Improving Photometric Redshift Prediction with Morphological Data in a Decision Tree Framework

Darcy Corson
Tufts University
05/08/2025

PART 1: BACKGROUND & MOTIVATION

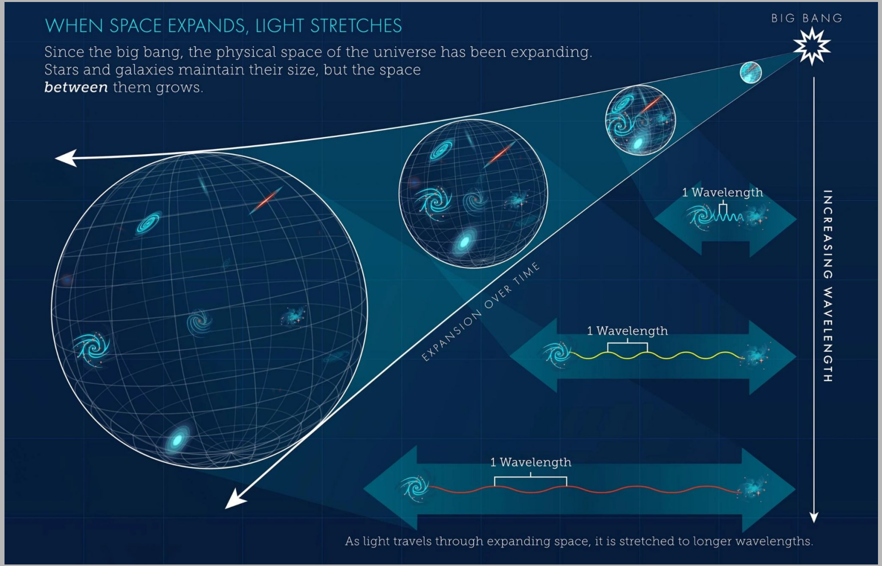
UNDERSTANDING REDSHIFT

Redshift (z) occurs when light wavelengths stretch due to cosmic expansion.

$$z = \frac{\lambda_{observed} - \lambda_{emitted}}{\lambda_{emitted}}$$

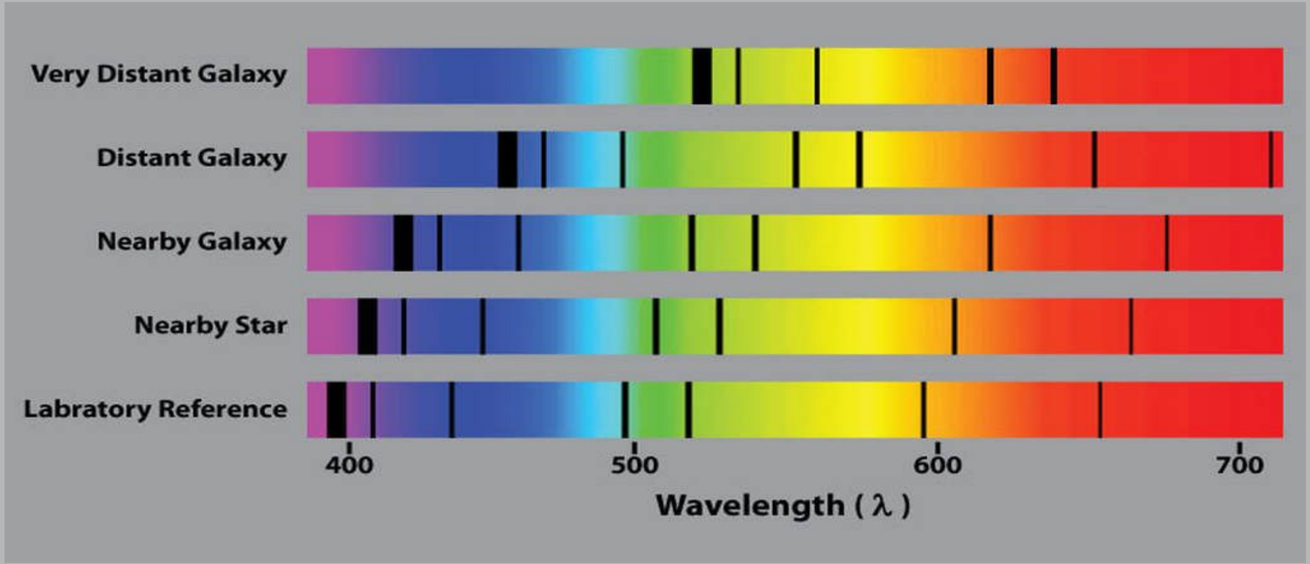
Hubble's Law connects redshift to distance.

Hubble’s Law: $v = H_0 \times d$



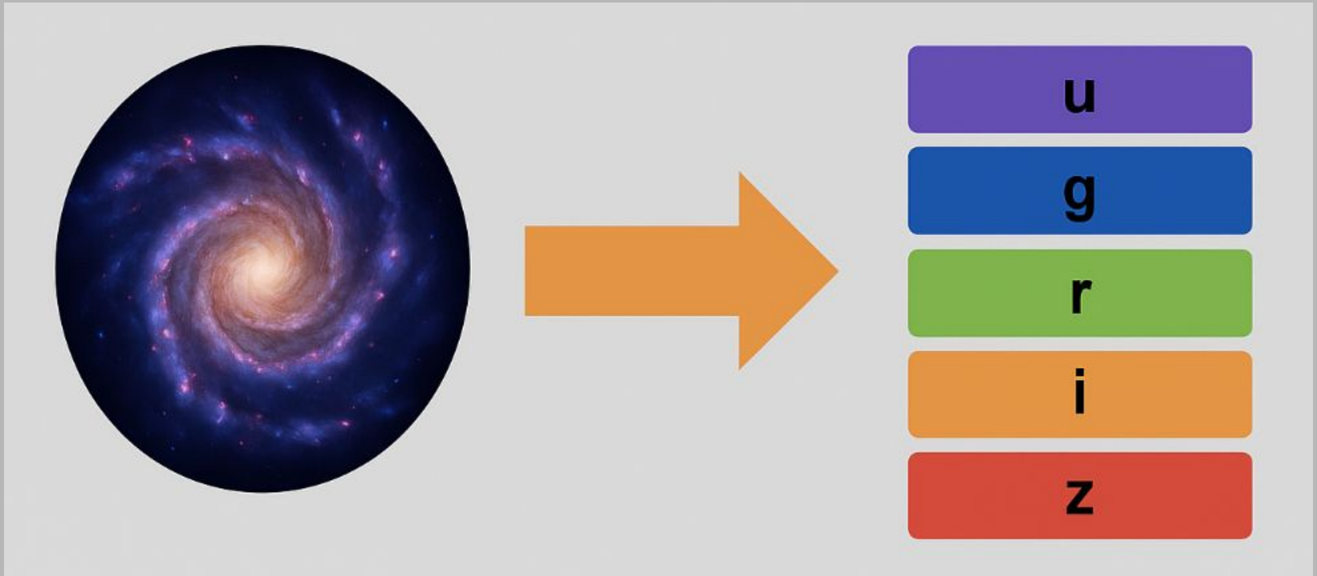
As galaxies move away from Earth, their light shifts toward the red end of the spectrum. This shift increases with distance from Earth.

SPECTROSCOPIC REDSHIFT DETERMINATION

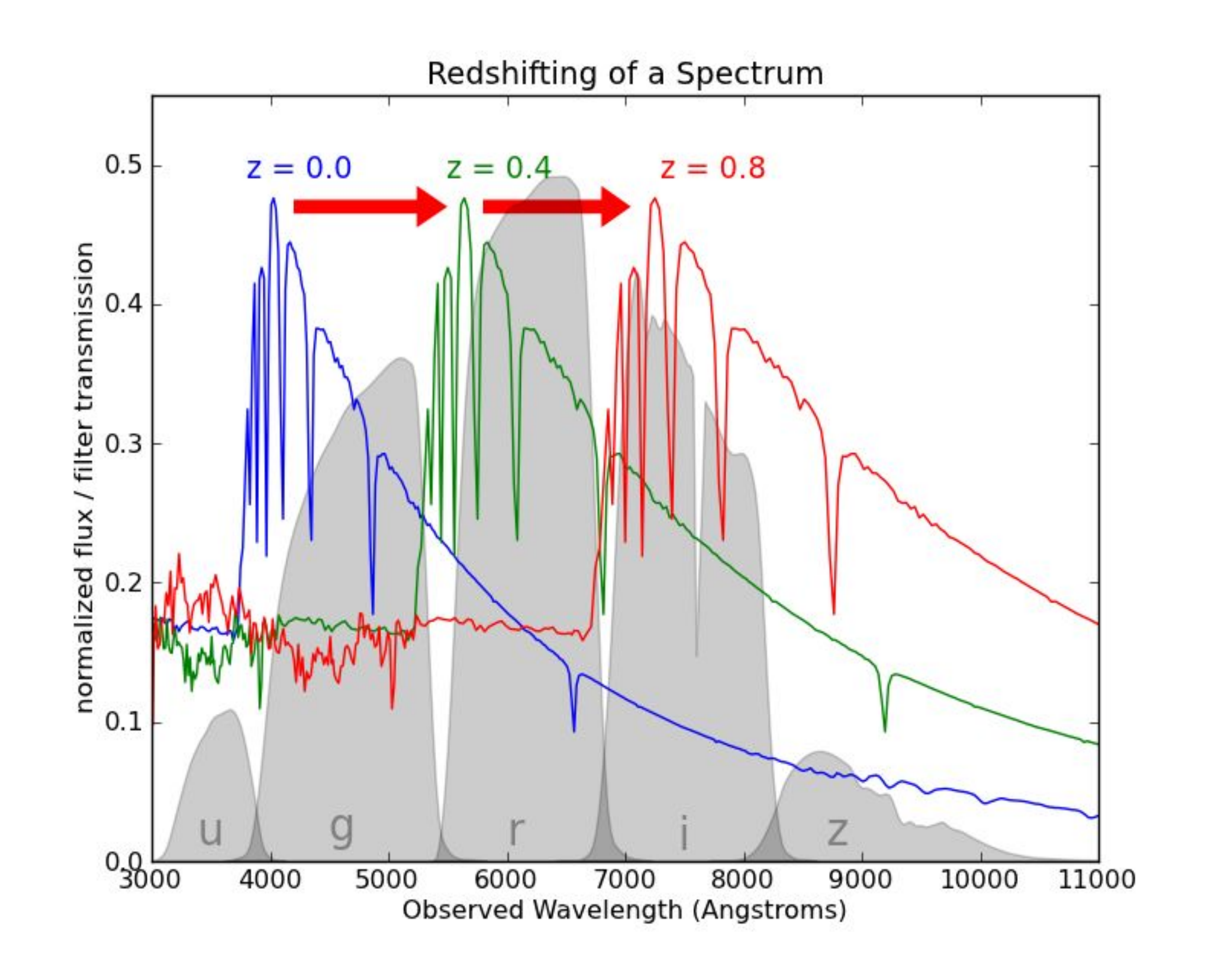


Spectroscopic redshift is the gold standard for measuring galaxy distances. The black lines show emission or absorption features in a galaxy's spectrum. For galaxies that are farther away, these features shift to longer (redder) wavelengths. This method is highly accurate, but is labor and resource intensive.

PHOTOMETRIC REDSHIFT DETERMINATION



Photometric redshift estimation measures galaxy distances using their colors across multiple wavelength bands. This technique analyzes light through filters like ugriz rather than collecting full spectra.



The spectrum shown here is that of the star Vega (α-Lyr) at three different redshifts. The SDSS ugriz filters are shown in grey for reference.

At redshift z = 0.0, the spectrum is bright in the u and g filter and dim in the i and z filters. At redshift z = 0.8, the opposite is the case. This demonstates the possibility of determining redshift from photometry alone.

