Black Hole Mass Prediction

Basil Turk
Assisted by: Dr. Raid Suleiman (rsuleiman@cfa.harvard.edu), Fatma Yousef (fatma.yousef@cfa.harvard.edu)
(Atomic and Molecular Physics Division)
Center for Astrophysics | Harvard & Smithsonian

Abstract

This research paper explores the use of classical machine learning models to predict the mass of black holes using data derived from the Sloan Digital Sky Survey's (SDSS) DR16 catalog. By leveraging key parameters such as FWHM (Full Width at Half Maximum) and luminosity measurements from various spectral lines, the model provides an accurate and efficient alternative to manual calculations of black hole mass. Four models were employed in this study: Linear Regression, Lasso Regression, Ridge Regression, and Random Forest Regression.

Keywords

Black hole mass prediction, DR16, SDSS, machine learning, astrophysics.

Introduction

The accurate estimation of black hole mass is fundamental for understanding their role in the evolution of galaxies and the dynamics of the universe. Current methods for calculating black hole mass often involve complex equations and manual computation, which can be prone to errors and are time-consuming. This research aims to address these challenges by employing machine learning techniques to predict black hole mass efficiently.

Dataset

The dataset was sourced from the Sloan Digital Sky Survey's (SDSS) Data Release 16 (DR16) catalog, a highly reliable and peer-reviewed astronomical dataset. The DR16 catalog contains detailed measurements of quasars, stars, and galaxies. For this study, we focused on the following key columns:

FWHM (Hβ, Mg II, C IV): These represent the Full Width at Half Maximum values for different emission lines, providing crucial information about the broadening of spectral lines due to the velocity of gas near the black hole.

LOGL5100, LOGL3000, LOGL1350: Luminosities at rest-frame wavelengths of 5100 Å, 3000 Å, and 1350 Å, respectively. These luminosities are used to estimate the accretion rate and indirectly infer black hole mass.

LOGMBH: The fiducial single-epoch black hole mass, which serves as the target variable for prediction.

Z_DR16Q: Redshift values, essential for understanding the cosmological distance and the quasar's properties.

SNR Columns (Signal-to-Noise Ratios): These include GAIA_G_FLUX_SNR, GAIA_BP_FLUX_SNR, which provide quality indicators of the measured fluxes from the Gaia mission.

These parameters were chosen due to their direct correlation with the black hole mass and their significance in established astrophysical models. The data underwent rigorous cleaning and preprocessing to handle missing or erroneous values. Given the precision of SDSS instruments and the peer-reviewed nature of the DR16 catalog, the dataset is considered highly accurate and suitable for scientific analysis.

Methodology

Preprocessing

The data underwent extensive cleaning, including handling missing values, scaling, and feature extraction.

Models

Four machine learning models were utilized: Linear Regression, Lasso Regression, Ridge Regression, and Random Forest Regression.

Evaluation Metrics

Mean Squared Error (MSE) and R-squared (R^2) were used to evaluate the performance of the models.

Results

The following table summarizes the performance of the four models on validation and test datasets:

Model	Validation MSE	Test MSE	Validation R^2	Test R^2
Linear Regression	0.6784841884351217	0.13527403322201947	-3.2106153760748617	0.173980349474819
Lasso Regression	4.4911255080405255	0.15428804347815173	-26.87154430769129	0.057875685979829616
Ridge Regression	0.6779386372367764	0.13527356100739032	-3.2072297315700033	0.17398323294442253
Random Forest Regression	0.038462007542412135	0.039681489296313363	0.7613080406098899	0.7576941476486986

Feature Importance Analysis:Using Random Forest Regression, we evaluated the relative importance of each feature in the prediction. Features such as FWHM (Mg II and Hβ) and luminosity (LOGL5100 and LOGL3000) exhibited the strongest correlation with the target variable (LOGMBH). The Signal-to-Noise Ratio (SNR) parameters showed less significant impact but were retained to ensure model robustness.

The correlation between features and their predictive influence suggests that emission line properties and luminosities remain the most critical predictors of black hole mass, aligning with established astrophysical theories.

Conclusion

The implementation of machine learning for black hole mass prediction provides a promising pathway for improving efficiency and accuracy in astrophysical studies. Random Forest Regression outperformed other models in this study, achieving the lowest MSE and highest R^2 scores. Future work will involve refining the models, expanding the dataset, and incorporating additional astrophysical features to enhance predictive power.

Acknowledgment: The author would like to thank Dr. Raid Suleiman and Fatma Yousef from the Atomic and Molecular Physics Division, Center for Astrophysics | Harvard & Smithsonian, for their invaluable support and guidance in preparing this paper.

References

[1] Lyke et al., 2020, Sloan Digital Sky Survey Data Release 16.
[2] Peterson, B. M., et al., 2004, ApJ, 613, 682.
[3] Vestergaard, M., & Peterson, B. M., 2006, ApJ, 641, 689.