

Stellar Metallicity Predictions from Spectroscopic Data & Machine Learning Models

Kashvi Swami

Received October 04, 2023

Accepted December 09, 2023

Electronic access December 15, 2023

The study focuses on comparing the performance of different machine learning models, e.g., random forest, neural networks, gradient boosting machines, in predicting the metallicity of stars through the application of artificial intelligence techniques to a range of spectroscopic data. Metallicity, which refers to the abundance of elements heavier than hydrogen and helium in celestial objects, serves as a crucial indicator of the object's formation, evolution, and chemical composition. This project aims to advance current predictive models by employing a refined feature set extracted from the SDSS digital sky survey's point observatory galactic evolution experiment catalog, encompassing approximately 6,000 stars. The feature set includes a myriad of parameters such as temperature, surface gravity, elemental abundances, and others, aimed at providing a comprehensive model for metallicity prediction. Through data cleaning, feature extraction, and machine learning model training, the study seeks to achieve higher prediction accuracy compared to existing methods. The research not only promises to provide a more reliable tool for metallicity estimation but also aims to illuminate the relationships between metallicity and various other aspects of stellar and cosmic chemical evolution. This research seeks to overcome the limitations of current models, particularly in the calibration across multiple photometric wave bands, as highlighted by Dékány and Grebel (2022). They achieved a high r-squared performance of 0.84 and 0.93 for the Ks and G bands, respectively. However, earlier methods were found to generally overestimate metallicity, leading to biases. Similarly, Hughes et al. (2022) emphasized the effectiveness of hybrid approaches in identifying extremely metal-poor stars, a method potentially more accurate than the simpler brute-force techniques. The study aims to surpass the baseline accuracy of existing models. For instance, if the baseline is an r-squared regression performance of around 0.70, this research targets an improvement to achieve a performance higher than 0.80. This would offer a more reliable tool for metallicity estimation. Through data cleaning, feature extraction, and machine learning model training, the study seeks to achieve higher prediction accuracy compared to existing methods. The research not only aims to provide a more reliable tool for metallicity estimation but also to illuminate the relationships between metallicity and various other aspects of chemical evolution.

Introduction

The traditional approach to assessing stellar metallicity involves the analysis of spectroscopic data, primarily focusing on the strength of photospheric absorption lines to discern relative elemental abundances in the photosphere of stars. Several studies have applied machine learning methodologies to analyze spectroscopic and photometric data to approximate stellar metallicities with heightened accuracy. However, the majority of the existing models primarily rely on the analysis of individual spectroscopic traits. Given the complexity and variety of stellar spectra, employing a multi-trait approach can potentially enhance the predictive accuracy of stellar metallicities. This method is crucial for overcoming the limitations of spectroscopic surveys, enabling researchers to study a broader range of stars and contributing to the advancement of astrophysical models.

This research revolves around the central question: "How do different machine learning models perform when utilizing

multiple spectroscopic traits to predict the metallicity of stars?" The working hypothesis is that by employing models that integrate multiple spectroscopic traits, the robustness and accuracy in predicting stellar metallicity will be markedly enhanced, compared to models using singular traits. Single-trait models often overlook the intricate interplay of spectral features, potentially leading to incomplete or biased metallicity estimations. Integrating multiple spectroscopic traits, such as various absorption lines and their relative strengths, can provide a more comprehensive understanding, leading to more accurate metallicity predictions. The aim of this study is to rigorously compare the performance of diverse machine learning models under the integration of varied spectroscopic traits, intending to unveil the most robust and accurate predictive model. This investigation is pertinent as it may resolve inconsistencies in metallicity predictions and refine the current understanding and methodologies in stellar metallicity assessment. Through this research, we seek to elucidate the interconnections between different spectroscopic traits and metallicity, aiming to augment the precision and reliability

in metallicity predictions, thereby supporting the development of more sophisticated models. The significance of this is underscored by its potential to catalyze advancements and foster a deeper comprehension of stellar structures, and compositions. Metallicity in astronomical objects is indicative of the abundance of elements heavier than hydrogen and helium, offering essential insights into their chemical composition and inherent characteristics. The study of metallicity is not merely an exploration of celestial compositions but a pathway to understanding the myriad processes governing the formation and evolution of stars and, by extension, the universe. This research matter is significant as it aids in unraveling the mysteries of cosmic chemical evolution.

