Draft version December 13, 2023
Typeset using LaTeX **modern** style in AASTeX631

Classification of Stars, Galaxies, and Quasars Utilizing Machine Learning via SDSS Photometric Data

Yijun Liu and Mingxuan Liu

ABSTRACT

Quasars are distant, luminous objects with emissions from supermassive black holes at galaxy cores, challenging to distinguish from stars. This study explores the classification of quasars, galaxies, and stars using supervised machine learning models, focusing exclusively on photometric data from the Sloan Digital Sky Survey (SDSS) (SDSS 2011). This research employs various models such as Decision Trees, Support Vector Machines, K-nearest neighbors, Random Forests, and XGBoost. Among these, XGBoost performed the best, achieving high scores in F1, recall, precision, and AUC_PR metrics, with an overall performance rate of 0.98. This performance surpasses previous studies using only SDSS data (Clarke et al. 2020; Peng et al. 2012). Notably, XGBoost's AUC_PR scores reach 1.00 for galaxies, 0.99 for stars, and 0.95 for quasars, demonstrating exceptional accuracy in classification. The study also highlights the advantages of machine learning in identifying classification patterns and methods, offering new perspectives beyond traditional astronomical approaches.

Keywords: Quasar Classification — SDSS Catalog — Machine Learning — Supervised Learning — XGBoost — Random Forest

1. INTRODUCTION

First discovered in 1963 through their radio emissions, quasars are extremely luminous and have enormous distances from Earth. Also known as Quasi-Stellar Objects (QSO), quasars have long posed a challenge in astrophysical classification due to their unresolved and indistinguishable images from stars when observed from Earth (Clarke et al. 2020). The radio emissions have been subsequently confirmed to originate from the accretion disks surrounding supermassive black holes at the cores of galaxies. Quasars are less commonly seen than stars and galaxies, yet a significant range of astronomy topics depend on the large sample of quasars (Clarke et al. 2020). Therefore, the need for precise and efficient classification methods between quasars and stars is imperative for furthering our understanding of supermassive black holes and the dynamics around Active Galactic Nuclei (AGN).

While spectroscopic data can offer definitive signatures to differentiate between quasars and stars, such observations are labor-intensive, time-consuming, and fail to provide comprehensive sky coverage because the spectroscopic observation is only limited to bright objects (Makhija et al. 2019). In contrast, photometric data is observationally cheap and easily accessible, due to many ongoing sky surveys inside/outside the Earth. Consequently, the prospect of leveraging solely photometric data for classifying quasars and stars appears particularly appealing, especially for informing future sky surveys or generating catalogs for follow-up spectroscopic observations. Moreover, we acknowledge the necessity to differentiate galaxies, given that quasars can appear as extended sources similar to galaxies when they are in close proximity to Earth. Therefore, our models will also seek to accurately classify galaxies alongside quasars and stars.

In the present study, we aim to develop supervised machine learning models for the classification of quasars, galaxies, and stars, only using photometric data from the Sloan Digital Sky Survey (SDSS). Different from traditional methodologies that combine data from multiple astronomy surveys, our study exclusively utilizes data from SDSS. This novel approach enables us to accurately classify objects from historical SDSS data releases as well as newly observed objects. Our study expands the range of machine learning models explored in this domain. We have investigated various machine learning models including Decision Trees, Random Forests, Support Vector Machines, K-Nearest Neighbors, and XGBoost, surpassing the breadth of models typically used in previous studies. Overall, our study contributes to quasar classification by training a variety of machine learning models solely on SDSS data, aiming to improve the accuracy and efficiency of classification for both historical and upcoming SDSS survey.

2. LITERATURE REVIEW

Various studies have employed machine learning techniques to categorize celestial objects using photometric data. For instance, Clarke et al. utilized the SDSS and WISE datasets with a Random Forest classifier, achieving a precision above 0.90, a recall of 0.87, and an F1 score of 0.89 for SDSS solely (Clarke et al. 2020). Similarly, Jin et al. employed Pan-STARRS and AllWISE databases with XGBoost and SVM classifiers for quasar selection, reaching an accuracy of 99.46% using specific color features (Jin et al. 2019). Peng et al. also focused on quasar classification using SDSS data and an SVM-based system, attaining an accuracy of 93.21% for the testing data (Peng et al. 2012). Makhija et al. explored the use of neural networks and a novel GAN to classify stars and quasars in the GALEX and SDSS catalogs, achieving an accuracy range between 91% and 100% (Makhija et al. 2019).

