Supplementary Material: Data Dictionary and Exploratory Plots

J. R. Muñoz Luque et al.

Machine learning terms

We follow with an explanation of some of the common terms used in machine learning and in the article An Efficient Framework for Automatic Carbon Star Detection Using Machine Learning Techniques on LAMOST DR9 and Gaia DR32.

Bagging: It is an ensemble learning technique useful to improve the accuracy and stability of machine learning models. It involves training multiple models on different random subsets of the training data. By averaging the models' predictions, this method reduces variance and helps to avoid overfitting.

Bagging Top Push PU: The ensemble method used by Y.-B. Li et al. (2018) is a semi-supervised machine learning model that involves multiple individual models using different random subsets of the training data (C. Du et al., 2016) that builds a ranking based on known carbon stars from previous studies. This model treats carbon star spectra as positive samples (P) and incorporates many unlabeled data (U). Negative samples are randomly selected from the U dataset, and the models in the ensemble are trained using both P and U samples. The final ranking prioritizes positive samples and places them at the top, using bagging as the ensemble technique. The algorithm scores all stars in the U dataset, ranking carbon star candidates ahead of other stars.

Binary model: It is a classification model that categorizes data into two distinct classes.

Class: In the context of machine learning, the class represents a categorical variable that indicates possible outcomes. In this study, there are two classes, called carbon star and non-carbon star.

Criterion: A hyperparameter used to split nodes and measure the quality of the split in decision trees. The values for tuning this hyperparameter available in `Scikit-Learn` are $\{Gini, Entropy, Logloss\}$. The index *Gini* and the index *Entropy* are based on the proportion of samples that belong to each class

in a node, measuring the purity of the nodes. Lower values mean greater node purity; for example, a value of 0 indicates that all samples in a node belong to the same class. On the other hand, *Logloss* measures the difference between the probabilities estimated by the model and the true labels of the sample class. The lower the value, the better, meaning 0 is a perfect classification.

Efficient Manifold Ranking: A machine learning model similar to *label propagation* utilizes anchors derived from clustering techniques like k-nearest neighbors to establish relationships, making it faster and more memory-efficient. This semi-supervised ML algorithm was used by J.-M. Si et al. (2015).

Feature fraction: This GBM hyperparameter is similar to the *Max Features* hyperparameter in other models; it is used by `LightGBM` as it controls the number of features from the dataset randomly sampled in each iteration to build a new tree in the ensemble. The features are sampled randomly at the beginning of each tree's construction, enhancing the diversity because each tree in the ensemble uses different sets of features.

Generalization: Refers to the ability of a model to identify and learn the correct underlying patterns in the training data that apply to unseen or new data. This capacity ensures that the model is not influenced by noise or spurious correlations present in the training dataset. A well-generalized model can reliably apply its learned insights to unseen data, maintaining robust performance across diverse datasets.

Highly correlated features: Occurs when the Pearson correlation coefficient between one or more pairs of features is strong, typically below -0.6 or above 0.6. This issue affects the interpretability of the models. In GBM, when highly correlated features are present without redundancy, it is essential to use specialized algorithms, such as SHAP (S. M. Lundberg & S.-I. Lee, 2017), to accurately assess feature importance and mitigate the impact of correlation on model interpretation. In fact, high correlation can lead to redundant information or mask individual contributions of features, complicating the understanding of model decisions. This problem is also known as multicollinearity.

Imbalanced dataset: This occurs when the distribution of classes is disproportionate, meaning one or more classes have far fewer observations compared to others. If not addressed properly, this imbalance can introduce bias into the models. To minimize this, techniques such as *stratified sampling* are used when dividing the data into training and testing sets, and balancing methods are applied to the training set.

Label Propagation: A semi-supervised machine learning model that utilizes a small number of labeled samples, referred to as "carbon stars," in conjunction with a large set of unlabeled samples (which do not have any assigned class). This combination forms a training dataset. The model analyzes the

labeled samples in comparison to the unlabeled ones, using the data to identify relationships based on distances between the samples, and produces a ranked list where the labeled samples receive the highest scores, and similar unlabeled samples are ranked accordingly. This semi-supervised machine learning algorithm was utilized by J.-M. Si et al. (2014).