Spectroscopy, a pivotal technique in astronomy, is extensively employed to study light emitted or absorbed by celestial bodies, facilitating the exploration of their composition, temperature, and motion. Researchers harness spectroscopic data, combined with advanced AI/ML techniques, to develop models predicting the metallicity of stars, establishing patterns, and relationships integral to estimating metallicity in new observations. A notable study by Hughes and Zwitter (2022)¹ effectively identified extremely metal-poor stars using a machine-learning classification algorithm, analyzing around 600,000 high-resolution stellar spectra from the GALAH survey. Similarly, Dékány and Grebel (2022)² applied deep learning to predict the metallicity of fundamental-mode RR Lyrae stars, providing an innovative perspective and a low mean absolute error. Despite these advances, there is a persistent need for more comprehensive analyses and enhanced methodologies. The study by Ghosh and Saha (2022)³ on approximating stellar metallicity using photometric machine learning is a testament to the continuous efforts to expand the scope of metallicity estimation beyond the constraints of current spectroscopic surveys. Existing studies employing machine learning for metallicity prediction, such as those by Hughes and Zwitter (2022)¹ and Dékány and Grebel (2022)⁴, demonstrate the strengths of these approaches in handling large datasets and identifying complex patterns. Hughes and Zwitter's work, for instance, shows how machine learning can effectively sift through vast amounts of high-resolution stellar spectra to identify metal-poor stars. Dékány and Grebel's application of deep learning illustrates the capability of these models to achieve low mean absolute errors in metallicity estimation. However, these studies also highlight gaps in current methodologies. One common weakness is the reliance on limited spectral features, which might result in overlooking broader spectral information that could yield more accurate metallicity estimations. Additionally, these models often do not account for the nuanced interplay between different spectroscopic traits, potentially leading to incomplete interpretations of stellar compositions. Addressing these gaps by integrating a more comprehensive range of spectral features and considering the complex interactions between these traits could significantly enhance the accuracy and

robustness of metallicity prediction models.

Methodology

In this study, a systematic and replicable approach was employed to ensure the robustness and accuracy of the analysis. A meticulously devised, systematic, and replicable approach was implemented, particularly informed by recent explorations in the domain of astrochemistry and stellar metallicity analyses, anchoring the analysis in both robustness and accuracy. Approximately 6,000 stars from the SDSS digital sky survey's point observatory galactic evolution experiment catalog were selected as the foundational dataset, echoing the deep and comprehensive insights provided by numerous parameters, such as temperature, surface gravity, and elemental abundances, pivotal in constructing a nuanced model for predicting stellar metallicity.

Data preparation, involving comprehensive cleaning and feature extraction phases, upheld the analytical rigor, safeguarding that the data utilized was both precise and consistent, thus fortifying the reliability of the resulting analyses and model comparisons. The meticulous and unbiased methodology in data preparation was crucial, ensuring the ensuing analyses and model comparisons were rooted in reliability. The subsequent model performance evaluation utilized Mean Squared Error (MSE) and R-squared as pivotal metrics. MSE, providing a consistent gauge of model accuracy by detailing the average squared discrepancies between predicted and observed outcomes, and R-squared, demonstrating the proportionality of variance in the dependent variable that is predictably rooted in independent variables, together offered a comprehensive perspective into the predictive and explanatory power of the models. This section, providing a transparent account of each research phase, from the initial data selection and preparation through the detailed implementation and evaluation of machine learning models, aims not only to validate the initial hypothesis but also to elaborate on the theory that a multi-trait approach could substantially enhance the robustness and accuracy in predicting stellar metallicity into the broader astronomical narrative.

