Galaxy Distance Estimation Using ML

Supervised Learning Approach

Samuel Ghalayini

Table of contents

01 Introduction

02 Scope

03 Data

04 Models

Introduction

- Motivation & Background
  - Interest sparked by data-driven astrophysics
  - Predict galaxy distances from SDSS photometry
  - Extend the cosmic distance ladder

Project Scope

- Spectroscopic redshift & Hubble's Law
  - Redshift z from spectral line shifts
  - Doppler Effect
  - Velocity v ≈ c × z
  - Distance $D = v / H_0$
  - Train ML models to predict z

Data Source

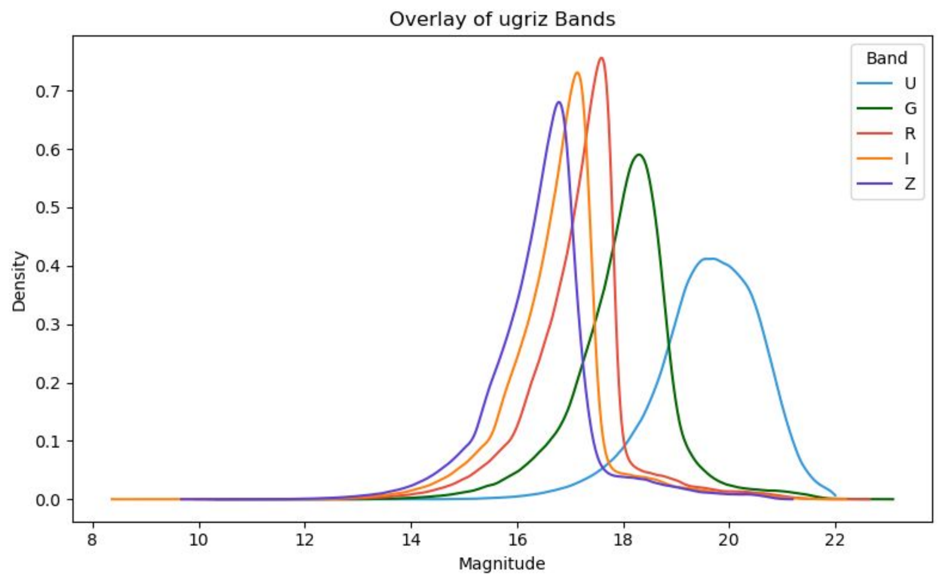
- Sloan Digital Sky Survey 18 via astroquery.sdss
  - Up to 500k rows per redshift bin
  - Stacked into ~2M galaxy entries
  - ugriz magnitudes, uncertainties, redshift

