Application of Supervised Machine Learning to the Classification of Variable Young Stars

Philip Carr

Mentor: Lynne Hillenbrand

SURF 2018

Outline

Outline

- Introduction and Motivation
  - What’s this project about?
  - Project scope
  - Why study young stars?
  - Why automate data processing?

Outline

- Introduction and Motivation
  - What’s this project about?
  - Project scope
  - Why study young stars?
  - Why automate data processing?
- Methods
  - Applying supervised machine learning to data
  - Software resources developed for applying machine learning
  - The general machine learning process

Outline

- Introduction and Motivation
  - What’s this project about?
  - Project scope
  - Why study young stars?
  - Why automate data processing?
- Methods
  - Applying supervised machine learning to data
  - Software resources developed for applying machine learning
  - The general machine learning process
- Results
  - Feature importances, best feature set, best classification results

Outline

- Introduction and Motivation
  - What’s this project about?
  - Project scope
  - Why study young stars?
  - Why automate data processing?
- Methods
  - Applying supervised machine learning to data
  - Software resources developed for applying machine learning
  - The general machine learning process
- Results
  - Feature importances, best feature set, best classification results
- Conclusions and Future Work
  - Evaluating current strategies and moving forward
  - Improvements and new research directions

Introduction: Looking Back in Time

https://en.wikipedia.org/wiki/Protoplanetary_disk#/media/File:HL_Tau_protoplanetary_disk.jpg

Introduction: Looking Back in Time

- Young stars are a window into the recent past
  - Indicate how solar system might have developed

https://en.wikipedia.org/wiki/Protoplanetary_disk#/media/File:HL_Tau_protoplanetary_disk.jpg

Introduction: Looking Back in Time

- Young stars are a window into the recent past
  - Indicate how solar system might have developed
- Best look into the various stages of stellar and planetary system evolution

https://en.wikipedia.org/wiki/Protoplanetary_disk#/media/File:HL_Tau_protoplanetary_disk.jpg

Introduction: Looking Back in Time

- Young stars are a window into the recent past
  - Indicate how solar system might have developed
- Best look into the various stages of stellar and planetary system evolution
- The better we understand other young star systems, the better we understand our own solar system

https://en.wikipedia.org/wiki/Protoplanetary_disk#/media/File:HL_Tau_protoplanetary_disk.jpg

Introduction: Looking Back in Time

- Young stars are a window into the recent past
  - Indicate how solar system might have developed
- Best look into the various stages of stellar and planetary system evolution
- The better we understand other young star systems, the better we understand our own solar system
- Young star systems exhibit variety of variability patterns

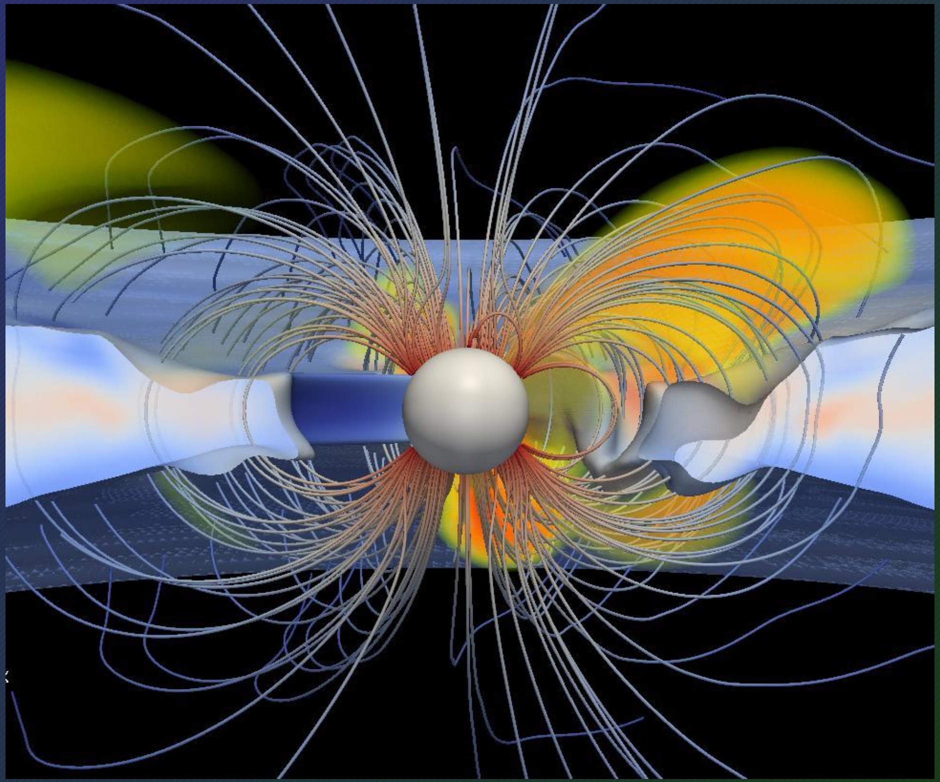
https://en.wikipedia.org/wiki/Protoplanetary_disk#/media/File:HL_Tau_protoplanetary_disk.jpg

Introduction: Looking Back in Time

- Young stars are a window into the recent past
  - Indicate how solar system might have developed
- Best look into the various stages of stellar and planetary system evolution
- The better we understand other young star systems, the better we understand our own solar system
- Young star systems exhibit variety of variability patterns
  - Can we classify young star variability?

https://en.wikipedia.org/wiki/Protoplanetary_disk#/media/File:HL_Tau_protoplanetary_disk.jpg