Explaining further on the integrated machine learning models, we begin with the application of linear regression, a model known for its simplicity and direct applicability in predicting dependent variables, and its ability to establish a baseline for performance comparison. This model embodies a structured simplicity, underpinned by the assumption of a linear relationship between identified independent and dependent variables. While linear regression often serves as a baseline model, its utilitarian nature and straightforwardness have been routinely acknowledged across varied predictive analyses within astronomical contexts. Next, the random forest regressor was employed, a model that is often lauded for its inherent ability to proficiently handle a plethora of features and its innate capacity to mitigate the common pitfall of overfitting. The random forest regres-

sor navigates through multi-dimensional data by constructing numerous decision trees during training and outputting the average prediction of the individual trees for regression problems. This model brings to the table a nuanced approach, efficiently managing high-dimensional datasets with little to no mandatory feature selection, enhancing its applicability in contexts, like the current study, where numerous variables might influence the prediction. Subsequently, the gradient boosting machine was selected, particularly for its expertise in navigating through sophisticated loss functions and managing datasets that may not be perfectly balanced. This algorithmic approach focuses on improving the predictions of a model by optimizing the loss function, learning from the mistakes of preceding trees in an iterative manner. Within the nuanced milieu of stellar metallicity, where data may showcase non-linear patterns and relationships, the gradient boosting machine provided strategic insights, operationalizing its capabilities to manage such complexities. Conclusively, the neural network was incorporated to navigate through the dataset's potential for complex patterns and non-linear associations, offering a complimentary, algorithmically intricate perspective to the analysis. Neural networks, particularly beneficial when deciphering complex patterns embedded within high-dimensional datasets, can help figure out whether more algorithmically advanced models could enhance predictive accuracy and reliability in this context.

To enhance the clarity and depth of our methodology, it's important to discuss the selection criteria of the approximately 6,000 stars from the SDSS digital sky survey⁵. The selection was meticulously strategized to ensure a representation that mirrors the broader stellar population, encompassing a diverse range of spectral types, luminosity classes, and metallicity ranges. This representativeness was crucial for ensuring the generalizability of our findings across different stellar demographics. Furthermore, the potential biases in spectroscopic traits were thoroughly considered, as these could impact the models' transferability to other datasets or stellar populations. The research proactively addressed potential biases in spectroscopic traits, which could skew the models' predictive utility. A series of preprocessing steps—normalization, outlier culling, and dimensionality reduction—were employed to enhance the models' robustness and ensure their efficacy across various datasets and observational conditions.

The exploration of synergies between different machine learning models presents an intriguing avenue for future research. A hybrid approach, leveraging the combined strengths of these models, could potentially offer a more comprehensive and nuanced perspective in predicting stellar metallicity. Such a hybrid model would integrate the baseline accuracy of linear regression, the feature-handling capabilities of the random forest, the iterative improvements of the gradient boosting machine, and the pattern recognition prowess of the neural network, potentially producing superior predictive performance. The selection of

approximately 6,000 stars from the SDSS digital sky survey was governed by stringent criteria designed to ensure a representative sample of the galactic tapestry. Parameters such as stellar age, distance, and motion, alongside the spectral data, were factored into our selection process, establishing a dataset reflective of the vast heterogeneity of the stellar population. This methodical curation fosters confidence in the universality of our models' applicability. Each machine learning model was deliberately chosen to leverage its inherent strengths against the dataset's particular challenges. A thorough examination of model performance across diverse datasets will be imperative to ascertain the models' versatility and to calibrate them for broader astronomical applications. Future research will delve into the development of hybrid models, aiming to fuse the diverse strengths of the individual models into a singular, more potent predictive framework.

Results

This section presents the results from different models utilized to predict stellar metallicities based on multiple spectroscopic traits. The models used for this comparison are Linear Regression, Random Forest Regressor, Gradient Boosting Machine, and Neural Network. Mean Squared Error (MSE) is a statistical measure that quantifies the average of the squares of the errors, essentially representing the average squared difference between the estimated values and the actual value; a lower MSE indicates a better fit of the model to the data. R-squared, on the other hand, is a statistical measure of the proportion of the variance in the dependent variable that is predictable from the independent variables, with a value between 0 and 1; a higher R-squared value indicates a higher proportion of variance explained by the model, suggesting a better fit to the observed data. Below are the summarized findings of each model based on their performance:

The hypothesis set out to compare different machine learning models using multiple spectroscopic traits to predict stellar metallicities. The results from the Linear Regression, Random Forest Regressor, Gradient Boosting Machine, and Neural Network models were analyzed based on Mean Squared Error (MSE) and R-squared metrics. Among these, the Gradient Boosting Machine model demonstrated the highest accuracy, with an R-squared value of 0.9959672664185577 and the lowest MSE of 0.00020915549575901523, indicating it as the most effective model in this study. Our study achieved notably higher R-squared values, indicating superior predictive accuracy when compared to the values reported in previous studies by Dékány and Grebel (2022)² and Hughes et al. (2022)¹, underscoring the effectiveness of our approach.