SPECTROSCOPIC VS. PHOTOMETRIC REDSHIFTS

SPECTROSCOPY

Extremely precise measurements of galaxy spectra. Requires significant telescope time per galaxy

THE CHALLENGE

Impossible to obtain spectra for trillions of faint galaxies. New methods needed.

PHOTOMETRY

Estimates redshift using broad-band filters. Enables study of vast galaxy populations.

SPECTROSCOPIC VS. PHOTOMETRIC REDSHIFTS

SPECTROSCOPY
Extremely precise measurements of galaxy spectra. Requires significant telescope time per galaxy

THE CHALLENGE
Impossible to obtain spectra for trillions of faint galaxies. New methods needed.

PHOTOMETRY
Estimates redshift using broad-band filters. Enables study of vast galaxy populations.

KEY CHALLENGES IN PHOTOMETRIC ESTIMATION

BIAS & SCATTER
Measurement errors vary with redshift and galaxy properties. Some galaxy types are harder to classify accurately.

COLOR DEGENERACY
Different galaxy types at different redshifts can appear similar in color. This creates ambiguity in measurements.

PART 2: DATA & METHODS

SLOAN DIGITAL SKY SURVEY (SDSS) DATA

Final data release of SDSS-IV, encompassing extensive photometric and spectroscopic observations of galaxies.

Data Acquisition

Queried SDSS DR17 using the CasJobs API with SQL.

Joined data across three key tables: `PhotoObj` (photometric), `SpecObj` (spectroscopic), and `galSpecExtra` (derived properties).

Selected galaxies with spectroscopic redshifts (z) between 0 and 0.4.

Used random sampling across z values to generate a table of 742,042 galaxies, avoiding galaxies with null and placeholder values (e.g., −9999) for relevant features.

Features Extracted

Photometric: Color indices (g−r, u−g, r−i, i−z)

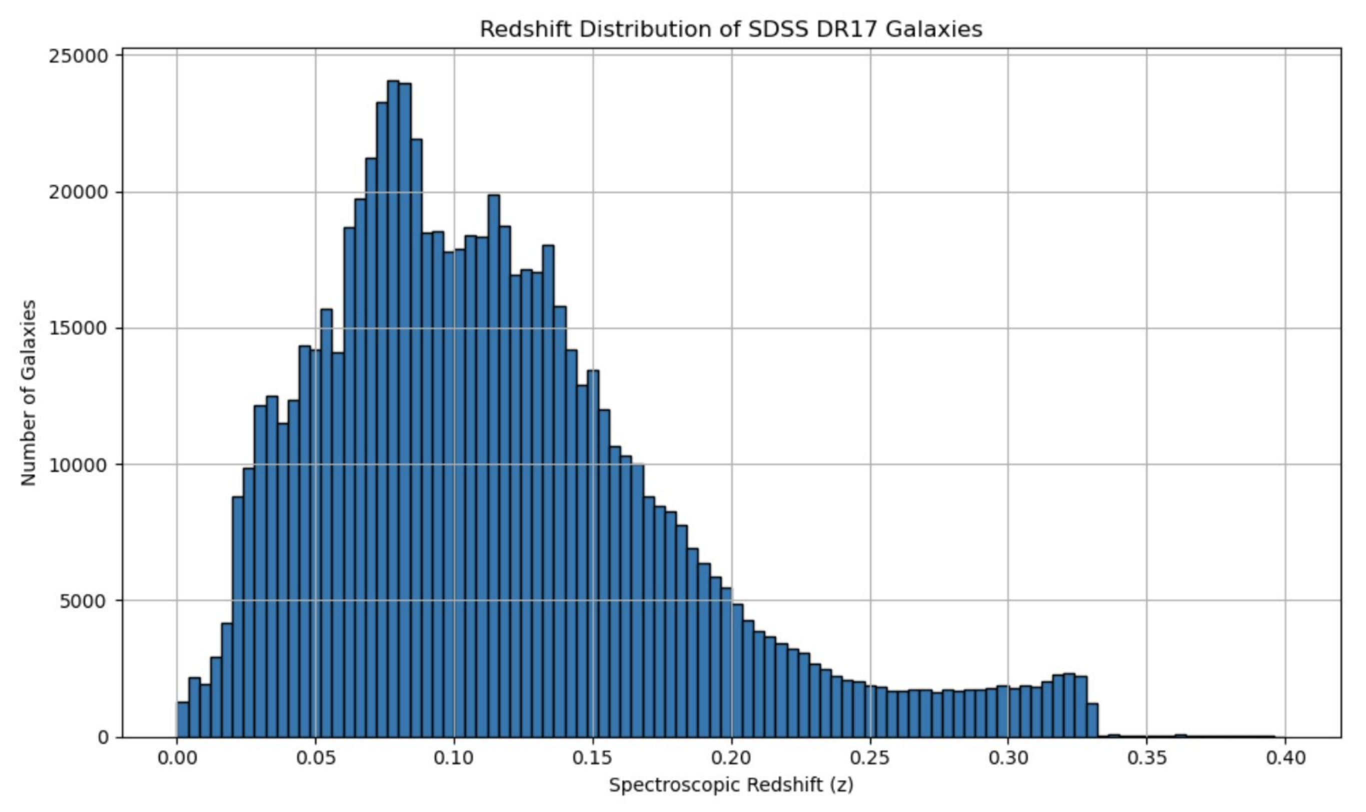
Morphological:
- Light profile shape (fracDeV_r)
- Axis ratios (expAB_r, deVAB_r)
- Stokes parameters (q_i, u_i)
- Galaxy size metrics (Petrosian radii and model radii)
- Log of star formation rate (logSFR)
- Derived petroR50_r/petroR90_r (compactness)

Data Preprocessing

Applied log-transformation to size-related features to normalize scale.

Verified redshift distribution and overall feature completeness prior to training.

DISTRIBUTION OF DATASET GALAXIES BY REDSHIFT



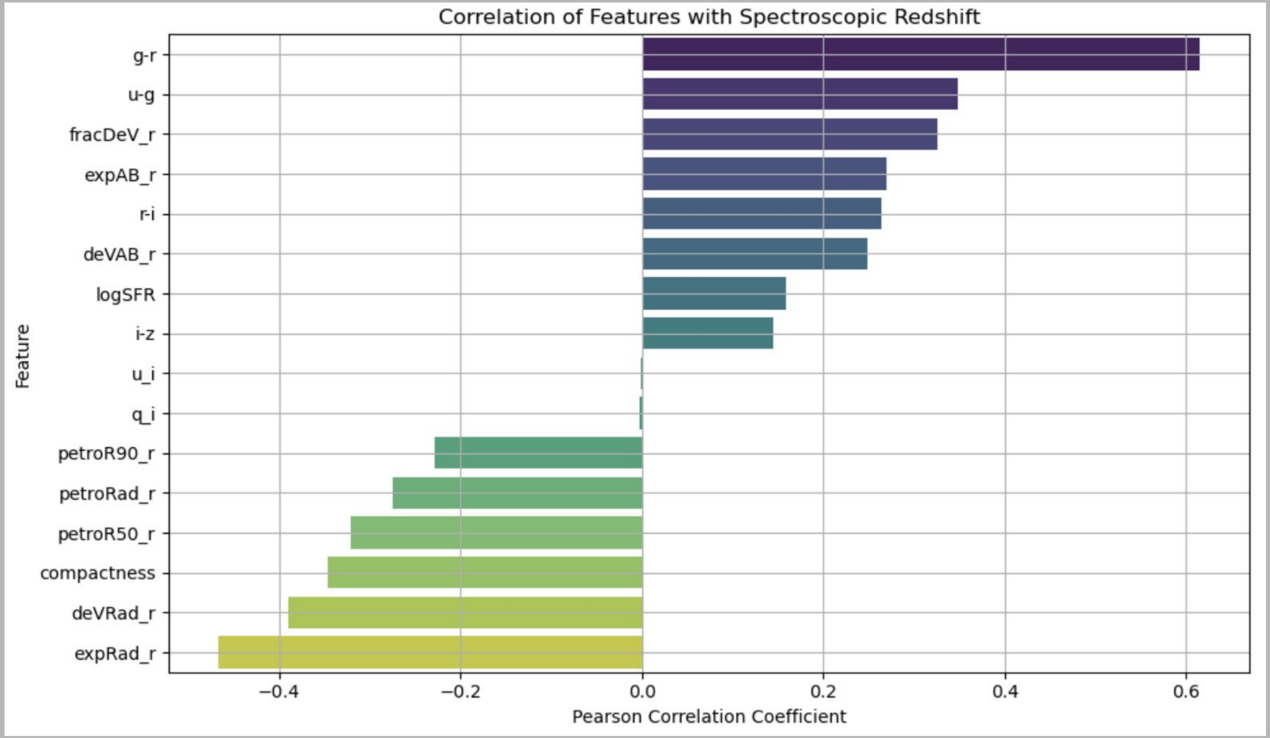
PHOTOMETRIC FEATURES

Feature	Description
u, g, r, i, z	Apparent magnitudes in ultraviolet to near-infrared filters
$g-r$, $u-g$, $r-i$, $i-z$	Color indices representing differences in brightness between bands

MORPHOLOGICAL FEATURES

Feature	Description
fracDeV_r	Fraction of light fit by a de Vaucouleurs profile (Sérsic index proxy)
expAB_r, deVAB_r	Axis ratios from exponential and de Vaucouleurs profile fits
q_i, u_i	Stokes parameters describing shape and orientation of galaxy light
petroR50_r, petroR90_r, petroRad_r	Petrosian radii for 50%, 90%, and total light (log-transformed)
deVRad_r, expRad_r	Effective radii from de Vaucouleurs and exponential models (log-transformed)
Compactness	Ratio of petroR50_r to petroR90_r, measures light concentration
logSFR	Logarithm of total star formation rate (log-transformed)

PRELIMINARY ASSESSMENT OF FEATURES' PREDICTIVE POWER



=== Correlation Table ===

Feature	Correlation with Redshift
g-r	0.615611
u-g	0.348688
fracDeV_r	0.326194
expAB_r	0.270476
r-i	0.264239
deVAB_r	0.248406
logSFR	0.158925
i-z	0.144754
u_i	-0.000818
q_i	-0.002920
petroR90_r	-0.228559
petroRad_r	-0.274119
petroR50_r	-0.320514
compactness	-0.346692
deVRad_r	-0.389776
expRad_r	-0.467408

MODEL 1: PHOTOMETRIC-ONLY DECISION TREE

FEATURE SET
Color indices derived from SDSS magnitudes (`u-g`, `g-r`, `r-i`, `i-z`)

TARGET VARIABLE
Spectroscopic redshift (`specz_redshift`), which the model is trained to predict

TRAIN/TEST SPLIT
80% training, 20% testing; fixed random seed (42)

TRAINING METHOD
Trained using `DecisionTreeRegressor` from scikit-learn

A range of tree depths (1 to 20) is is evaluated to balance model complexity and predictive performance
(Best Depth = 12)

MODEL 2: COMBINATION DECISION TREE

FEATURE SET
Color indices derived from SDSS magnitudes (`u-g`, `g-r`, `r-i`, `i-z`)

Structural properties (`compactness`, `fracDeV_r`, `deVRad_r`, `expRad_r`, `petroRad_r`, `petroR50_r`, `petroR90_r`, `expAB_r`, `deVAB_r`, `q_i`, `u_i`, `logSFR`)

TRAIN/TEST SPLIT
80% training, 20% testing; fixed random seed (42)

TARGET VARIABLE
Spectroscopic redshift (`specz_redshift`), which the model is trained to predict

TRAINING METHOD
Trained using `DecisionTreeRegressor` from scikit-learn

A range of tree depths (1 to 20) is is evaluated to balance model complexity and predictive performance
(Best Depth = 14)

