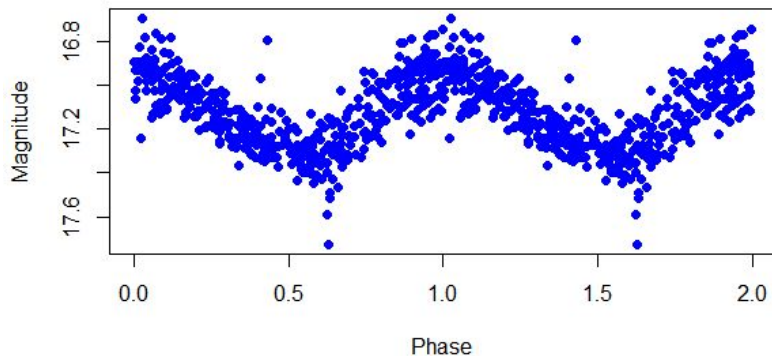


Lightcurves

Brooke Leverton, Kevin Multani, Rachel DeGardner,
Rachel Zilinskas, and Yao Shi
with mentor David Jones



Outline



- The Lightcurve Problem
 - Tree Classification Methods
 - Feature Selection
 - Results
-

Introduction

PROBLEM:

Classify different types of stars

- Data is collected for a large number of stars.
- The data are reduced to features which are then used for classification.

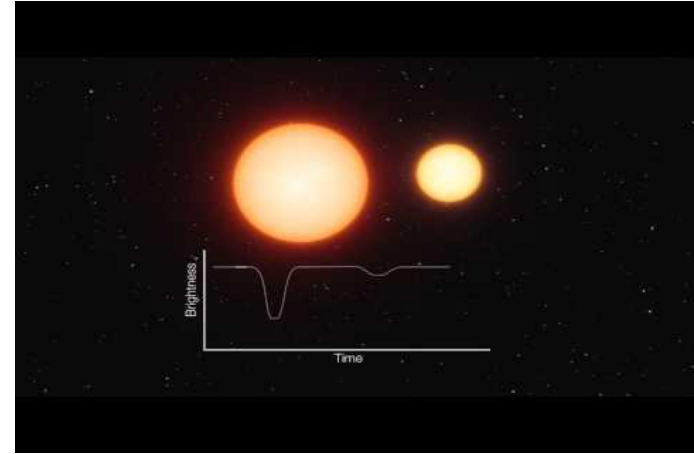
TARGET:

Classification accuracy based on three basic features provided by Catalina Real-Time Transients Survey (CRTS)^[1] is **65.1%**

<https://www.eso.org>

<https://www.spacetelescope.org>

Eclipsing Binaries



Pulsating Stars

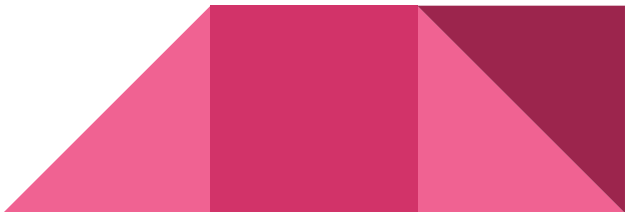


Solution

DATA:

- Raw lightcurves with three basic features

PROCESS:

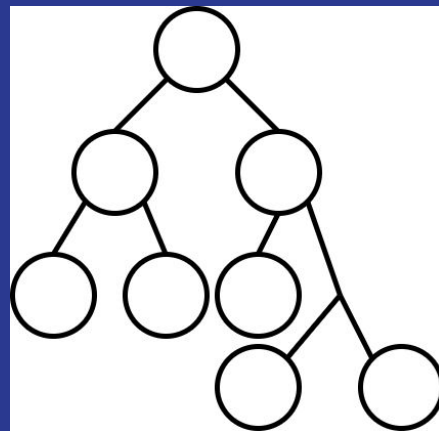
- Compute additional features (*FATS* package in Python)
 - Implement algorithm for classification (*randomForest* in R)
- 

