Efficient Classification of Pulsars

Syd Rothman

January 22, 2020

Target audience

Proposal to a lab planning to complete large data collection in the coming year on the best model for eliminating noise from pulsar candidate datasets.

Current practices:

- Survey-specific models
- Manual inspection

2019

Challenges

Pulsar candidate data volume is rising significantly. Most of these candidates are noise.

Lab Implications:

- More time spent identifying valid candidates than examining them
- Limits potential for real-time classification

Solution

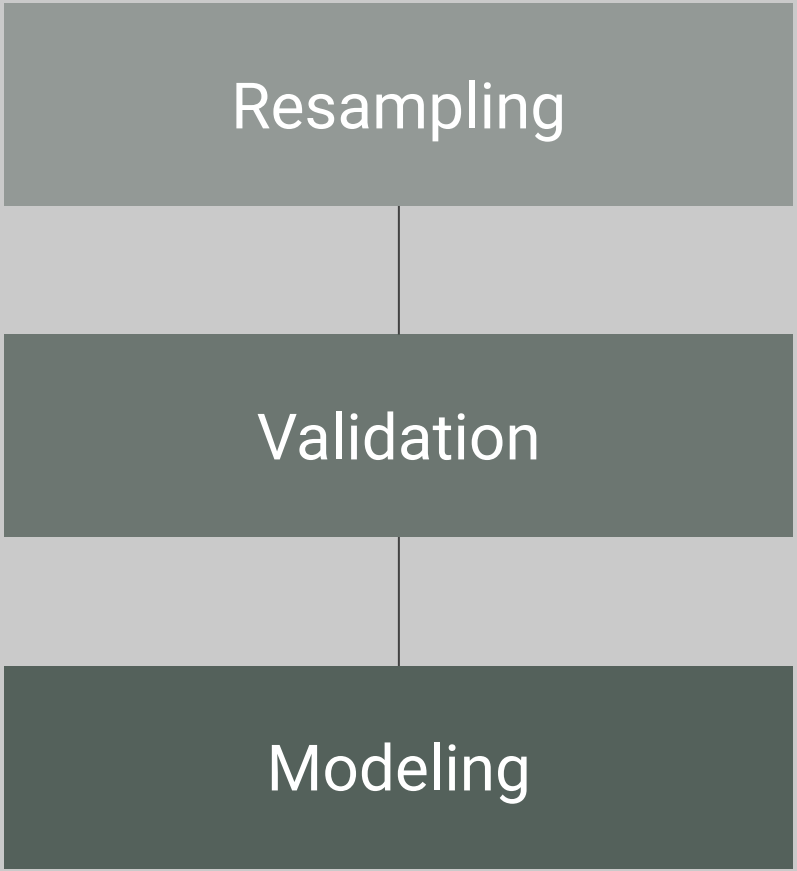
Focus on creating models effective at identifying noise in big data that can be applied as a initial candidate filter.

Lab Applications:

- Survey-specific models can then be applied to the refined data set
- Manual inspection time and telescope time can then be reallocated

Methods

This project used a combination of undersampling and SMOTE to address a large data class imbalance. Train_test_split and kfold validation were both used to ensure model integrity.



Model Performance

Models Used	• Logistic Regression, Support Vector Machine, Naive Bayes, KNN, Linear SVC, Gradient Descent, Decision Tree, Random Forest
Mechanics	• Looping function through kfolds that accepts a model and a scoring metric as arguments
Metrics Used	• Recall score, precision score, accuracy score, f1 score,
Next Steps	• GridSearch to refine parameter tuning • Run comparisons of data resampled differently

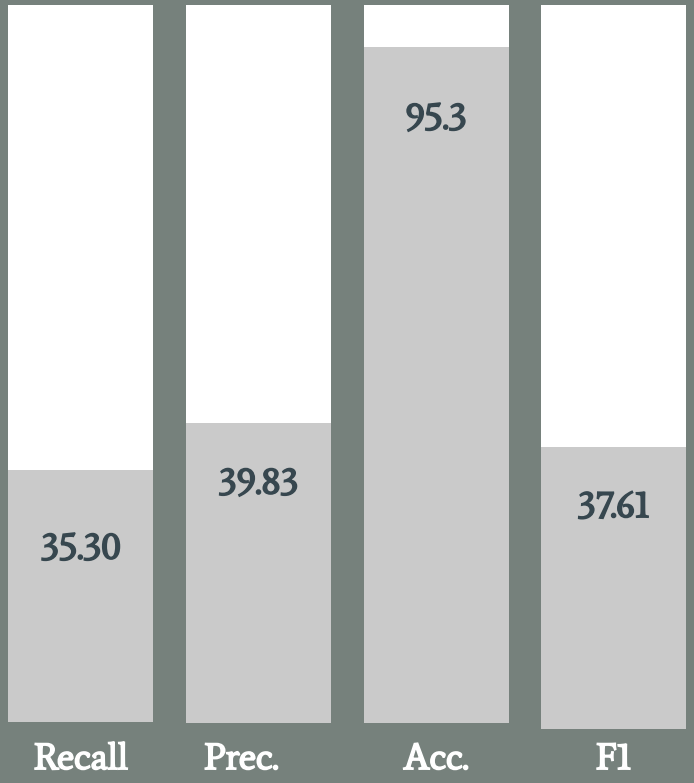
Top Performing Model

Random Forest

Highest mean accuracy score over resampled and validated data at 95.3

Client Implications:

- Highest mean score over kfold validation in 3 of 4 metrics
- Low recall and precision scores, and by extension low f1 score, are likely due to data resampling



Thank You