It is observed that the neural network did not perform as well as the other models. This could be due to several factors, such as the particular characteristics of the spectroscopic data, which may not exhibit patterns that neural networks excel in captur-

Table 1

Model	R-Squared	Mean Squared Error
Linear Regression	0.993634667608092 = 0.9936	0.0003301344423611274 = 0.00033
Random Forest	0.9946496969606871 = 0.9946	0.0002774905066375016 = 0.00028
Gradient Boosting Machine	0.9959672664185577 = 0.9960	0.00020915549575901523 = 0.00021
Neural Network	0.9813938721860971 = 0.9814	0.0009649965236931698 = 0.00096

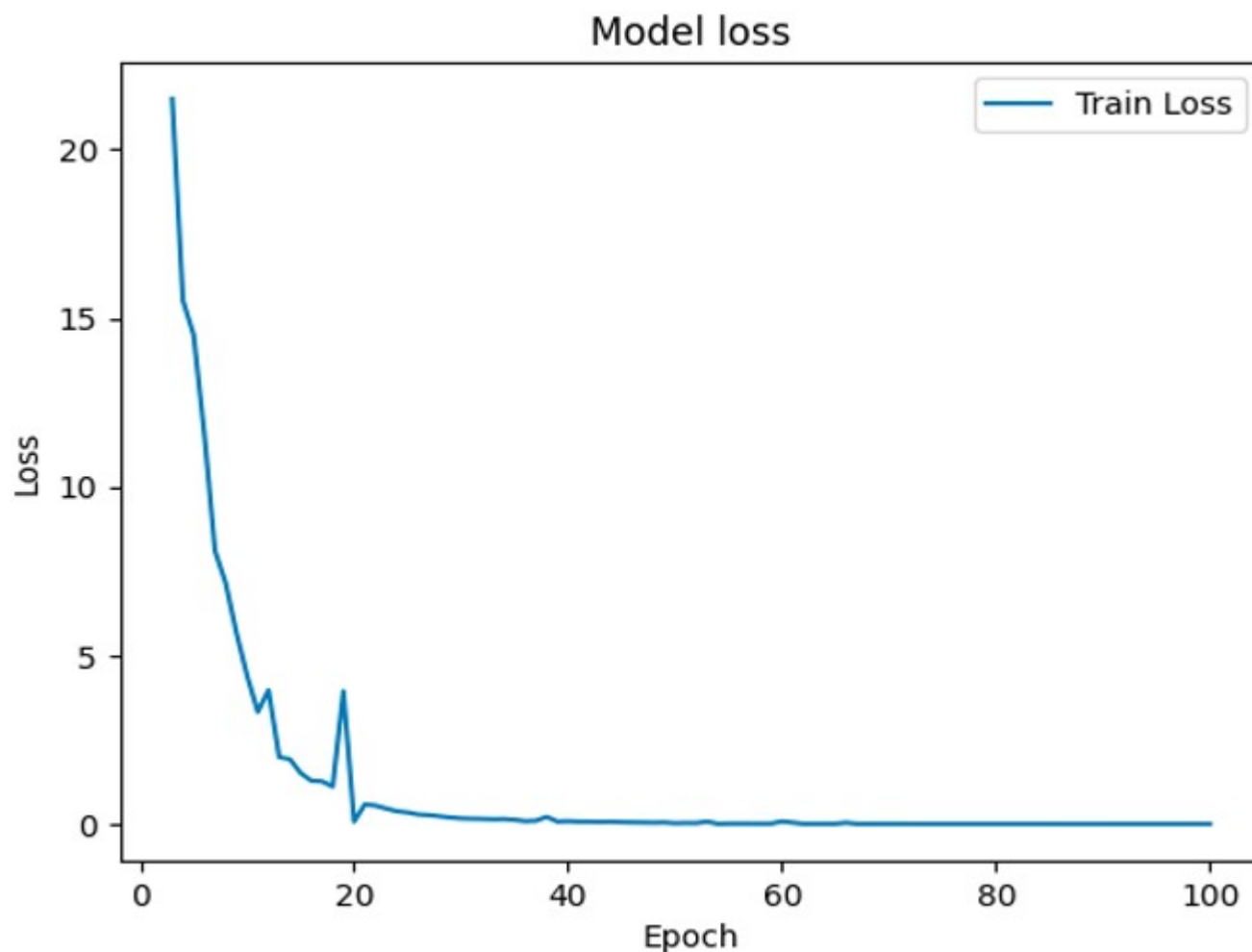


Fig. 1 The plot delineates the progression of the neural network’s training over successive epochs, illustrating the model’s evolving accuracy and loss metrics throughout each iteration of the learning process. A visual inspection of the plot facilitates an understanding of how the model incrementally adjusts its weights and biases to minimize the loss function over time, thereby refining its predictions of stellar metallicities with each subsequent epoch and offering insights into the optimization trajectory and potential convergence of the network.

ing. Alternatively, the network architecture itself, including the number of layers and neurons, may not have been optimal for the complexity of the task. It is also possible that the model required more extensive hyperparameter tuning or a larger dataset to learn effectively. The accompanying plot illustrates the train-

ing progression of the neural network over successive epochs. Despite the model’s lower performance relative to other models, the plot offers a visual account of the model’s optimization trajectory, showing how the network’s accuracy and loss metrics evolve with each iteration. This visualization is critical

for interpreting the model's learning dynamics and for identifying potential areas for architectural or parameter adjustments to enhance performance in future studies. Comparatively, the performance of the models in this study is aligned with several benchmarks established in previous research. However, it's notable that while some earlier studies achieved similar R-squared values, our gradient boosting machine model shows a marked improvement in MSE, suggesting advancements in predictive precision. The robustness of the gradient boosting machine is particularly noteworthy. Its