Mapping the Universe Through Data:

Stellar Mass Prediction with COSMOS-Web

DATA 495 Data Science Capstone
James McIntyre
Professor: John Cook

Project Description:

- Predict galaxy stellar mass
- Use COSMOS-Web catalog
- Inputs: multi-band photometry and environment
- Sample: 25k galaxies, 312 features
- GOAL: Link light to galaxy growth

JWST's Pillars of Creation: A high-resolution infrared view of star-forming regions within the Eagle Nebula that reveals newborn stars hidden within cosmic dust.

COSMOS-Web (COSMOS2025)

Catalogue Overview

- Released by the Institut d’Astrophysique de Paris in 2025
- Combines JWST, Hubble, Subaru, and VISTA observations
- Over 784,000 galaxies observed from 2004–2024
- Each record includes photometric data, redshift, and derived physical parameters
- Distributed in FITS format (Flexible Image Transport System) for astronomical data

A window into the early universe: JWST’s SMACS 0723 Deep Field captures light from galaxies that formed shortly after the Big Bang.

Working Dataset for Analysis

- 25,000 galaxies selected for manageability and processing performance
- 312 variables covering photometry, morphology, and model-derived parameters
- Merged the photometry and LePhare parameter tables
- Converted from FITS to a Pandas DataFrame using Astropy
- Data includes flux, magnitude, signal-to-noise, and derived values like stellar mass and star formation rate

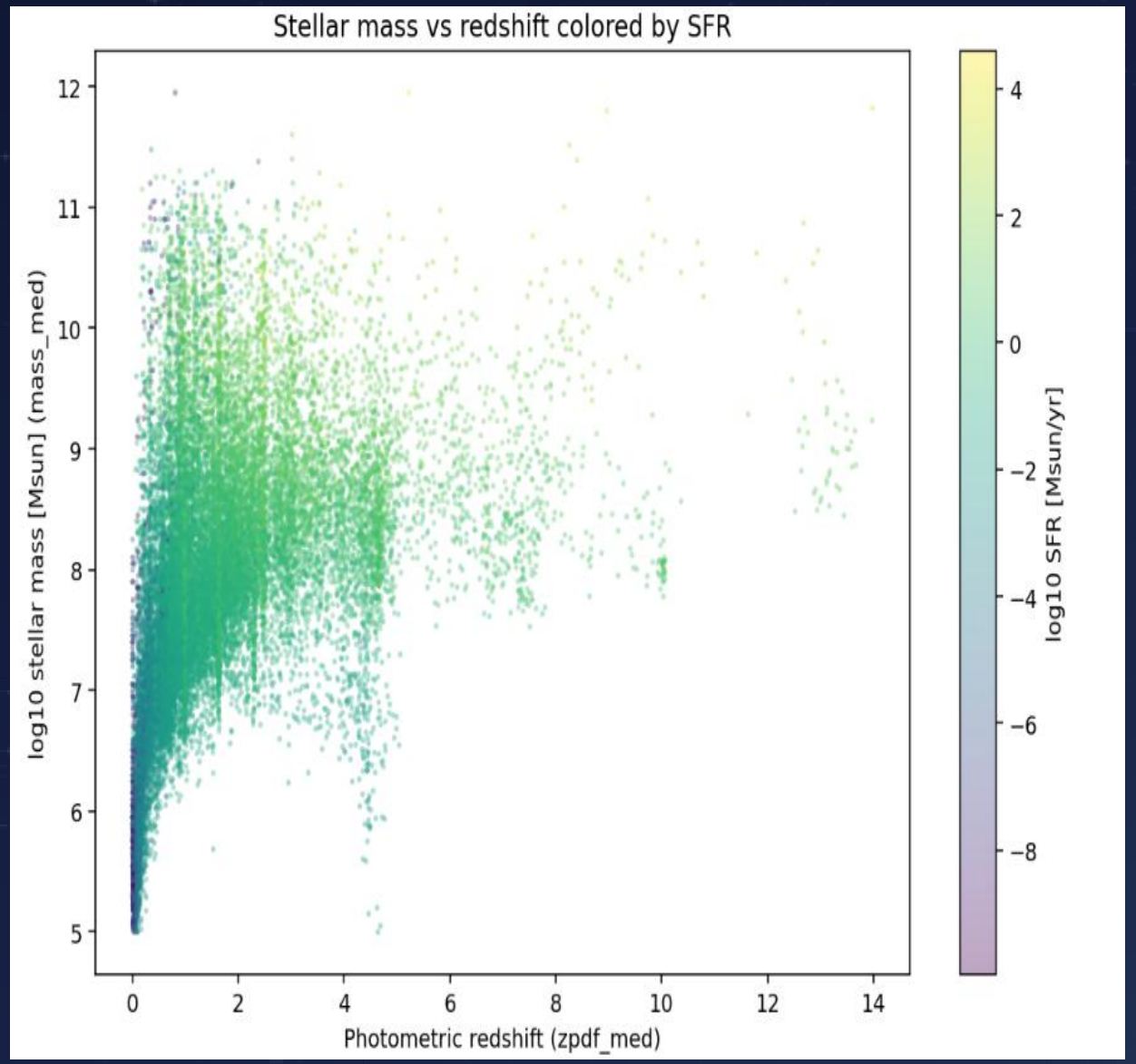
Feature	Description
Rows	25,000 Galaxies
Columns	312 Variables
Format	FITS → Pandas DataFrame
Years Covered	2004--2024
Key Variable	Steller Mass