MODEL 1: PHOTOMETRIC-ONLY DECISION TREE

FEATURE SET
Color indices derived from SDSS magnitudes (`u-g`, `g-r`, `r-i`, `i-z`)

TARGET VARIABLE
Spectroscopic redshift (`specz_redshift`), which the model is trained to predict

TRAIN/TEST SPLIT
80% training, 20% testing; fixed random seed (42)

TRAINING METHOD
Trained using `DecisionTreeRegressor` from scikit-learn

A range of tree depths (1 to 20) is is evaluated to balance model complexity and predictive performance
(Best Depth = 12)

MODEL 2: COMBINATION DECISION TREE

FEATURE SET
Color indices derived from SDSS magnitudes (`u-g`, `g-r`, `r-i`, `i-z`)

Structural properties (`compactness`, `fracDeV_r`, `deVRad_r`, `expRad_r`, `petroRad_r`, `petroR50_r`, `petroR90_r`, `expAB_r`, `deVAB_r`, `q_i`, `u_i`, `logSFR`)

TRAIN/TEST SPLIT
80% training, 20% testing; fixed random seed (42)

TARGET VARIABLE
Spectroscopic redshift (`specz_redshift`), which the model is trained to predict

TRAINING METHOD
Trained using `DecisionTreeRegressor` from scikit-learn

A range of tree depths (1 to 20) is is evaluated to balance model complexity and predictive performance
(Best Depth = 14)

MODEL 3: MORPHOLOGICAL-ONLY DECISION TREE

FEATURE SET
Structural properties (`compactness`, `fracDeV_r`, `deVRad_r`, `expRad_r`, `petroRad_r`, `petroR50_r`, `petroR90_r`, `expAB_r`, `deVAB_r`, `q_i`, `u_i`, `logSFR`)

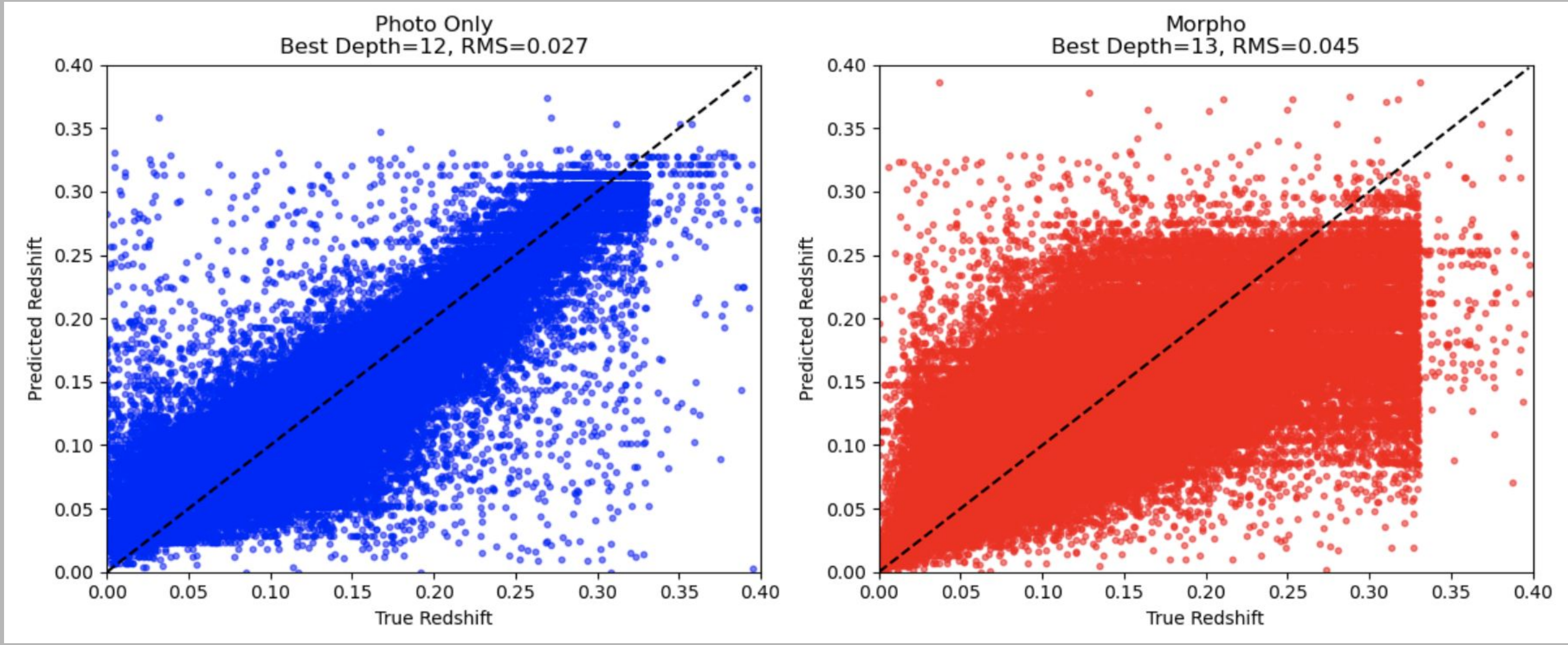
TARGET VARIABLE
Spectroscopic redshift (`specz_redshift`), which the model is trained to predict

TRAIN/TEST SPLIT
80% training, 20% testing; fixed random seed (42)

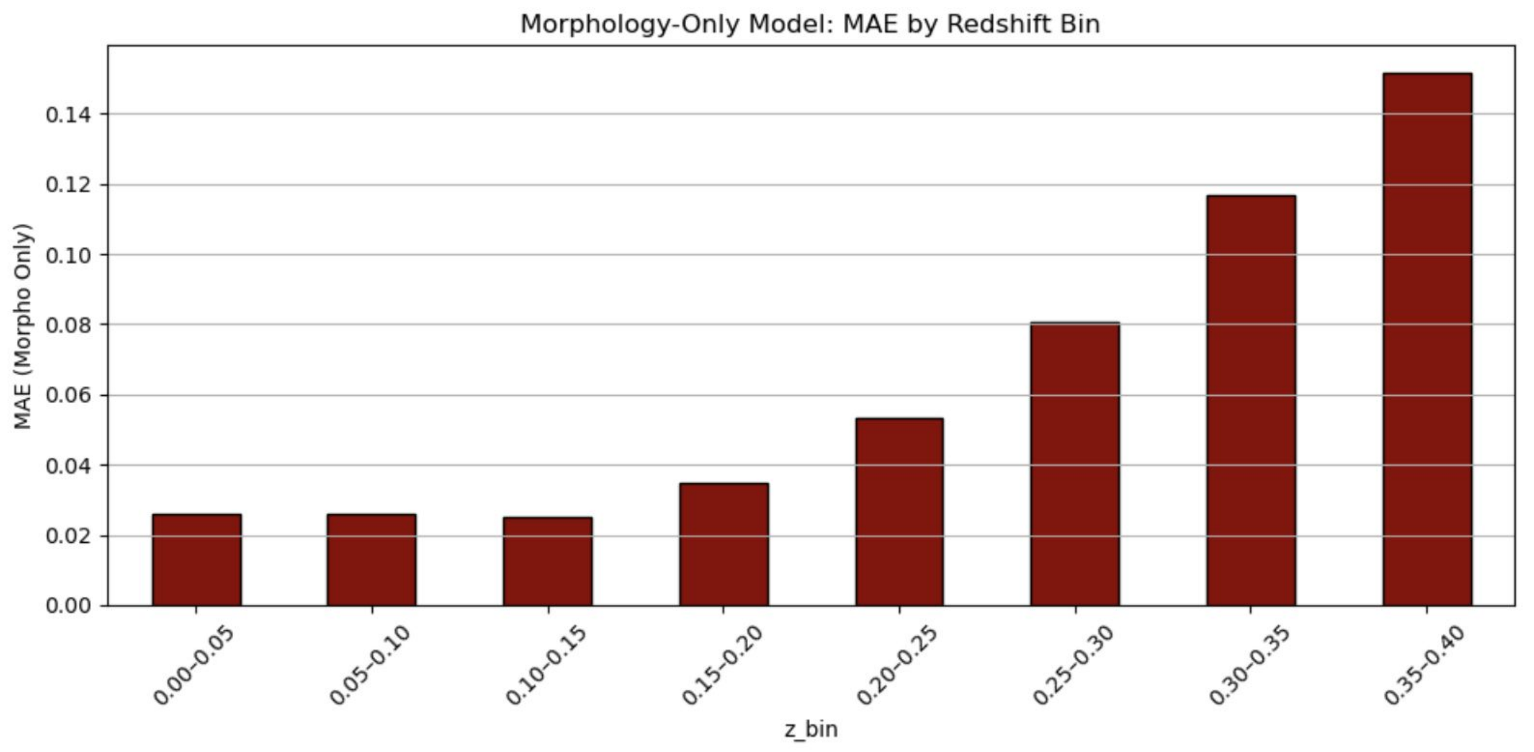
TRAINING METHOD
Trained using `DecisionTreeRegressor` from scikit-learn

A range of tree depths (1 to 20) is is evaluated to balance model complexity and predictive performance
(Best Depth = 13)

Performance Comparison (RMS): Photometric-Only vs. Morphological-Only



PERFORMANCE OF MORPHOLOGICAL-ONLY MODEL: MAE PER REDSHIFT BIN



METRICS USED TO EVALUATE MODEL PERFORMANCE

METRIC	EXPLANATION
RMS Error	Measures the average magnitude of the prediction error, giving greater weight to larger errors.
MAE	Measures the average absolute difference of the prediction error.
R^2 Score	Coefficient of determination; measures how well the model explains variance in the data. Ranges from 0 (no explanatory power) to 1 (perfect fit).
Number of Significant Errors	Counts predictions where the absolute error exceeds a meaningful threshold ($0.03 \leq \lvert\Delta z\rvert \leq 0.05$), indicating notable deviation from true redshift.
Number of Catastrophic Errors	Counts extreme mispredictions ($\lvert\Delta z\rvert \geq 0.05$), which are especially problematic for scientific inference.
Percentage of Catastrophic Errors Made Tolerable	Quantifies proportion of catastrophic errors in the photometry-only model that were reduced below the catastrophic threshold in the combined model, indicating successful correction by added features.

PART 3: HYPOTHESES & RESULTS

HYPOTHESIS 1:

An optimized decision tree model that integrates both photometric and morphological galaxy features will outperform a photometric-only model in predicting spectroscopic redshift over the redshift range (0.0, 0.4).

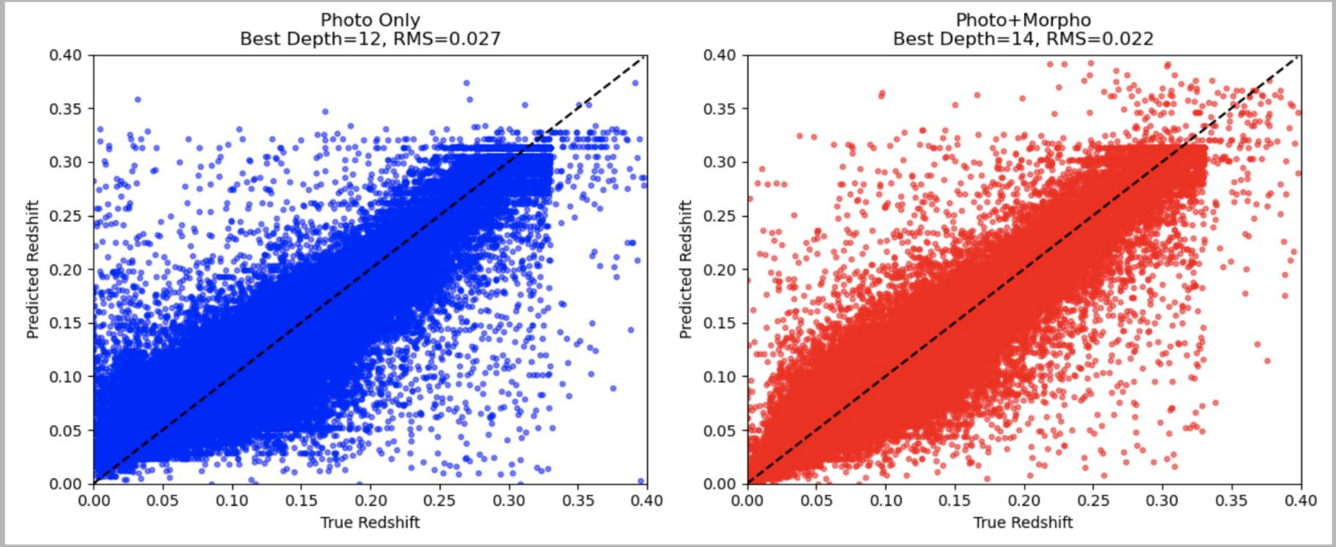
Rationale:

Morphological attributes may capture physical properties that evolve with redshift and are not fully captured by color alone.