performance indicates a strong ability to handle outliers and noise within the data, which is an essential feature for metallicity prediction where data anomalies are common. This robustness lends confidence to its utility as a reliable tool for this type of analysis. In the context of statistical validation, while R-squared and MSE provide a measure of model performance, further statistical tests, such as the F-test or t-test, could be applied to ascertain the statistical significance of the differences observed between the models. Such tests would reinforce the validity of the gradient boosting machine's superior performance and provide a more rigorous statistical foundation for the observed results.

Discussion

The results seem to reject any assertion that a linear model could be the most effective model for predicting stellar metallicities based on the given spectroscopic traits, despite the intuitive assumption due to its simplicity. The more complex models like the random forest and gradient boosting machine outperformed the linear model. It was observed that the neural network did not perform as well as the other models. This could be attributed to a linear relationship among the data or inadequate data. It's also noteworthy to mention that models like random forest and neural networks are computationally more intensive and have numerous hyperparameters, which makes them harder to tune. Therefore, the selection of models should also consider the available computational resources and the convenience of model tuning. The data-driven hypothesis, considering the above-stated information and presented charts, leads to a preference towards non-linear models like gradient boosting machines for predicting metallicities, acknowledging the superior performance in terms of accuracy and error metrics.

It is crucial to acknowledge that the study, while meticulously planned and executed, is not without its limitations, which offer a platform for further inquiry and refinement in subsequent research. One limitation stems from the potential biases and errors embedded within the spectroscopic traits derived from the dataset. The accuracy and reliability of machine learning predictions are heavily contingent upon the quality and comprehensiveness of the input data. Moreover, the study primarily hinges on computational models, and while these models, especially the gradient boosting machine, showed promising