Most of the previous studies have employed more than one survey data, which makes it impossible to apply the model to newly observed objects of a single survey. In addition, although Peng et al. used only the SDSS survey data, they only exper-

imented with the SVM model, but not the other models which may provide better performance. Therefore, our study can contribute to the field by utilizing only the SDSS photometric data and exploring more machine learning models to classify stars, galaxies, and quasars.

3. METHODS

This section will introduce the machine learning models that have been used in this study. The dataset used is constituted of continuous variables. Since there are three different categories for each data point, this study mainly employs supervised machine-learning models that are capable of multi-class classification. All the models are implemented via the python scikit-learn library (Pedregosa et al. 2011).

3.1. *Decision Tree*

The Decision Tree model is a tree structure where each internal node represents a feature test, each branch a feature value, and each leaf a class label. Paths from root to leaf represent classification rules. Its non-parametric nature makes it possible to focus on the high-dimensional data and it allows for multi-class classification.

3.2. *Support Vector Machine*

Support Vector Machine (SVM) can find the optimal hyperplane to separate different classes in the feature space. This is particularly beneficial for this study as it effectively handles the complex boundaries that can exist between different types of astronomical objects in the feature-rich SDSS dataset.

3.3. *K-Nearest Neighbors*

K-Nearest Neighbors (KNN) identifies the 'k' closest data points to a given sample and classifies it based on the majority class among these neighbors. This approach is suitable for this study as similar types of astronomical objects usually have similar feature values, making KNN effective for classifying stars, quasars, and galaxies.

3.4. *Random Forest*

The Random Forests classifier is an ensemble machine learning technique that creates multiple decision trees at training time and delivers the class that is the mode of the classes in classifying tasks. This algorithm was selected for our dataset because it can substantially reduce variance and overfitting on the costs of sightly increasing the bias. Variance and overfitting are common problems in decision tree models.

3.5. *XGBoost*

XGBoost (eXtreme Gradient Boosting) is an ensemble technique. It sequentially constructs multiple decision trees, with each subsequent tree focusing on addressing the errors made by its predecessors. As it combines multiple weak learners (decision trees) into a strong one, this model improves the ability to generalize well. XGBoost

also includes L1 (Lasso) and L2 (Ridge) regularization, which helps in reducing overfitting in Decision Tree and Random Forest. The iterative correction of residuals in XGBoost can also lead to a more accurate model than standard Decision Tree or Random Forest. Therefore, XGBoost is advantageous for its high efficiency and performance, making it a potent tool in predictive modeling, especially in situations with complex and large datasets.

4. RESULTS

4.1. *Data description*

We utilized the publicly accessible data from the Sloan Digital Sky Survey (SDSS) Data Release 8 (DR8). SDSS, using a 2.5-m-wide-angle optical telescope at Apache Point Observatory, has created the most detailed three-dimensional maps of the Universe ever made and has been one of the most successful surveys in the history of astronomy (SDSS 2011). The eighth data release, following the Data Releases 1-7 of SDSS-I/II is available as of January 2011 (SDSS 2011). SDSS DR8 contains both the imaging data taken by the SDSS imaging camera and spectral data taken by the SDSS spectrograph (SDSS 2011).

Each object in SDSS DR8 has measurements of photometric fluxes through u, g, r, i, and z filters, and a spectrally confirmed classification among star, galaxy, and quasar (SDSS 2011). In addition, each entry of the database also comes with technical details about the object's name, coordinates, and telescope systematics. For our project, we aim to utilize solely the photometric data from SDSS DR8 for machine learning-based classification of stars, galaxies, and quasars. Our dataset contains 84,962 rows and 24 columns, reaching a total dimensionality of 2,038,848. In this dataset, the numbers of stars, galaxies, and quasars are highly imbalanced. There are 39,872 stars, 39,609 galaxies, and 5,516 quasars in total.