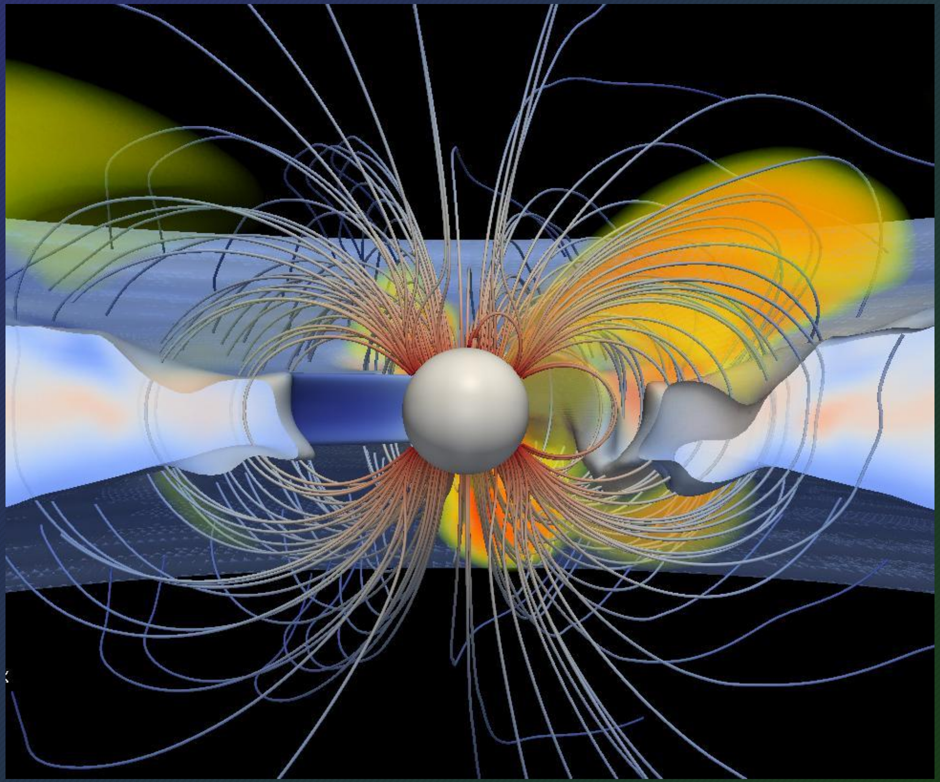
Motivation: Young Stars



http://cerere.astropa.unipa.it/progetti_ricerca/HPC/research.htm

Motivation: Young Stars

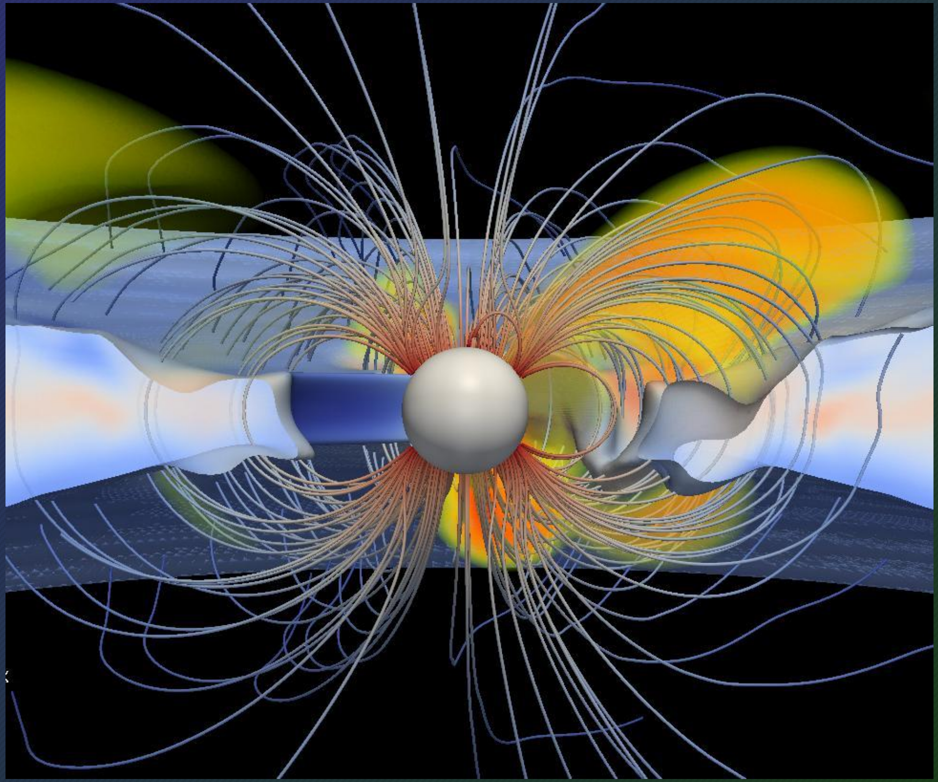
- Millions of years old
  - 1/1000th of star's lifetime



