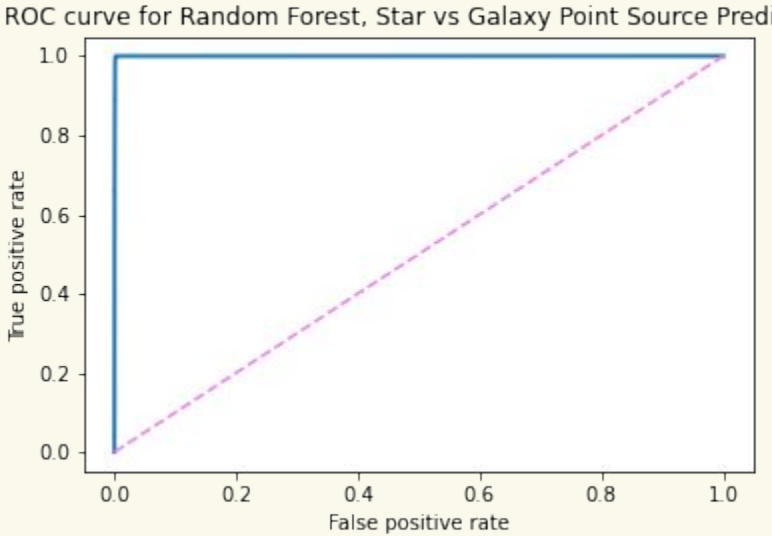
Validation Score	>0.999
Test Data Score	0.9983



Final Random Forest Model (Binary Class Only)