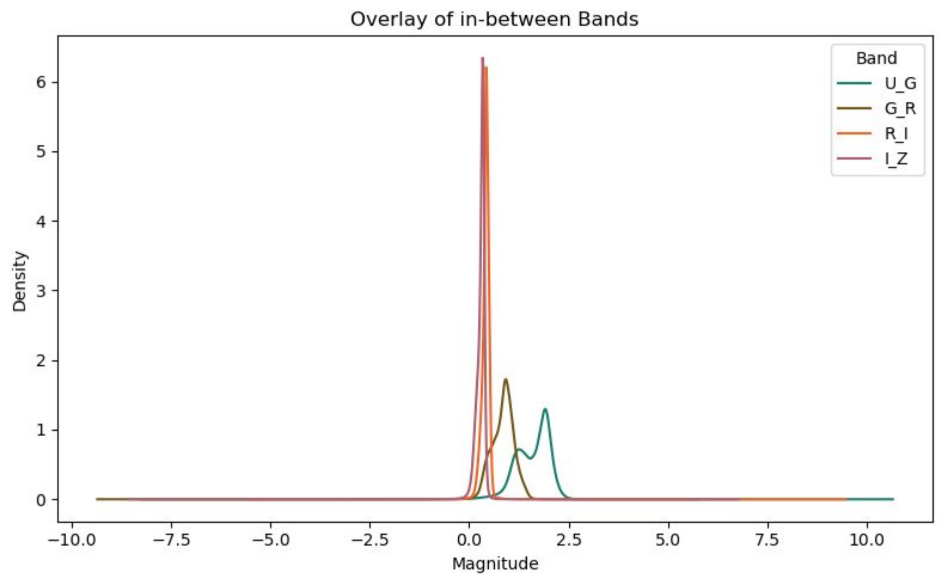
Data Cleaning

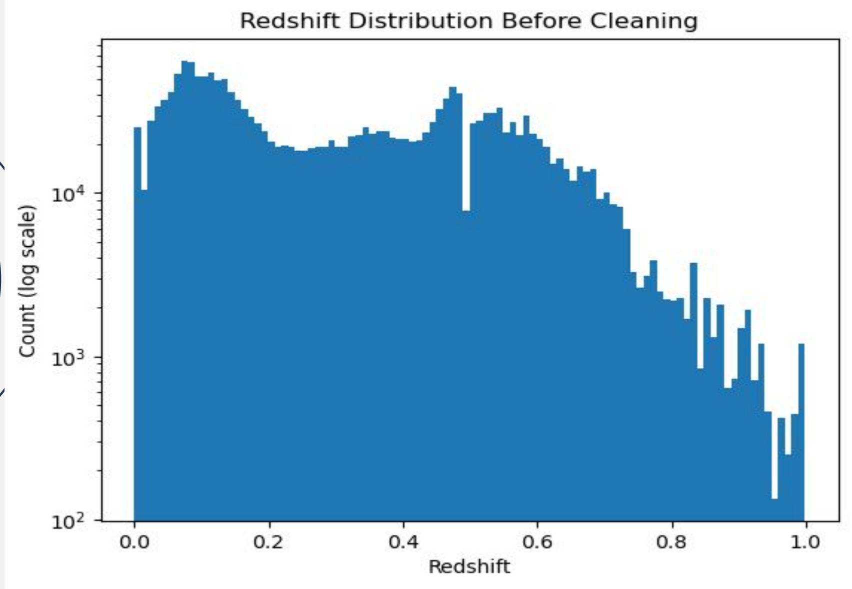
- Key steps
  - Replace -9999 with NaN and drop
  - Exclude $z < 0.01$ or $z > 1.0$
  - Require err < 0.2 mag (S/N > 5)
  - Remove fainter than 5σ depth
  - Create u−g, g−r, r−i, i−z indices

