




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

Darcy Corson
Tufts University
05/08/2025



Thank you for the opportunity to share my project, Improving Photometric Redshift Prediction with Morphological Data in a Decision Tree Framework. This presentation is organized into four sections. I'll begin with the project's Background and Motivation, introducing key concepts related to redshift and outlining current challenges in redshift estimation. I'll then describe the project's Data and Methods, including data sources, feature extraction, and the modeling pipeline. Next, I'll present the project's Hypotheses and Results, focusing on the impact of incorporating morphological features into photometric redshift prediction. Finally, I'll conclude with a short discussion of Future Research Directions.

A cosmic background image featuring a dense field of galaxies. In the center, a large, bright, blue-toned spiral galaxy is prominent, surrounded by numerous smaller, distant galaxies in various colors (yellow, orange, blue) and shapes. The background is a deep black space filled with these celestial objects.

PART 1: BACKGROUND & MOTIVATION

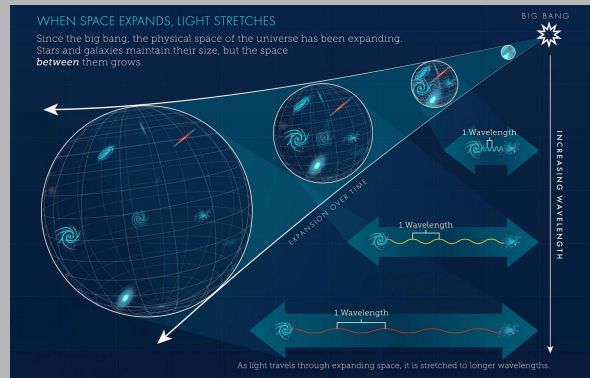
UNDERSTANDING REDSHIFT

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

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

Hubble's Law connects redshift to distance.

$$\text{Hubble's Law: } v = H_0 \times d$$



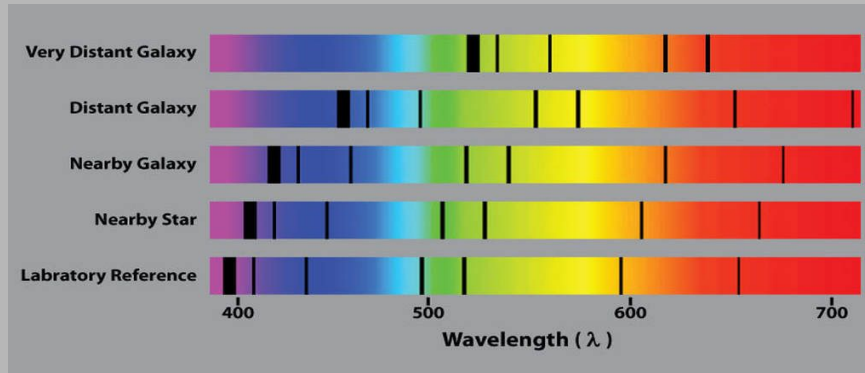
As galaxies move away from Earth, their light shifts toward the red end of the spectrum. This shift increases with distance from Earth.

Redshift, denoted by the symbol z , refers to the stretching of light's wavelength as it travels through the expanding fabric of space. This stretching shifts the light toward longer wavelengths—specifically, toward the red end of the electromagnetic spectrum. The greater the distance between a galaxy and Earth, the more its light is redshifted by the time it reaches us.

Hubble's Law builds on this phenomenon by describing a linear relationship between a galaxy's recessional velocity and its distance from Earth. In essence, galaxies that are farther away are moving away faster—an observation that supports the theory of an expanding universe.

The graphic on this slide visually illustrates this principle. As space expands over time, the light emitted by galaxies is increasingly stretched. Distant galaxies not only appear fainter, but also show more significant redshifts compared to nearby galaxies. This fundamental relationship enables astronomers to use redshift as a proxy for distance, allowing us to map the large-scale structure of the cosmos and understand the universe's expansion history.

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.

Spectroscopic redshift is widely regarded as the gold standard for determining how far away a galaxy is. Every galaxy emits a unique spectral signature—a kind of fingerprint—comprising distinct features across the electromagnetic spectrum. These signatures provide detailed insight into a galaxy's chemical composition, physical conditions, and dynamical state.

One of the most important elements of a spectral signature is its emission lines. These lines are generated when electrons in atoms or ions transition from higher to lower energy states, releasing photons at precise, known wavelengths. Emission lines are especially prominent in regions of active star formation, where energetic radiation from young stars ionizes the surrounding gas. As the ionized gas cools and recombines, it emits light at characteristic wavelengths, creating well-defined emission features in the spectrum.

As light from a galaxy travels through expanding space, these spectral features shift toward longer, redder wavelengths. By comparing the observed positions of these lines to their known rest-frame values in laboratory conditions, we can calculate how much the light has been stretched. This measurement yields the galaxy's redshift value with exceptional precision.

Because of its reliance on high-resolution spectral data, this method is extremely accurate. However, it is also labor- and resource-intensive, requiring long exposure times and specialized instruments. Due to these intensive resource requirements, spectroscopic measurements, despite their accuracy, remain practical only for limited, targeted samples of galaxies.

PHOTOMETRIC REDSHIFT DETERMINATION

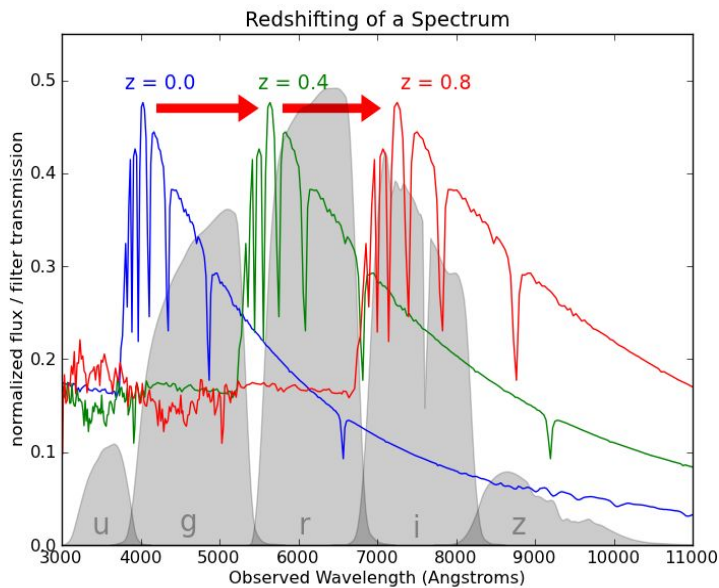


Photometric redshift estimation measures galaxy distances using their colors across multiple wavelength bands. This technique analyzes light through filters like ugriz rather than collecting full spectra.

Photometric redshift determination is a computational method for estimating galaxy redshifts using spectral energy distributions (SEDs). Unlike spectroscopic approaches, photometric methods use broadband photometry. This technique measures a galaxy's brightness across several broad filters—such as the SDSS ugriz system—each covering a wide wavelength range. The result is a set of flux measurements, which serve as coarse but efficient proxies for the galaxy's full spectral profile.

By observing how a galaxy's light is distributed across these filters, we can detect broad patterns indicative of redshifted features.

There are two primary approaches to estimating redshift photometrically. The first is template fitting, which compares observed color indices and fluxes to a library of simulated SEDs at different redshifts. The algorithm searches for the redshift at which the template best matches the observed data. The second approach is machine learning, where models are trained on large datasets of galaxies with known spectroscopic redshifts. Once trained, these models can predict redshifts for new observations with impressive speed and scalability.



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 demonstrates the possibility of determining redshift from photometry alone.

Photometric redshift estimation works by tracking how a galaxy's light shifts through broad filters (u, g, r, i, z) as redshift increases.

- The colored curves (blue for $z=0.0$, green for $z=0.4$, red for $z=0.8$) show how a galaxy's spectrum appears at different redshifts.
- The gray curves represent the sensitivity of each filter (u, g, r, i, z) to different wavelengths of light.
- As redshift increases, the galaxy's spectral features (like the 4000 Å break) move rightward, passing through different filters.
- Photometric redshift estimation infers the galaxy's redshift by analyzing its relative brightness across these filters — effectively capturing how much light fell into each band due to the shift.

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.

The trade-off between spectroscopic precision and photometric scalability defines a central challenge in modern astronomy.

Spectroscopic methods deliver highly accurate redshift measurements, yet their intensive observational requirements restrict their use to limited, carefully chosen samples of galaxies. In contrast, photometric methods offer the advantage of scalability, making them indispensable for analyzing large galaxy populations that characterize contemporary surveys. Nevertheless, the lower precision inherent to photometric estimation highlights a clear need for methodological improvements.