http://cerere.astropa.unipa.it/progetti_ricerca/HPC/research.htm

Motivation: Young Stars

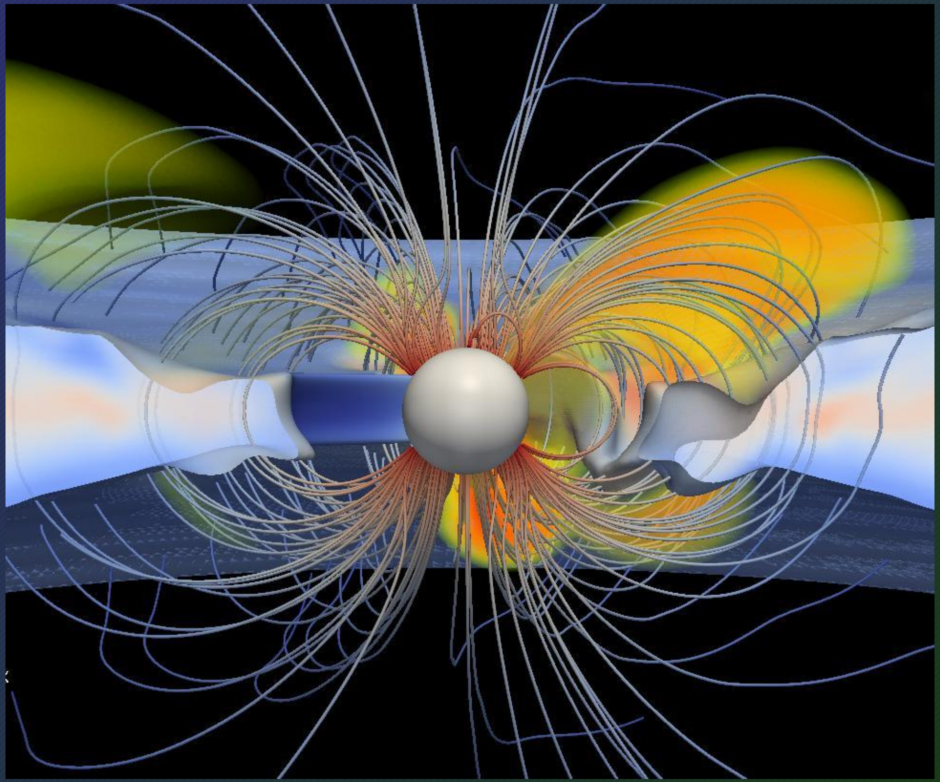
- Millions of years old
  - 1/1000th of star's lifetime
- Just beginning some forms of nuclear fusion in cores
  - e.g. lithium burning, deuterium burning
  - Have yet to reach hydrogen burning



http://cerere.astropa.unipa.it/progetti_ricerca/HPC/research.htm

Motivation: Young Stars

- Millions of years old
  - 1/1000th of star's lifetime
- Just beginning some forms of nuclear fusion in cores
  - e.g. lithium burning, deuterium burning
  - Have yet to reach hydrogen burning
- Highly dynamic
  - Variable flux
  - Accretion disk interaction



http://cerere.astropa.unipa.it/progetti_ricerca/HPC/research.htm

Motivation: Automation

Motivation: Automation

- Age of Time Domain Astronomy

Motivation: Automation

- Age of Time Domain Astronomy
- Dataset increases
  - Surveys increase in size and scope
  - Data resolution (cadence) increases

Motivation: Automation

- Age of Time Domain Astronomy
- Dataset increases
  - Surveys increase in size and scope
  - Data resolution (cadence) increases
- More increasingly, astronomical work once done by-hand becomes clearly more efficient/precise by computer
  - LSST, ZTF

Motivation: Automation

- Age of Time Domain Astronomy
- Dataset increases
  - Surveys increase in size and scope
  - Data resolution (cadence) increases
- More increasingly, astronomical work once done by-hand becomes clearly more efficient/precise by computer
  - LSST, ZTF
- Machine learning

Motivation: Machine Learning

Motivation: Machine Learning

- What is machine learning?
  - Type of artificial intelligence
  - Uses statistical methods to infer patterns of given data
  - Applied in various ways to reach conclusions or perform tasks often more efficiently or practically impossible to do by-hand

Motivation: Machine Learning

- What is machine learning?
  - Type of artificial intelligence
  - Uses statistical methods to infer patterns of given data
  - Applied in various ways to reach conclusions or perform tasks often more efficiently or practically impossible to do by-hand
- How can machine learning help us learn more about young stars?
  - Supervised learning
    - Allows for the classification of data according to known labels