4.2. *Preprocessing and Feature Extraction*

The dataset queried from the SDSS database contains contains telescope systematic information such as plate number, run number, camera number, and object names. We removed these features since how the celestial objects are observed does not change the physical property of objects. In addition, we also removed each object's coordinate information. Based on the cosmology principle, our universe is homogeneous and isotropic, and therefore the distribution of quasars does not depend on which direction we observe the sky.

After removing these irrelevant features, we have 11 features left: the photometric fluxes through each filter and the spectrally confirmed object classification. For SDSS survey, there are five different filters: u (3551Å), g (4686Å), r (6166Å), i (7480Å), and z (8932Å), representing wavelengths from ultraviolet to infrared regimes. For the flux through each band, there are two ways to estimate the magnitudes: the Point Spred Function (PSF) magnitude and the Petrosian magnitude. The PSF magnitude represent the response of the imaging system to a singular point source such as a

star, and therefore is better suited to measure the brightness of point sources. The Petrosian magnitude, on the other hand, is better suited for galaxy (extended source) because it measures galaxy fluxes within a circular aperture whose radius is defined by the shape of the azimuthally averaged light profile (SDSS 2011). The photometric data are continuous and don't need normalization because the values are very close to each other.

Based on these photometric magnitudes, we explicitly calculated the color indices $u - g$, $g - r$, $r - i$, and $i - z$, following the conventional astronomy methodology. Calculating the ANOVA p-value for each remaining feature, we decided to remove $r - i$ and $i - z$ colors for both magnitudes because their p-values are much larger than the other features. At the end, the remaining 16 features include: the original magnitude through each filter for both PSF and Petrosian, the $u - g$ and $g - r$ color indices for both PSF and Petrosian, and the classification target.

4.3. *Exploratory Data Analysis*

As discussed in section 4.2, the features remained in the dataset effectively forms a two dimensional color-color space; that is, $u - g$ color against $g - r$ color. Therefore, it is tempting to plot the distribution of stars, galaxies, and quasars on a color-color diagram (Figure 1) to examine their distinguishability based on color characteristics alone.

In the diagram, the left panel illustrates the star and quasar distributions, indicating a relatively clear separation between them in this color space. This distinction aligns with physical expectations due to their different physical processes, resulting in different emission lines. The middle panel compares galaxies and quasars, showing a more blended distribution. This overlap reflects the physical reality where quasars are active galactic nuclei residing within galaxies. The right panel, including all three object types, further demonstrates their overlap in the 2D color space, underscoring the challenge of distinguishing them based solely on conventional astronomical methods. This complexity necessitates the application of machine learning algorithms to detect patterns and relationships not immediately apparent through astrophysical theory, thus enabling accurate classification of these astronomical objects.

4.4. *Modeling*

Initially, we examine the model performance for all models mentioned in the Method3 with no hyperparameter specified and 20 Monte Carlo runs with 70-30 train-test split. The evaluation metrics include precision, recall, f1, and AUC_PR due to the imbalanced data discussed in 4.1 Data Description, with weighted averages for different classes. We also report the time taken for one run in each model.

The initial model performances are shown as below:

Since Random Forest and XGBoost's performances exceed all the rest of the models in all metrics, we decided to proceed with Random Forest and XGBoost for future hyperparameter tuning and feature selection. Although Random Forest has the best

