HIGH SCHOOL EDITION
Journal of Student Research

QUAS(AI)R: A Novel Machine Learning Algorithm to Predict
X-Ray Brightness in Active Galactic Nuclei

Deepthi Kumar[1] and Tony Rodriguez[#]

[#]Advisor

ABSTRACT

Active Galactic Nuclei (AGNs) are a compact region at the center of galaxies that emit more energy than the rest of the galaxy itself. They emit light across the electromagnetic spectrum, from radio waves to optical light to high-energy X-rays. AGNs indicate the existence of highly energetic phenomena in the nucleus of the galaxy. Although AGNs were identified 60 years ago, our knowledge about their physical properties is limited. Quasars, a subtype of AGNs, provide some of the most intense forms of X-rays, which are among the most energetic light known. Furthering our understanding of X-rays in the dynamic environments of quasars will add to our understanding of how to benefit from their use on Earth. In my project, I study the X-ray brightness in quasars to develop six types of regression-based machine learning models for the X-ray brightness predictions. These six models were Stochastic Gradient Descent (SGD), Random Forest, Ridge, Lasso, Bayesian and the baseline linear regression model, built on the scikit-learn Python package. The training/testing split on the MILLIQUAS dataset was 80/20 percent, and each model was tuned on model-specific hyperparameters. Benchmarked with the normalized mean absolute error (NMAE), the top three performing models were the Bayesian (0.022%), Ridge (0.180%), and Lasso (0.183%), with the baseline NMAE at 0.284%. With this, we can learn more about the evolution of galaxies in the early Universe and understand how these dynamic environments came to be.

Introduction

Active Galactic Nuclei (AGN) present unique observational signatures that cover the full electromagnetic spectrum over more than twenty orders of magnitude in frequency. They are energetic astrophysical sources powered by accretion onto supermassive black holes in galaxies. Quasars are very bright, distant and active supermassive black holes that are millions to billions of times the mass of the Sun. Quasars' light outshine that of all the stars in its host galaxy combined. This project develops different types of regression-based machine learning models to predict the observed X-ray brightness emission of quasars in the X-ray spectra given the optical, near infrared and redshift properties. These six models were Stochastic Gradient Descent (SGD), Random Forest, Ridge, Lasso, Bayesian and the baseline linear regression model. X-ray brightness predictions for quasars will enable researchers to learn more about the patterns of these energies and predictable methods to harness them. The X-ray spectroscopy of high redshift quasars can help determine the structural evolution of quasars and better understand the evolution of galaxies across cosmic time. Furthermore, this project utilized supervised machine learning for regression to aid with the X-ray brightness predictions. Each model was refined with model-specific hyperparameter tuning with an overall training/testing split of 80/20. Each model generated a mean absolute error (MAE), which tells us the average size of the error/mistake in prediction. The normalized mean absolute error (NMAE) value was then taken from the MAE. However, the NMAE was more useful because it provided a more balanced value in order to help compare the accuracy of each model against each other. The MILLIQUAS dataset was used in this study, which comprises over 400,000 type-I quasars with numerical data for each quasar.

Background

Active Galactic Nuclei (AGNs) are a compact region at the center of galaxies that emit more energy than the rest of the galaxy itself. They have extremely high luminosities (up to Lbol ≈ 1048 ergs s−1), making them the most powerful non-explosive sources in the Universe. Therefore, they are also visible up to high redshifts (z = 7.1).

AGNs are also known as active supermassive black holes (SMBH) that emit jets and winds - actively gaining mass. Supermassive black holes are classically defined as black holes with a mass above 100,000 solar masses.

All quasars are AGNs but not all AGNs are quasars. A quasar is an AGN that is viewed from a particular angle. In 1962, the first quasar was discovered because it looked extremely different from any other celestial object qualitatively. Furthermore, in the graph of wavelength vs. flux (brightness) for stars, we see dips, however, in the same graph for quasars we see spikes. The dips represent absorption, while the spikes represent emission. There is no other area in astrophysics that relies more heavily on multi-wavelength studies than the field of AGNs.