Data Description

- Subset of COSMOS-Web: 25,000 galaxies, 312 fields pulled from JWST photometry and LePhare outputs
- Most columns are numeric brightness measures by filter, their errors, and simple shape/size stats
- A small set are derived properties we care about, like stellar mass and photo-z
- Bright sources show high flux and better signal to noise; faint ones sit in the long tail
- Coverage is strongest in NIRCam bands; F770W is shallower and has more blanks

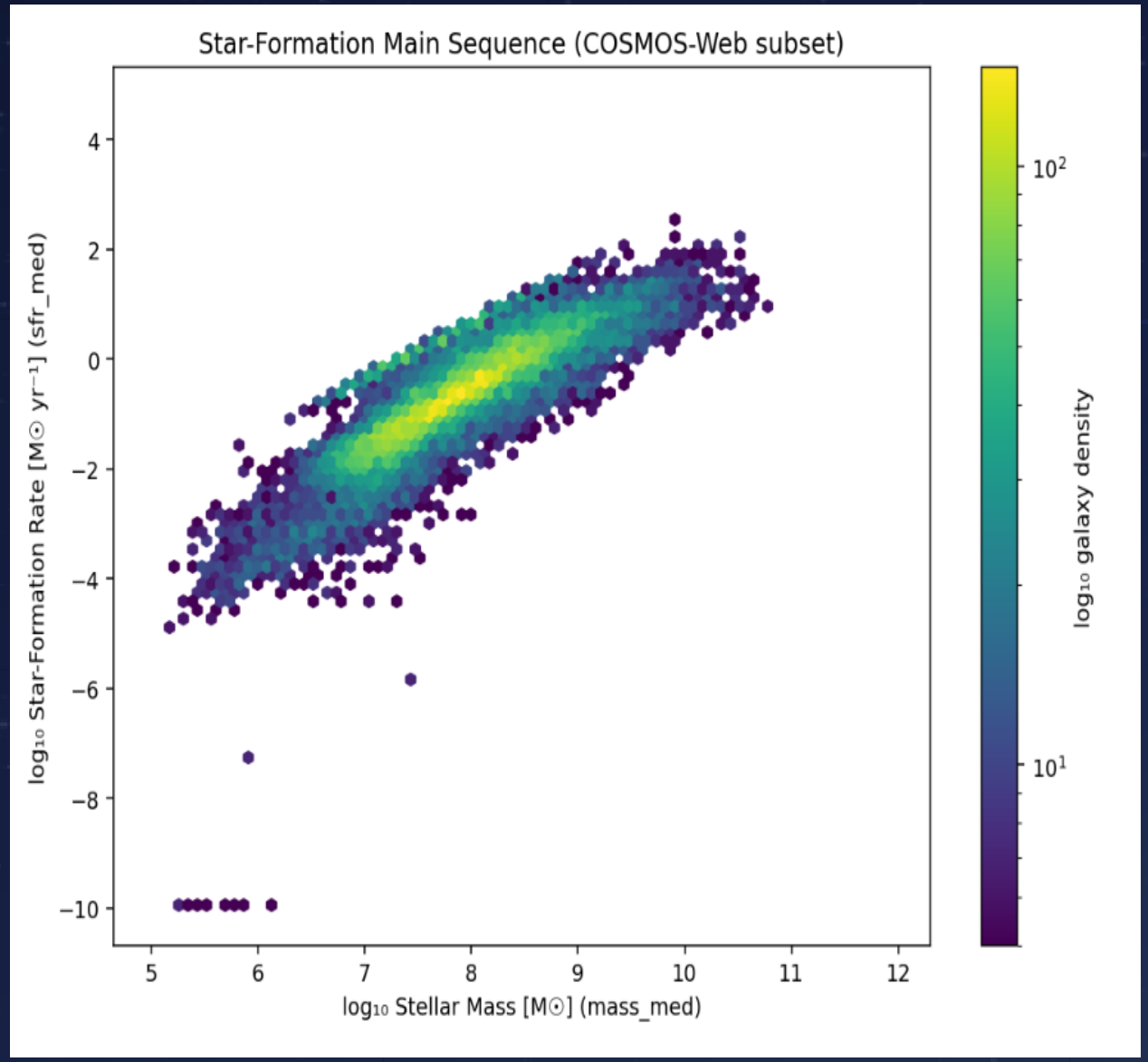
Stellar Mass vs Redshift (colored by Star Formation Rate)

- Shows how stellar mass changes across redshift in the COSMOS-Web subset
- Most galaxies sit at low redshift with a wide spread in mass
- Higher-redshift objects are fewer and show more scatter
- Color scale highlights patterns in star-formation activity
- Confirms a noisy relationship that needs modeling, not simple correlation