Motivation: Machine Learning

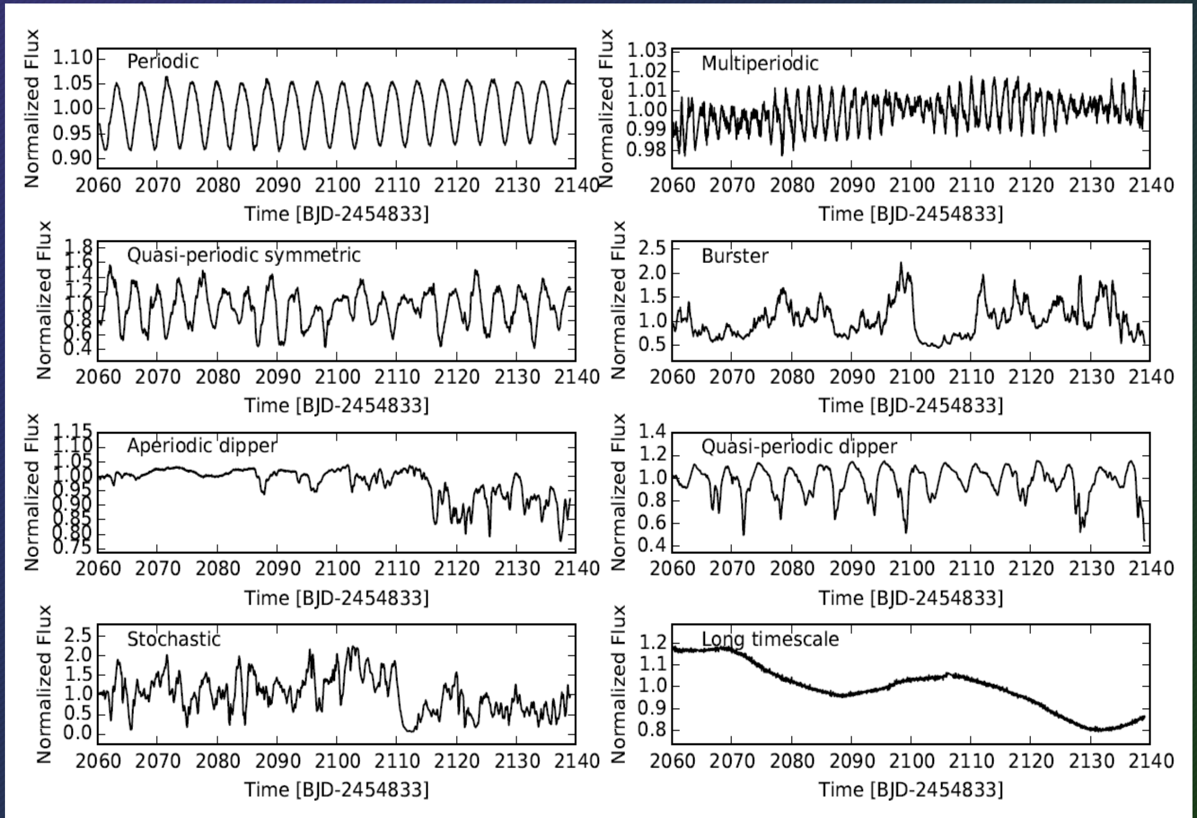
- What is machine learning?
  - Type of artificial intelligence
  - Uses statistical methods to infer patterns of given data
  - Applied in various ways to reach conclusions or perform tasks often more efficiently or practically impossible to do by-hand
- How can machine learning help us learn more about young stars?
  - Supervised learning
    - Allows for the classification of data according to known labels
- Previous examples of machine learning applied in astronomy

Motivation: Machine Learning

- What is machine learning?
  - Type of artificial intelligence
  - Uses statistical methods to infer patterns of given data
  - Applied in various ways to reach conclusions or perform tasks often more efficiently or practically impossible to do by-hand
- How can machine learning help us learn more about young stars?
  - Supervised learning
    - Allows for the classification of data according to known labels
- Previous examples of machine learning applied in astronomy
  - Classifying other types of variable stars (Richards, J. W. et al. 2011)
  - Applying recurrent neural networks and feature engineering for prediction and classification of stellar properties (Hinners et al. 2018)