Classification Trees

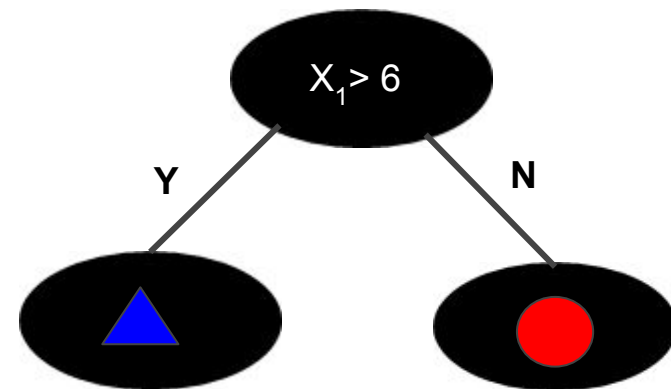
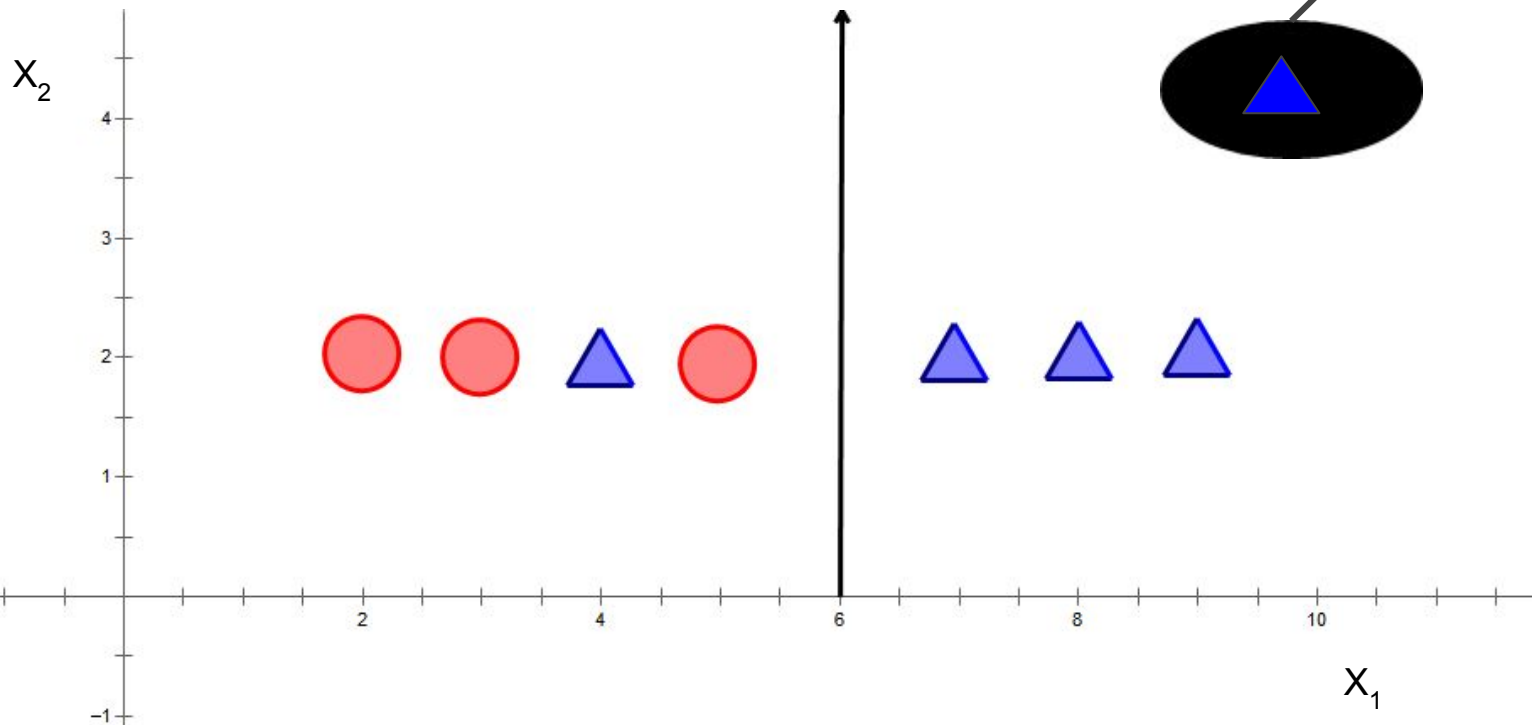
- A hierarchy of binary decisions to assign labels to different objects

Advantages: Simplistic and can be interpreted easily.

Disadvantages: Not very accurate and can be unstable.