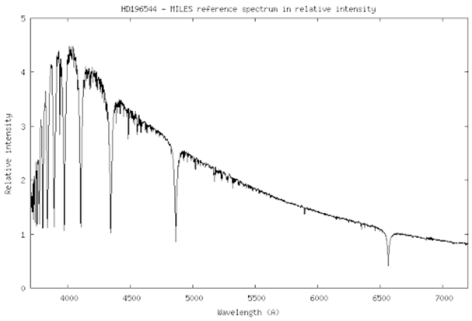


Figure 1. Graph of wavelength vs. brightness for stars, Source: hantsastro.org.uk

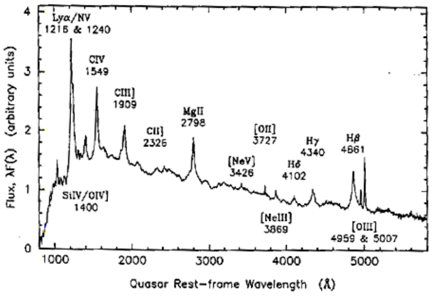


Figure 2. Graph of wavelength vs. brightness for quasars, Source: NASA

Recent international campaigns which monitor specific AGNs across various wavebands, are the key contributors to most of our knowledge of correlations between continuum (overall shape) and emission line properties. Emission lines emphasize the fact that glowing hot gas emits lines of light, while absorption lines happen when cool atmospheric gas absorbs the same lines of light. A similar pattern is illustrated through stars versus quasars. Stars keep getting colder as they move from the center - they are absorbing heat from the center, while in quasars the super heated gas is on the outside of the supermassive black holes so it has to emit the energy, thus why quasars are such energetic objects in the X-ray spectra.

The AGN contains a supermassive black hole (SMBH) surrounded by an accretion disk. The accretion disk feeds both the black hole and dusty torus. Through the Hubble Space Telescope we can see the jets of ionized material being launched out of both ends of the AGN in radio images. The accretion disk is formed by rapidly rotating gas which slowly spirals onto a central gravitating body. This is how all the matter is making its way to the black hole. AGNs are powered by the extracted gravitational energy of inflating matter.

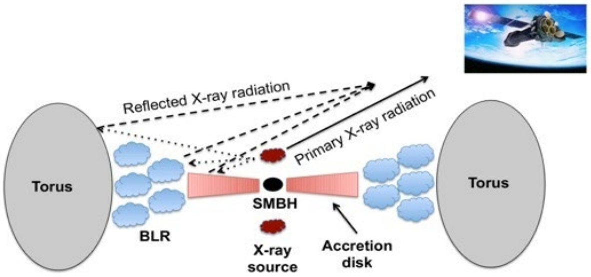


Figure 3. Schematic representation of X-ray production in AGN, Source: isdc.unige.ch

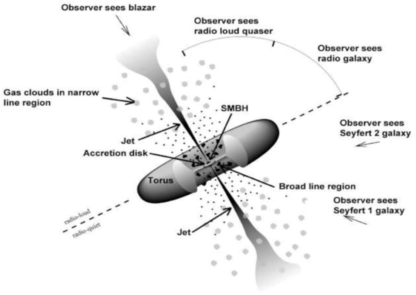

Observer sees blazar
Observer sees radio loud quaser
Observer sees radio galaxy
Gas clouds in narrow line region
SMBH
Observer sees Seyfert 2 galaxy
Jet
Accretion disk
Torus
Broad line region
Observer sees Seyfert 1 galaxy
radio-loud
radio-quiet
Jet

Figure 4. A model of the AGN from different angles, Source: NASA

The electromagnetic spectrum is the full range of waves that transfer energy, organized by wavelength. It comprises different types of light/radiation. This spectrum ranges from radio waves to optical light to high-energy X-rays. On the electromagnetic spectrum, the human eye can only see visible light. On the other hand, AGNs emit the entire spectrum, due to the complex interactions between the accretion, magnetic fields, scattering, and jet activity properties. One key attribute of the dataset is “Z” or the redshift, which is the light when an object is moving away from us - as opposed to the blueshift which is the light when an object is moving towards us. Spectroscopy is the study of the absorption and emission of light and other radiation by objects, which is helpful when studying the physical properties of AGN. Hubble’s law states that redshifts in the spectrum of distant galaxies (and hence their speeds of recession) are proportional to their distance - velocity is proportional to distance because the universe is expanding, which is the basis for quasars.