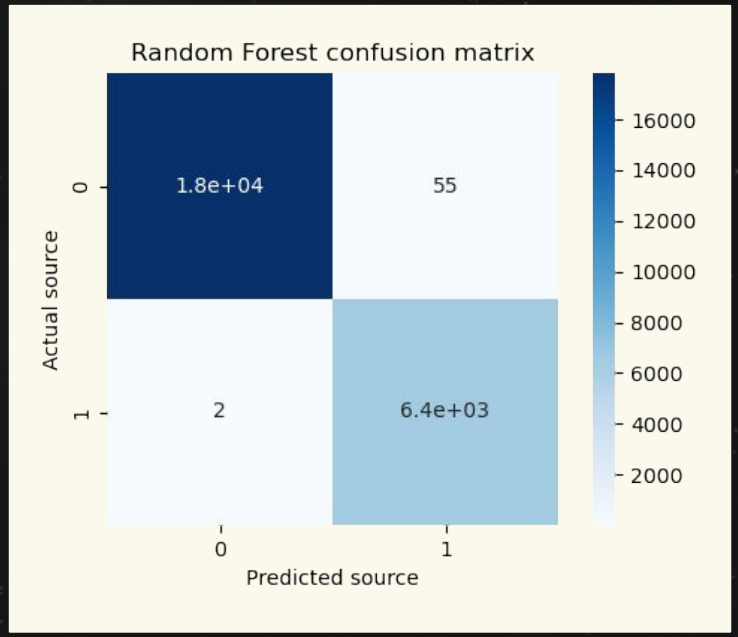
Baseline model



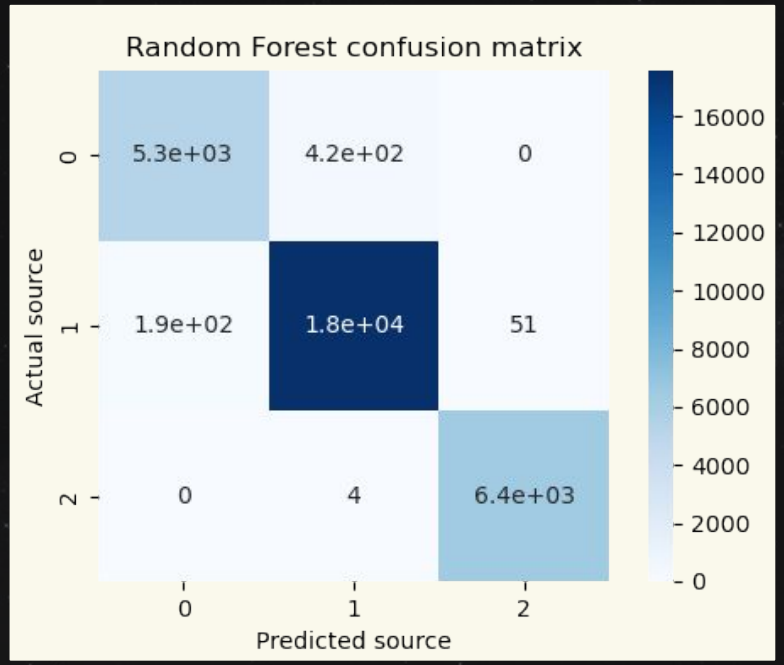


Final Random Forest Model (Binary Class Only)

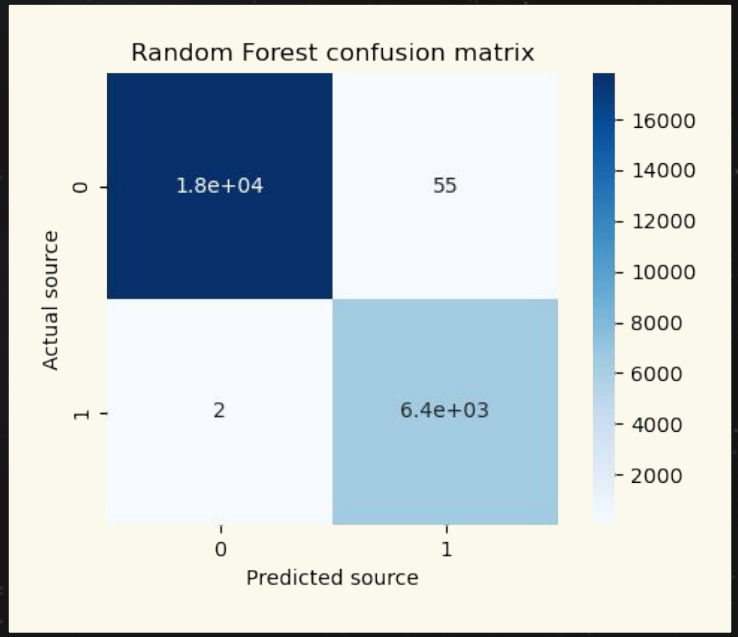
Binary



MultiClass



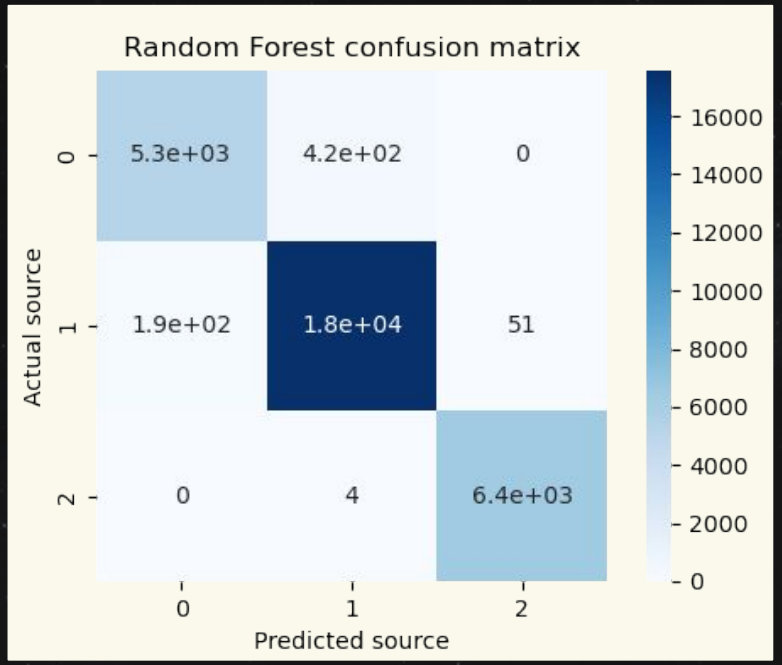
Binary



57 sources misclassified!