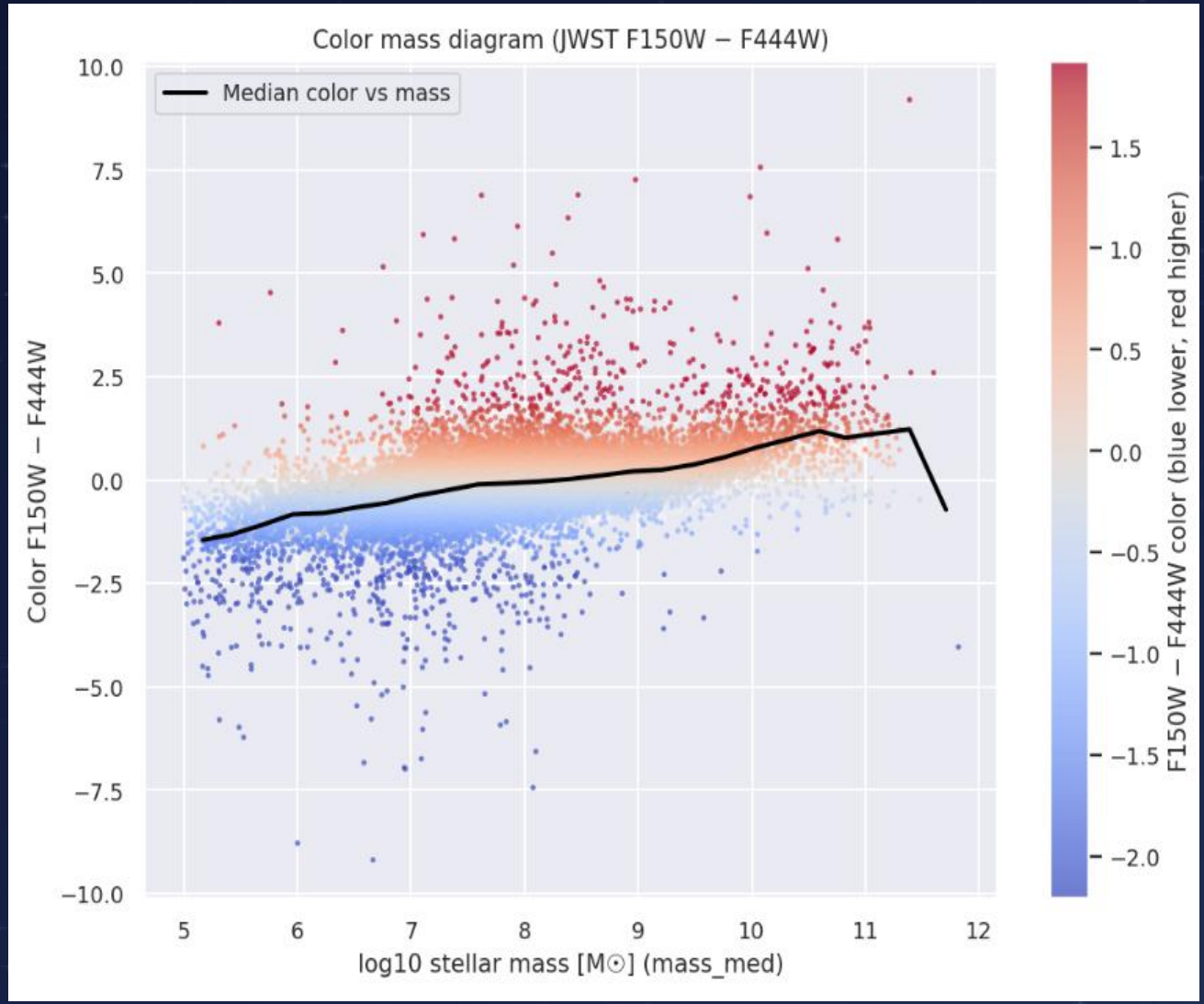
Star-Formation Main Sequence

- Shows how star-formation rate scales with stellar mass in the dataset
- Clear rising sequence: more massive galaxies form stars at higher rates
- Hexbin density highlights the high-population ridge line
- Outliers reveal non–star-forming or dust-obscured galaxies
- Supports idea that mass can be modeled from photometry + environment



Color–Mass Diagram (F150W – F444W)

- Connects JWST photometry to galaxy stellar mass
- Clear trend: redder colors align with higher-mass systems
- Shows separation between younger blue galaxies and older red ones
- Supports why photometric bands help predict stellar mass
- Good sanity check that the dataset behaves as expected physically



WEBB SPACE TELESCOPE

Modeling Approach

Data

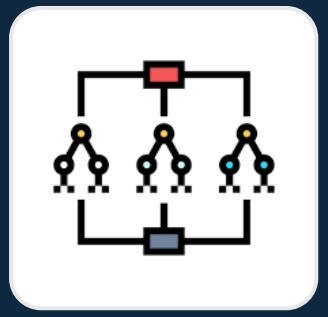
Train

Evaluate

Predict

Ridge Regression

- Fast baseline model
- Handles many features well
- Good for overall trend detection