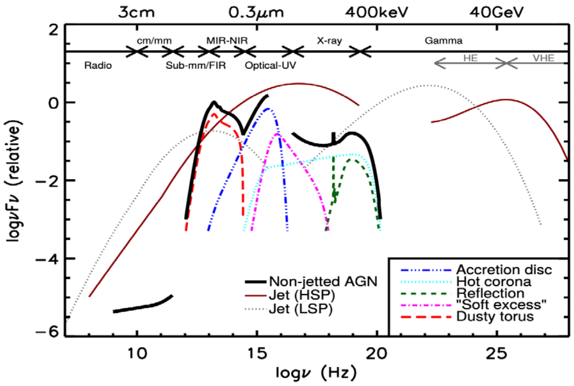


Figure 5. Schematic representation of the spectral energy distribution (SED) of AGN. The primary emission emanating from the accretion disk of the AGN, reaches its highest intensity in the ultraviolet region. Source: Harrison (2014)

Previous Studies

Prior studies have been done towards predicting X-ray brightness for quasars. However, they used less sophisticated datasets and models. In my project, I used six different regression-based models, while previous studies have only been able to use two to three max, with models such as Random Forest and just the baseline or ‘vanilla’ linear regression mode. Additionally, other studies have done brightness prediction for various energetic objects in the universe, including pulsars, XRB, etc., while my project focuses solely on quasars. My project also uses the most updated version of the MILLIQUAS quasar catalog dataset (updated in 2023), which prior studies have not been able to utilize. Instead, they have primarily worked with radio-selected AGN. The original way to find AGN was through radio selection, however, before we were only able to work with ‘radio loud’ AGN, not ‘radio quiet’ AGN. The primary difference between the two types of AGN is that the former is characterized by powerful radio emission.

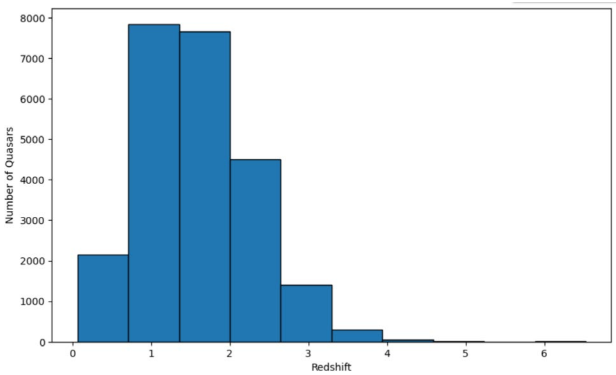
Dataset

This study uses the MILLIQUAS (The Million Quasars) dataset to predict the X-ray brightness of quasars in the X-ray spectra. The MILLIQUAS dataset is a compendium of 907,144 type-I QSOs and AGN. However, due to the fact that not all AGN are quasars, only the data for quasars was used. There were 23940 rows and 18 columns, overall producing 430920 total readings of quasar data. I used three different types of light properties available through the dataset, for predicting the X-ray brightness - the redshift, near infrared and optical light. I used three different wavelengths of near infrared light - short, medium and large wavelengths. For optical I used two different types of magnitudes - the red magnitude and the green magnitude, which conveys how bright the quasar is in either red or green optical light. To split the data across training and testing, I used 80% of the data for training and 20% of the data for testing. To help clean through the data I dropped all null values for readings of light. These null values most likely occurred due to the fact that the specific reading was not verified/taken for the certain quasar.