HYPOTHESIS 1:

An optimized decision tree model that integrates both photometric and morphological galaxy features will outperform a photometric-only model in predicting spectroscopic redshift over the redshift range (0.0, 0.4).

Rationale:
Morphological attributes may capture physical properties that evolve with redshift and are not fully captured by color alone.



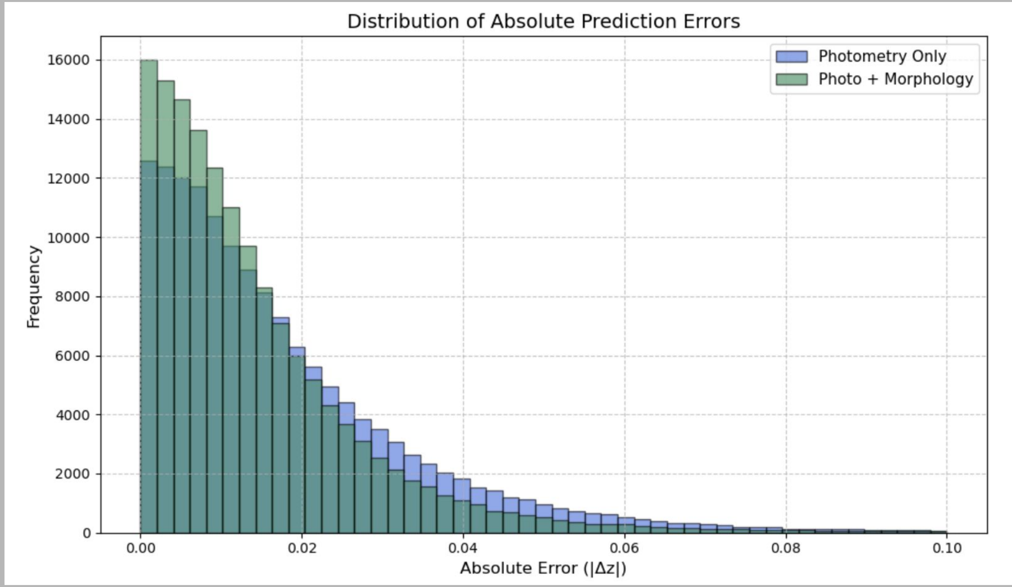
PHOTOMETRIC-ONLY
RMS ERROR: 0.027499
MSE: 0.000756
R^2 SCORE: 0.824859

COMBINED
RMS ERROR: 0.022175
MSE: 0.000492
R^2 SCORE: 0.886115

HYPOTHESIS 1:

An optimized decision tree model that integrates both photometric and morphological galaxy features will outperform a photometric-only model in predicting spectroscopic redshift over the redshift range (0.0, 0.4).

Rationale:
Morphological attributes may capture physical properties that evolve with redshift and are not fully captured by color alone.



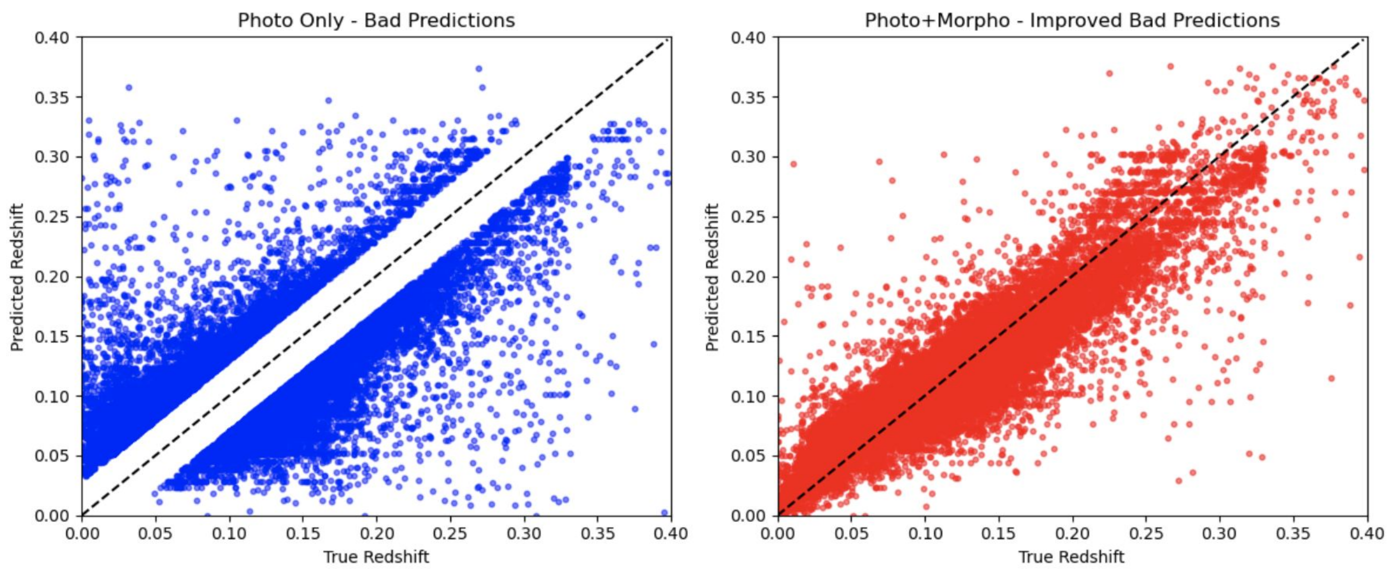
The difference in the models' prediction errors is statistically significant ($p < 0.05$).

t-statistic: 81.6820
p-value: 0.000000

HYPOTHESIS 1:

An optimized decision tree model that integrates both photometric and morphological galaxy features will outperform a photometric-only model in predicting spectroscopic redshift over the redshift range (0.0, 0.4).

Rationale:
Morphological attributes may capture physical properties that evolve with redshift and are not fully captured by color alone.



Number of bad predictions ($|\Delta z| \geq 0.03$) improved at all by Combination Model::
23650 of 27360 (86.4%)

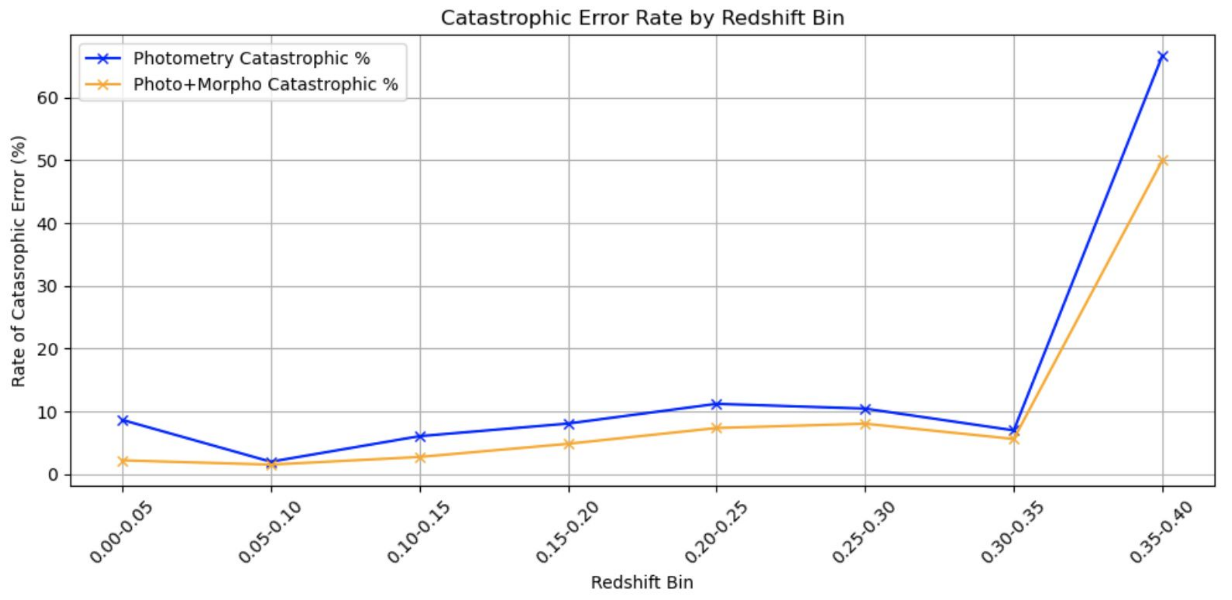
Number of bad predictions ($|\Delta z| \geq 0.03$) made tolerable ($|\Delta z| \leq 0.03$) by Combination Model:
18020 of 27360 (65.9%)

HYPOTHESIS 1:

An optimized decision tree model that integrates both photometric and morphological galaxy features will outperform a photometric-only model in predicting spectroscopic redshift over the redshift range (0.0, 0.4).

Rationale:

Morphological attributes may capture physical properties that evolve with redshift and are not fully captured by color alone.



PHOTOMETRIC-ONLY CATASTROPHIC PREDICTIONS ($|\Delta z| \geq 0.05$): 6371

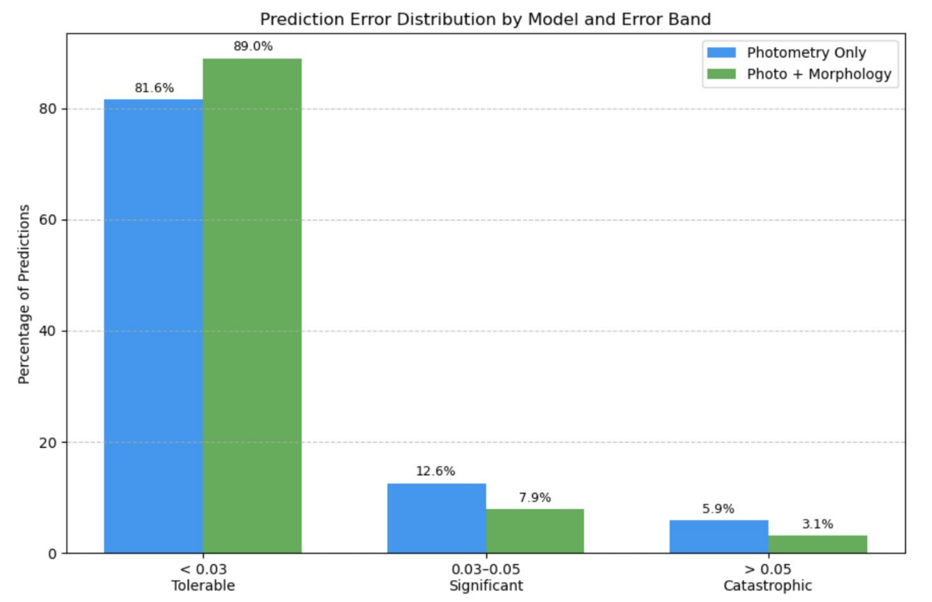
COMBINED CATASTROPHIC PREDICTIONS ($|\Delta z| \geq 0.05$): 3281