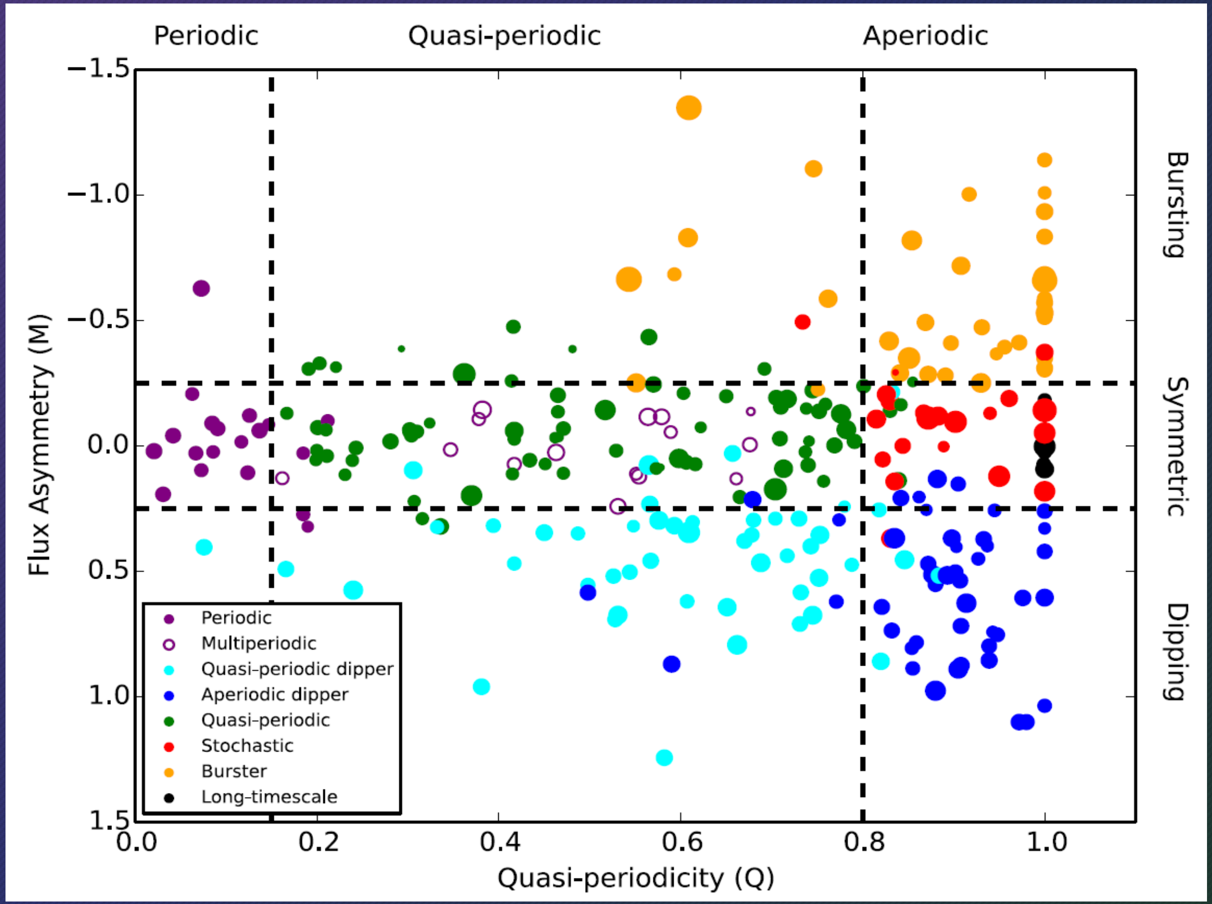
Machine Learning: Data Organization

- About the Data
  - Light curves
  - Obtained from Kepler K2 mission
  - Data were labelled (supervised learning)
  - 8 variability types (Cody & Hillenbrand 2018)
    - Periodicity
    - Flux Asymmetry



Adapted from Cody, A. M. & Hillenbrand, L. A. 2018

Motivation: Young Stars + Automation



Methods: Technical Organization and Data Organization

Methods: Technical Organization and Data Organization

- Jupyter Notebooks running Python 3.6
- Python source code

Methods: Technical Organization and Data Organization

- Jupyter Notebooks running Python 3.6
- Python source code
- Data directory containing 274 labelled light curves used in machine learning applications

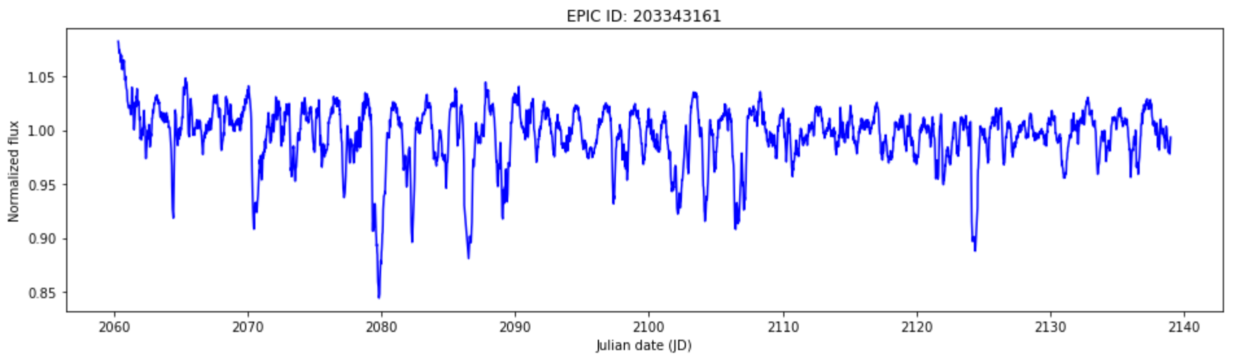
Methods: Technical Organization and Data Organization

- Jupyter Notebooks running Python 3.6
- Python source code
- Data directory containing 274 labelled light curves used in machine learning applications
- Besides the machine learning notebooks, several Notebooks created for assistance with data processing, data verification, and feature development

Methods: Technical Organization and Data Organization

- Jupyter Notebooks running Python 3.6
- Python source code
- Data directory containing 274 labelled light curves used in machine learning applications
- Besides the machine learning notebooks, several Notebooks created for assistance with data processing, data verification, and feature development
- Reading in Data
  - Data read in through Python
  - Option to remove first n days of observation from original light curve

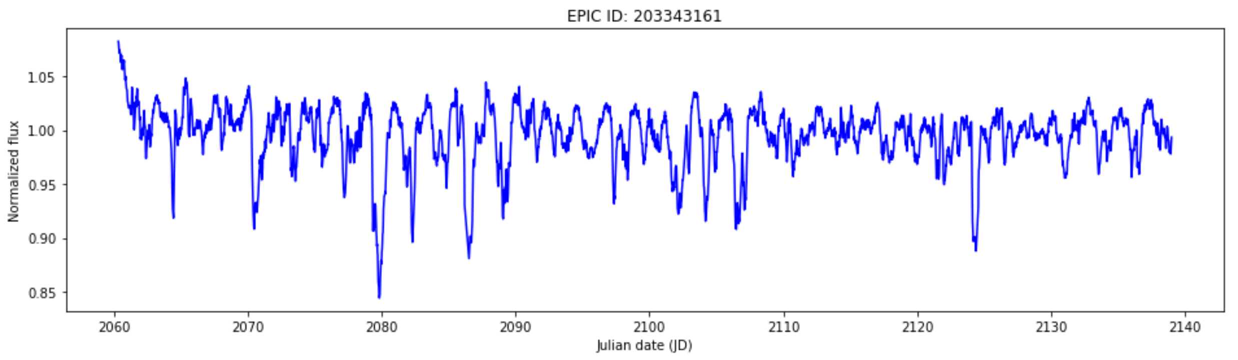
Machine Learning: Feature Selection



Feature	Value
Normalized Flux Amplitude	0.0869
Timescale	0.237
Quasi-periodicity	0.771
Flux Asymmetry	0.621