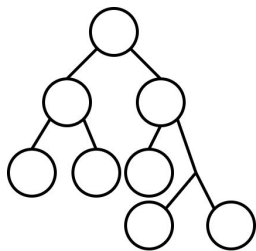


Classification Trees

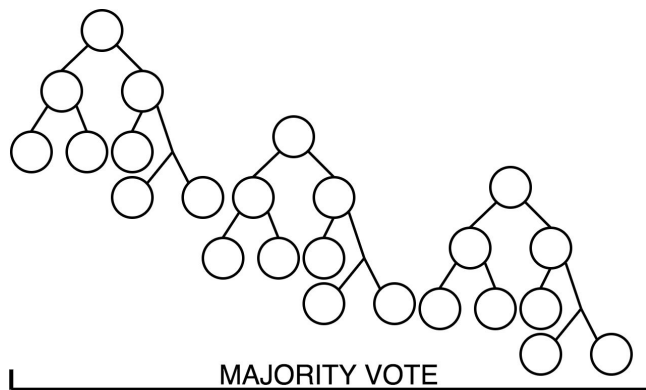


Bootstrap Aggregation (Bagging)

Tree



Bagging



- **Advantages:** Reduces the variance of prediction
- **Disadvantages:** Trees highly correlated, causing bias



Random Forests

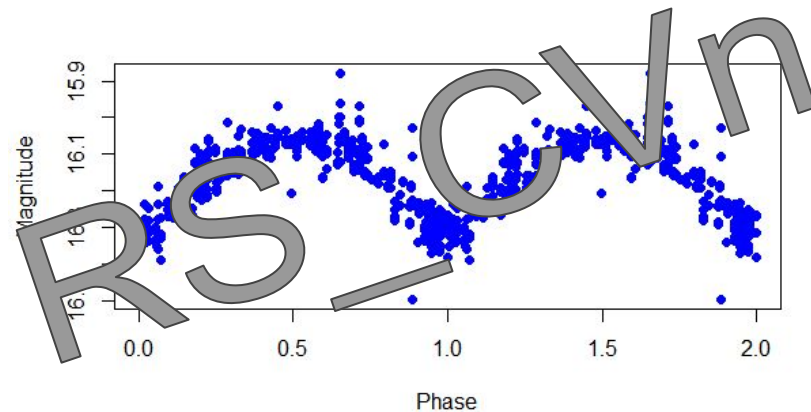
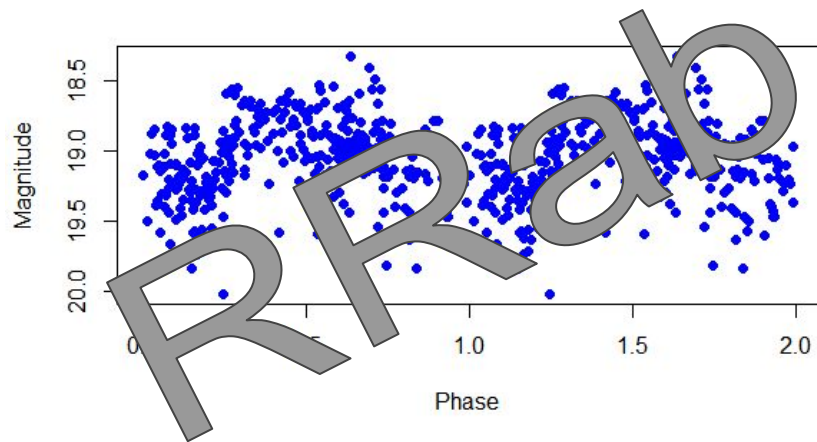
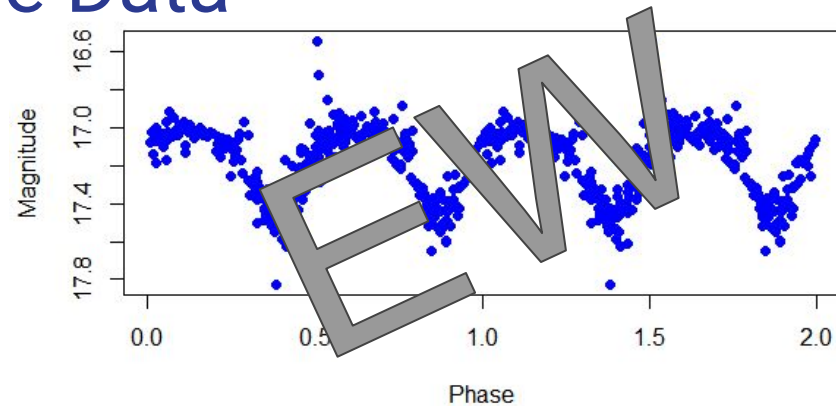
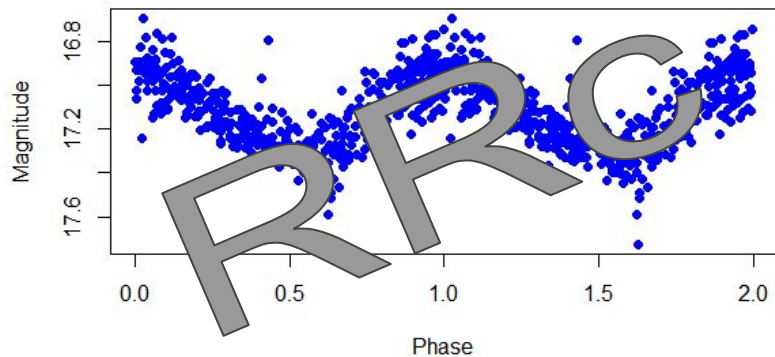
Principle:

- Does bagging, but also randomly selects choice of features at each decision node.
- This decorrelates the trees.
- The final class is chosen by majority voting among the trees.

R Package randomForest:

- Helps identify the features that are most important for classification
- Number of features randomly selected at each node and number of trees can be altered^[3]

Lightcurve Data



Features

CRTS Features:

- Mean magnitude, period, and range for each observed star.
- Random Forest classification accuracy is 65.1%.

Feature Analysis for Time Series (FATS)

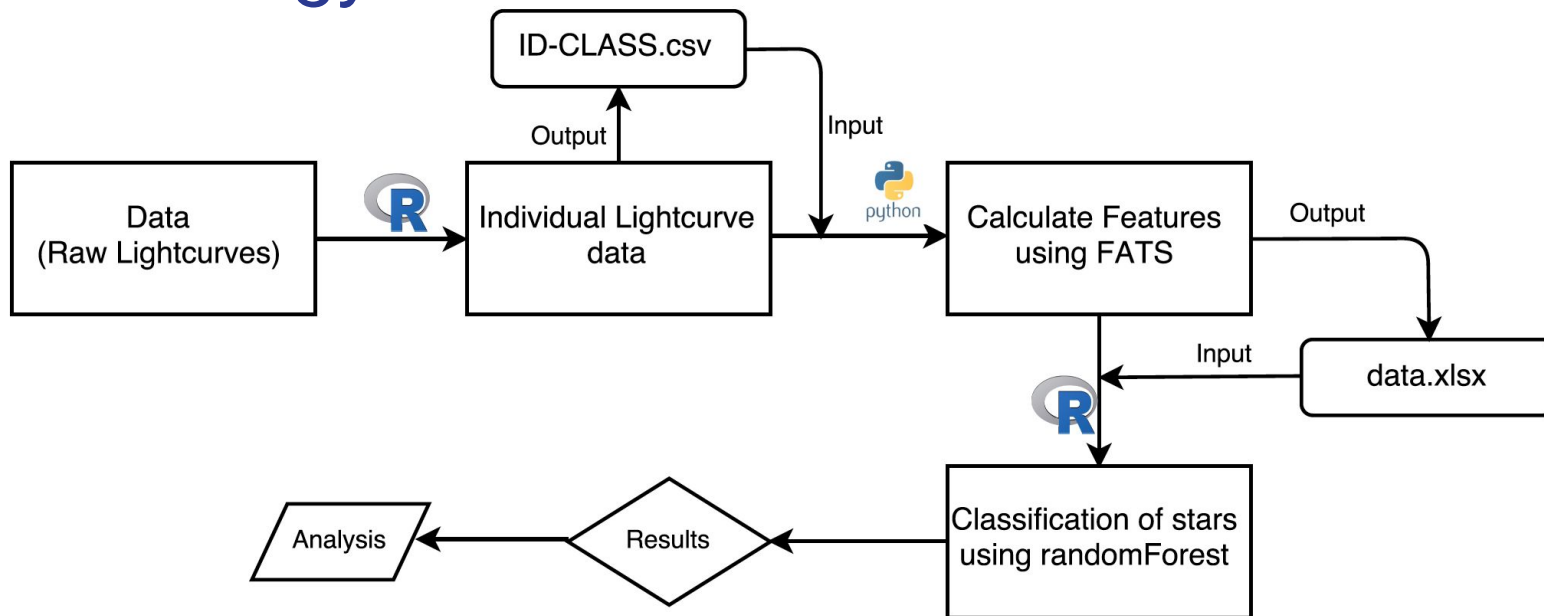


Institute for
Applied Computational Science
HARVARD SCHOOL OF ENGINEERING AND APPLIED SCIENCES