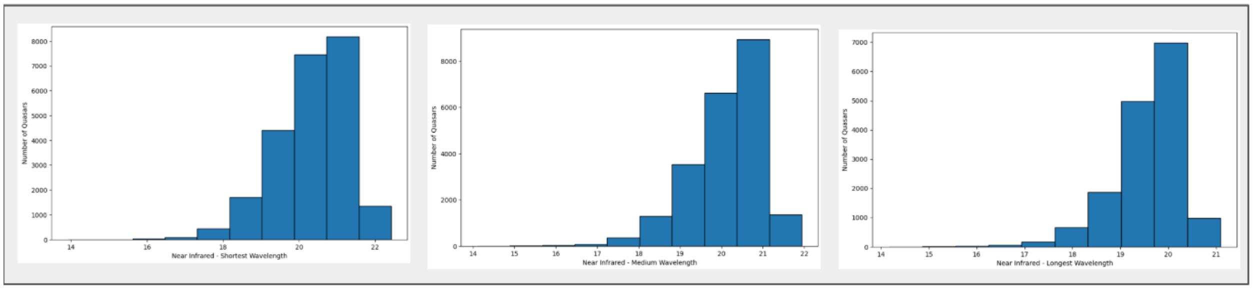
Table 1. Guide to Features

Feature	Explanation

RA	Right Ascension, Declination: Coordinates on the sky (longitude vs latitude)
DEC	
NAME	Name
TYPE	X-ray bright quasars
Z	Redshift: the larger the value, the farther away it is.
ZCITE	Where redshift information is coming from
SC_EP_8_FLUX	X-ray flux (brightness)
SC_EP_8_FLUX_ERR	Error in X-ray flux
gmag	Green magnitude (how bright it is in green optical light)
e_gmag	Error in green magnitude
rmag_x	Red magnitude
e_rmag	Error in Red magnitude
imag	Near infrared (shortest wavelength)
e_imag	Error in near infrared (shortest wavelength)
zmag	Near infrared (medium wavelength)
e_zmag	Error in near infrared (medium wavelength)
ymag	Near infrared (longest wavelength)
e_ymag	Error in near infrared (longest wavelength)



Graph 1. Visualization of redshift in relation to the amount of quasars



Graph 2. Visualization of the infrared data for all three wavelengths

A major aspect of this project was feature identification, which is illustrated in the above figures. Redshift is a crucial feature in this dataset, which is light when an object is moving further away. Another feature that is important is the log X-Ray Flux, which was chosen after noticing the strong correlation as demonstrated on the histogram on the right versus the left. Infrared magnitudes of short, medium, and long were all demonstrated as correlated with the presence of quasars through the histograms. Finally, through these histograms, additional features of red and green magnitude were discovered as important.

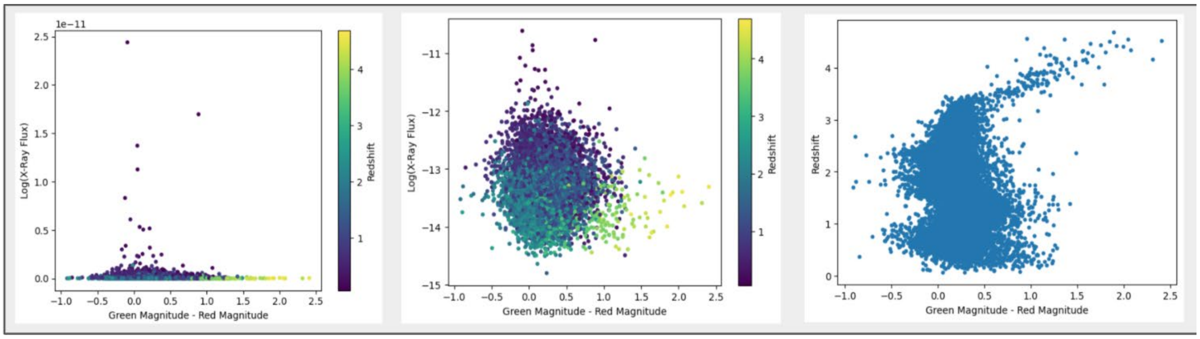


Figure 6. The color plots for the difference between the green and red magnitudes. A property called "color" is defined as the difference between magnitudes. Figure 6a shows the color plots with respect to redshift, as a function of the normal X-ray flux. Figure 6b shows this same data, however as the log(X-ray flux). Figure 6c shows the color as a function of redshift. The small range of color at the top and large range of color at the bottom of figure 6b, shows that it is much harder to predict X-ray flux of fainter quasars.

Methods

Data Preprocessing

After loading the MILLIQUAS dataset into my Google Collab notebook, I dropped all 'null' values in the dataset - values which were left blank, and therefore, do not tell us any valuable reading for the quasars. I also had to do data cleaning, mainly to ensure that the X-ray flux error was always greater than three and that the error in any magnitude is less than 0.2. Lastly, I performed data visualization by creating histograms and scatterplots for comparisons of the data attributes.

Model Development and Evaluation

After researching various aspects of AGNs and quasars, including relevant attributes for their X-ray brightness prediction properties, I decided on six different regression based models. These models were Stochastic Gradient Descent (SGD), Random Forest, Ridge, Lasso, Bayesian and the baseline linear regression model, each built on the scikit-learn Python package. I split and shuffled the dataset into a training set and a testing set, the former containing 80% of the dataset and the latter containing 20% of the dataset. I ensured to use a computer to support the Google Collab Python notebook, so that I would be able to import the relevant models. I trained and tested each type of model on both the training and testing dataset, and kept reshuffling the data and repeating these steps for a total of four trials. I tried to observe key characteristics of each of the models during training and testing, for performance scalability. Every model, other than the baseline linear regression model was tuned on model-specific hyperparameters (i.e. 'alpha'), in order to provide the best results. Models were evaluated based on mean absolute error (MAE) and normalized mean absolute error (NMAE). Lastly, I analyzed each of the regression-based models' results and ranked them in their X-ray brightness predictions performance in the descending order of performance, by their normalized mean absolute error (NMAE).