HYPOTHESIS 1:

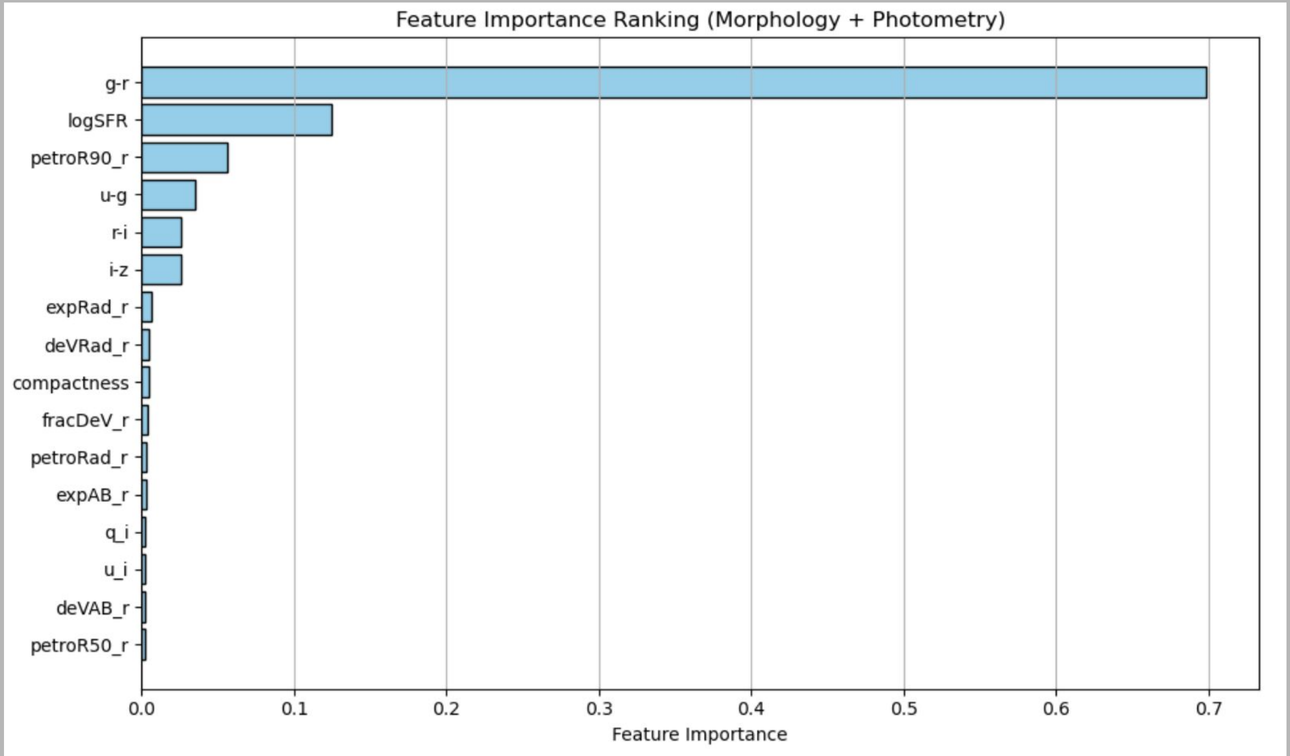
An optimized decision tree model that integrates both photometric and morphological galaxy features will outperform a photometric-only model in predicting spectroscopic redshift over the redshift range (0.0, 0.4).

Rationale:

Morphological attributes may capture physical properties that evolve with redshift and are not fully captured by color alone.



RELATIVE IMPORTANCE OF COMBINATION MODEL'S FEATURES



	Feature	Importance
1	g-r	0.698112
15	logSFR	0.124727
9	petroR90_r	0.056446
0	u-g	0.035404
2	r-i	0.026172
3	i-z	0.025749
6	expRad_r	0.005937
5	deVRad_r	0.004503
4	compactness	0.004299
10	fracDeV_r	0.003680
7	petroRad_r	0.002820
11	expAB_r	0.002635
13	q_i	0.002483
14	u_i	0.002415
12	deVAB_r	0.002395
8	petroR50_r	0.002223

HYPOTHESIS 2:

The improvement in performance obtained from including morphological features will be even more pronounced in the low-redshift regime (0.0, 0.05), where structural differences are more readily observed and better resolved.

Rationale:
At low redshift, galaxies are brighter and spatially better resolved, enabling morphological metrics to serve as stronger predictors of redshift.

z-VALUES & RELATIVE DISTANCE OF OBJECTS FROM EARTH

Near (lower z):
Objects with low redshift values (e.g., $z < 0.1$) are relatively close to Earth. Examples include nearby galaxies in the Local Group or the Virgo Cluster.

Middle (moderate z):
Objects with moderate redshift values (e.g., $0.1 < z < 1$) are at intermediate distances. These might include galaxies in more distant clusters or groups.

Far (higher z):
Objects with high redshift values (e.g., $z > 1$) are located at great distances, often in the early universe. These could include very distant galaxies, quasars, or even the remnants of the early universe.

Data Acquisition Modifications for Hypothesis 2 Evaluation

Selected galaxies with spectroscopic redshifts (z) between 0 and 0.05.

Used random sampling across z values to generate a table of 100,460 galaxies, avoiding galaxies with null and placeholder values (e.g., −9999) for relevant features.

MODEL 1: PHOTOMETRIC-ONLY DECISION TREE

FEATURE SET
Color indices derived from SDSS magnitudes (`u-g`, `g-r`, `r-i`, `i-z`)

TARGET VARIABLE
Spectroscopic redshift (`specz_redshift`), which the model is trained to predict

TRAIN/TEST SPLIT
80% training, 20% testing; fixed random seed (42)

TRAINING METHOD
Trained using `DecisionTreeRegressor` from scikit-learn

A range of tree depths (1 to 20) is is evaluated to balance model complexity and predictive performance
(Best Depth = 9)

MODEL 2: COMBINATION DECISION TREE

FEATURE SET
Color indices derived from SDSS magnitudes (`u-g`, `g-r`, `r-i`, `i-z`)

Structural properties (`compactness`, `fracDeV_r`, `deVRad_r`, `expRad_r`, `petroRad_r`, `petroR50_r`, `petroR90_r`, `expAB_r`, `deVAB_r`, `q_i`, `u_i`, `logSFR`)

TRAIN/TEST SPLIT
80% training, 20% testing; fixed random seed (42)

TARGET VARIABLE
Spectroscopic redshift (`specz_redshift`), which the model is trained to predict

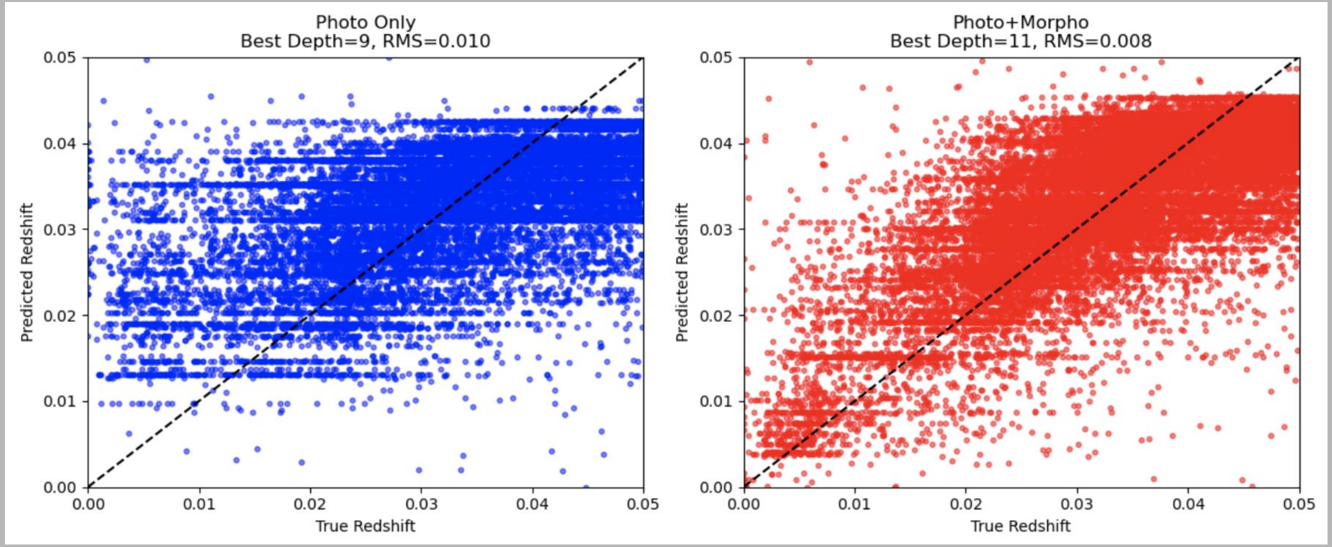
TRAINING METHOD
Trained using `DecisionTreeRegressor` from scikit-learn

A range of tree depths (1 to 20) is is evaluated to balance model complexity and predictive performance
(Best Depth = 11)

HYPOTHESIS 2:

The improvement in performance obtained from including morphological features will be even more pronounced in the low-redshift regime (0.0, 0.05), where structural differences are more readily observed and better resolved.

Rationale:
At low redshift, galaxies are brighter and spatially better resolved, enabling morphological metrics to serve as stronger predictors of redshift.



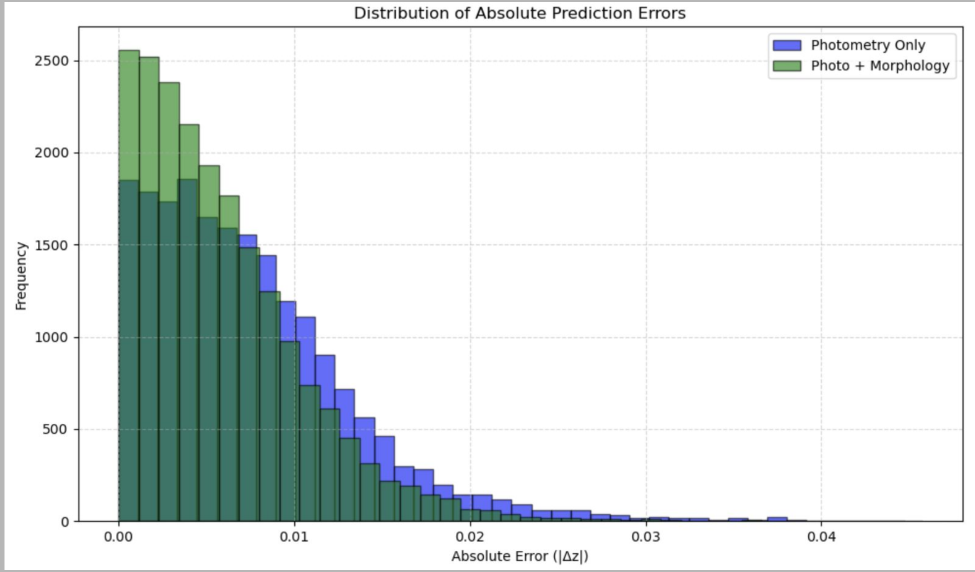
PHOTOMETRIC-ONLY
RMS ERROR: 0.009516
MSE: 0.000091
R^2 SCORE: 0.285124

COMBINED
RMS ERROR: 0.007525
MSE: 0.000057
R^2 SCORE: 0.553015

HYPOTHESIS 2:

The improvement in performance obtained from including morphological features will be even more pronounced in the low-redshift regime (0.0, 0.05), where structural differences are more readily observed and better resolved.

Rationale:
At low redshift, galaxies are brighter and spatially better resolved, enabling morphological metrics to serve as stronger predictors of redshift.