The goal of this project is therefore to advance photometric redshift estimation by integrating supplementary galaxy information, thereby aiming to retain the scalability of photometry while approaching the accuracy traditionally associated with spectroscopy.

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.

However, enhancing photometric redshift techniques requires addressing two significant limitations inherent to these methods.

First is the challenge of bias and scatter: prediction accuracy is not uniform across all galaxy types or redshifts. Errors vary depending on intrinsic galaxy properties, such as brightness, morphology, and star-formation activity, with faint or irregular galaxies posing particular classification difficulties. Second is the issue of color degeneracy, where galaxies with vastly different properties and redshifts can exhibit similar broadband colors. For example, a dusty, red galaxy at low redshift might closely resemble a young, blue galaxy at high redshift, complicating accurate distance estimations.

To mitigate these challenges, this research specifically incorporates morphological data into a photometric model. By combining structural characteristics with traditional photometric measurements, the aim is to reduce ambiguity, minimize bias, and enhance overall redshift prediction accuracy.

A cosmic background image featuring a dense field of galaxies. In the center, a large, bright, blue-toned spiral galaxy is prominent, surrounded by numerous smaller, distant galaxies in various shapes and colors (yellow, orange, blue). The background is a deep black space filled with stars and faint galaxy structures.

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 ($\log\text{SFR}$)
- Derived $\text{petroR50}_r/\text{petroR90}_r$ (compactness)

Data Preprocessing

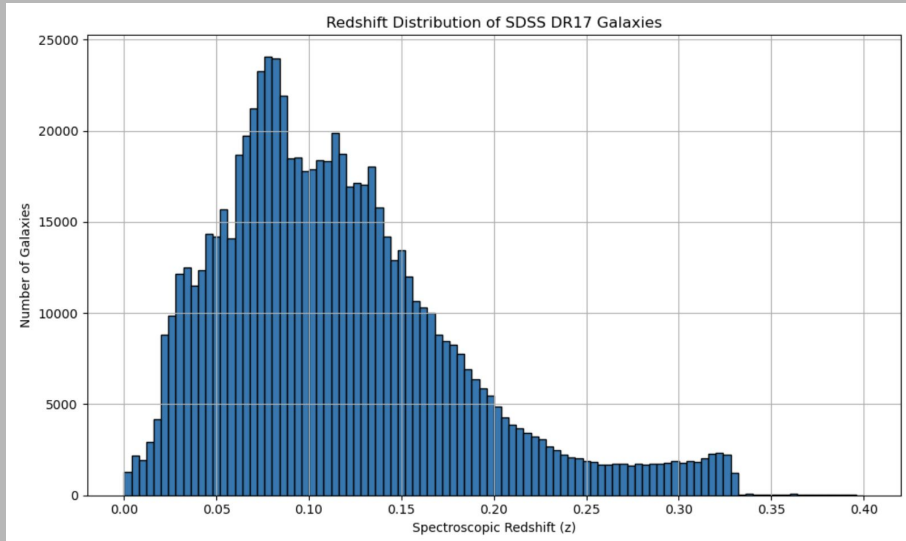
Applied log-transformation to size-related features to normalize scale.

Verified redshift distribution and overall feature completeness prior to training.

For this analysis, I leveraged the final public release of SDSS-IV (DR17), which includes extensive photometric and spectroscopic measurements for over a million galaxies. I retrieved observations via the CasJobs API, employing SQL joins across three principal tables: PhotoObj for broad-band photometry features, SpecObj for spectroscopic redshift data, and galSpecExtra for derived physical properties. I restricted the sample to galaxies with known spectroscopic redshifts within the interval $0 \leq z \leq 0.4$, excluding records containing null or placeholder values for features of interest. A uniform random sampling across this redshift range yielded a representative dataset of approximately 742,000 galaxies.

For each galaxy, I extracted both photometric and morphological features.

DISTRIBUTION OF DATASET GALAXIES BY REDSHIFT



The histogram displayed here characterizes the resulting redshift distribution of the data set. A pronounced peak exists between $z \approx 0.05$ and $z \approx 0.15$, with a gradual taper toward $z = 0.4$. This natural sampling density obviated the need for stratification; models trained on the entire dataset demonstrated superior generalization compared to a downsized, stratified subset of 60,000 galaxies (10 per bin across six bins), which exhibited degraded performance, likely due to reduced sample size. Consequently, I employed the full, randomly sampled catalogue for all subsequent analyses.

PHOTOMETRIC FEATURES

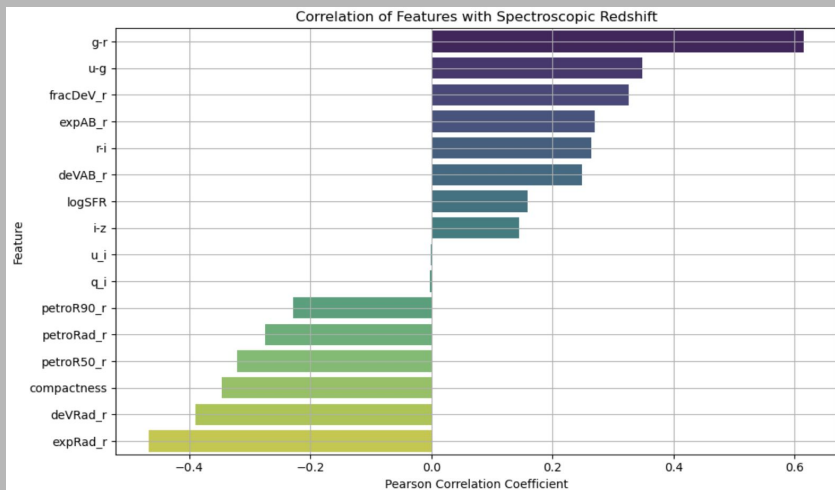
Feature	Description
u, g, r, i, z	Apparent magnitudes in ultraviolet to near-infrared filters
$g - r, u - g, r - i, i - z$	Color indices representing differences in brightness between bands

MORPHOLOGICAL FEATURES

Feature	Description
fracDeV_r	Fraction of light fit by a de Vaucouleurs profile (Sérsic index proxy)
$\text{expAB}_r, \text{deVAB}_r$	Axis ratios from exponential and de Vaucouleurs profile fits
q_i, u_i	Stokes parameters describing shape and orientation of galaxy light
$\text{petroR50}_r, \text{petroR90}_r, \text{petroRad}_r$	Petrosian radii for 50%, 90%, and total light (log-transformed)
$\text{deVRad}_r, \text{expRad}_r$	Effective radii from de Vaucouleurs and exponential models (log-transformed)
Compactness	Ratio of petroR50_r to petroR90_r , measures light concentration
$\log\text{SFR}$	Logarithm of total star formation rate (log-transformed)

These tables summarize the features that the models that I created used to predict redshift. On the top, photometric features are listed—specifically the magnitudes in the SDSS u, g, r, i , and z bands, as well as color indices like $g-r$ and $u-g$, which capture the relative brightness between bands. On the bottom are morphological features that describe a galaxy's shape, size, and structure. For example, fracDeV_r indicates how much of a galaxy's light is fit by a de Vaucouleurs profile, which is typically associated with elliptical galaxies. Features like expAB_r and deVAB_r give axis ratios, while petroR50_r and petroRad_r describe galaxy sizes. Several features, including size metrics and star formation rate ($\log\text{SFR}$), were log-transformed to normalize their distributions.