Random Forest

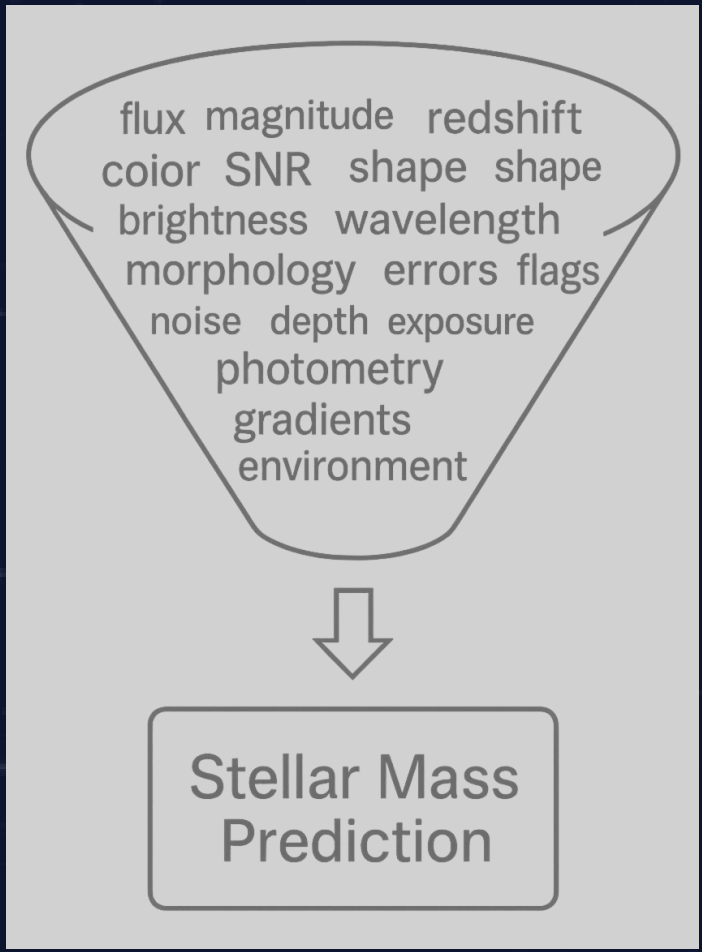
- Captures nonlinear patterns
- Works well with mixed feature types
- Reduces noise by averaging many trees

Histogram Gradient Boosting

- Learns from mistakes step-by-step
- Strong performance on complex data
- Usually the most accurate of the three