Labeled sample: It is the data sample collected to develop a supervised ML model. The labeled sample is commonly split into training and test datasets. The first is used to train the model, while the second is used to evaluate its performance.

Learning rate: A hyperparameter used in GBM models. It is crucial in machine learning as it determines the step size for each iteration, controlling each tree's contribution to the ensemble. A higher Learning Rate can make the algorithm converge faster, but it also increases the risk of missing the optimal solution. On the other hand, a lower learning rate can cause the algorithm to converge slowly, but it reduces the risk of missing the optimal solution. According to T. Hastie et al. (2009), it is recommended to set values no greater than 0.1. The Learning Rate acts as a *regularization* technique, helping to mitigate the risk of overfitting.

Majority class: The class with the most observations in a dataset.

Max depth: It is a hyperparameter that refers to the maximum number of edges allowed in a tree, counting from the root to the leaves. This hyperparameter helps to deal with the model bias-variance trade-off (P. M. Domingos, 2000). In a shallow tree, the model has high bias and low variance, capturing low-order interactions between features and making the model prone to underfitting. Conversely, a tree with a higher Max Depth can learn more intricate interactions, having low bias but high variance, becoming a more complex model and prone to overfitting. As a result, this hyperparameter is useful for controlling overfitting but needs to be tuned to find a balance between bias and variance.

Max features: A hyperparameter that controls the number of features randomly selected for analysis when finding the best split in a decision tree, which helps reduce the number of features used per split. Besides, it controls the randomness construction of the tree, helping to deal with overfitting. If the value is less than 1, this hyperparameter is treated as a proportion of features. The total number of features is calculated using this proportion, rounding down to obtain an integer. The available values are those commonly used in`Scikit-Learn`: $None$, which means that all characteristics are used, $\frac{\log_2(n)}{n}$, and $\frac{\sqrt{n}}{n}$, being the last analyzed by R. Genuer et al. (2008) in random forest models. This last value can be applied to decision trees because random forests are an ensemble of trees, where n is the number of features in the analysis.

Minority class: The class with fewer observations in a dataset represents the minority class in binary problems. In multiclass problems, there may be one or more minority classes.

Multicollinearity: This phenomenon occurs when there are highly correlated features.

N-estimators: In a GBM, this hyperparameter controls how many times the boosting algorithm will iterate, adding a new tree to the ensemble in each iteration. This hyperparameter is also related to the Learning Rate, with an existing trade-off between them. For example, a higher Learning Rate leads to smaller N-estimators.

Node: In a decision tree, a node represents a point where the data are divided based on a specific feature and a chosen *criterion*. The process starts at the root node, which is the initial point where the first division occurs. Subsequent points where additional divisions take place are referred to as intermediate nodes.

Overfitting: It is a generalization problem when the model learns the training data very well, but its performance decreases in the data that the model has not seen before. In a learning rate curve plot, it can be detected when there is a much better learning rate at training than at validation.

Principal Components Analysis: A statistical algorithm to reduce the number of variables in the dataset, while preserving as much information as possible.

Production sample: A data sample that is not used during training or testing. It is entirely new to the PM. It contains no labels and is used only to run the PM and identify new candidates.

Ranking: In machine learning, ranking refers to the process of sorting results based on their scores, with the highest scores appearing at the top of the list. The output is a ranked list, for example, produced by semi-supervised ranking models.

Regularization: It is a technique used to manage model complexity. In this context of logistic regression, complexity refers to the number of features in the model. The method chosen for this study is L_2 regularization, as it is computationally efficient and particularly effective when dealing with highly correlated features. L_2 regularization modifies the loss function by adding a term that represents the sum of the squares of the estimated parameters (excluding the intercept). This term is scaled by a factor to control its impact. As a result, L_2 reduces large coefficients in the model, shrinking them towards zero without making them exactly zero. This not only prevents overfitting but also enhances

the model's ability to generalize, improving its performance on unseen data.

Semi-supervised models: Machine learning models that use a few labeled observations and a huge number of unlabeled observations during training.

Stratified sampling: This is a technique commonly used with imbalanced datasets to ensure that the sampled data maintain the same class proportions as the original dataset.

Supervised classification models: Machine learning models that require all training observations to be labeled, indicating to which class each observation belongs.

Unsupervised classification models: Machine learning models that do not require labeled observations to work.