PRELIMINARY ASSESSMENT OF FEATURES' PREDICTIVE POWER

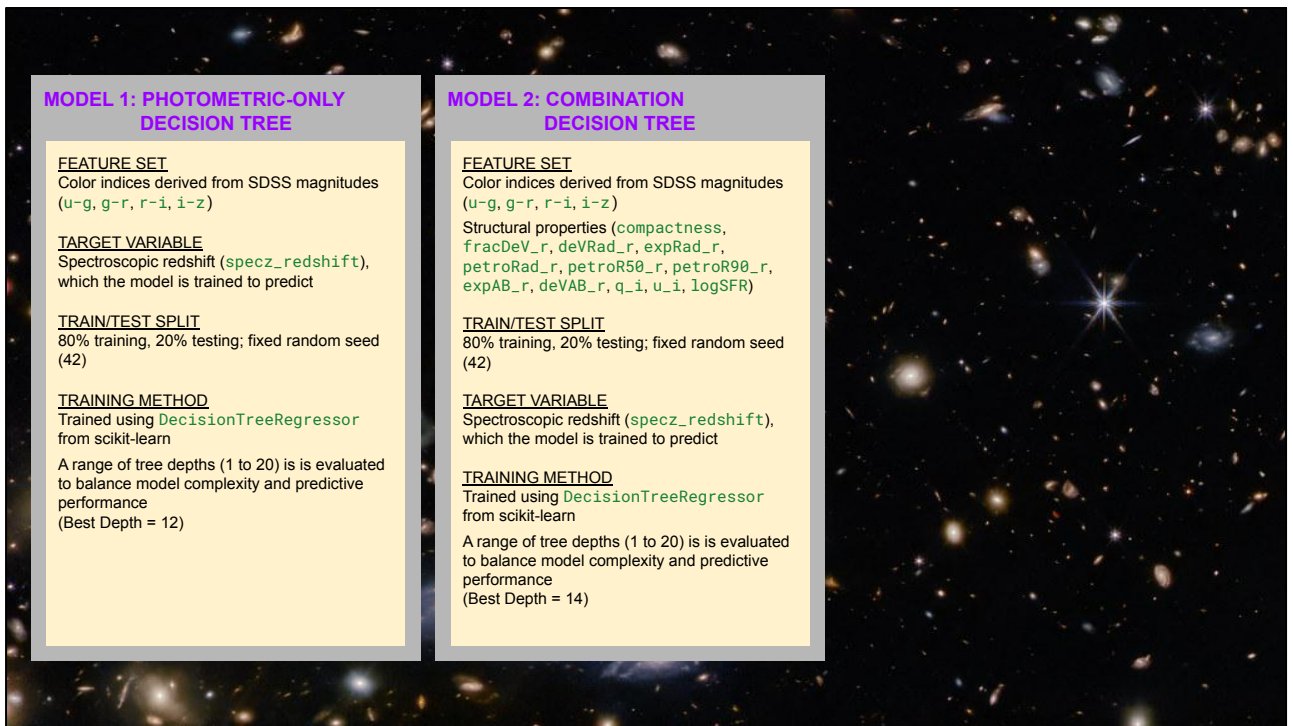


=== Correlation Table ===

Feature	Correlation with Redshift
g-r	0.615611
u-g	0.348688
fracDeV_r	0.326194
expAB_r	0.278476
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

This slide shows the Pearson correlation coefficients between each input feature and the true spectroscopic redshift values. Features at the top of the bar chart exhibit the strongest positive correlations, meaning they tend to increase with redshift. In contrast, features at the bottom are negatively correlated, suggesting that galaxy size decreases with increasing redshift. This is consistent with expectations from cosmological structure evolution.

While correlation reveals the strength of linear relationships, it doesn't capture complex or nonlinear interactions. Thus, this chart provides a useful but partial view into what the decision tree might learn. For example, some low-correlation features may still carry valuable nonlinear patterns that the model can exploit.



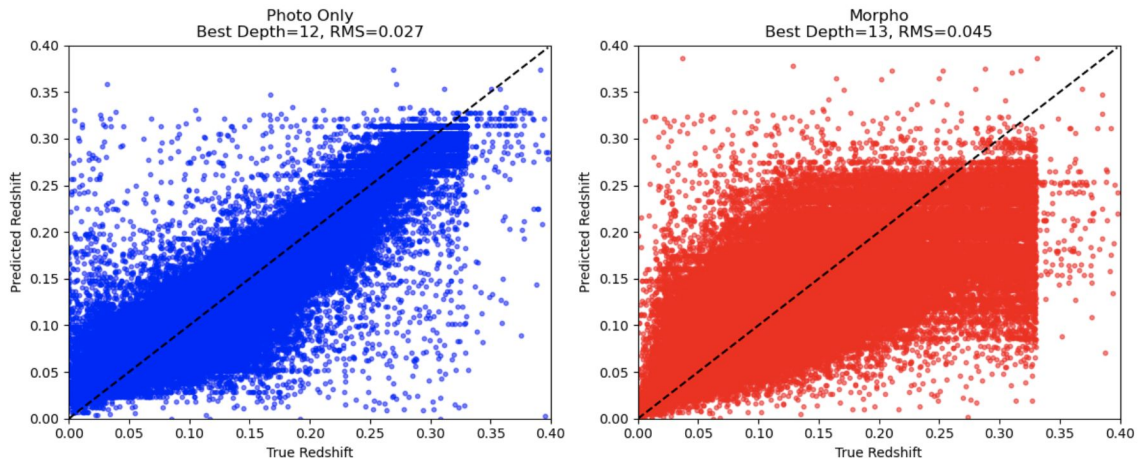
I implemented two decision-tree regressors to predict spectroscopic redshift, selecting this algorithm for its interpretability, modest computational demands, and intrinsic capacity to capture nonlinear feature interactions without additional scaling. The first model—my baseline—relies exclusively on photometric colour indices (u-g, g-r, r-i, i-z) derived from SDSS broadband magnitudes. These indices represent the prevailing approach in large-scale surveys, owing to their ready availability and minimal preprocessing requirements. I trained this photometric-only tree using an 80/20 train-test split (random seed 42) and performed grid tuning over depths 1–20, identifying an optimal depth of 12.

Building on this foundation, the second model integrates morphological descriptors alongside the same colour indices. In addition to u-g, g-r, r-i, and i-z, I incorporated structural parameters such as de Vaucouleurs fraction, axis ratios, Stokes parameters, compactness, and logarithmic star-formation rate. These supplementary features encode physical and morphological information—bulge-to-disk profiles, inclination, intrinsic shape, and stellar activity—that colour alone cannot fully capture. Under identical training conditions, this combined model achieved its best performance at a tree depth of 14.

MODEL 1: PHOTOMETRIC-ONLY DECISION TREE	MODEL 2: COMBINATION DECISION TREE	MODEL 3: MORPHOLOGICAL-ONLY DECISION TREE
<p><u>FEATURE SET</u> Color indices derived from SDSS magnitudes (u-g, g-r, r-i, i-z)</p> <p><u>TARGET VARIABLE</u> Spectroscopic redshift (<code>specz_redshift</code>), which the model is trained to predict</p> <p><u>TRAIN/TEST SPLIT</u> 80% training, 20% testing; fixed random seed (42)</p> <p><u>TRAINING METHOD</u> Trained using <code>DecisionTreeRegressor</code> from scikit-learn A range of tree depths (1 to 20) is evaluated to balance model complexity and predictive performance (Best Depth = 12)</p>	<p><u>FEATURE SET</u> Color indices derived from SDSS magnitudes (u-g, g-r, r-i, i-z) Structural properties (<code>compactness</code>, <code>fracDeV_r</code>, <code>deVRad_r</code>, <code>expRad_r</code>, <code>petroRad_r</code>, <code>petroR50_r</code>, <code>petroR90_r</code>, <code>expAB_r</code>, <code>deVAB_r</code>, <code>q_i</code>, <code>u_i</code>, <code>logSFR</code>)</p> <p><u>TARGET VARIABLE</u> Spectroscopic redshift (<code>specz_redshift</code>), which the model is trained to predict</p> <p><u>TRAIN/TEST SPLIT</u> 80% training, 20% testing; fixed random seed (42)</p> <p><u>TRAINING METHOD</u> Trained using <code>DecisionTreeRegressor</code> from scikit-learn A range of tree depths (1 to 20) is evaluated to balance model complexity and predictive performance (Best Depth = 14)</p>	<p><u>FEATURE SET</u> Structural properties (<code>compactness</code>, <code>fracDeV_r</code>, <code>deVRad_r</code>, <code>expRad_r</code>, <code>petroRad_r</code>, <code>petroR50_r</code>, <code>petroR90_r</code>, <code>expAB_r</code>, <code>deVAB_r</code>, <code>q_i</code>, <code>u_i</code>, <code>logSFR</code>)</p> <p><u>TARGET VARIABLE</u> Spectroscopic redshift (<code>specz_redshift</code>), which the model is trained to predict</p> <p><u>TRAIN/TEST SPLIT</u> 80% training, 20% testing; fixed random seed (42)</p> <p><u>TRAINING METHOD</u> Trained using <code>DecisionTreeRegressor</code> from scikit-learn A range of tree depths (1 to 20) is evaluated to balance model complexity and predictive performance (Best Depth = 13)</p>