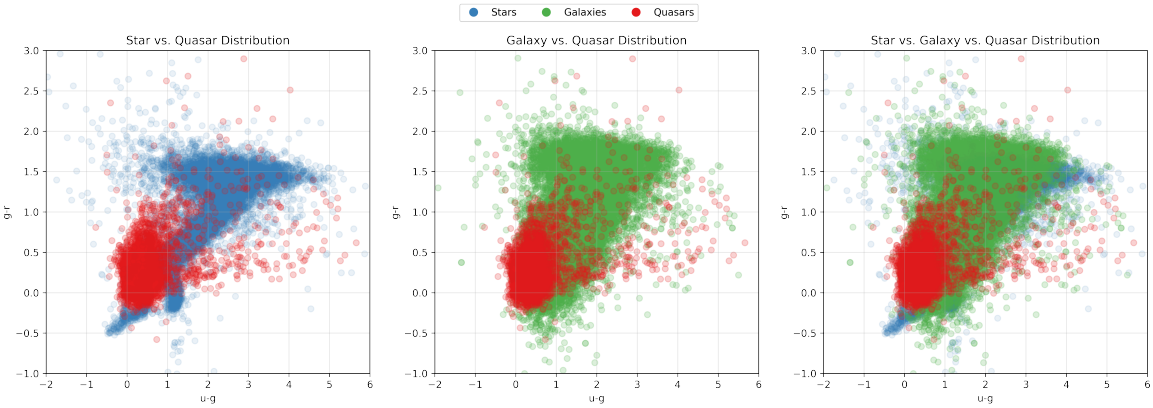


Figure 1: 2D color-color diagram. It shows the distribution of stars vs. quasars, galaxies vs. quasars, and stars vs. galaxies vs. quasars.

Table 1: Initial Model Performance Comparison

Model	Precision	Recall	F1	AUC_PR	Time(s)
Random Forest	0.978815	0.979077	0.978862	0.980403	16.241741
XGBoost	0.978248	0.978420	0.978311	0.980200	0.951049
KNN	0.976121	0.976458	0.976128	0.970993	1.431806
Decision Tree	0.963952	0.963919	0.963932	0.929209	1.305646
SVM	0.957271	0.957742	0.957419	0.891113	189.938522

AUC_PR is short for Area Under the Precision-Recall Curve. Time is the time taken for one run of the model in terms of seconds.

overall performance, it is almost 16 times more expensive than XGBoost. Therefore, we proceed with both Random Forest and XGBoost.

The hyperparameters for Random Forest and XGBoost are tuned through GridSearchCV in scikit-learn library (Pedregosa et al. 2011). For both models, the scoring criteria is the weighted F1 score, and the cross-validation is set to 5. For Random Forest, we choose to tune n_estimators (200), max_depth (10), min_samples_split (4), and min_samples_leaf (2); for XGBoost, we choose to tune n_estimators (100), learning_rate (0.5), max_depth (10), reg_alpha (1), and reg_lambda (10). The final best parameter values are included in the brackets for each parameter.

4.5. *Feature Selection*

In addition to the filter method described in the 4.6 Preprocessing and Feature Extraction, this study also performs recursive feature elimination (RFE). RFE works by recursively removing the least important features and building the model on the remaining features. The RFE is implemented through the scikit-learn library with scoring criteria of weighted F1 and crossing validation set to three (Pedregosa et al.

2011). Figure 2 shows the relationship between number of features selected and the F1 score for both models.

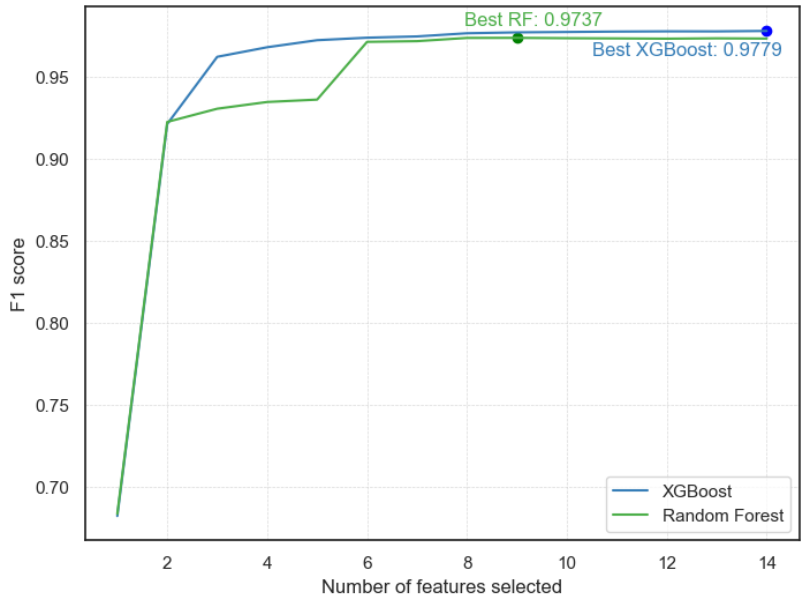


Figure 2: RFE for XGBoost and Random Forest

We will be using the features that yield the best score for both Random Forest and XGBoost, which are all features for XGBoost and 9 features for Random Forest.