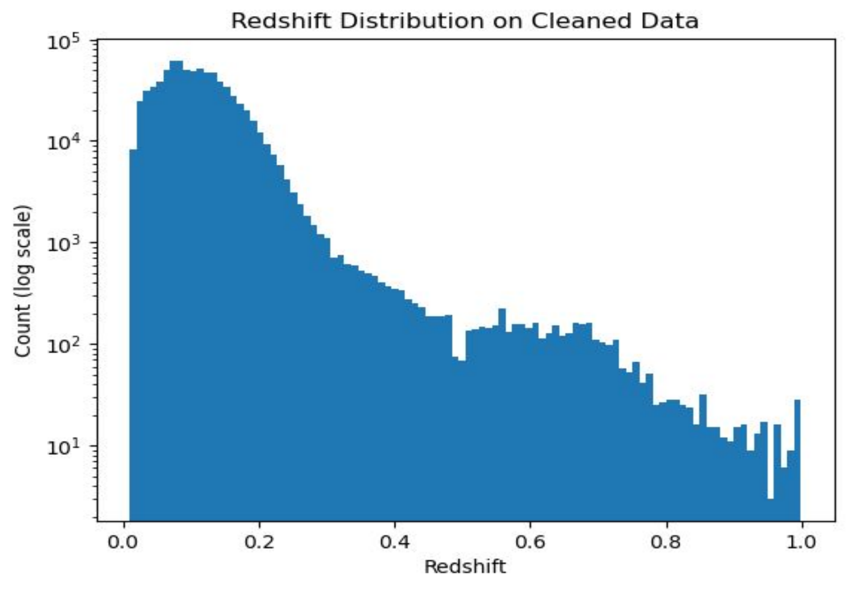
Exploratory Data Analysis

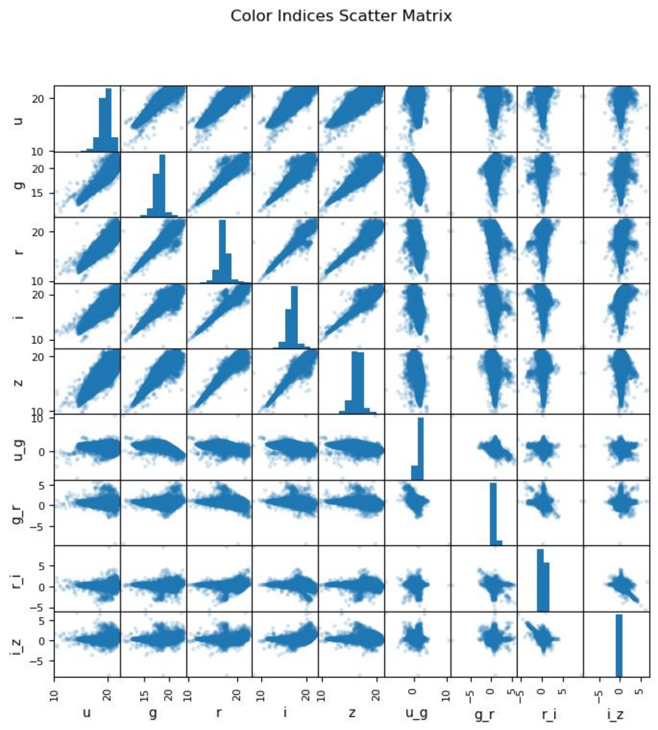
- Highlights
  - ugriz KDEs show spectral sequence
  - Pairplot: strong band correlations
  - Color–color clouds reveal variance
  - Cleaned redshifts


Overlay of ugriz Bands
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
8
10
12
14
16
18
20
22
Magnitude
Band
U
G
R
I
Z

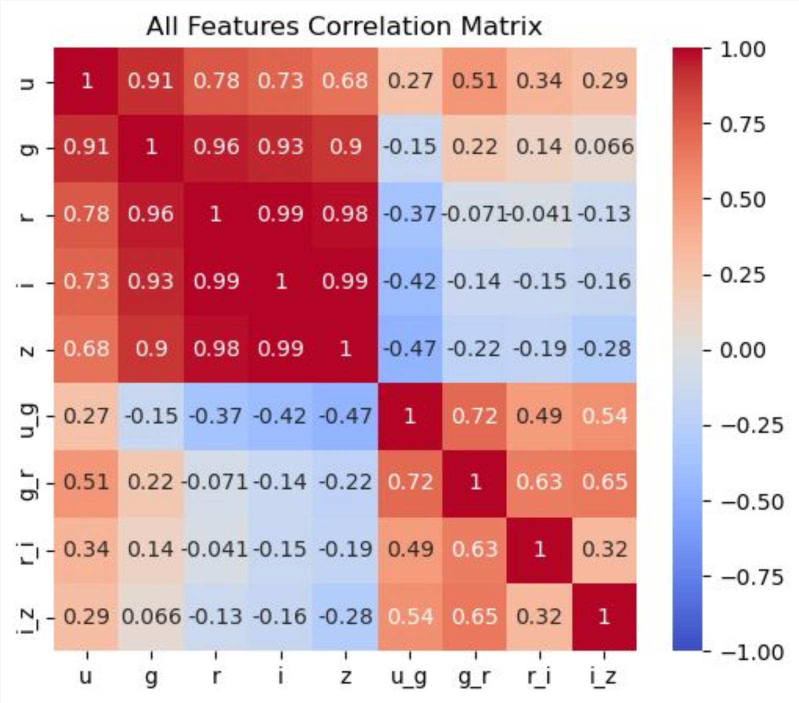

Overlay of in-between Bands
Density
0
1
2
3
4
5
6
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
Magnitude
Band
U_G
G_R
R_I
I_Z