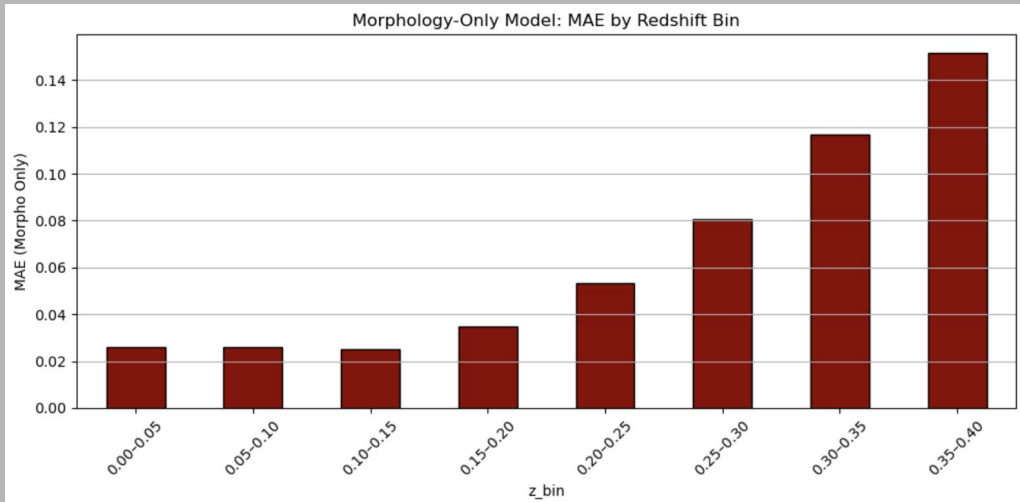
To isolate the predictive contribution of galaxy structure, I constructed a third decision-tree model using only morphological features. This morphology-only regressor does not aim to surpass the photometric baseline; rather, it quantifies how much redshift signal resides in structural characteristics alone. Maintaining the same 80/20 split and depth-search procedure, I found the optimal tree depth to be 13. The resulting performance underscores the intrinsic value of morphology: while insufficient on its own to replace colour information, structural parameters nevertheless encode non-redundant redshift cues, particularly for well-resolved, low-redshift galaxies.

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



The resulting performance of this morphology-only model underscores the intrinsic value of morphology: while insufficient on its own to replace colour information, structural parameters nevertheless encode non-redundant redshift cues.

PERFORMANCE OF MORPHOLOGICAL-ONLY MODEL: MAE PER REDSHIFT BIN



The morphology-only decision tree exhibits strong predictive performance at low redshift (approximately $z = 0.0\text{--}0.2$), where the mean absolute error remains relatively low. This reflects the fact that nearby galaxies are imaged with sufficient resolution and signal-to-noise for structural parameters—such as light-profile shape, size metrics, and compactness—to carry meaningful information about their distances. Beyond $z \approx 0.2$, however, the model's error increases sharply. At these higher redshifts, galaxy morphologies become less distinct: fainter surface brightness, reduced angular size, and more frequent irregular or merging systems all conspire to degrade the reliability of structural descriptors. Consequently, the rising MAE confirms that morphological features alone are most informative for nearby galaxies and lose predictive power as objects recede and imaging quality diminishes.

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 ($0.03 \leq \Delta z \leq 0.05$), indicating notable deviation from true redshift.
Number of Catastrophic Errors	Counts extreme mispredictions ($ \Delta z \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.

To assess model performance comprehensively, I employ both continuous and categorical metrics. Root-mean-square (RMS) error and mean absolute error (MAE) quantify the average magnitude of prediction deviations, with RMS placing greater weight on large outliers and MAE treating all errors uniformly. The coefficient of determination (R^2) indicates the proportion of redshift variance explained by the model, with values approaching 1 denoting stronger explanatory power. Beyond these standard statistics, I track specific error bands that are astrophysically meaningful: “significant errors” ($0.03 < |\Delta z| \leq 0.05$), which represent moderate misestimates, and “catastrophic errors” ($|\Delta z| > 0.05$), which can misclassify galaxies into entirely incorrect cosmic epochs. Finally, I quantify “errors made tolerable,” capturing cases where the inclusion of morphological features rescues a previously catastrophic photometric-only prediction to within the acceptable threshold ($|\Delta z| \leq 0.03$). Collectively, these metrics evaluate not just average predictive accuracy, but also the model’s resilience in avoiding severe redshift misclassifications.

A cosmic background image featuring a dense field of galaxies. In the center, a large, bright, blue-toned spiral galaxy is prominent, surrounded by numerous smaller, distant galaxies in various shapes and colors (blue, yellow, orange). The background is a deep black space filled with these celestial bodies.

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.

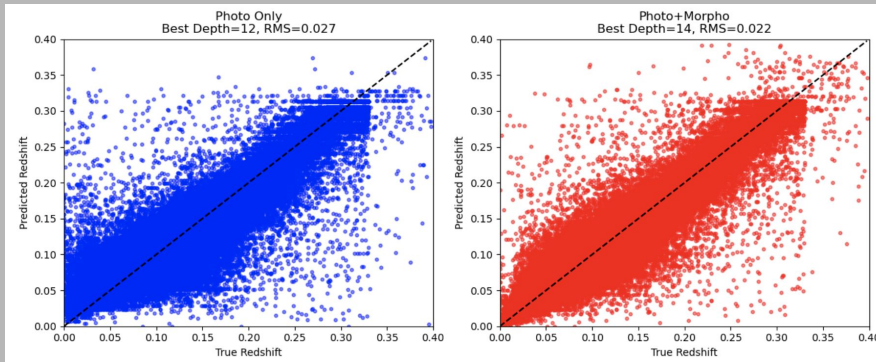
I hypothesize that an optimized decision-tree regressor incorporating both photometric and morphological galaxy features will outperform a model relying solely on photometric inputs across the redshift interval (0, 0.4). The underpinning rationale for this hypothesis is that morphological descriptors encode physical processes (e.g., mergers, bulge growth, disk fading) that evolve with cosmic time. While broad-band colour indices primarily trace stellar population age and dust content, the addition of structural features offers complementary information about a galaxy's dynamic state and internal morphology. By integrating both feature classes, I expect the combined model to capture redshift-related patterns more fully, yielding measurably higher predictive accuracy.

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

R² SCORE: 0.824859

COMBINED

RMS ERROR: 0.022175

MSE: 0.000492

R² SCORE: 0.886115

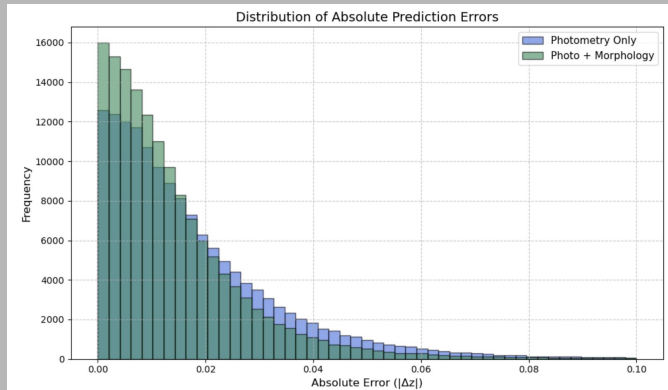
A direct comparison of the photometric-only and combined models demonstrates the value of adding morphological information. The photometric-only model exhibits a root-mean-square error (RMSE) of 0.0275, mean-squared error (MSE) of 0.000756, and an R^2 of 0.825, with a clear tendency to underestimate at higher redshifts, as evidenced by the scatter around the 1:1 line. In contrast, the combined model yields an RMSE of 0.0222, MSE of 0.000492, and R^2 of 0.886. Visually, its residuals cluster more tightly around the 1:1 line indicating that morphological features materially improve the model's fit—both statistically and practically—throughout the full redshift range.

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

The histogram in the center of the slide shows the distribution of absolute redshift prediction errors for each model. In this plot, the photometric-only model is shown in blue, and the combined model is shown in green.

You'll notice a clear leftward shift in the green distribution—indicating that the combined model produced more predictions with lower absolute error. This visual trend suggests that adding morphological features improves the model's precision, reducing the frequency of large errors.

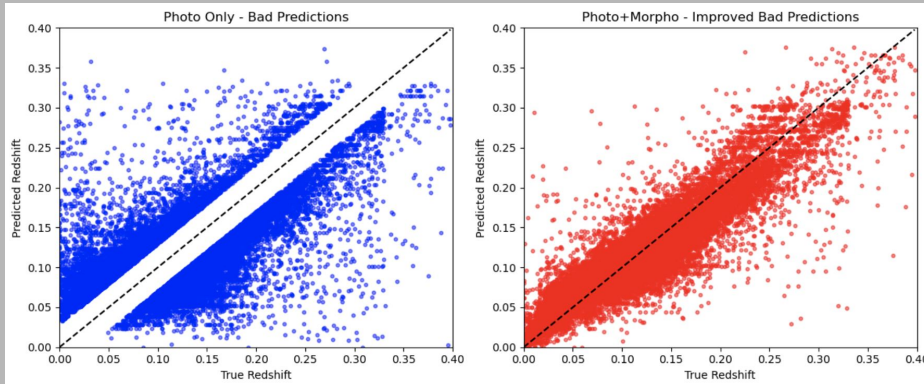
To determine whether the observed reduction in absolute residuals is statistically significant, I conducted a paired t-test comparing the photometric-only and the combined models on the same test set. The resulting t-statistic of 81.68 and associated p-value below 10^{-6} decisively reject the null hypothesis of equal mean errors. This finding confirms that the addition of morphological parameters confers a genuine, reproducible improvement in redshift estimation accuracy rather than reflecting random variation.