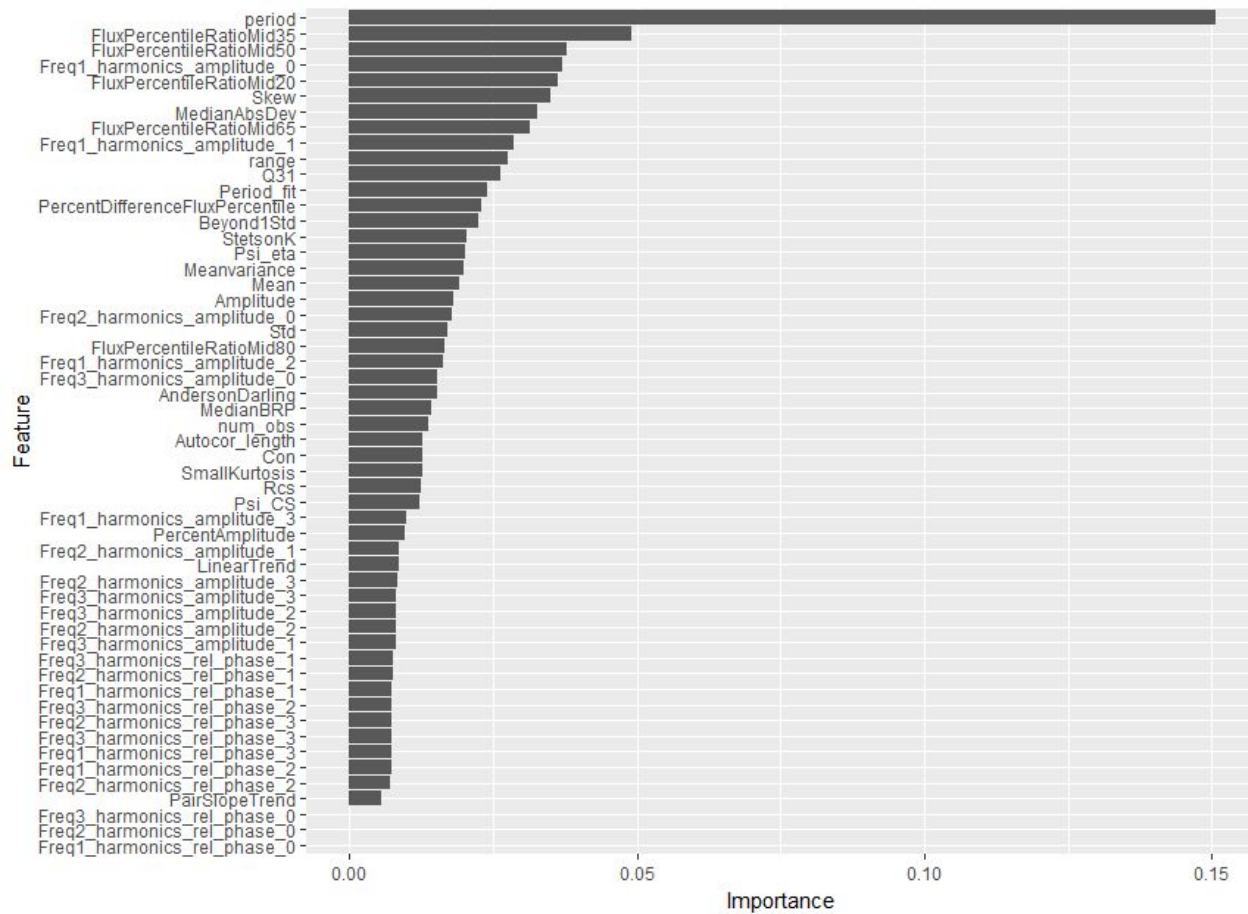


- A library coded in Python that standardizes feature extractions for time series data, such as lightcurve data.
- Created by Isadora Nun, Pavlos Protopapas, and many contributors^[4].
- The raw lightcurve are inputted and it computes more than 50 new features.