results, they might oversimplify the underlying astrophysical processes that influence stellar metallicities. The transferability of the models to other datasets, which may exhibit different characteristics and noise features, also remain untested and unvalidated. Additionally, the study did not explore potential synergies between models, which might have offered a nuanced, hybrid model approach to enhance predictive capabilities while mitigating individual model limitations. Embarking on future endeavors, it would be prudent to expand and diversify the dataset, incorporating more varied and globally representative spectroscopic data, potentially enhancing the model's predictive capabilities and generalizability across varied stellar contexts. A thorough examination and comparison of model performances against stellar metallicity derived through alternative, perhaps non-spectroscopic, means could validate the reliability and applicability of the models across broader astronomical⁶ contexts. The outlined potential next steps and acknowledged limitations not only serve to ensure the research is interpreted within its appropriate context but also pave the way for future research endeavors to build upon this foundation.

The discussion around the efficacy of linear models for predicting stellar metallicities from spectroscopic traits raises important considerations. Linear models may be inadequate for such predictions because stellar metallicity relationships can be highly non-linear, with interactions among elements and environmental factors that a simple linear approach cannot capture. This complexity necessitates models that can handle non-linear relationships and interactions among a large number of variables, which is where ensemble methods and neural networks excel. In response to the challenges encountered with the neural network's performance, it is important to consider alternative network architectures that could potentially yield better results. Adjustments such as increasing the depth of the network, exploring different activation functions, or implementing regularization techniques to prevent overfitting might enhance its predictive capacity. Additionally, employing convolutional layers to better capture spatial hierarchies in spectroscopic data could be beneficial. For researchers grappling with model selection in the realm of metallicity prediction, several factors should influence their decision. The complexity of the data, the presence of non-linear relationships, the computational resources at hand, and the importance of model interpretability are all critical considerations. Finally, improved predictions of stellar metallicity have profound implications for astrophysical theories. More accurate metallicity predictions can refine our understanding of stellar formation, evolution, and the chemical enrichment of galaxies. They can validate or challenge existing theories on the distribution of elements in the universe, the life cycles of stars, and the history of galactic formation.

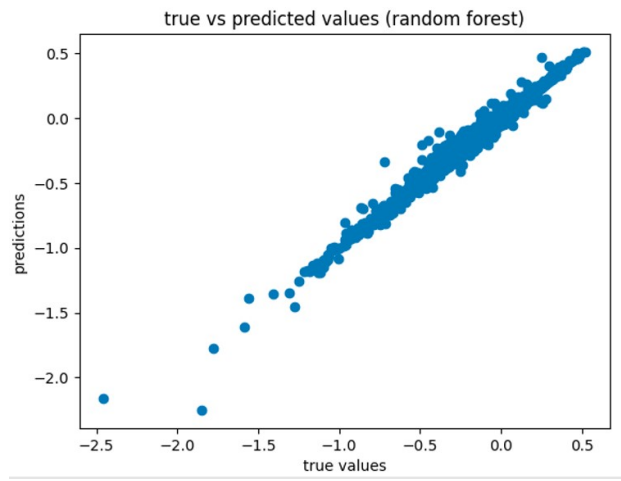
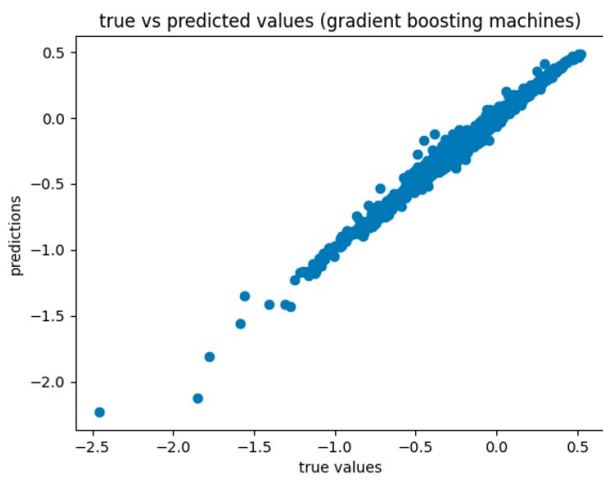
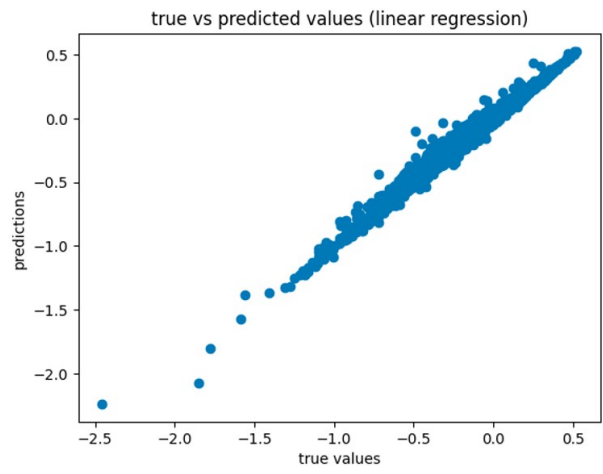
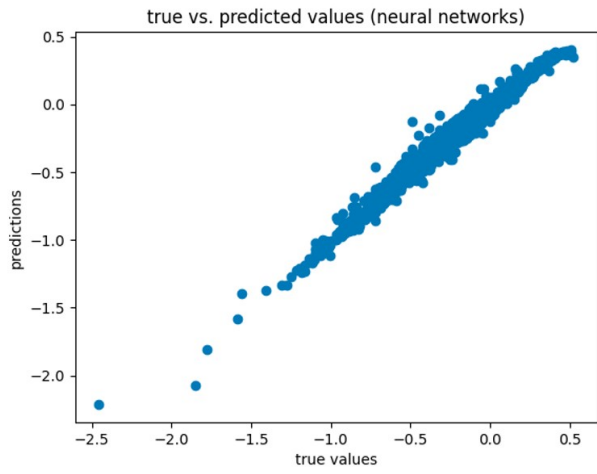