4.6. *Results*

Firstly, we compared the models' overall performances in XGBoost and Random Forest in Table 2. Both F1, precision, recall, and AUC_PR are stable across both models, with a small 95% confidence interval. They are all performing well in terms of their high scores across all metrics.

Table 2: Comparison of XGBoost and Random Forest Overall Performance.

Model	**F1 Score**	**Precision**	**Recall**	**AUC-PR**
XGBoost	0.980 (0.97941, 0.98029)	0.980 (0.97935, 0.98023)	0.980 (0.97958, 0.98043)	0.980 (0.97923, 0.98124)
Random Forest	0.974 (0.97385, 0.97487)	0.974 (0.97375, 0.97477)	0.975 (0.97422, 0.97519)	0.973 (0.97167, 0.97397)

The values are averaged from 10 Monte Carlo runs. The brackets under the averaged value define the 95% confidence interval from the Monte Carlo runs.

The Figure 3 shows the feature importance of the features in XGBoost and Random Forest. Since Random Forest only has 9 features after RFE, some of the features do not have values from Random Forest. The psfMag_i feature has the highest feature importance in both models.

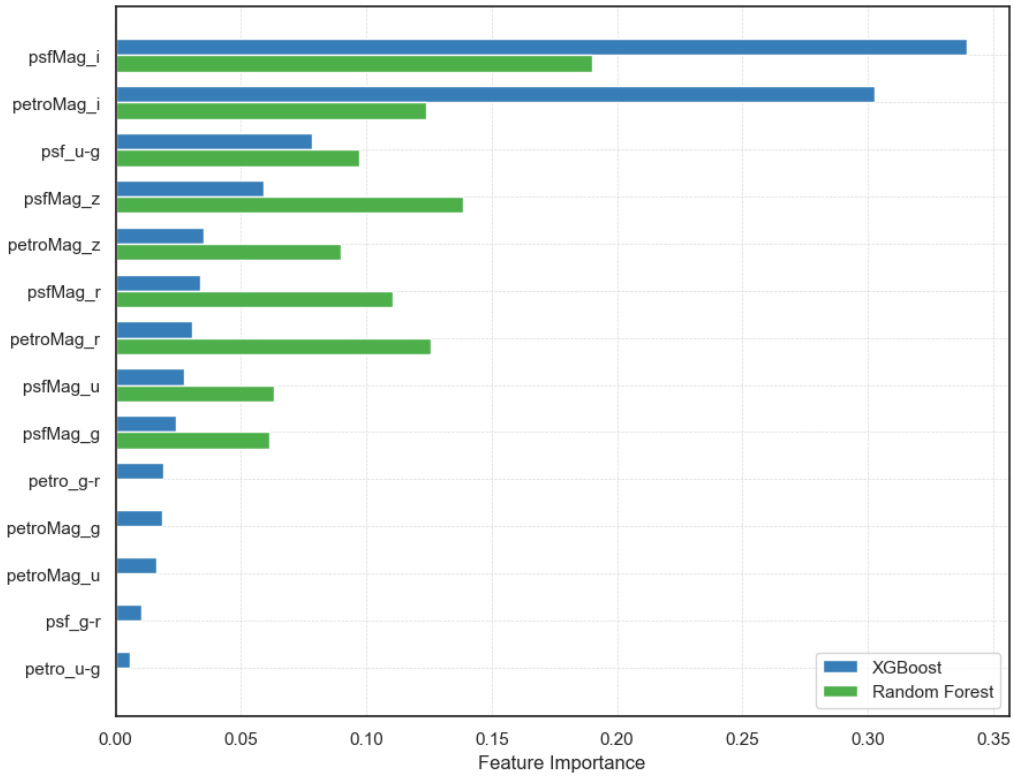
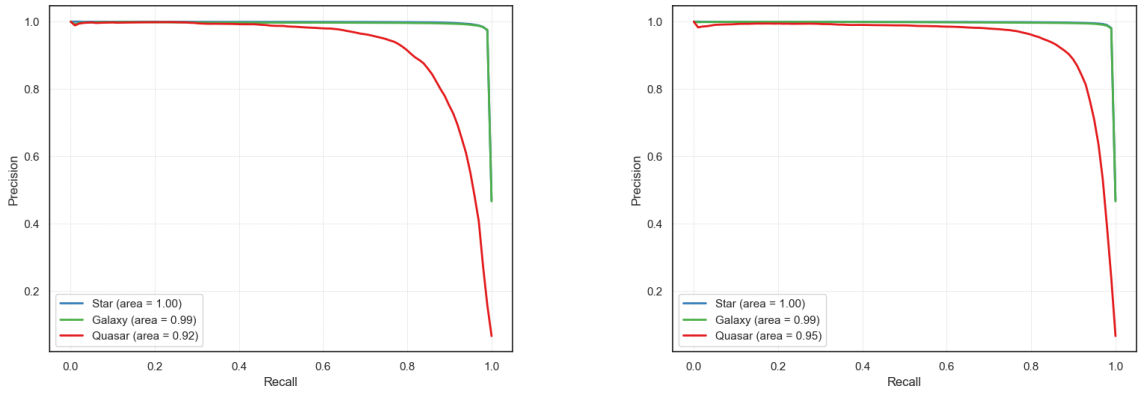


