Classifying BL Lacs with FERMI Data

Authors: Saaniya Bhushan, Malcolm Ehlers, Cindy Xiong
Advisor: Peter Freeman

Introduction

- The Fermi Gamma-Ray Space Telescope is used to detect high-energy photons produced by astronomical objects, among which are gamma-ray-emitting BL Lacertae objects, or BL Lacs. BL Lacs have beamed jets of matter pointed directly towards the Earth, which are streamed from black holes.
- However, BL Lacs are difficult to identify due to their spectra's similarities to other classes of astronomical objects.
- **The goal of this analysis is to accurately classify BL Lacs, as well as determine the predictive variables most useful for said classification.**

Data

- These data are from the Fermi LAT 10-Year Point Source Catalog (Ajello et al., 2020).
- It was collected by the Large Area Telescope, an instrument aboard the Fermi satellite.
- Each datum is a point in space which appears to produce light, which cannot be categorized by appearance alone.
- There are 4283 data in total, 1226 (28.6%) of which are BL Lacs and 3057 (71.4%) of which are not BL Lacs.
- The predictor variables in our data set include
  - Two recording the object's position in the sky
  - One recording signal strength
  - Two recording the variability of the signal
  - Fifteen for recording features of the spectrum
- The categorical response variable `source_type` recorded values of `BLL` or `NOT_BLL`

References

Ajello, M., et al. 2020,The Fourth Catalog of Active Galactic Nuclei Detected by the Fermi Large Area Telescope, online at https://arxiv.org/pdf/1905.10771.pdf

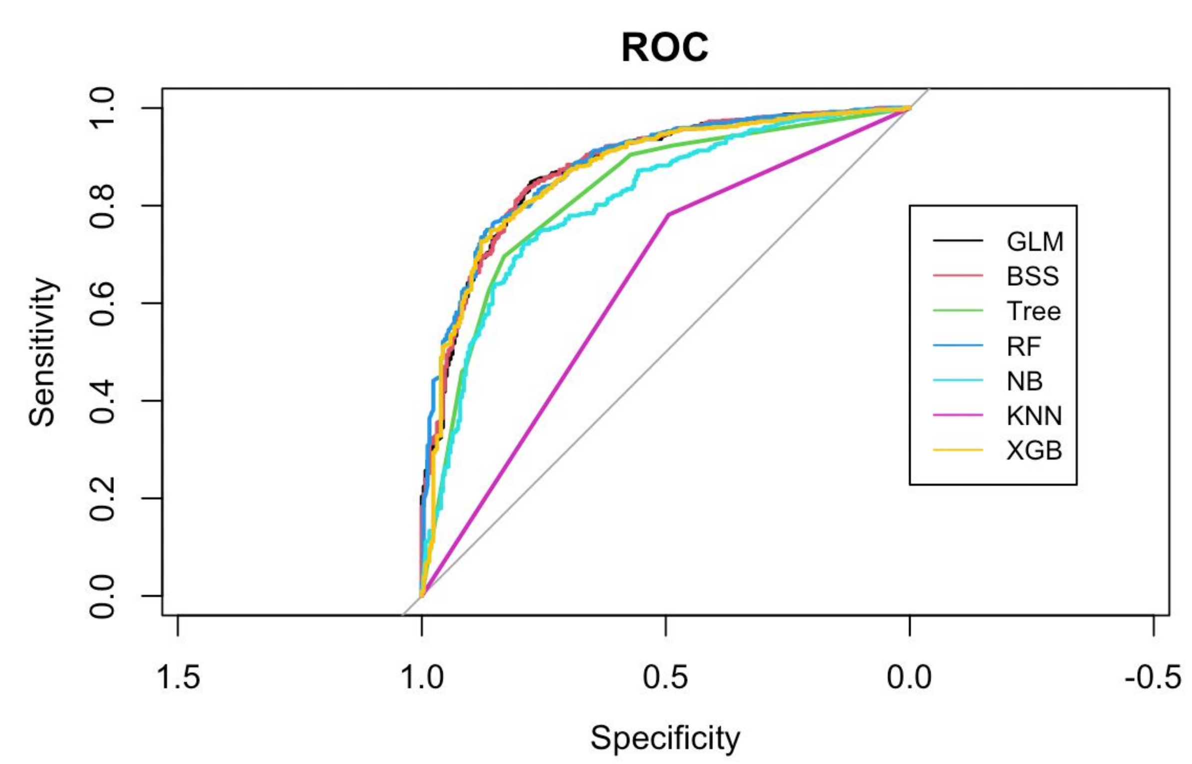
Freeman, P. E. 2021, online at https://github.com/pefreeman /36-290/tree/master/PROJECT_DATASETS/FERMI

NASA. (2020, December 17). FERMILPSC - Fermi LAT 10-Year Point Source Catalog (4FGL-DR2). Retrieved September 26, 2021, from https://heasarc.gsfc.nasa.gov/W3Browse/fermi/fermilpsc.html.