Machine Learning: Feature Selection

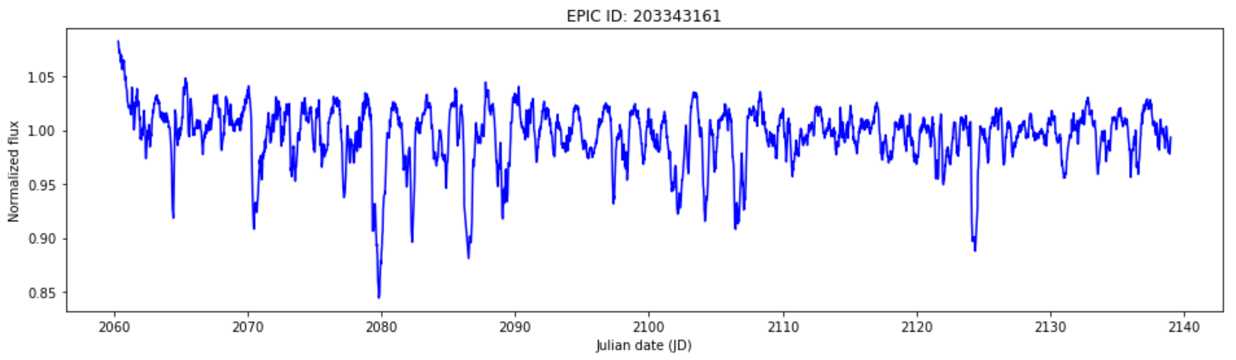
- About feature selection
  - Quantitative data indicating some kind of statistical and/or physical behavior
  - Supervised learning algorithm only sees feature data, not necessarily raw data



Feature	Value
Normalized Flux Amplitude	0.0869
Timescale	0.237
Quasi-periodicity	0.771
Flux Asymmetry	0.621

Machine Learning: Feature Selection

- About feature selection
  - Quantitative data indicating some kind of statistical and/or physical behavior
  - Supervised learning algorithm only sees feature data, not necessarily raw data
- Feature selection
  - Specialized features
    - Based largely off of classification system (including periodicity, flux asymmetry, timescale, flux amplitude, etc.)
  - Public feature libraries
    - FATS
    - feets



Feature	Value
Normalized Flux Amplitude	0.0869
Timescale	0.237
Quasi-periodicity	0.771
Flux Asymmetry	0.621

Machine Learning: Training a Classifier

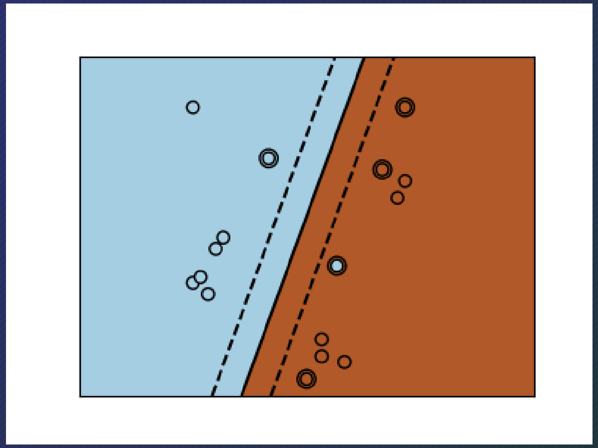
Machine Learning: Training a Classifier

- Instances of Supervised Learning Classification Algorithm
- Given features and labels, classifier trains on the data
  - Different classifiers utilize different techniques to detect patterns in feature data
  - Training ultimately establishes relationship between features and labels
  - Training cannot be done on entire set of labelled data
    - 70% used as training data proportion

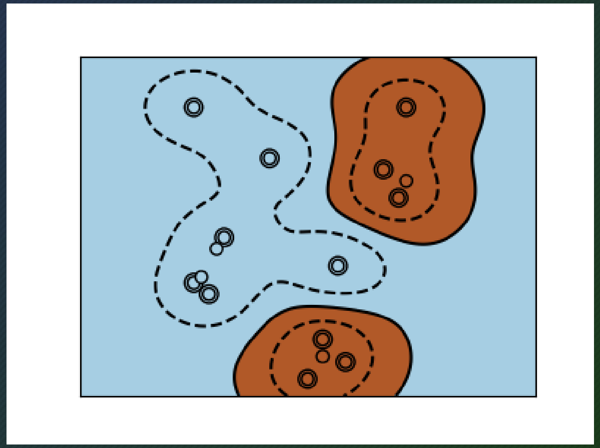
Machine Learning: Training a Classifier

- Instances of Supervised Learning Classification Algorithm
- Given features and labels, classifier trains on the data
  - Different classifiers utilize different techniques to detect patterns in feature data
  - Training ultimately establishes relationship between features and labels
  - Training cannot be done on entire set of labelled data
    - 70% used as training data proportion
- Once trained, classifier can make predictions based on how it learned from training data

Machine Learning: Optimization



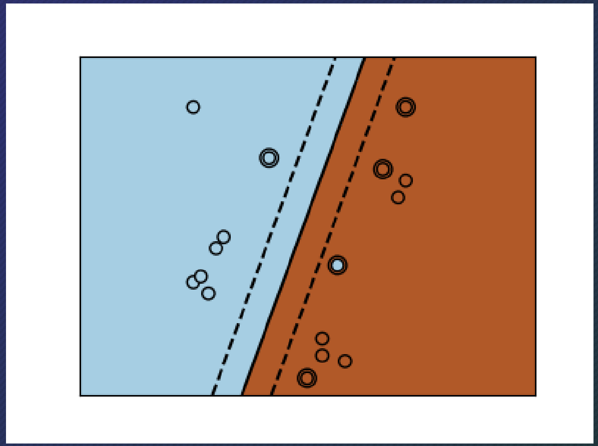
http://scikit-learn.org/stable/auto_examples/svm/plot_svm_kernels.html#sphx-glr-auto-examples-svm-plot-svm-kernels-py