Results and Discussion

The baseline linear regression model had a MAE of 0.397% and a NMAE of 0.284%. However, the NMAE value is more useful because it provides a more balanced value in order to help compare the accuracy of various other models. It is important to note that the NMAE can be regarded as the accuracy after subtracting by 1. As this is the 'vanilla' linear regression model, no further hyperparameters could be tuned in order to make this model more efficient.

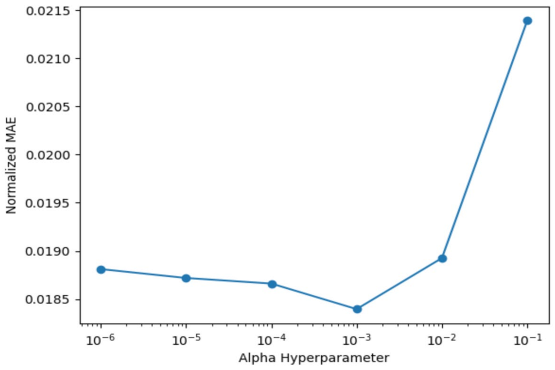
Overall, the Stochastic Gradient Descent (SGD) regression model had the highest training time, and had results that were greatly fluctuating during hyperparameter tuning. The initial NMAE was 0.355%, and the hypertuned NMAE was around 0.263%. The SGD model was hypertuned based on the learning rate by a scale of 1.

The Random Forest Regressor had an initial NMAE of 0.294%, and a hypertuned NMAE of 0.264%, showing that hypertuning the model did not majorly impact its performance. The Random Forest model was hypertuned on the 'max depth' parameter, and overall showed a trend that an increase in depth led to an increase in model performance.

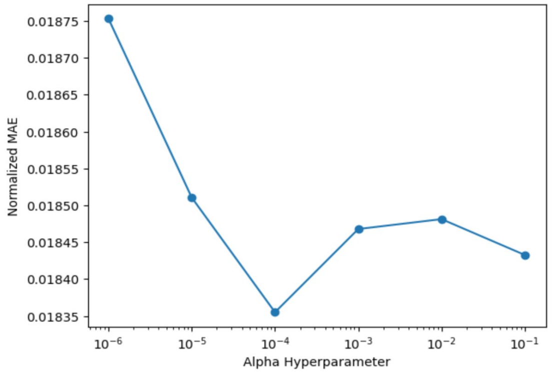
The Ridge Regression model was one of the models that performed the best. It had an initial NMAE of 0.248% and a hypertuned NMAE of 0.180%, showing a significant increase in model performance post hypertuning. This Ridge Regression model was hypertuned on the parameter 'lambda', and showed little fluctuation in the grid search plot in terms of NMAE.

The Lasso Regression model, was also a high performing model, achieving an initial NMAE of 0.221%, and a hypertuned NMAE of 0.183%. This model was hypertuned on the parameter ‘alpha’, and showed a general pattern of an increase in error with a higher ‘alpha’ value.

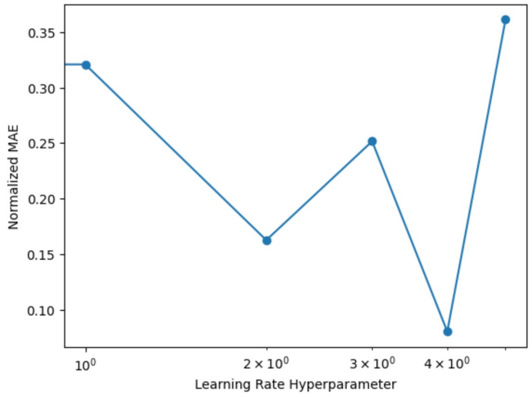
The highest performing model was the Bayesian Regression model. It achieved an initial NMAE of 0.249% and a hypertuned NMAE of 0.022%. This model was also hypertuned on the parameter ‘alpha’, and overall showed a significant increase in performance. Furthermore, since most of the models were relatively similar in training time, the better performance of the Bayesian model is something that can be scaled to larger amounts of data.