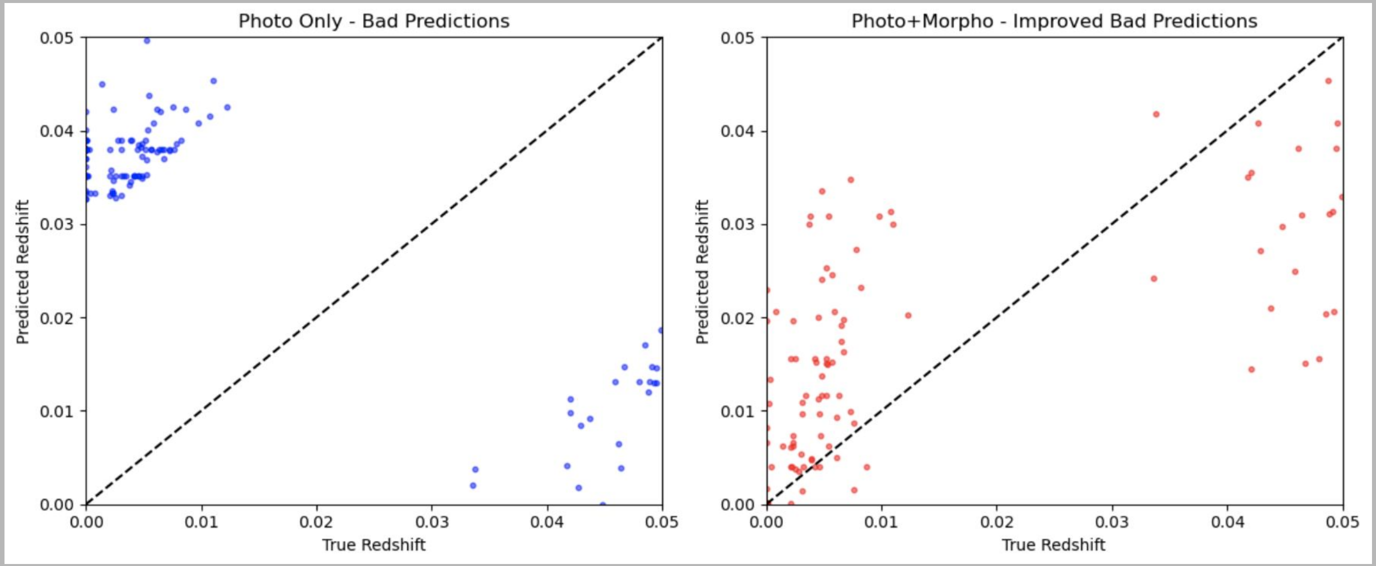
The difference in the models' prediction errors is statistically significant ($p < 0.05$).

t-statistic: 38.4998
p-value: 0.000000

HYPOTHESIS 2:

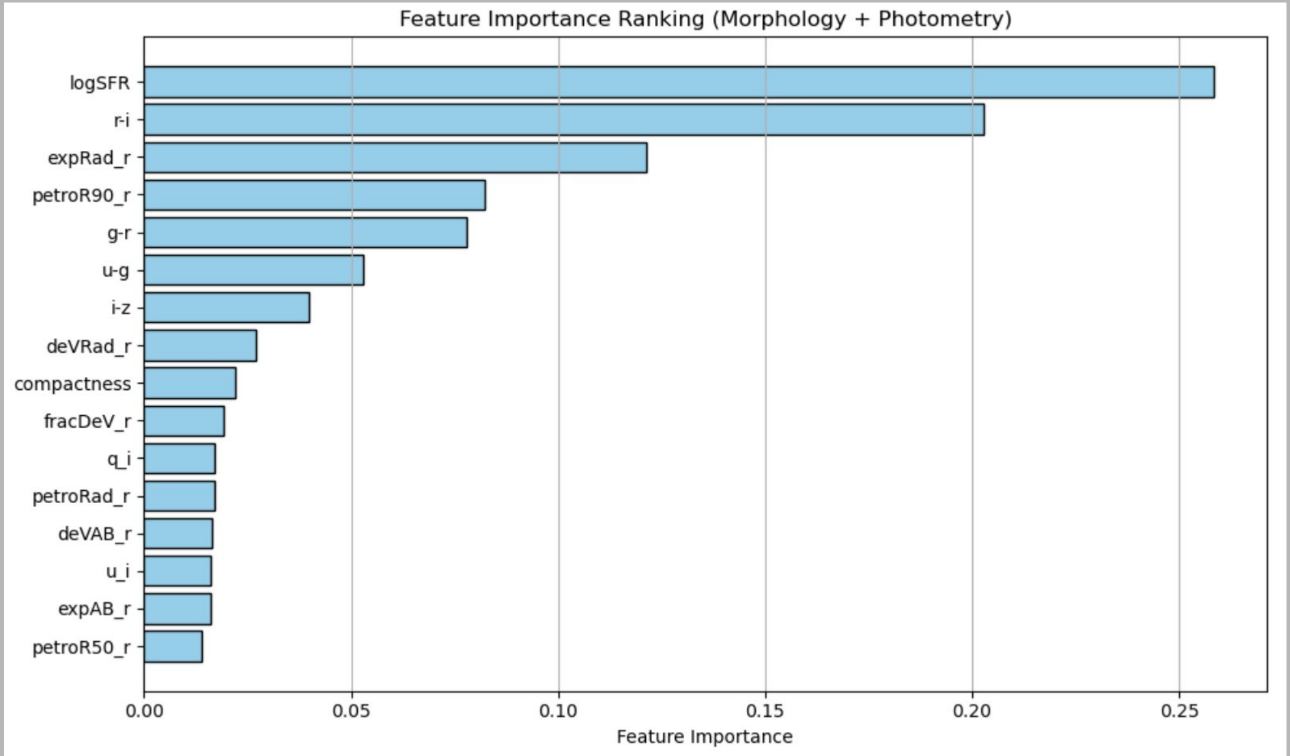
The improvement in performance obtained from including morphological features will be even more pronounced in the low-redshift regime (0.0, 0.05), where structural differences are more readily observed and better resolved.

Rationale:
At low redshift, galaxies are brighter and spatially better resolved, enabling morphological metrics to serve as stronger predictors of redshift.



Number of bad predictions ($|\Delta z| \geq 0.03$) made tolerable ($|\Delta z| \leq 0.03$) by Combination Model: 129 of 133 (97.0%)

RELATIVE IMPORTANCE OF COMBINATION MODEL'S FEATURES AT LOW z



	Feature	Importance
15	logSFR	0.258270
2	r-i	0.202875
6	expRad_r	0.121381
9	petroR90_r	0.082399
1	g-r	0.077938
0	u-g	0.052939
3	i-z	0.039829
5	deVRad_r	0.026910
4	compactness	0.022086
10	fracDeV_r	0.019270
13	q_i	0.017063
7	petroRad_r	0.017023
12	deVAB_r	0.016313
14	u_i	0.015992
11	expAB_r	0.015936
8	petroR50_r	0.013776

PART 4: FUTURE WORK

IDEAS FOR FUTURE WORK

- **Ensemble and Hybrid Models** — Experiment with gradient-boosted trees, random forests, and neural networks, as well as stacked ensembles, to capture complementary patterns and mitigate overfitting.

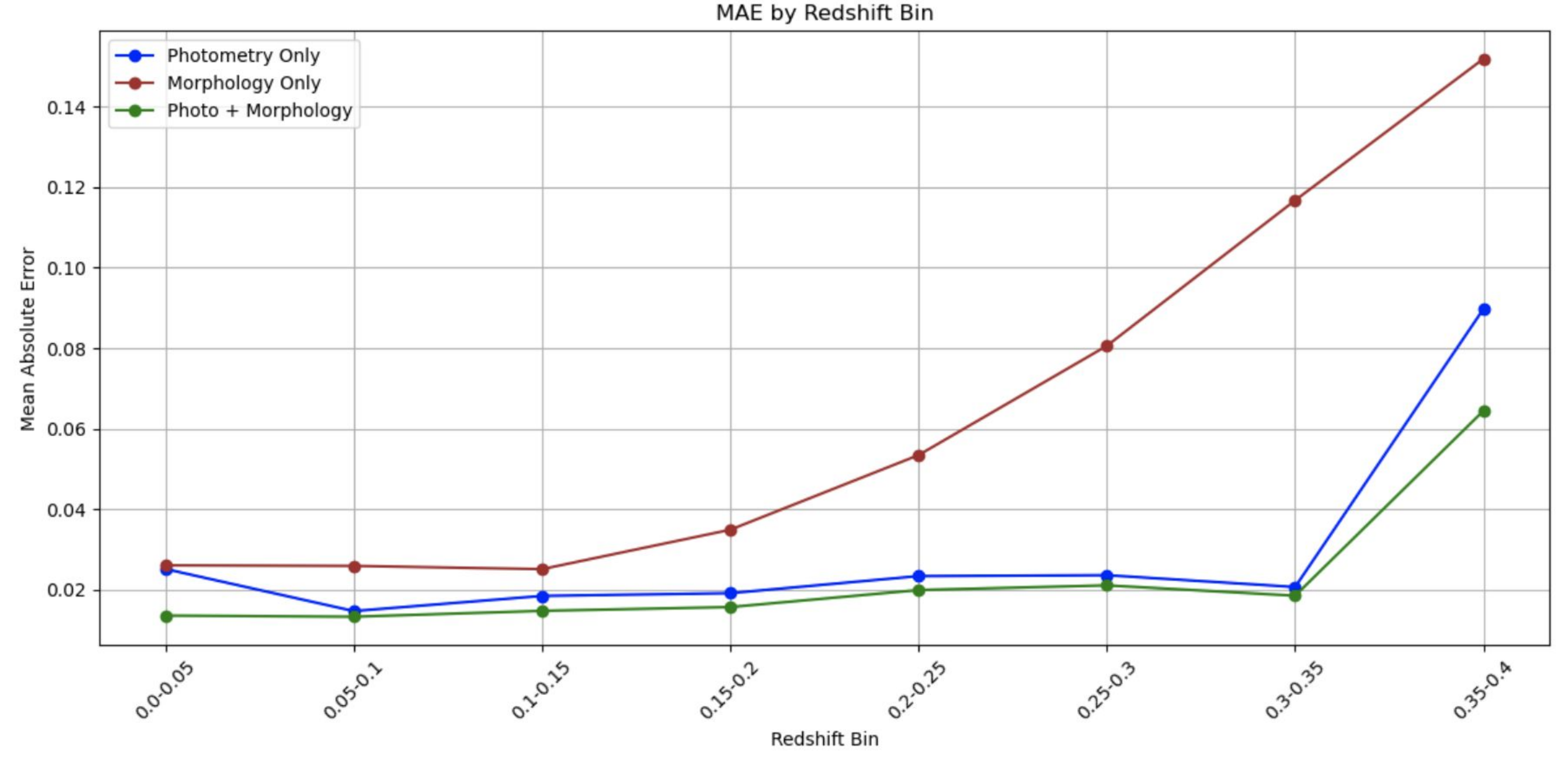
- **Expanded Feature Sets** — Integrate near-infrared (NIR) and ultraviolet (UV) photometry alongside environmental metrics (e.g., local galaxy density, cluster membership) to resolve lingering degeneracies and improve robustness.

- **Deep and High-Redshift Surveys** — Apply the combined photometric + morphological methodology to LSST, Euclid, and JWST datasets to assess performance under diminished morphological resolution and extended redshift ranges.