Redshift Distribution Before Cleaning
Count (log scale)
10^4
10^3
10^2
0.0
0.2
0.4
0.6
0.8
1.0
Redshift


Redshift Distribution on Cleaned Data
Count (log scale)
10^5
10^4
10^3
10^2
10^1
0.0
0.2
0.4
0.6
0.8
1.0
Redshift



Color Indices Scatter Matrix



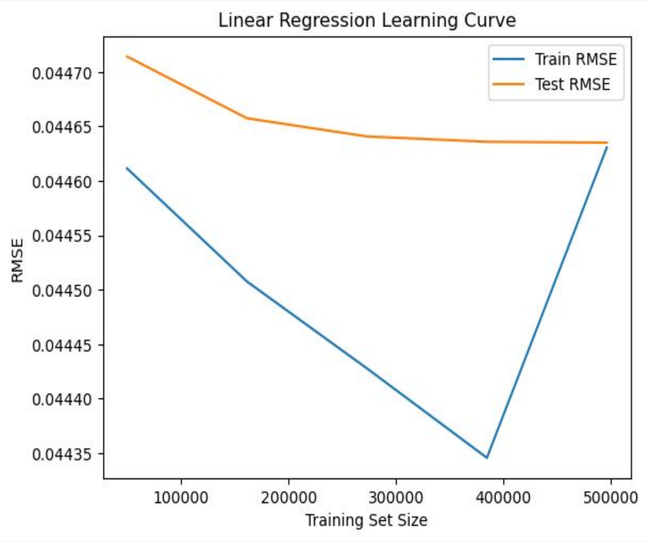
All Features Correlation Matrix

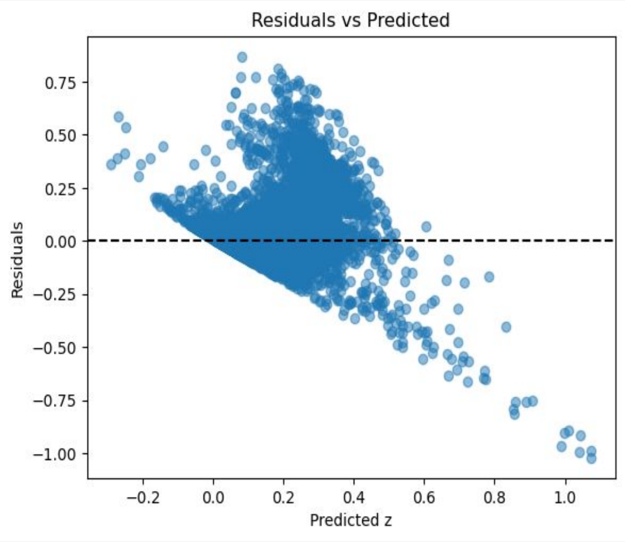
	u	g	r	i	z	u_g	g_r	r_i	i_z
u	1	0.91	0.78	0.73	0.68	0.27	0.51	0.34	0.29
g	0.91	1	0.96	0.93	0.9	-0.15	0.22	0.14	0.066
r	0.78	0.96	1	0.99	0.98	-0.37	-0.071	-0.041	-0.13
i	0.73	0.93	0.99	1	0.99	-0.42	-0.14	-0.15	-0.16
z	0.68	0.9	0.98	0.99	1	-0.47	-0.22	-0.19	-0.28
u_g	0.27	-0.15	-0.37	-0.42	-0.47	1	0.72	0.49	0.54
g_r	0.51	0.22	-0.071	-0.14	-0.22	0.72	1	0.63	0.65
r_i	0.34	0.14	-0.041	-0.15	-0.19	0.49	0.63	1	0.32
i_z	0.29	0.066	-0.13	-0.16	-0.28	0.54	0.65	0.32	1