WEBB SPACE TELESCOPE

Ridge Regression



Pros

- Fast and Stable
- Handles Correlated Inputs
- Good for Broad Trends

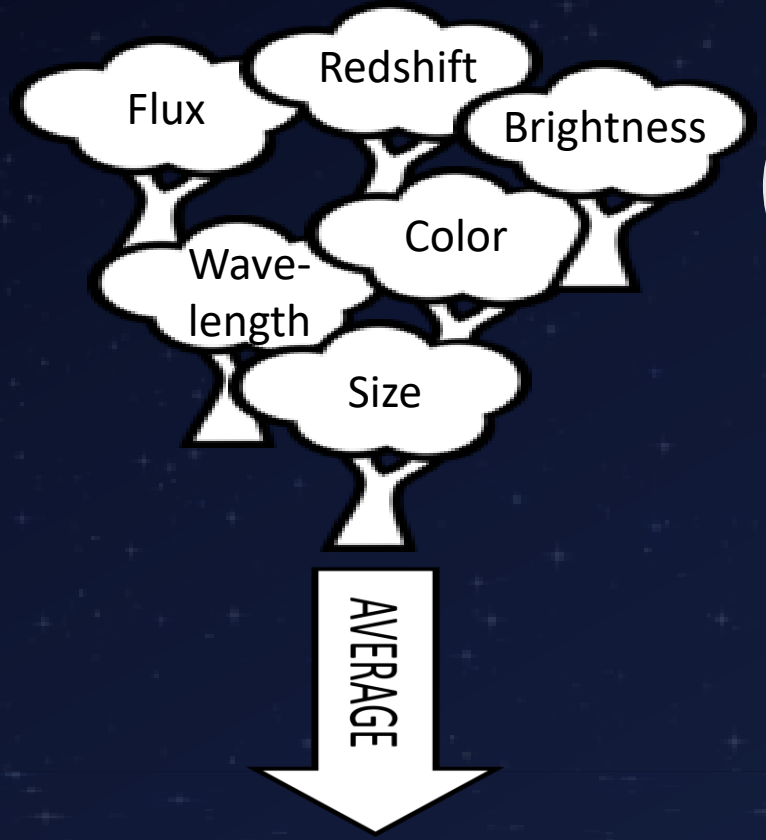
Cons

- Misses Non-Linear Patterns
- Can Underfit Complex Datasets
- Less Accurate than Tree Models

Metric	Value
R2	0.9458