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%)

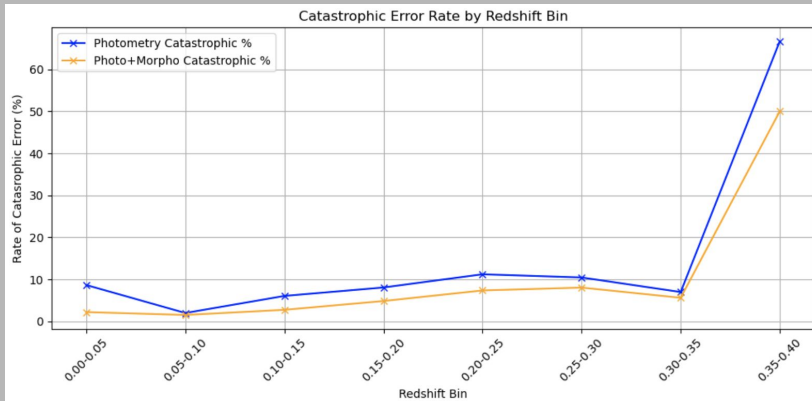
Focusing on mispredictions whose magnitudes exceeded 0.03, I isolated the 27 360 cases in which the photometric-only model failed to meet this threshold. Incorporation of morphological features improved 86.4 % of these instances (23 650 galaxies), and in 65.9 % of cases (18 020 galaxies), reduced the error below the critical 0.03 cutoff. These results illustrate that structural information not only improves average performance metrics but plays a critical role in rescuing otherwise poor predictions, thereby enhancing model robustness in challenging regions of parameter space.

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

Catastrophic errors—defined as $|\Delta z| \geq 0.05$ —pose serious risks for scientific analysis.

Looking first at the line plot, we see the catastrophic error rate by redshift bin for both models. Across nearly every bin, the combined model—shown in orange—reduces the fraction of catastrophic errors compared to the photometric-only model (in blue). This performance gap is especially clear in the lower redshift bins ($z < 0.15$), and again at the highest bin (0.35–0.40), where catastrophic errors spike in both models due to resolution and signal limitations.

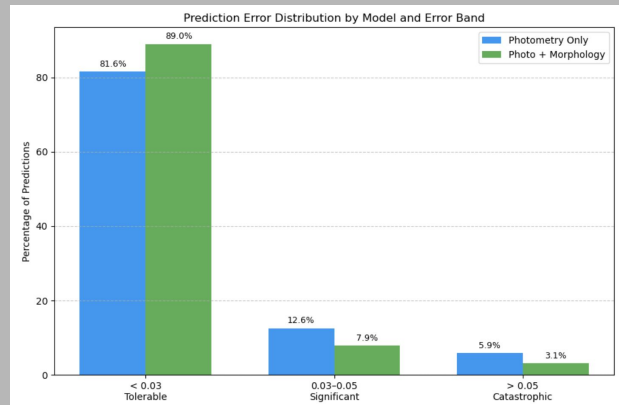
The summary statistics on the right quantify this improvement. The photometric-only model produced 6,371 catastrophic errors, whereas the combined model with morphological features reduced this number by nearly half, down to 3,281 catastrophic errors. This reinforces the value of integrating structural galaxy information—particularly in preventing the most damaging types of predictive failure.

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).

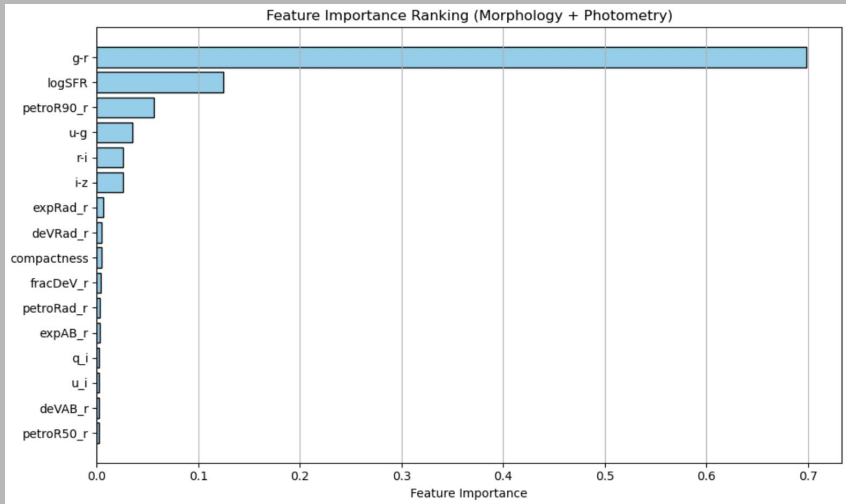
Rationale:

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



A categorical breakdown across three error bands—tolerable ($|\Delta z| < 0.03$), significant ($0.03 \leq |\Delta z| \leq 0.05$), and catastrophic ($|\Delta z| > 0.05$)—further illustrates model reliability gains. The combined model attains an 89.0 % tolerable prediction rate versus 81.6 % for photometry alone, while the proportion of significant errors falls from 12.6 % to 7.9 %. Most notably, the catastrophic error rate is cut from 5.9 % to 3.1 %. This shift toward low-error outcomes across the distribution emphasizes that morphological information systematically bolsters both average performance and worst-case reliability.

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

Examining feature importance in the combined model reveals that the photometric $g-r$ colour index dominates, contributing approximately 70 % of the total reduction in mean-squared error—a testament to the enduring power of broad-band colours. The star-formation proxy log SFR emerges as the second most influential feature ($\approx 12\%$), confirming that morphological tracers of physical activity carry substantial, non-redundant signal. Additional structural metrics, such as Petrosian radii, exponential-profile scale lengths, and de Vaucouleurs-profile fractions, contribute smaller yet meaningful shares. These results validate that, while photometry remains essential, morphology provides a complementary enhancement—particularly for specific galaxy subpopulations—thus validating the logic of Hypothesis 1.

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.

Building on the consistent improvement observed with combined photometric and morphological features, I noted that the gains in predictive accuracy were particularly pronounced at lower redshifts. Morphological descriptors—such as light concentration, axis ratios, and bulge-to-disk profile fractions—are inherently more reliable for nearby galaxies, where spatial resolution and signal-to-noise ratio are highest. Preliminary analyses, including lower mean absolute errors and higher improvement percentages for $z < 0.1$, reinforce the expectation that structural information may yield its greatest benefit in this regime. This rationale underpins Hypothesis 2, which tests whether the incorporation of morphological features provides maximal predictive gains at low 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.

It is essential to interpret redshift values as proxies for cosmological distance and data quality.

Low-redshift galaxies ($z < 0.1$) reside within the Local Group and nearby clusters such as Virgo; at these distances, telescope resolution permits detailed recovery of morphological features including spiral arms and central bulges.

Intermediate-redshift galaxies ($0.1 < z < 1$) remain observable but their structures begin to suffer surface-brightness dimming and reduced angular extent.

High-redshift objects ($z > 1$) are observed at early cosmic epochs, often appearing faint and unresolved, which limits morphological characterization to coarse shape or brightness gradients.

Given this degradation with distance, I anticipated that morphological features would contribute most substantially in the low-redshift domain, where data quality supports confident structural measurements.

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 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 evaluated to balance model complexity and predictive performance
(**Best Depth = 11**)

To evaluate Hypothesis 2, I selected galaxies with spectroscopic redshifts in the interval $0.0 \leq z \leq 0.05$, constructing a balanced evaluation set of 100,460 nearby objects via random sampling. I then trained two decision-tree regressors under identical conditions:

- The **photometric-only model** used colour indices ($u-g$, $g-r$, $r-i$, $i-z$) with an 80/20 train-test split (seed 42), achieving an optimal tree depth of 9.
- The **combined model** incorporated the same colour indices plus morphological descriptors (de Vaucouleurs fraction, axis ratios, Stokes parameters, compactness, log SFR), yielding an optimal depth of 11.