Models

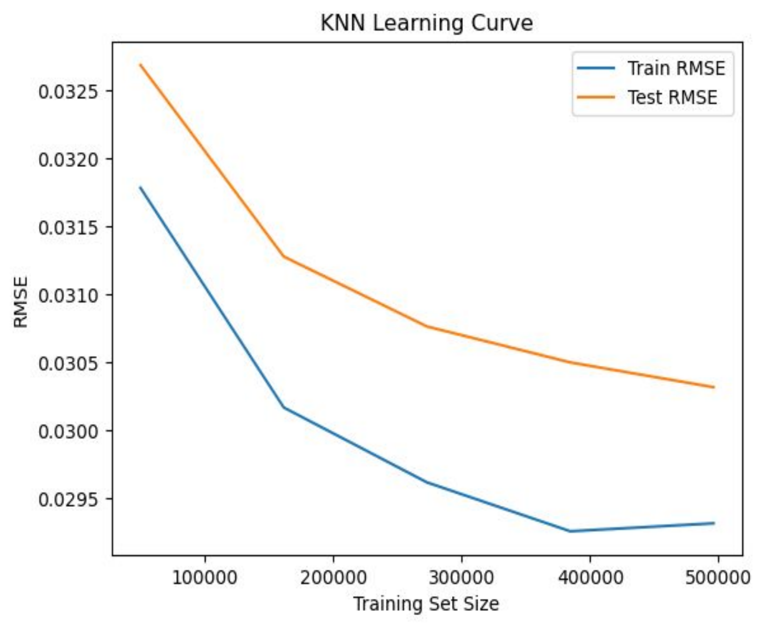
- Algorithms
  - Linear Regression (feature selection)
  - KNN (k=30)
  - Random Forest (RandomizedSearchCV)
  - XGBoost (Boosting)