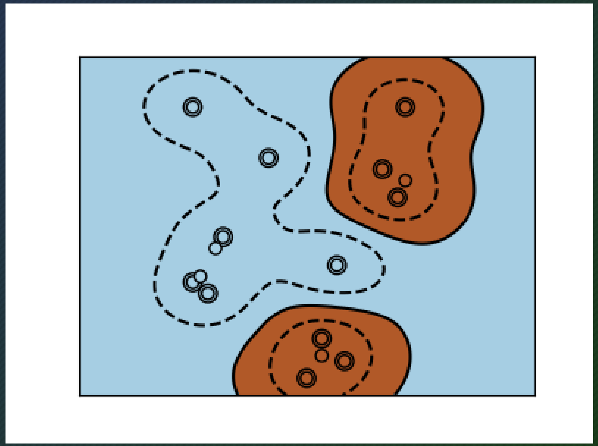
http://scikit-learn.org/stable/auto_examples/svm/plot_svm_kernels.html#sphx-glr-auto-examples-svm-plot-svm-kernels-py

Machine Learning: Optimization

- Hyperparameters
  - Parameters that modify different aspects of machine learning algorithm
  - Can have significant impact on how well classifier learns from data
  - Example:
    - Linear vs. RBF kernel



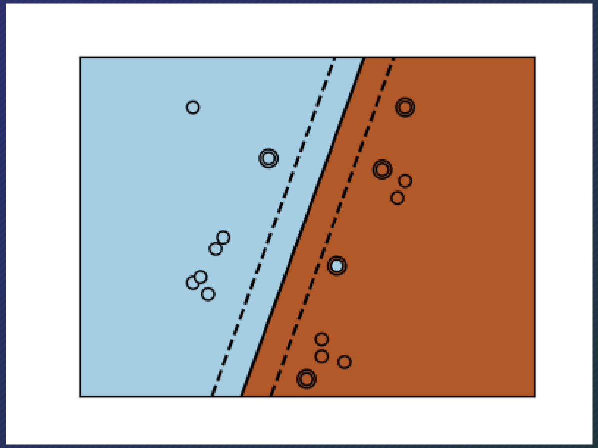
http://scikit-learn.org/stable/auto_examples/svm/plot_svm_kernels.html#sphx-glr-auto-examples-svm-plot-svm-kernels-py



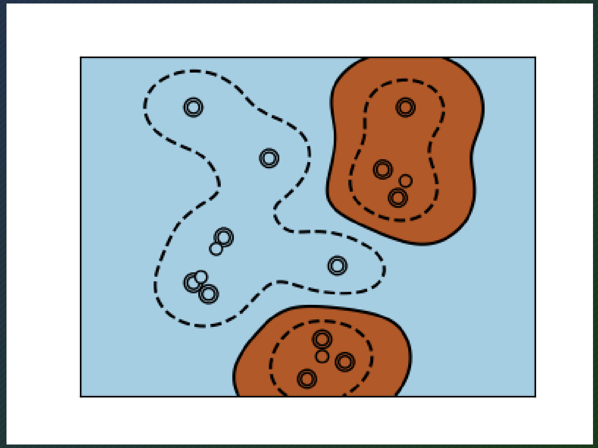
http://scikit-learn.org/stable/auto_examples/svm/plot_svm_kernels.html#sphx-glr-auto-examples-svm-plot-svm-kernels-py

Machine Learning: Optimization

- Hyperparameters
  - Parameters that modify different aspects of machine learning algorithm
  - Can have significant impact on how well classifier learns from data
  - Example:
    - Linear vs. RBF kernel
- Sklearn implements two optimization methods
  - Randomized Search
  - Grid Search



http://scikit-learn.org/stable/auto_examples/svm/plot_svm_kernels.html#sphx-glr-auto-examples-svm-plot-svm-kernels-py



http://scikit-learn.org/stable/auto_examples/svm/plot_svm_kernels.html#sphx-glr-auto-examples-svm-plot-svm-kernels-py