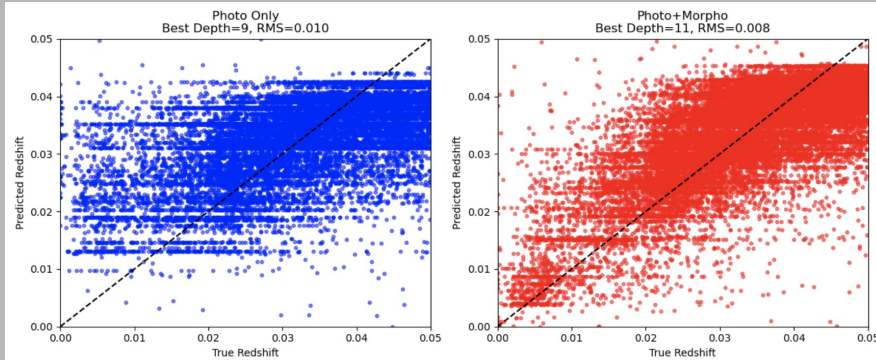
By comparing their performance on this low-redshift test set, the specific contribution of morphology in the regime where structural data are of highest fidelity is isolated.

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

R² SCORE: 0.285124

COMBINED

RMS ERROR: 0.007525

MSE: 0.000057

R² SCORE: 0.553015

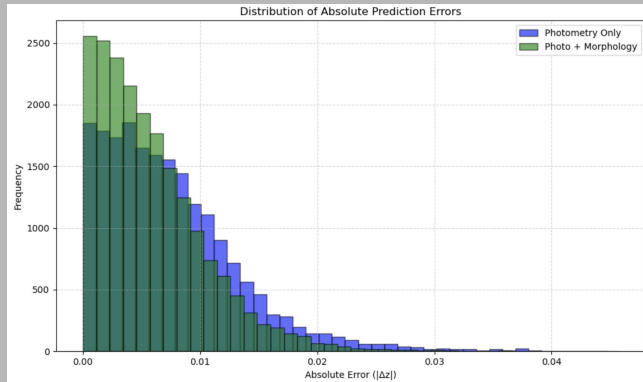
The left-hand scatter plot shows the photometric-only decision tree (optimal depth 9) plotted against true redshift values; here the root-mean-square error is 0.00952, the mean-squared error is 0.000091, and the R^2 score is 0.285, revealing considerable dispersion around the unity line. By contrast, the right-hand plot displays the combined photometric + morphological model (optimal depth 11), which achieves a lower RMS error of 0.00753, an MSE of 0.000057, and an R^2 of 0.553. The markedly tighter clustering of points along the diagonal demonstrates that structural features contribute substantial, non-redundant information—particularly when galaxies are sufficiently resolved—thereby validating Hypothesis 2.

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

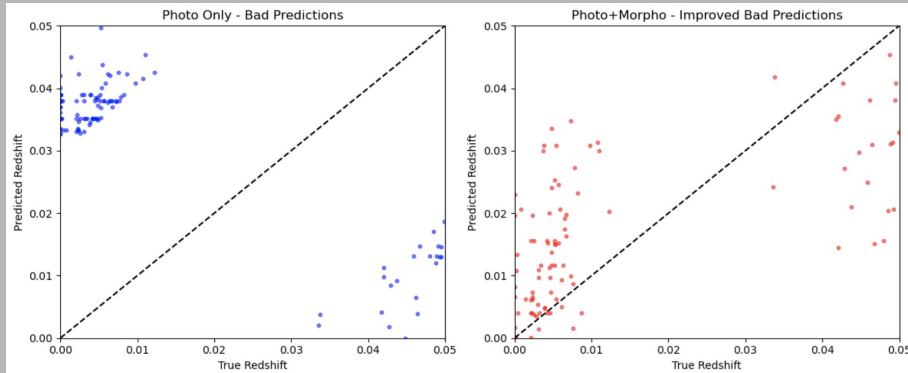
To further validate Hypothesis 2, I compared the distribution of absolute prediction errors for the photometric-only and the combined models within $0.0 \leq z \leq 0.05$. The histogram reveals that the combined model's error distribution (green) is shifted markedly to the left relative to the photometric-only model (blue), indicating a greater proportion of very small residuals. To assess statistical significance, I performed a paired t-test on the absolute errors: the resulting t-statistic of 38.5 (with $p \ll 10^{-6}$) confirms that the reduction in error variance is not due to chance. Thus, adding morphological features not only lowers the average error but also tightens the overall spread of predictions, demonstrating a genuine precision gain at low redshift.

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.

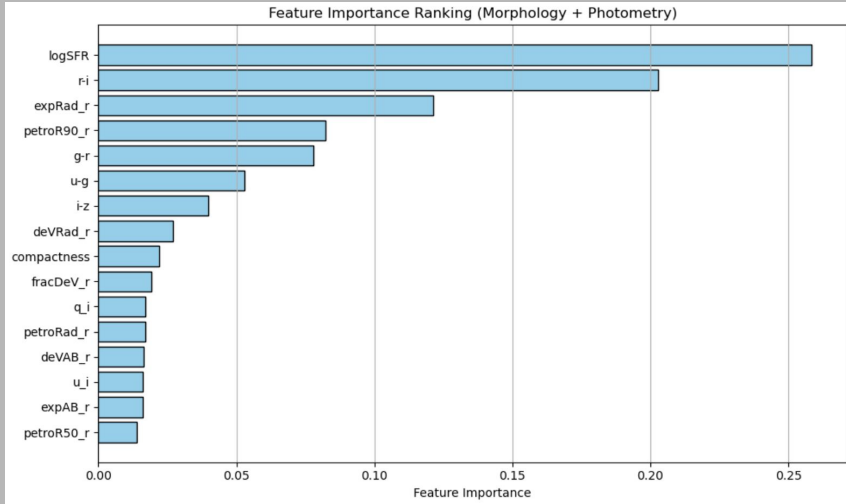


Number of bad predictions ($|\Delta z| \geq 0.03$) made tolerable ($|\Delta z| \leq 0.03$) by Combination Model: 129 of 133 (97.0%)

Predictions improved using morphology: 12313 out of 20092 (61.3%)

Next, I isolated instances in which the photometric-only model committed significant errors among the $0.0 \leq z \leq 0.05$ subset. Of the 133 such failures, the combined model successfully corrected 129 cases—an improvement rate of 97 %. In the accompanying scatter plots, the left panel shows these outliers for the photometric-only model (points well off the diagonal), while the right panel depicts the same galaxies under the combined model, almost all now clustered within the tolerable error band. This dramatic rescue of extreme mispredictions highlights the critical role that morphological information can play in rectifying otherwise intractable errors for nearby galaxies.

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

Finally, I examined the relative importance of each feature in the low-redshift combined decision tree. Notably, logSFR emerges as the most influential morphological predictor, underscoring the value of physical galaxy properties beyond broadband colours. The next highest ranks are occupied by the r-i color index and the exponential profile radius, indicating that both photometric and structural parameters contribute meaningfully. Additional morphological descriptors—such as Petrosian radius, compactness, and de Vaucouleurs radius—also feature prominently. Together, this ranking confirms that morphology provides non-redundant, and in some cases leading, information for redshift estimation in the low-z regime.

A cosmic background image featuring a dense field of galaxies. In the center, a large, bright, blue-toned spiral galaxy is prominent. Surrounding it are numerous smaller, distant galaxies in various shapes and colors, including yellow, orange, and blue, set against a deep black space. A horizontal grey band is superimposed across the middle of the image.

PART 4: FUTURE WORK

IDEAS FOR FUTURE WORK

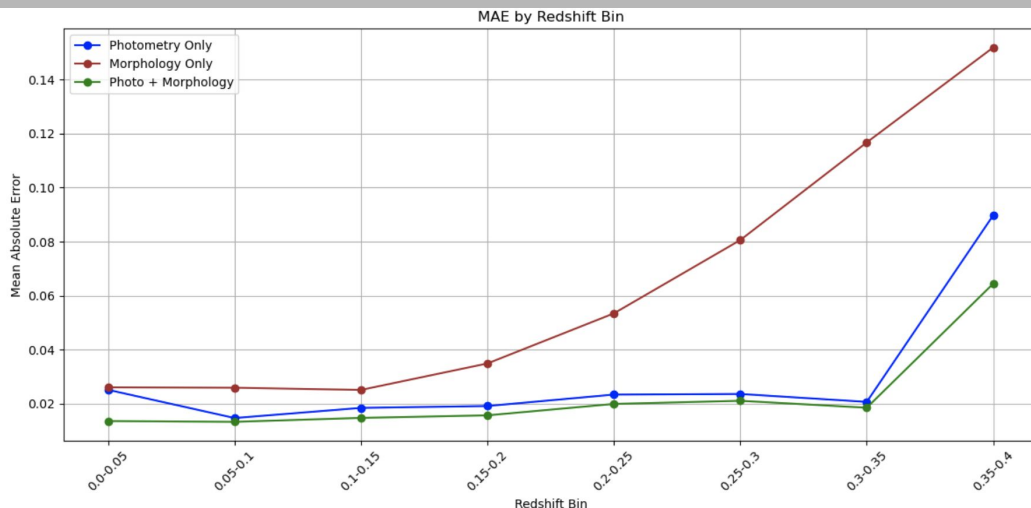
- **Ensemble and Hybrid Models** — Experiment with gradient-boosted trees, random forests, and neural networks, as well as stacked ensembles, to capture complementary patterns and mitigate overfitting.
- **Expanded Feature Sets** — Integrate near-infrared (NIR) and ultraviolet (UV) photometry alongside 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.