Figure 3: Feature Importance for Random Forest and XGBoost. The features are ranked by XGBoost's feature importances.

Since this study aims at a multi-class classification, it is important to understand the model's performance within each class. Figure 4 shows how each model performs within each class (star, galaxy, and quasar). In both models, star has reached a 100% classification, and galaxy has reached 99%. Nevertheless, quasar has the worst performance with 95% in the XGBoost and 92% in Random Forest.



(a) Random Forest (b) XGBoost

Figure 4: AUC_PR by Class (Star, Galaxy, and Quasar)

5. DISCUSSION

Our research demonstrates that machine learning models can uncover patterns and relationships beyond astrophysical theories, achieving high accuracy in classifying stars, galaxies, and quasars. We found that the XGBoost model has the best overall performance as well as the performance within classes. The model is also stable with the input data. However, comparing the performance within each class, quasar seems to be the most difficult to classify.

Our results not only provide a model to quickly and accurately classify objects, but also provide us the insights into astronomy through the lens of machine-learning. Notably, as discussed in the Results section, the feature importance rankings for the Random Forest and XGBoost models (Figure 3) reveal intriguing insights. The XGBoost model prioritizes original magnitudes over explicitly calculated color indices, suggesting its exploration in novel, non-linear combinations of the original features. This approach significantly outperforms traditional methods, which is also demonstrated in Figure 1 where different object categories are indistinguishable in the 2D color-color space. This result underscores the necessity of integrating machine learning into astronomy, especially for problems where analytical solutions have been elusive.

Moreover, future work should focus on improving the classification of the quasar class, since classifying the galaxy and star has already achieved nearly perfect performances. Looking forward, there is potential to enhance our model's performance. Figure 5 shows the distribution of stars, galaxies, and quasars across redshift z, which measures a celestial object's recession speed and thus serves as a proxy for distance. This figure highlights the varying redshifts among these objects, with quasars typically at higher redshifts and greater distances, often seen in the younger universe. However, redshift is determined via spectroscopy, which violates our principle to classify objects solely based on photometric data. Therefore, given redshift's potential as a distinguishing feature, our future goal is to employ machine learning to predict the photometric redshift for each object, thereby refining our classification accuracy.