Analysis

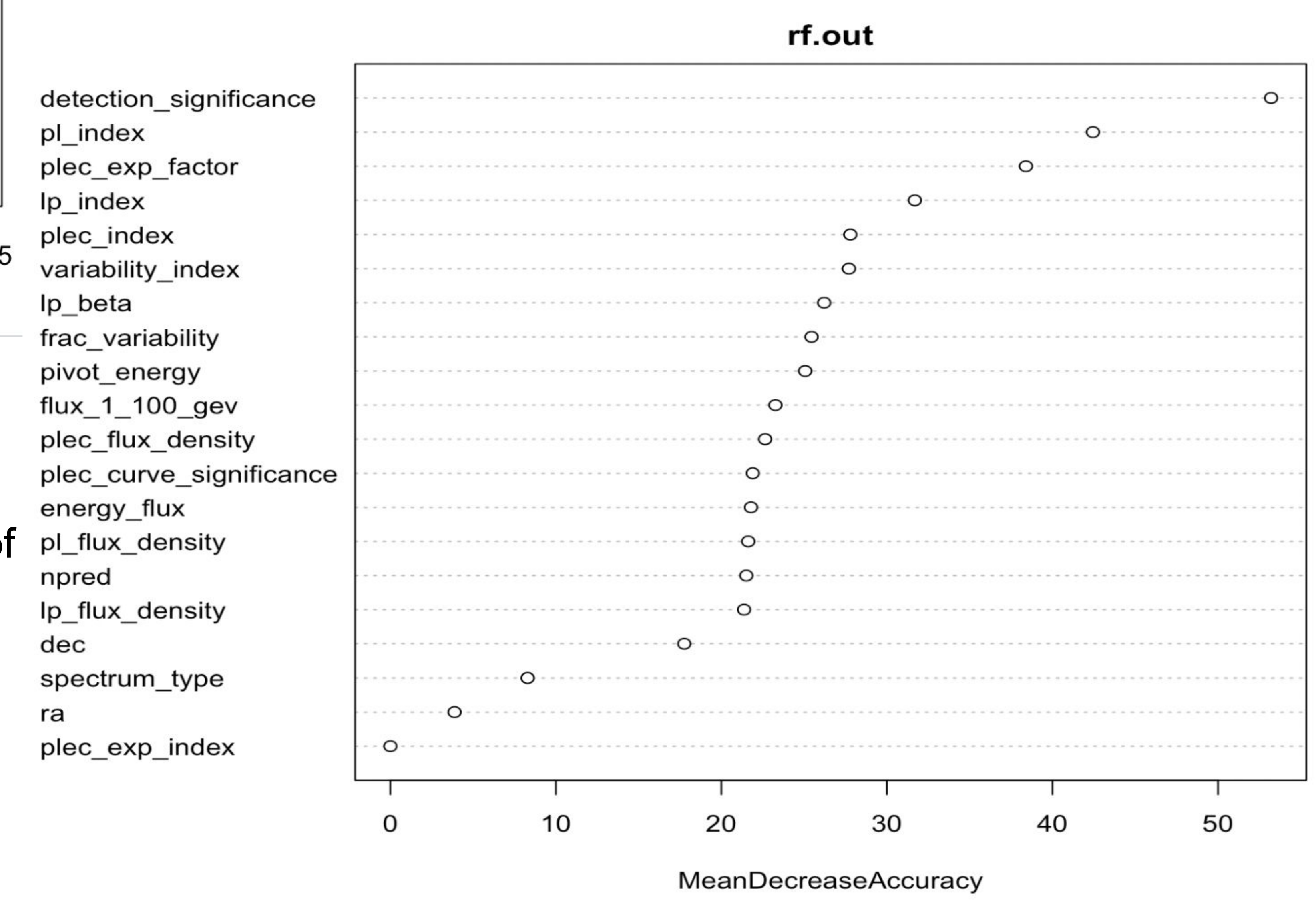
- We attempt to learn a number of classifiers.
  - The methods we use are Logistic Regression, Forward Subset Selection, Classification Tree, Random Forest, Gradient Boosting, K Nearest Neighbors (KNN), and Naive Bayes.
- We use Area Under Curve (AUC) as our metric of model quality.
  - AUC is a measure of how well a model performs over a range of classification thresholds.
  - Possible AUC values range from 0 to 1, and higher values indicate better predictive ability.
- The highest AUCs are from logistic regression, logistic regression with forward selection, and random forest.
  - Logistic regression models the probability that we would sample data of one of the classes, given specific values of the predictor variables.
  - Forward Selection chooses the optimal predictor variables by adding them one at a time and performing logistic regression at each step.
  - Random forest aggregates decision tree constructed using subsets of predictor variables and bootstrapped samples of the dataset.
- We select the Random Forest for our final predictions because it has the highest AUC value.
  - Optimal threshold for random forest: 0.709.
  - MCR: 0.1988

Model	AUC
Logistic Regression	0.876
Forward Selection	0.877
Classification Tree	0.822
Random Forest	**0.879**
Gradient Boosting	0.869
KNN	0.638
Naive Bayes	0.807



- **Above** is the receiver operating characteristic (ROC) curve for all of the classifiers, showing the tradeoff between the ability to identify members of one class well versus members of another class.
- For the Random Forest classifier, the sensitivity is 0.839 (i.e., the classification accuracy for actual BLLs) and the specificity (i.e., the accuracy for non-BLLs) is 0.785.

- **Below** is the variable importance plot for the Random Forest classifier.
- We can see that the most important variable is the one that records signal strength.



Confusion Matrix for Random Forest	Actual BLL	Actual NOT_BLL
Predicted BLL	214	126
Predicted NOT_BLL	41	459

Conclusion

We were successfully able to construct a classifier using the Random Forest model that classifies objects in both classes relatively well, with an MCR 10% lower than if we were to simply classify all objects as non-BLLs.