MAE	0.1262
RMSE	0.2602

WEBB SPACE TELESCOPE

Random Forest



Pros

- Captures Nonlinear Patterns
- Handles Mixed Feature Types
- Reduces Noise

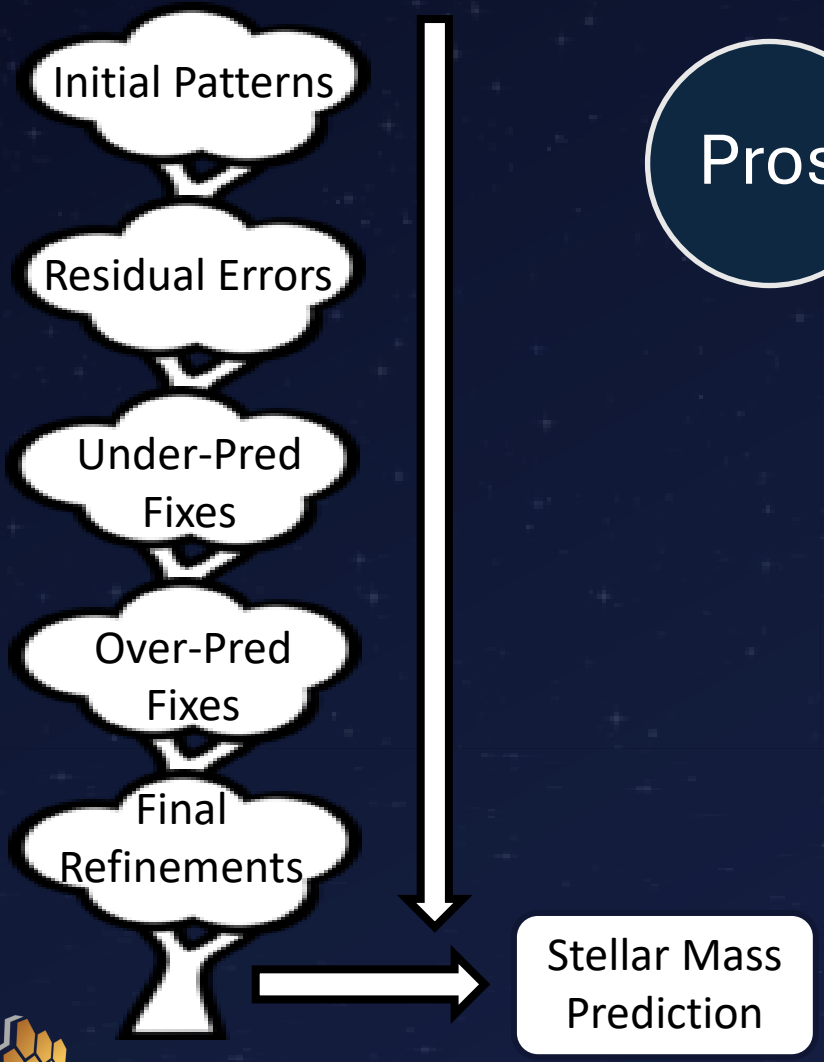
Cons

- Can Overfit w/o Tuning
- Harder to Interpret
- Slower to Train on Larger Datasets