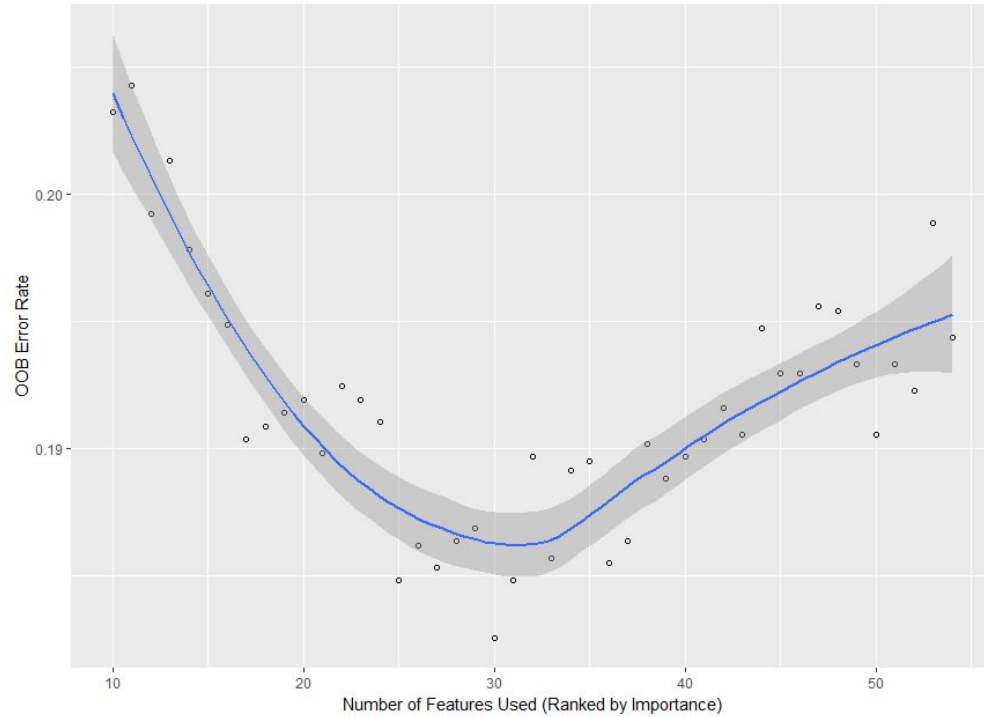
Methodology



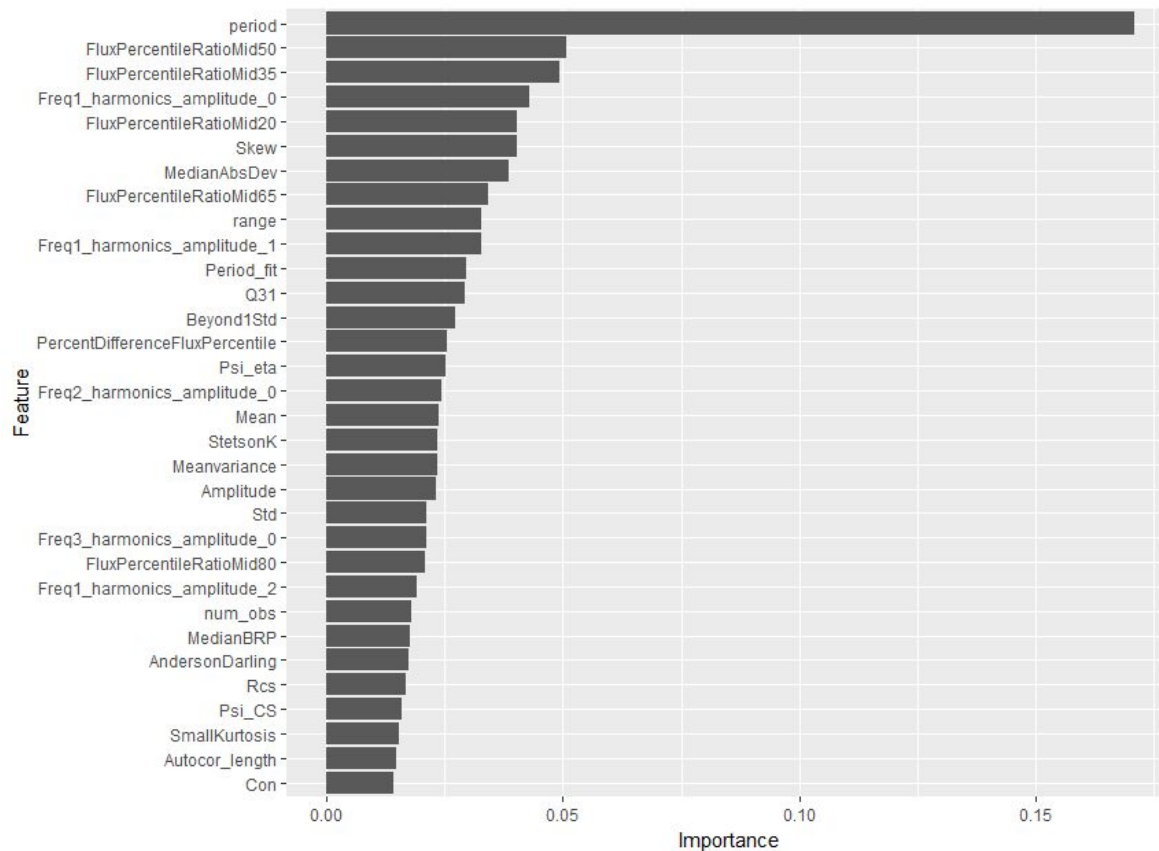
Feature Importance



Out of Bag Error Rate vs. Number of Features Used



Selected Feature Importance



Results

Accuracy for Star Classifications	
Accuracy to beat	65.1%
Training data	81.43%
Testing data	81.59%

Secondary Goal - Eclipsing Binaries Correctly Classified as Eclipsing Binaries	
Accuracy to beat	67.5%
Training data	89.54%
Testing data	90.60%

Moving Forward

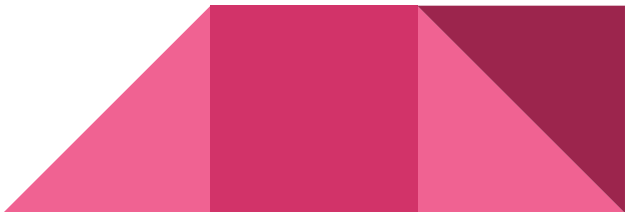
Limitations and Future Work

- Study was limited to periodic star classification
 - Extension to include aperiodic stars
- Extend study to explore other classifiers
 - Support Vector Machine
 - Boosted Trees
- Further feature analysis for optimal combination
- Unsupervised clustering methods
 - Finding new classes



References

- [1] Drake, A. J., M. J. Graham, S. G. Djorgovski, M. Catelan, A. A. Mahabal, G. Torrealba, D. Garc a- lvarez, C. Donalek, J. L. Prieto, R. Williams, S. Larson, E. Christen