Recall = >0.9997

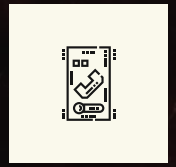
MultiClass



665 sources misclassified!

Recall = 0.9708

Looking Ahead

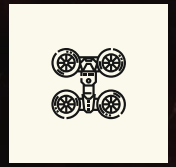


Train and Tune MultiClass Models

Inspect MisClassifications

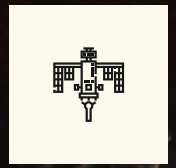
Try Stacking or Iterative Classifiers

Summary



Mapping the Cosmos is Important

Don’t know the Full Impact or All Applications Yet



Random Forest was Best Performer

ROC Score of 0.9983 vs EDA Logistic Regression Score 0.77

Model Can be Improved

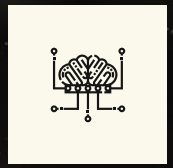
Tune for Multiclass and try stacking or voting classifiers

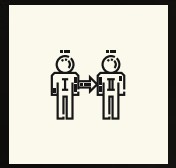
References

- Sloan Digit Sky Survey https://www.sdss.org/dr17/
- Identifying galaxies, quasars, and stars with machine learning: A new catalogue of classifications for 111 million SDSS sources without spectra A. O. Clarke, A. M. M. Scaife, R. Greenhalgh and V. Griguta A&A, 639 (2020) A84 DOI: https://doi.org/10.1051/0004-6361/201936770

Thanks!

Any questions?