Adjusted Box Plots

The adjusted box plot (M. Hubert & E. Vandervieren, 2008; M. Maechler et al., 2024; R Core Team, 2024) is a visual tool for descriptive statistics that provides insight into the distribution's location, scale, skew, and outliers. This type of plot consists of a central box and two lines (whiskers) that extend along the axis of symmetry. The line inside the box represents the median, indicating the central tendency of the distribution. The properties of the median are discussed in A. M. Law & W. D. Kelton (1991) and D. Peña (2002), while its application in astronomy is explored in E. D. Feigelson & G. J. Babu (2012). The length of the box, defined by the first and third quartiles, and the distance between the whiskers' ends reflect the scale, specifically the interquartile range and the range excluding outliers, respectively. The shift of the median from the box's center is indicative of skewness, and any points beyond the whisker ends are outliers. In contrast to standard box plots, which may fail or incorrectly identify outliers in skewed distributions, the adjusted version avoids assumptions about data distribution by optimizing the whisker lengths to be robust to skewness. Moreover, the notch within the box represents the confidence interval of the median, approximately at a 95% confidence level, as detailed in Y. Benjamini (1988). The median confidence intervals are useful for comparisons, although this descriptive test is not definitive if the intervals overlap, as noted in N. Schenker & J. F. Gentleman (2001).

References

Benjamini, Y. 1988, Opening the Box of a Boxplot, The American Statistician, 42, 257, doi: 10.1080/00031305.1988.10475580

Domingos, P. M. 2000, A Unifeid Bias-Variance Decomposition and its Applications, in International Conference on Machine Learning. https://api.semanticscholar.org/CorpusID:15534779

Du, C., Luo, A., Yang, H., Hou, W., & Guo, Y. 2016, An Efficient Method for Rare Spectra Retrieval in Astronomical Databases, , 128, 034502, doi: 10.1088/1538-3873/128/961/034502

Feigelson, E. D., & Babu, G. J. 2012, Modern statistical methods for astronomy with R applications, 1st edn. (Cambridge; Cambridge University Press)

Genuer, R., Poggi, J.-M., & Tuleau, C. 2008, Random Forests: some methodological insights, https://arxiv.org/abs/0811.3619

Hastie, T., Tibshirani, R., & Friedman, J. 2009, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Series in Statistics (Springer)

Hubert, M., & Vandervieren, E. 2008, An adjusted boxplot for skewed distributions, Computational Statistics Data Analysis, 52, 5186, doi: https://doi.org/10.1016/j.csda.2007.11.008

Law, A. M., & Kelton, W. D. 1991, Simulation Modeling and Analysis, 2nd edn. (McGraw Hill International Editions, Industrial Engineering Series)

Li, Y.-B., Luo, A.-L., Du, C.-D., et al. 2018, Carbon Stars Identified from LAMOST DR4 Using Machine Learning, , 235, 42, doi: 10.3847/1538-4365/aaa415

Lundberg, S. M., & Lee, S.-I. 2017, A unified approach to interpreting model predictions, in Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17 (Red Hook, NY, USA: Curran Associates Inc.), 4768–4777. https://arxiv.org/abs/1705.07874

Maechler, M., Rousseeuw, P., Croux, C., et al. 2024, robustbase: Basic Robust Statistics. http://robustbase.r-forge.r-project.org/

Peña, D. 2002, Análisis de Datos Multivariantes (McGraw -Hill/Interamericana de España, S.A.U.)

R Core Team. 2024, R: A Language and Environment for Statistical Computing, https://www.R-project.org/

Schenker, N., & Gentleman, J. F. 2001, On Judging the Significance of Differences by Examining the Overlap Between Confidence Intervals, The American Statistician, 55, 182, doi: 10.1198/000313001317097960

Si, J.-M., Luo, A.-L., Li, Y.-B., et al. 2014, Search for carbon stars and DZ white dwarfs in SDSS spectra survey through machine learning, Science China Physics, Mechanics, and Astronomy, 57, 176, doi: 10.1007/s11433-013-5374-0

Si, J.-M., Li, Y.-B., Luo, A.-L., et al. 2015, Identifying Carbon stars from the LAMOST pilot survey with the efficient manifold ranking algorithm, Research in Astronomy and Astrophysics, 15, 1671, doi: 10.1088/1674-4527/15/10/005