Sen, V. Belokurov, S. E. Koposov, E. Beshore, A. Boattini, A. Gibbs, R. Hill, R. Kowalski, J. Johnson, and F. Shelly. "The Catalina Surveys Periodic Variable Star Catalog." The Astrophysical Journal Supplement Series 213.1 (2014): 9. Web.
- [2] Richards, Joseph W., Dan L. Starr, Nathaniel R. Butler, Joshua S. Bloom, John M. Brewer, Arien Crellin-Quick, Justin Higgins, Rachel Kennedy, and Maxime Rischard. "On Machine-Learned Classification Of Variable Stars With Sparse And Noisy Time-Series Data." The Astrophysical Journal 733.1 (2011): 10. Web.
- [3] Breiman, Leo, and Adele Cutler. "Random Forests." Random Forests. N.p., n.d. Web. 17 May 2017.
<<https://www.stat.berkeley.edu/~breiman/RandomForests/>>.
- [4] Nun, Isadora, Pavlos Protopapas, Brandon Sim, Ming Zhu, Rahul Dave, Nicolas Castro, and Karim Pichara. "FATS: Feature Analysis for Time Series." [1506.00010] FATS: Feature Analysis for Time Series. N.p., 31 Aug. 2015. Web. 17 May 2017.



Special Thanks!

- David Jones
- Sujit Ghosh
- Thomas Gehrman