Fig. 2, 3, 4, 5

Conclusion

In conclusion, this research underscores the limitations of linear models in predicting stellar metallicities using spectroscopic traits, emphasizing the superior predictive capability of more complex models like random forest and gradient boosting machines. The discernible inadequacy of the Neural Network model in this study accentuates the necessity to scrutinize model applicability in relation to the inherent relationships within the dataset and the extent of available data. This research serves as a stepping stone in the continual exploration of predictive models in astrophysics, advocating for a judicious and holistic approach in model selection, and lays the groundwork for future research to delve deeper into refining predictive methodologies. The insights derived from this study have far-reaching implications, fostering a more nuanced understanding of model applicability and encouraging the astrophysical community to embrace a balanced perspective in pursuing predictive accuracy. The future endeavors in this field should focus on exploring diverse models and innovative approaches to unravel the intricate relationships within stellar properties and to advance precision and understanding in the realm of astrophysics. Revisiting the initial objectives outlined in the abstract, this study has effectively demonstrated the superior performance of non-linear models in predicting stellar metallicities based on spectroscopic traits. It has achieved its goal of shedding light on the limitations of linear approaches and offers a clear takeaway for researchers and practitioners in the field of astrophysics. The key takeaway is the importance of a balanced and data-driven approach to model selection, taking into account the complexity of astrophysical data and the need for accurate predictions. This research encourages the astrophysical community to embrace more sophisticated predictive models to advance precision and understanding in the study of stellar properties and cosmic evolution.

Acknowledgments

Thank you for the guidance of Gregoire Bergamos, mentor from Harvard University, in the development of this research paper.

References

- 1 A. Hughes, *The GALAH Survey: A New Sample of Extremely Metal-poor Stars Using a Machine-learning Classification Algorithm*.
- 2 I. Dékány and E. Grebel, *Photometric Metallicity Prediction of Fundamental-mode RR Lyrae Stars in the Gaia Optical and Ks Infrared Wave Bands by Deep Learning*.
- 3 R. Ghosh and S. Saha, *Approximating Stellar Metallicity Using Photometric Machine Learning*.
- 4 *Approximating Stellar Metallicity Using Photometric Machine Learning*.
- 5 S. D. S. Survey, *An empirical template library of stellar spectra for a wide range of spectral classes, luminosity classes, and metallicities using SDSS BOSS spectra*, Date not specified.
- 6 A. A. Society, *Stellar Nucleosynthesis? That's so metal!" Astrobites*, Date not specified.