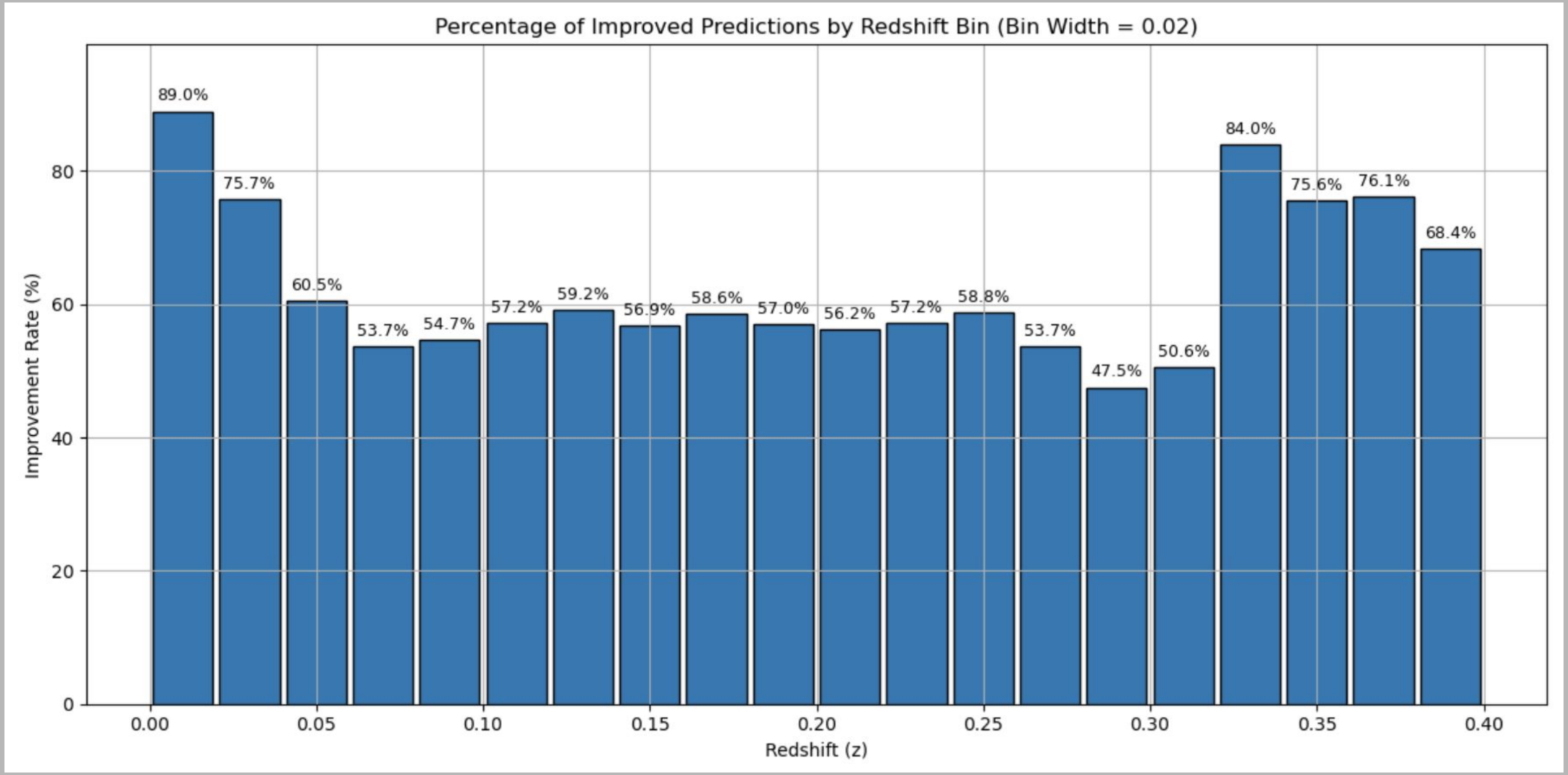
- **Domain Adaptation Across Surveys** — Pursue transfer-learning and domain-adaptation strategies to generalize models across diverse instruments and survey depths, enabling harmonized redshift catalogs.

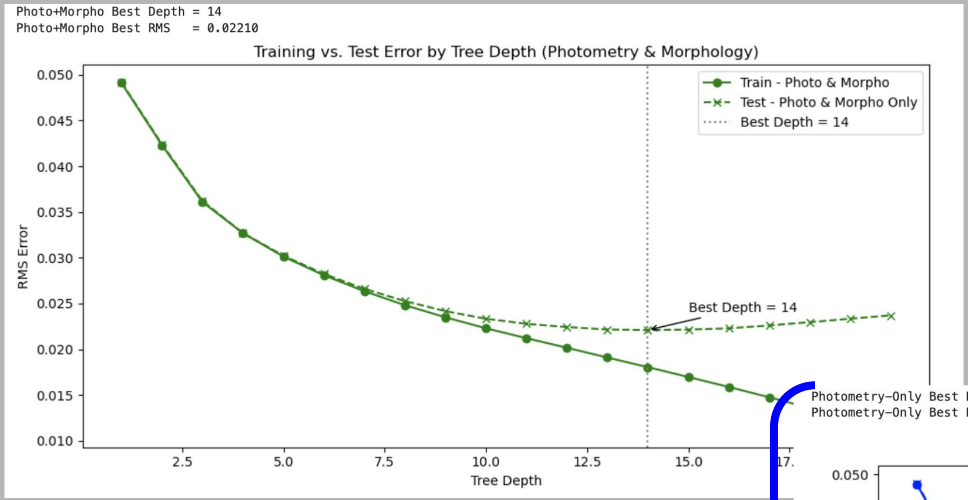
SUPPORT SLIDES

Performance Comparison (MAE per Redshift Bin): Photometric-Only vs. Morphologic-Only vs. Combined

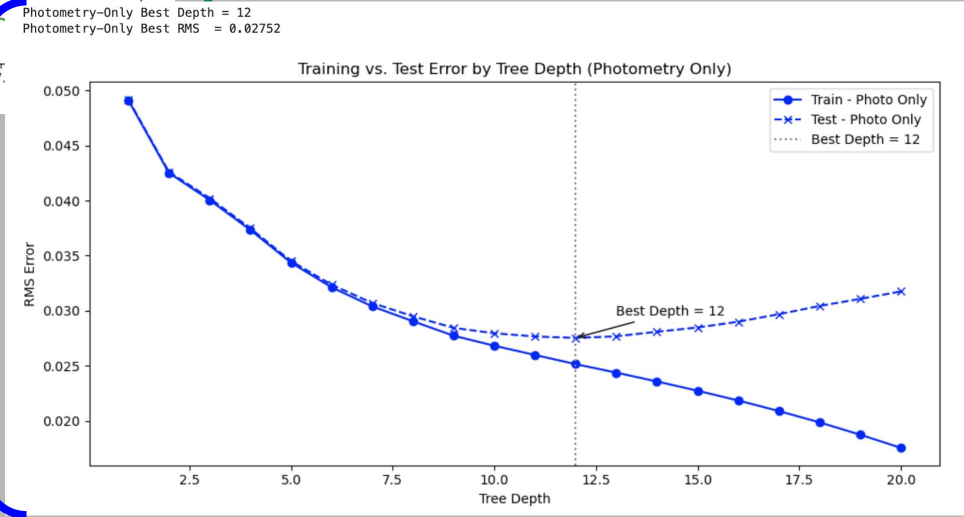


PERCENTAGE IMPROVED PER BIN BY COMBINATION MODEL

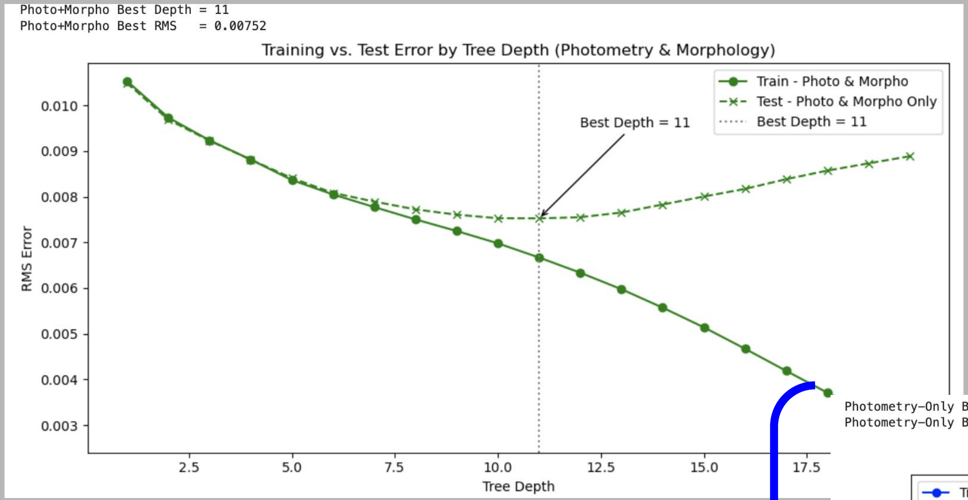




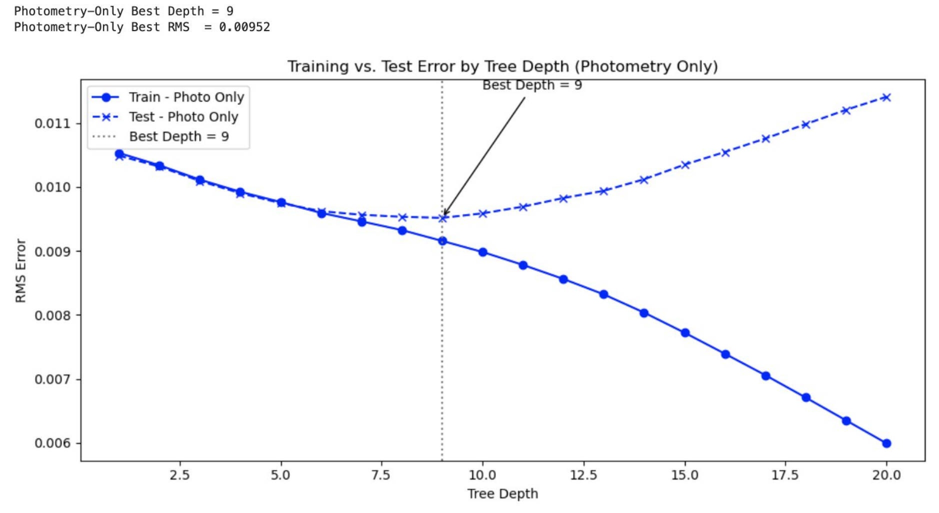
The optimal depth for the model that considers photometric and morphological features is 14, yielding a RMS error of 0.0221.



The optimal depth for the model that considers only photometric features is 14, yielding a RMS error of 0.0275.