Metric	Value
R2	0.9892
MAE	0.0769
RMSE	0.166

WEBB SPACE TELESCOPE

Histogram Gradient Boosting (HGBO)



Pros

- Learns from Errors
- Strong Accuracy
- Fast with Binning

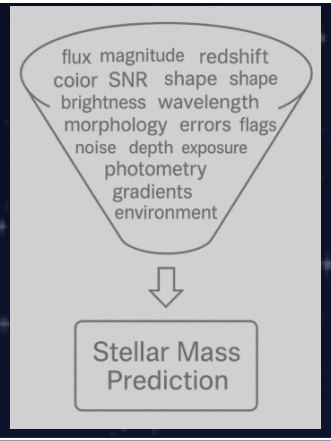
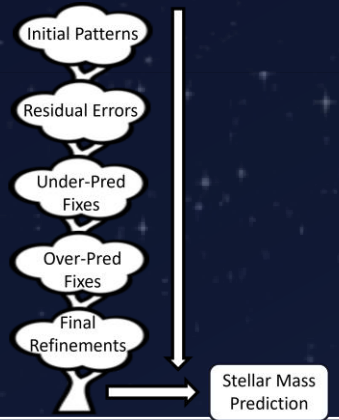
Cons

- Can Overfit w/o Tuning
- Needs Tuning
- Harder to Interpret

Metric	Value
R2	0.9895
MAE	0.0786
RMSE	0.1143

WEBB SPACE TELESCOPE

Model Comparison

Ridge Regression	Random Forest	Histogram Gradient Boosting
		
R2: 0.9458→Ranked 3rd	R2: 0.9892→Ranked 2nd	R2: 0.9895→Ranked 1st
Strong baseline model that captures overall trends.	Handles nonlinear relationships and mixed features well.	Most accurate model, excelling on complex patterns.

WEBB SPACE TELESCOPE

Key Findings

- Stellar mass increases with redshift but shows large scatter.
- More massive galaxies follow a strong star-formation sequence.
- Color differences (F150W–F444W) separate young blue galaxies from older red ones.
- Photometry alone provides strong predictive power for stellar mass.
- Histogram Gradient Boosting achieved the highest accuracy overall.
- Dataset patterns align with established galaxy evolution trends.

Recommended Next Steps

- Use Histogram Gradient Boosting as the primary model for predicting stellar mass from photometric features.
- Expand the feature set to include additional JWST bands or environment metrics to improve accuracy.
- Train models on a larger portion of the COSMOS-Web survey to validate scalability and robustness.
- Explore model explainability tools to identify which photometric bands contribute most to stellar-mass predictions.

Lessons Learned

- Data cleaning and feature preparation had a strong impact on model performance.
- Photometric color indices (like F150W–F444W) reveal key physical trends that guide modeling.
- Gradient-based models can outperform simpler methods even with noisy astrophysical data.
- Visualizations were essential for understanding galaxy behavior before modeling.
- Small differences in R^2 values can still meaningfully separate model performance.

References

- Shuntov, M., Ilbert, O., Mehta, V., Kartaltepe, J. S., Finkelstein, S. L., Weaver, J. R., Koekemoer, A. M., & COSMOS-Web Team. 2025. COSMOS2025 catalog. ApJS. https://arxiv.org/abs/2506.03243
- Hausen, R., & Robertson, B. E. 2020. Morpheus: pixel-level galaxy morphology. ApJS 248, 20. https://doi.org/10.3847/1538-4365/ab8862
- Peng, Y., Lilly, S. J., Kovač, K., Bolzonella, M., Pozzetti, L., Renzini, A., Zamorani, G., … Zucca, E. 2010. Mass and environment drive galaxy evolution. ApJ 721, 193–221. https://doi.org/10.1088/0004-637X/721/1/193