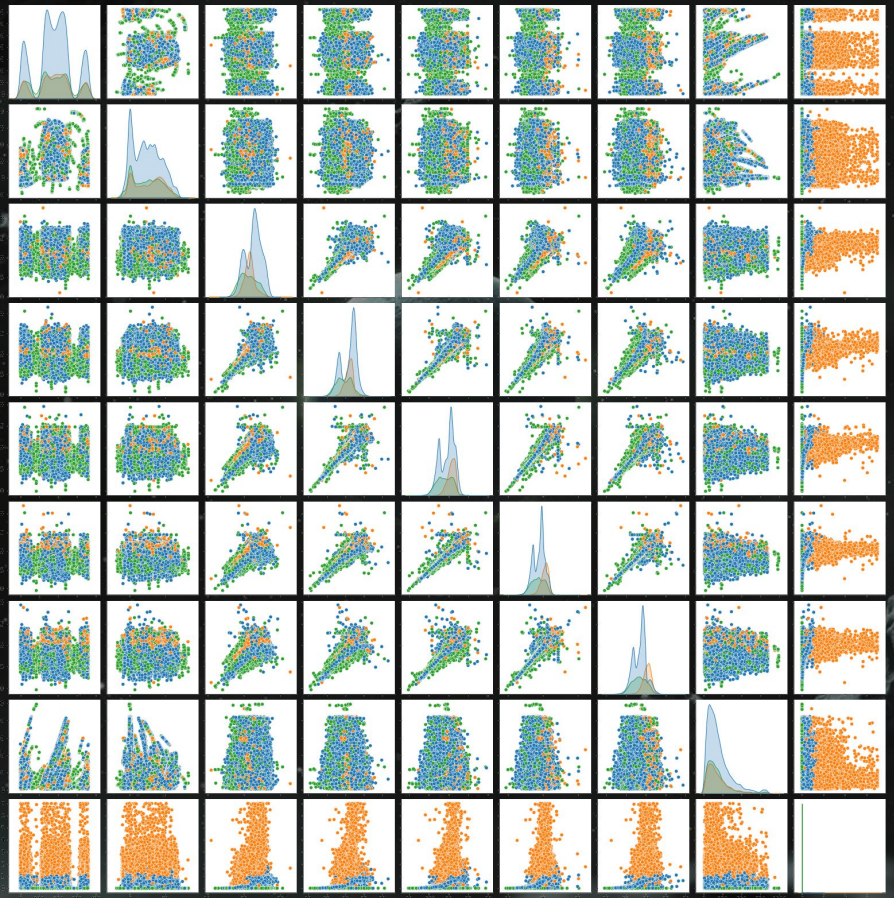




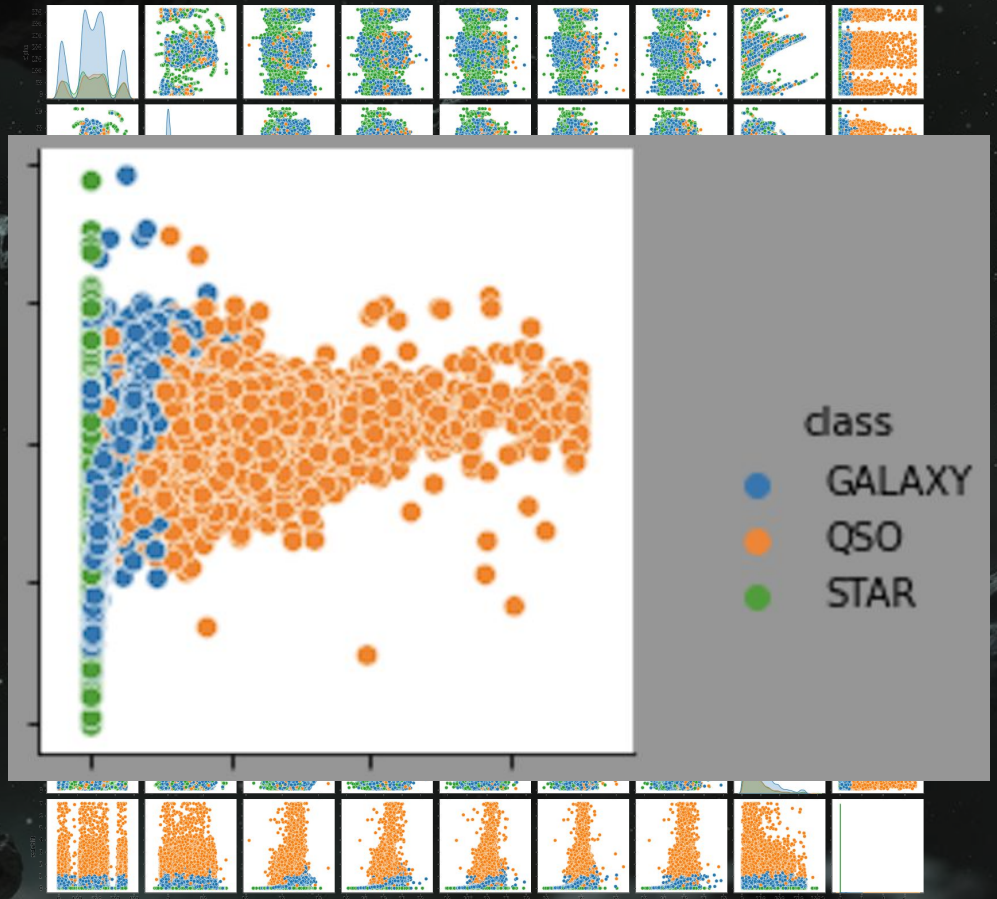
CREDITS: This presentation template was created by **Slidesgo,** including icons by **Flaticon,** and infographics & images by **Freepik**

Final Random Forest Parameters (Binary Class Only)

- max_features': 'sqrt'
- min_samples_leaf': 4
- n_estimators': 944,
- 'scaler': MinMaxScaler()



Baseline Pair Plot



Redshift