Looking forward, a number of promising avenues could further enhance photometric redshift estimation. First, ensemble techniques—such as gradient-boosted trees or stacked generalization combining decision trees with algorithms like random forests—may capture complementary patterns and reduce overfitting. Second, incorporating additional observational data, including near-infrared and ultraviolet photometry, as well as environmental metrics (e.g., local galaxy density or cluster membership), could break remaining degeneracies and improve robustness. Third, extending the methodology to deeper and higher-redshift surveys (such as LSST, Euclid, or JWST data) will test its scalability and effectiveness when morphological resolution degrades.

A cosmic background image featuring a dense field of galaxies. In the center, a large, bright, yellowish-white spiral galaxy is prominent, surrounded by numerous smaller, distant galaxies in various colors (blue, orange, white) and shapes. The background is a deep black space filled with these celestial objects.

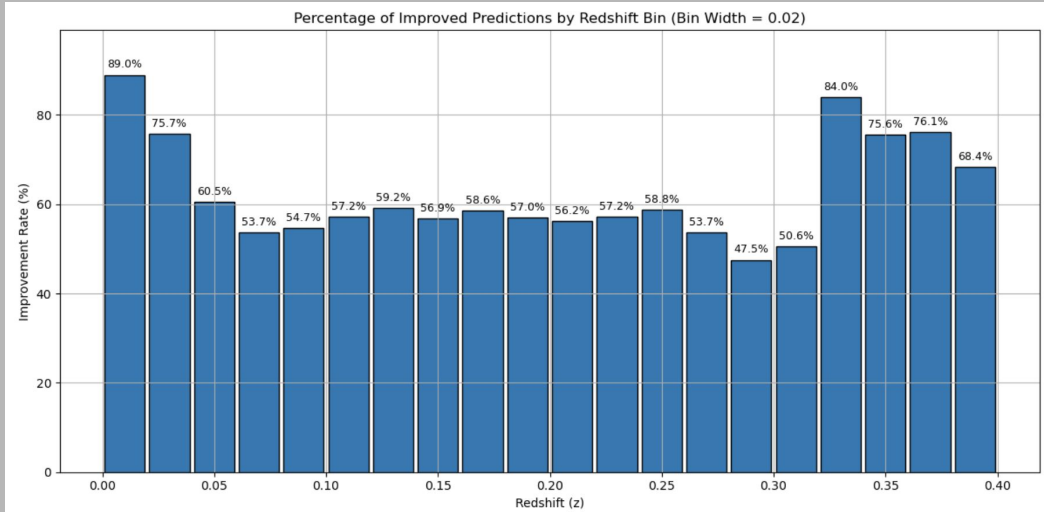
SUPPORT SLIDES

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



This graph illustrates how model performance, measured by Mean Absolute Error (MAE), varies across redshift bins for three different feature sets. In the lowest redshift bin (0.0–0.05), the combined model (green) achieves noticeably lower MAE than both the photometry-only (blue) and morphology-only (red) models. This indicates that morphological features contribute meaningful, complementary information at low redshift—likely because galaxy structure is more well-resolved, with higher-quality imaging and clearer shape indicators. However, as redshift increases, the morphology-only model's error rises sharply, suggesting that structural features become less reliable at greater distances due to fainter light, reduced resolution, and more ambiguous profiles. Overall, the graph supports the conclusion that morphological features are most helpful for redshift prediction at low redshift, where physical structure is better preserved in the data.

PERCENTAGE IMPROVED PER BIN BY COMBINATION MODEL



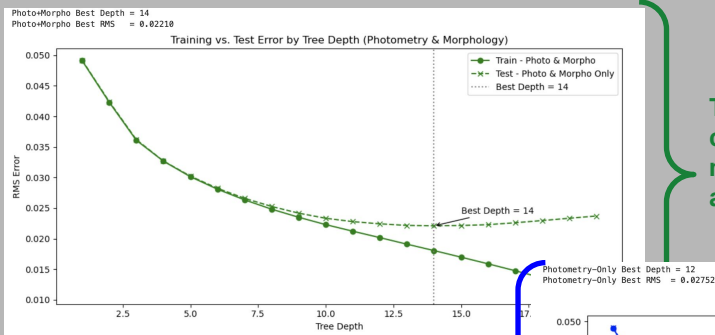
This slide shows the percentage of predictions improved by adding morphological features, grouped by redshift bin ($\Delta z = 0.02$).

As expected, improvement rates are highest at low redshift, with 89% of predictions improving in the 0.00–0.02 bin. This strongly supports our hypothesis: morphological features are especially helpful when galaxies are well-resolved and structural measurements are most reliable.

In the mid-range bins ($z \approx 0.05$ –0.30), improvement stabilizes around 50–60%, indicating that while morphology still contributes, its incremental value is reduced, likely due to declining image resolution and measurement uncertainty.

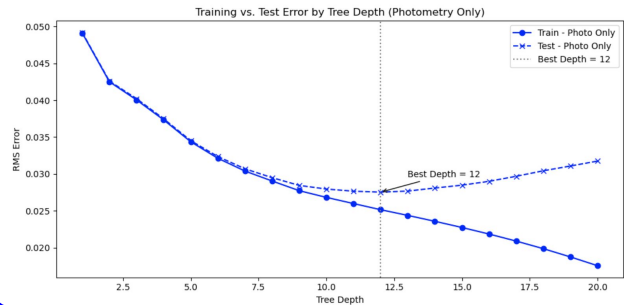
At higher redshifts ($z > 0.32$), we observe a secondary rise in improvement rates, peaking at 84%. However, this is likely an artifact of low sample counts in these bins, rather than a meaningful resurgence of morphological signal. Sparse data in these redshift ranges can introduce noise in the improvement rate calculation, so this pattern should be interpreted cautiously. I did isolate this range of redshift values and evaluate the combination model specifically against them, but didn't note performance better than photometric-only.

In summary, this slide confirms that morphological features are most beneficial at low redshift, while highlighting the need to account for sample size effects when analyzing performance trends at the high- z end.



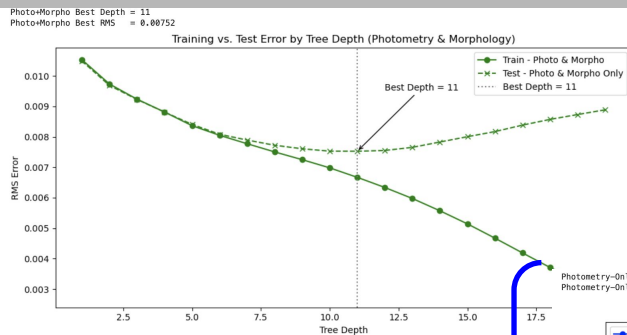
The optimal depth for the model that considers photometric and morphological features is 14, yielding a RMS error of 0.0221.

The optimal depth for the model that considers only photometric features is 12, yielding a RMS error of 0.0275.