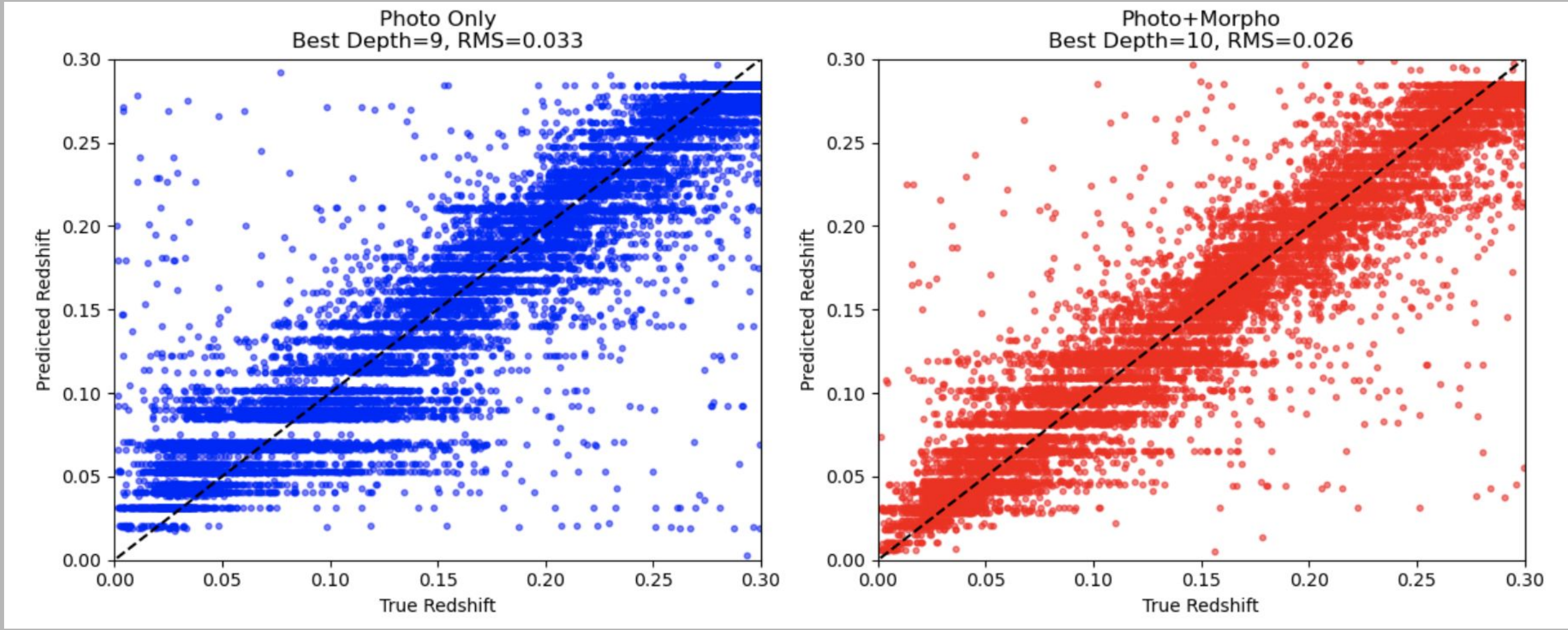


The optimal depth for the model that considers photometric and morphological features is 11, yielding a RMS error of 0.00752.

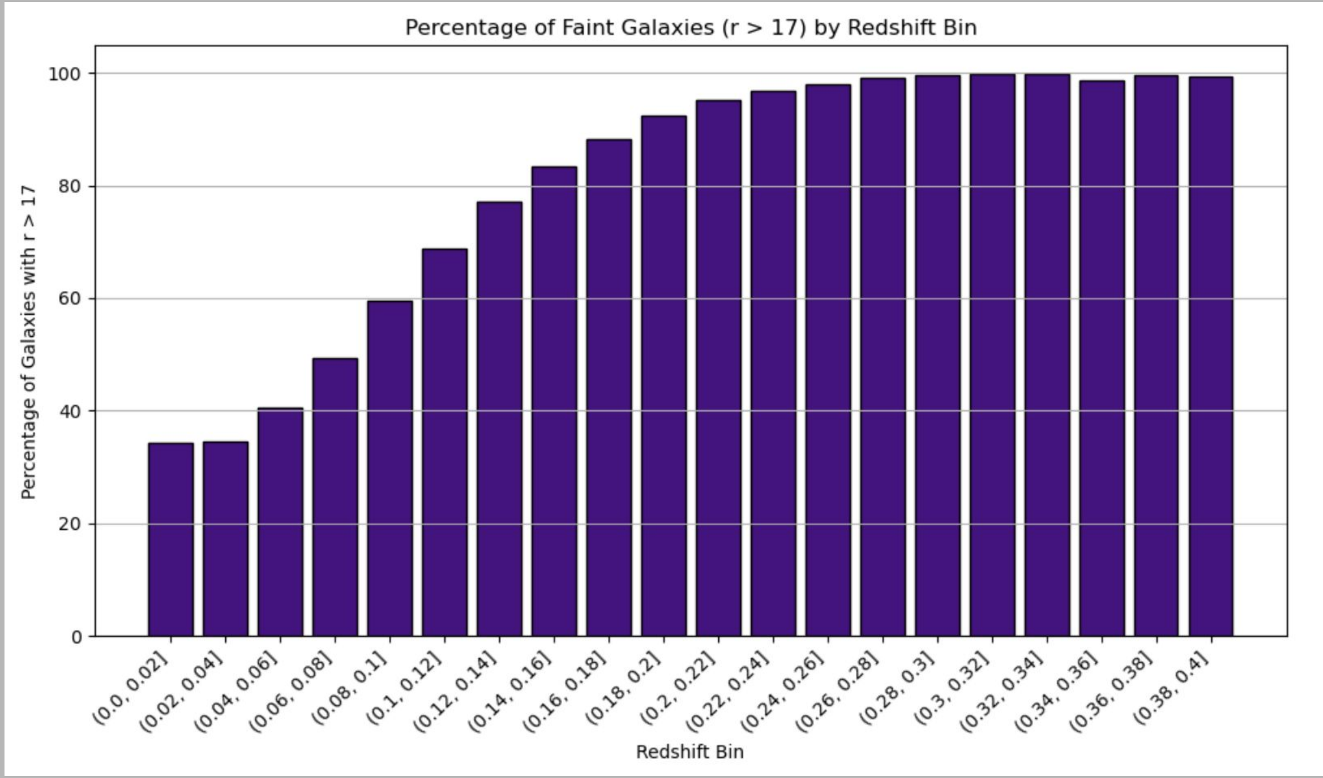


The optimal depth for the model that considers only photometric features is 9, yielding a RMS error of 0.00952.

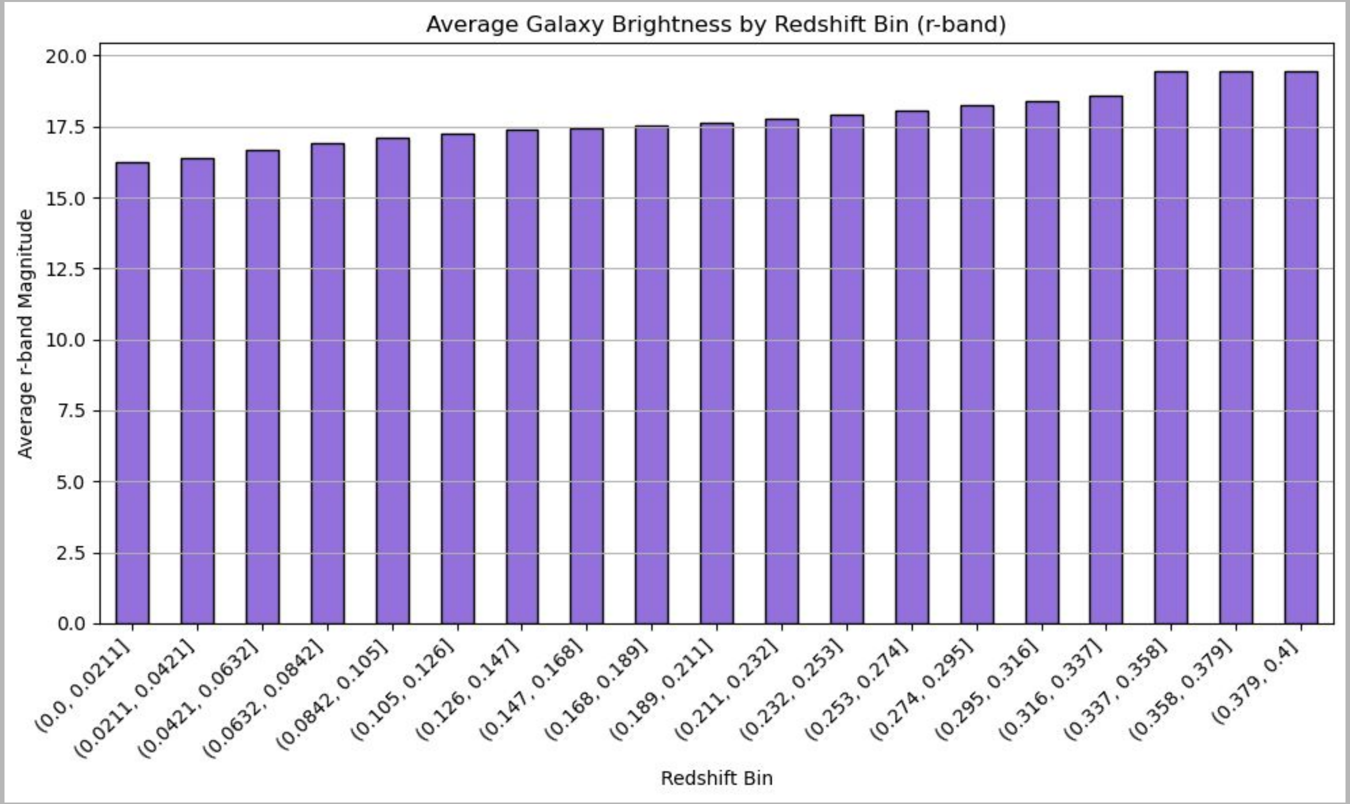
60,000 Galaxies, Stratified Sampling (6 z-value bins)



Many faint galaxies are located at mid/low red shifts.



AVERAGE BRIGHTNESS OF GALAXIES BY REDSHIFT BIN



PHOTOMETRIC COLOR INDICES

These indices are calculated by subtracting the magnitudes between two SDSS filters, providing insights into the spectral energy distribution of galaxies.

- **u–g**: Difference between ultraviolet (u) and green (g) bands. Sensitive to recent star formation and the presence of young, hot stars.
- **g–r**: Difference between green (g) and red (r) bands. Indicates the age of the stellar population; lower values suggest younger, bluer stars, while higher values indicate older, redder stars.
- **r–i**: Difference between red (r) and near-infrared (i) bands. Useful for distinguishing between different types of galaxies and stellar populations.
- **i–z**: Difference between near-infrared (i) and infrared (z) bands. Helps in identifying very red objects, such as distant galaxies or those with significant dust content.

MORPHOLOGICAL RADII (Log Transformed)

These parameters describe the size and light distribution of galaxies, often transformed logarithmically to normalize their distributions.

- **deVRad_r**: Scale radius from the de Vaucouleurs profile fit in the r-band. Represents the effective radius containing half the total light for elliptical galaxies.
- **expRad_r**: Scale radius from the exponential profile fit in the r-band. Represents the effective radius for disk-dominated galaxies like spirals.
- **petroRad_r**: Petrosian radius in the r-band, defining the aperture within which the Petrosian flux is measured. It provides a consistent way to measure galaxy sizes across different types.
- **petroR50_r**: Radius containing 50% of the Petrosian flux in the r-band. Indicates the concentration of light towards the center.
- **petroR90_r**: Radius containing 90% of the Petrosian flux in the r-band. Used alongside petroR50_r to assess the light concentration and galaxy morphology.

MORPHOLOGICAL STRUCTURE & SHAPE

These features capture the structural characteristics and orientation of galaxies.

- **fracDeV_r**: Fraction of the galaxy's light in the r-band best fit by a de Vaucouleurs profile. Values close to 1 suggest elliptical galaxies; values near 0 indicate disk-like structures.
- **expAB_r**: Axis ratio (minor/major axis) from the exponential profile fit in the r-band. Reflects the ellipticity of disk components.
- **deVAB_r**: Axis ratio from the de Vaucouleurs profile fit in the r-band. Reflects the ellipticity of bulge components.
- **q_i**: Stokes parameter Q in the i-band, representing the difference in intensity between horizontal and vertical polarization components. Used to analyze galaxy shapes and orientations.
- **u_i**: Stokes parameter U in the i-band, representing the difference in intensity between polarization components at +45° and -45°. Complements q_i in shape analysis.

STAR FORMATION RATE & COMPACTNESS

- **logSFR**: Logarithm of the star formation rate, typically measured in solar masses per year. Derived from emission lines or model fits, it indicates the current rate at which a galaxy forms new stars.
- **compactness:** A derived parameter, calculated as the ratio of petroR50_r to petroR90_r. It quantifies how concentrated a galaxy's light is towards its center, aiding in morphological classification.

WHAT IS A DECISION TREE?