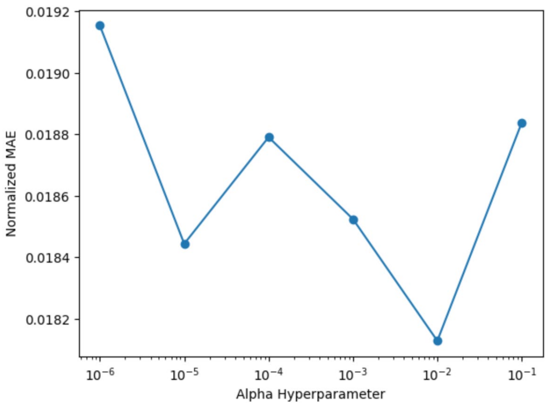
Graph 3. The hyperparameter search for the Lasso Regression model through a Grid Search Plot.



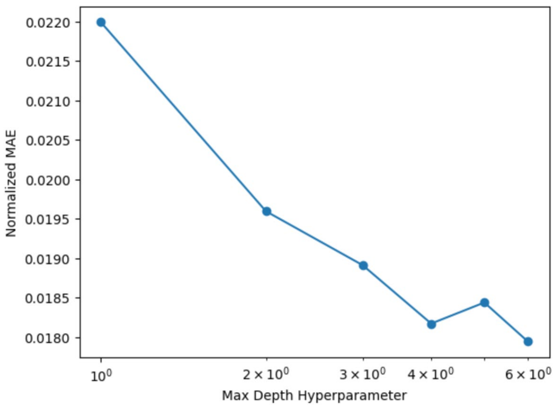
Graph 4. The hyperparameter search for the Bayesian Regression model through a Grid Search Plot.



Graph 5. The hyperparameter search for the Stochastic Gradient Descent (SGD) Regression model through a Grid Search Plot.



Graph 6. The hyperparameter search for the Ridge Regression model through a Grid Search Plot.



Graph 7. The hyperparameter search for the Random Forest Regression model through a Grid Search Plot.

On the whole, adding complexity to each of the models led to a better regression score - lower NMAE. Furthermore, these models show that predicting the X-ray flux of quasars is not straight forward, so adding more complexity to each of the models will aid in increasing accuracy. Comparing the performance of the models to the data visualization plots (Figure 6), a correlation can be identified. The X-ray faint end of quasars has greater diversity - as demonstrated in Figure 6, which is where these six models seem to have difficulty. It is also important to note that all the quasars with a high X-ray flux tend to have a small range of color. Upcoming telescopes, which can observe millions of the faint end of quasars, will reveal various galaxies in the early universe, alongside interactions between black holes and surrounding material, which we have not seen before. This will ultimately be shining light on how galaxies form and evolve through Cosmic time.

Conclusion

This study tested six different regression-based Machine learning models to find the most performant model in the MILLI-QUAS dataset. These six models were Stochastic Gradient Descent (SGD), Random Forest, Ridge, Lasso, Bayesian and the Baseline Linear Regression model. Each of these models developed in this project demonstrate significant strides towards better methods of X-ray flux (X-ray brightness) prediction so as to improve the efficiency of predicting the X-ray flux of a quasar and harnessing its energy for usage on earth. To train and test each of the models, an 80/20 split of the MILLI-QUAS dataset was used. Additionally, each model was tuned on model-specific hyperparameters. These hyperparameters include 'alpha' for Lasso and Bayesian regression, 'lambda' for Ridge regression, 'max_depth' for the Random Forest regressor and the learning rate for SGD regression. The Grid Search Plots for each model were also graphed in order to analyze the effects of tuning the hyper-parameters on model performance. Overall, the top three performing models based on the Normalized Mean Absolute Error (NMAE) were the Bayesian (0.022%), Ridge (0.180%), and Lasso (0.183%), with the baseline NMAE at 0.284%. After thorough analysis of model performance, the best performing model was the Bayesian Regression model. This may be due to the fact that Bayesian Regression employs prior belief or knowledge about the complex and/or ambiguous data to "learn" more about it which provided a more accurate prediction of the AGNs' X-ray brightness. The performance of these specialized models shows promise for future experiments with a fine-tuned Reinforcement Learning model through astrophysicist feedback and additional explain-ability through the integration of OpenAPI. With this, we can learn more about the evolution of galaxies in the early Universe and understand how these dynamic environments came to be.