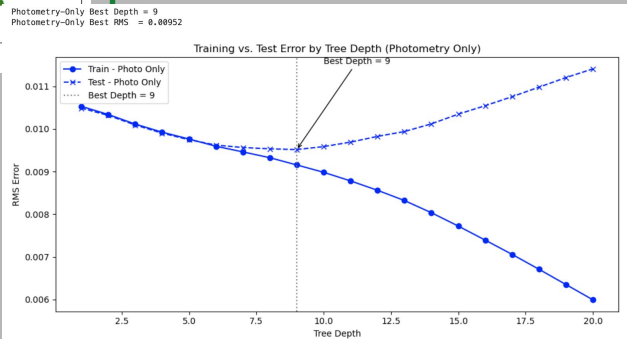
Whole range of z

These plots show RMS error across tree depths for Decision Tree models trained with (top-left) and without (bottom-right) morphological features. The optimal depth for the photo-only model is 12, yielding an RMS error of 0.0275. When morphology is included, the optimal depth shifts to 14, reducing RMS error to 0.0221



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

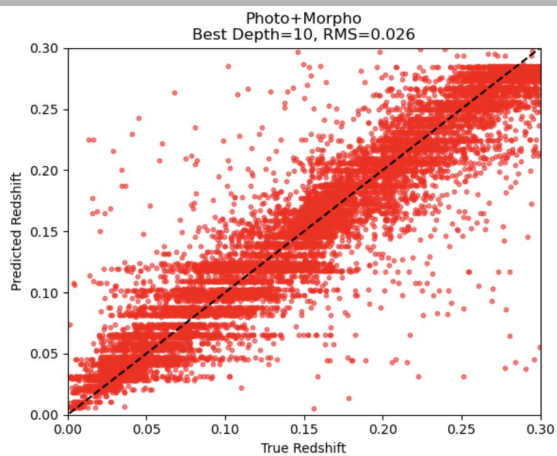
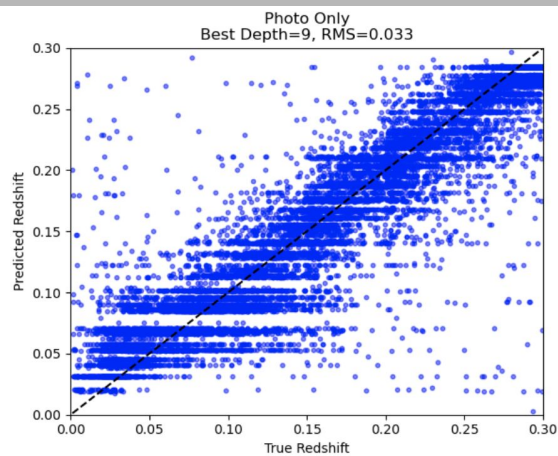
The optimal depth for the model that considers only photometric features is 9, yielding a RMS error of 0.00952.



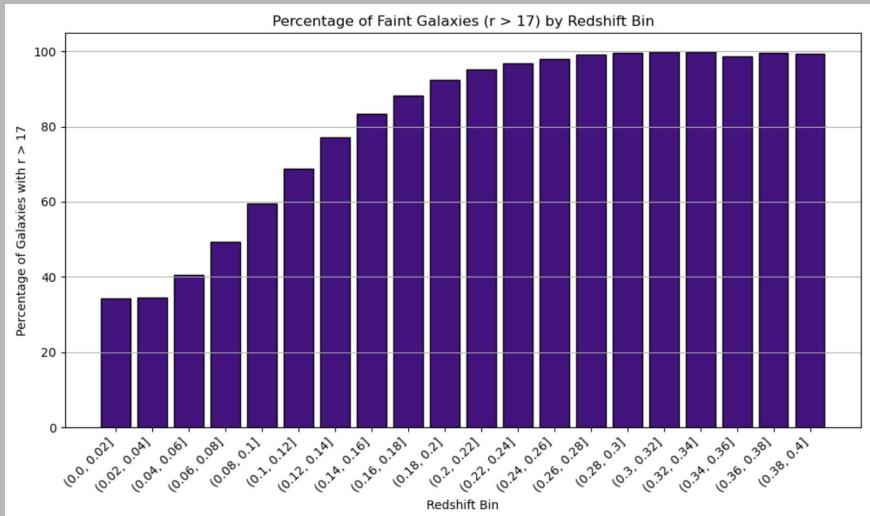
LOW z

These plots show RMS error across tree depths for Decision Tree models trained with (top-left) and without (bottom-right) morphological features. The optimal depth for the photo-only model is 9, yielding an RMS error of 0.00952. When morphology is included, the optimal depth shifts to 11, reducing RMS error to 0.00752

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



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

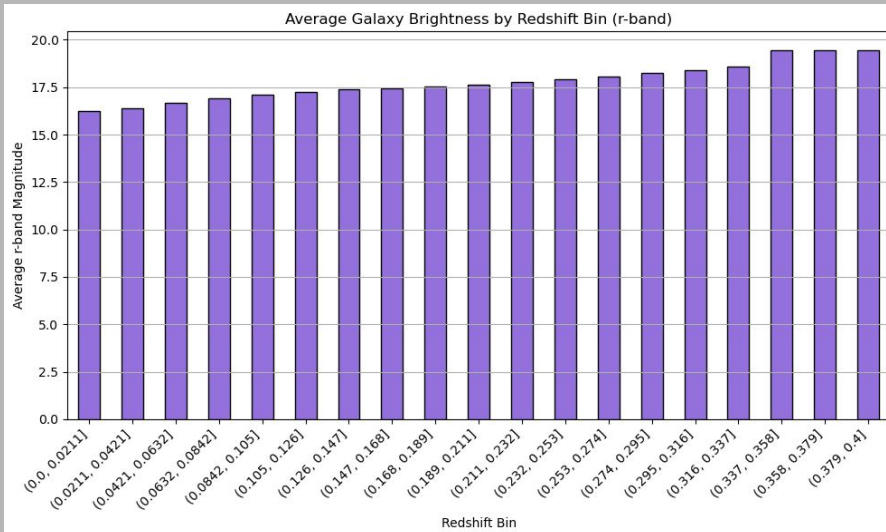


“Dim” is defined as $r > 17$.

In the SDSS photometric system, **p.r** (the r-band magnitude) is commonly used as a proxy for overall brightness, since it lies near the center of the optical spectrum and has relatively low atmospheric absorption. Lower magnitudes correspond to brighter galaxies.

Not all dim galaxies are far away — galactic dimness depends on more than just distance. While distant galaxies tend to be fainter due to the expansion of the universe and light travel time, some nearby galaxies also appear dim because they are intrinsically small, have low surface brightness, or are obscured by dust. This means dim galaxies exist at a wide range of redshifts, and their faintness doesn't always imply they are far away.

AVERAGE BRIGHTNESS OF GALAXIES BY REDSHIFT BIN



In the SDSS photometric system, **p.r** (the r-band magnitude) is commonly used as a proxy for overall brightness, since it lies near the center of the optical spectrum and has relatively low atmospheric absorption. Lower magnitudes correspond to brighter galaxies

This histogram presents the number of faint galaxies observed across various redshift intervals. As we move from left to right along the x-axis, we're essentially looking further back in cosmic time/at galaxies that are farther away from Earth. The distribution indicates that while a significant number of faint galaxies are located at lower redshifts, there's a notable presence extending to higher redshifts. This suggests that faint galaxies are not exclusively distant; some are relatively nearby but intrinsically dim due to factors like low stellar mass or limited star formation activity.

At $r > 21$:

- **Morphological measurements** start becoming unreliable around $r > 21$, depending on resolution and SNR.
- **Photometric redshift** estimates still function but lose precision at high magnitudes due to increased noise.
- beyond the **practical limit** for spectroscopic surveys like SDSS.

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

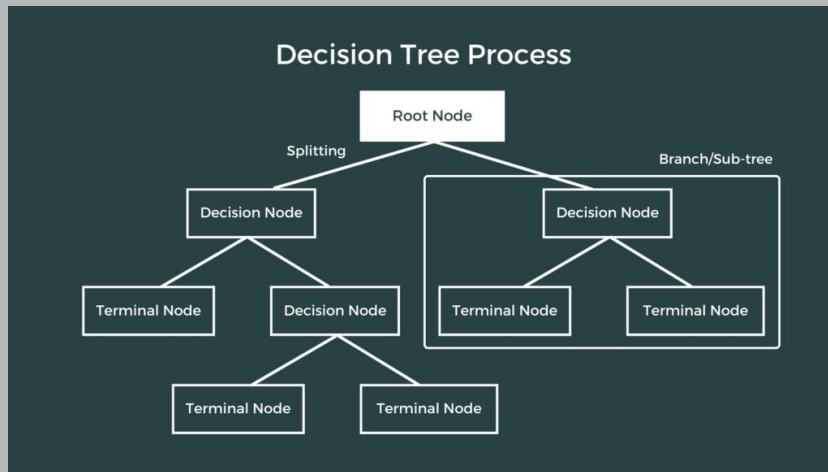
These features capture the structural characteristics and orientation of galaxies.

- **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 $+45^\circ$ 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?



A decision tree is a supervised learning algorithm used for both classification and regression tasks. It works by recursively splitting the training data based on input feature values to minimize prediction error. At each node, the model selects the feature and threshold that best separates the data according to a loss function—such as mean squared error for regression or Gini impurity for classification. These splits continue until a stopping condition is met, such as reaching a maximum tree depth or a minimum number of samples per leaf. The resulting tree structure consists of decision rules that lead to terminal nodes (leaves), where the final prediction is made—typically the average value (for regression) or majority class (for classification) of the training samples in that node. Because each path through the tree represents a set of if-then rules, decision trees offer interpretable models that can capture complex, nonlinear relationships without requiring feature scaling.