Linear Regression

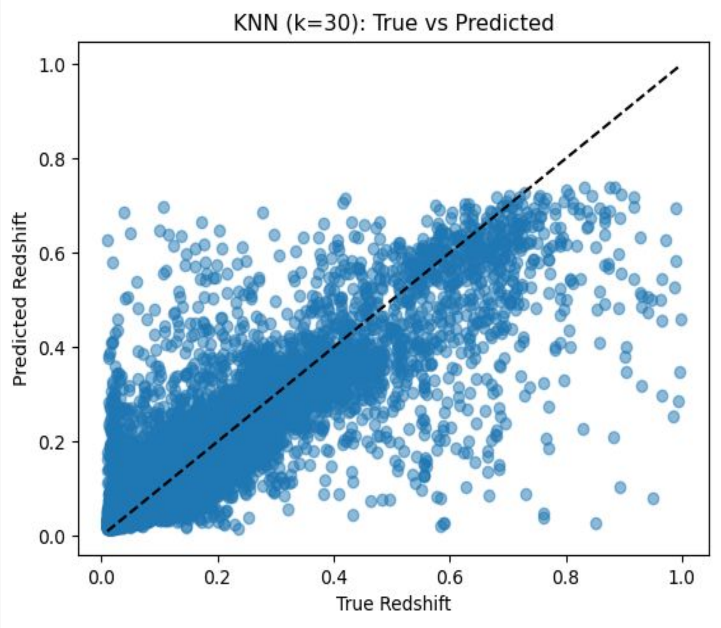




K-Nearest Neighbors



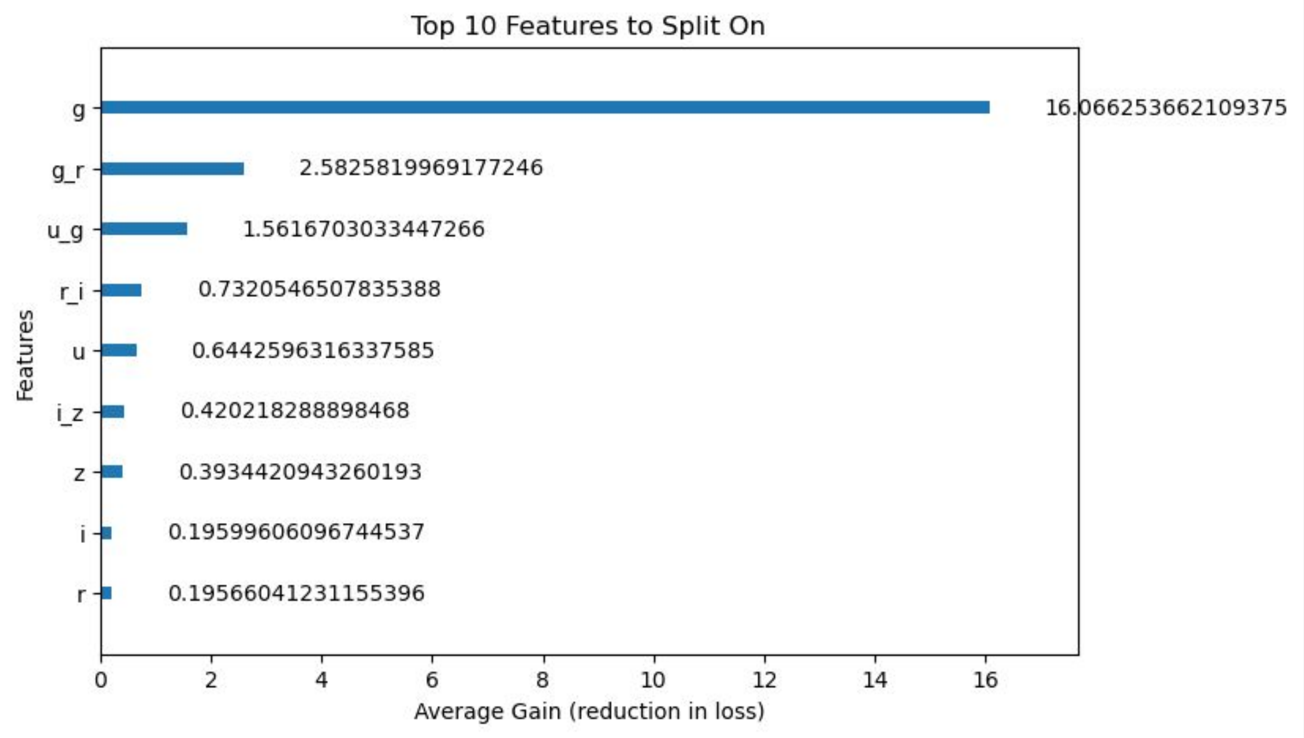
KNN Learning Curve



KNN (k=30): True vs Predicted

Random Forest and Notebook

XGBoost


Top 10 Features to Split On
g
g_r
u_g
r_i
u
i_z
z
i
r
16.066253662109375
2.5825819969177246
1.5616703033447266
0.7320546507835388
0.6442596316337585
0.420218288898468
0.3934420943260193
0.19599606096744537
0.19566041231155396
Features
0
2
4
6
8
10
12
14
16
Average Gain (reduction in loss)

Results & Performance

Model	MAE	RMSE	R^2	Time (s)
Linear Regression	0.0263	0.0444	0.6079	0.0283
KNN (k=24)	0.0180	0.0311	0.8076	5.1731
Random Forest	0.0374	0.0522	0.4565	353.6183
XGBoost	0.0176	0.0307	0.8126	1.9744

Discussion & Next Steps

- Takeaways
  - Data cleaning drives performance
  - XGBoost best balance of speed & accuracy
  - Future: more bands, cloud hyperparameter tuning
- Github: https://github.com/sam-ghala/galaxy_distance