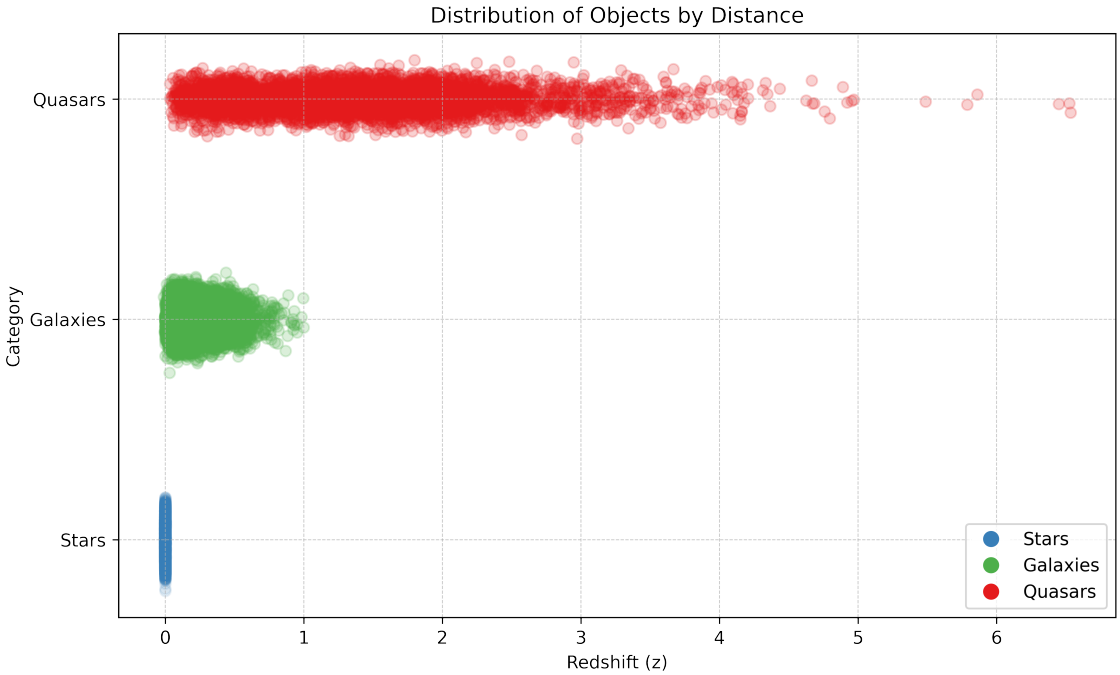


Figure 5: Jittered distribution of stars, galaxies, and quasars along redshift (z). Stars mostly have 0 redshift, and galaxies have redshift limited to 1, representing the local universe. Quasar, however, can have much higher redshift indicating their enormous distances from Earth.

APPENDIX

A. CONTRIBUTION

We contributed equally to this study.

Yijun Liu: Performed machine learning in the cleaned dataset, including the initial model selection, hyperparameter tuning, recursive feature elimination, and final models. I also drafted the final report.

Mingxuan Liu: Queried the dataset from SDSS DR8, cleaned and processed the data based on domain knowledge, conducted exploratory data analysis, and drafted the final report.

B. CODE RELEASE

The code and dataset used for this project are publicly released on GitHub at https://github.com/Mingxuan-Liu/Star-Type-Classifier under the MIT license.

REFERENCES

Clarke, A. O., Scaife, A. M. M., Greenhalgh, R., & Griguta, V. 2020, Astronomy & Astrophysics, 639, A84, doi: 10.1051/0004-6361/201936770

Jin, X., Zhang, Y., Zhang, J., et al. 2019, Monthly Notices of the Royal Astronomical Society, 485, 4539, doi: 10.1093/mnras/stz680

Makhija, S., Saha, S., Basak, S., & Das, M. 2019, Astronomy and Computing, 29, 100313, doi: https://doi.org/10.1016/j.ascom.2019.100313

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, J. Mach. Learn. Res., 12, 2825–2830

Peng, N., Zhang, Y., Zhao, Y., & Wu, X.-b. 2012, Monthly Notices of the Royal Astronomical Society, 425, 2599, doi: 10.1111/j.1365-2966.2012.21191.x

SDSS. 2011, Sloan Digital Sky Survey Data Release 8 (DR8), [https://www.sdss3.org/dr8/]