Acknowledgments

I would like to thank Inspirit AI and Tony Rodriguez for their help and guidance throughout this project. I would also like to thank the creators of the MILLIQUAS catalog, for making the data publicly available.

References

Beckmann, V., & Shrader, C. (2012). *Active Galactic Nuclei*. Wiley.

Dewangan, G. C. (2017, October 25). *X-ray Emission from Active Galactic Nuclei*. Chandra X-ray Center. Retrieved October 29, 2023, from
https://cxc.harvard.edu/ciao/workshop/oct17_pune/agn_chandra_workshop.pdf

ESA/Hubble. (n.d.). *Active Galactic Nucleus*. ESA/Hubble. Retrieved November 15, 2023, from
https://esahubble.org/wordbank/active-galactic-nucleus/

Fruscione, A. (n.d.). *An Introduction to Active Galactic Nuclei in the X-Rays*. HEASARC. Retrieved February 19, 2024, from https://heasarc.gsfc.nasa.gov/docs/xrayschool-2005/talks/fruscione_agn.pdf

Gamma Editorial Team. (2021, July 2). *Electromagnetic Spectrum 101: Radio, Microwave, and Infrared – Gamma Scientific*. Gamma Scientific. Retrieved December 29, 2023, from https://gamma-sci.com/2021/07/02/electromagnetic-spectrum-101-radio-microwave-and-infrared/

Goswami, T., & Sinha, G. R. (Eds.). (2022). *Statistical Modeling in Machine Learning: Concepts and Applications*. Elsevier Science.
Padovani, P. (2017, October 23). *Active Galactic Nuclei at All Wavelengths and from All Angles*. Frontiers. Retrieved November 13, 2023, from https://www.frontiersin.org/articles/10.3389/fspas.2017.00035/full

Ricci, C. (2011). *AGN in the X-ray band*. ISDC. Retrieved November 12, 2023, from https://www.isdc.unige.ch/~ricci/Website/AGN_in_the_X-ray_band.html

Science Direct. (2020, May 26). Efficient Fermi source identification with machine learning methods. *Science Direct*, *32*. https://doi.org/10.1016/j.ascom.2020.100387

Springel, V. (2019, March 9). Simulating the joint evolution of quasars, galaxies and their large-scale distribution. *Arxiv*, 42. Retrieved October 26, 2023, from https://arxiv.org/pdf/astro-ph/0504097.pdf

Bennett, C. L. et al. First-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Preliminary Maps and Basic Results. Astrophys. J. Suppl. 148, 1–27 (2003).

Colberg, J. M. et al. Clustering of galaxy clusters in cold dark matter universes. Mon. Not. R. Astron. Soc. 319, 209–214 (2000).

Press, W. H. & Schechter, P. Formation of Galaxies and Clusters of Galaxies by SelfSimilar Gravitational Condensation. Astrophys. J. 187, 425–438 (1974).

White, S. D. M. Formation and evolution of galaxies: Les houches lectures. In Schaefer, R., Silk, J., Spiro, M. & Zinn-Justin, J. (eds.) Cosmology and Large-Scale Structure (Dordrecht: Elsevier, astro-ph/9410043, 1996).

Padovani, P. (2017, June 12). Active Galactic Nuclei: what's in a name? *Arxiv*, 56. Retrieved January 12, 2024, from https://arxiv.org/pdf/1707.07134.pdf

Zuo W., Wu X.-B., Liu Y.-Q., et al. (2012) The Correlations between Optical Variability and Physical Parameters of Quasars in SDSS Stripe 82. Astrophys J 758: 104. doi:10.1088/0004- 637X/758/2/104

Worseck G., Prochaska J. X. (2011) GALEX Far-ultraviolet Color Selection of UV-bright High-redshift Quasars. Astrophys J 728: 23. doi:10.1088/0004-637X/728/1/23

Wilms J., Allen A., McCray R. (2000) On the Absorption of XRays in the Interstellar Medium. Astrophys J 542: 914-924. doi:10.1086/317016

Weymann R. J., Carswell R. F., Smith M. G. (1981) Absorption lines in the spectra of quasistellar objects. Annu Rev Astron Astr 19: 41-76. doi:10.1146/annurev.aa.19.090181.000353