Machine Learning: Performance Metrics

Machine Learning: Performance Metrics

- Feature Importances

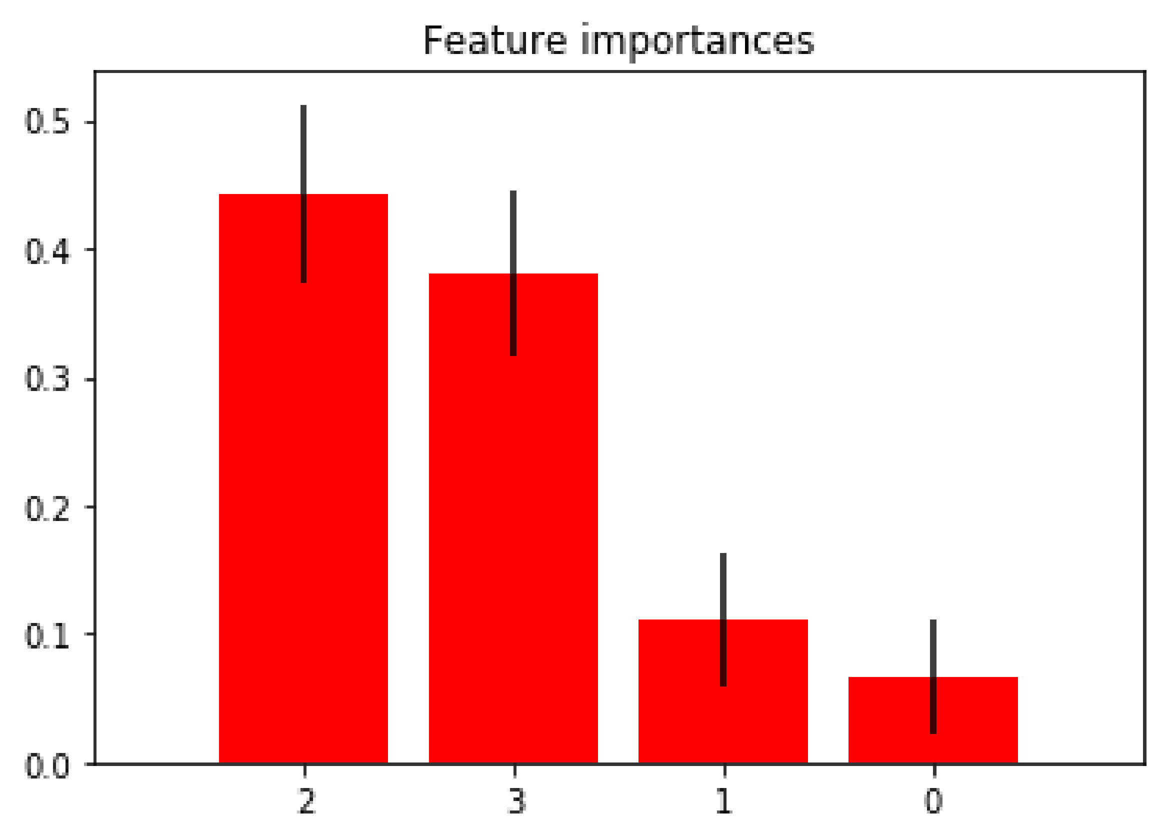
Machine Learning: Performance Metrics

- Feature Importances
- Score (weighted accuracy)
  - number correctly predicted / total
- Balanced accuracy
  - number correctly predicted of a certain label / total of a certain label averaged over all labels

Machine Learning: Performance Metrics

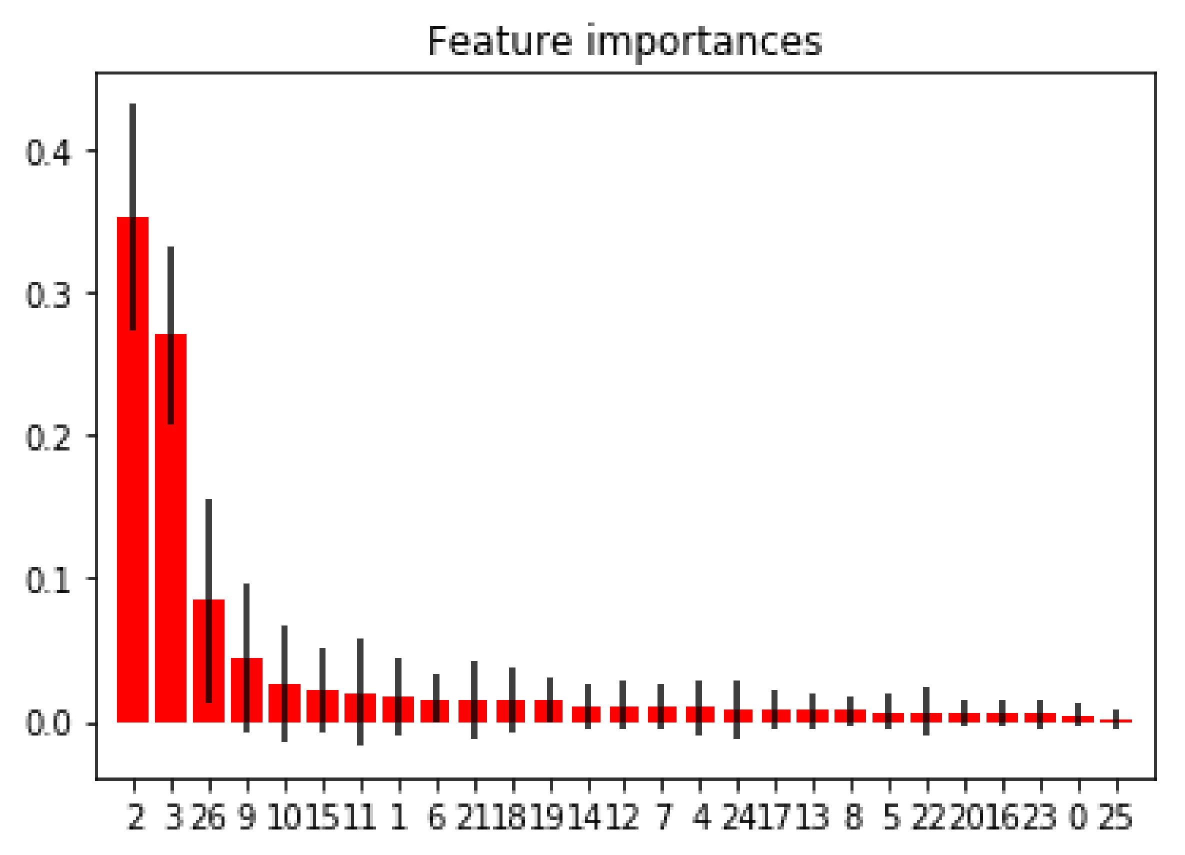
- Feature Importances
- Score (weighted accuracy)
  - number correctly predicted / total
- Balanced accuracy
  - number correctly predicted of a certain label / total of a certain label averaged over all labels
- Confusion matrices
  - Show detailed information about how well objects are classified
  - Show ways in which classifier confuses different objects in prediction

Results: Feature Importances



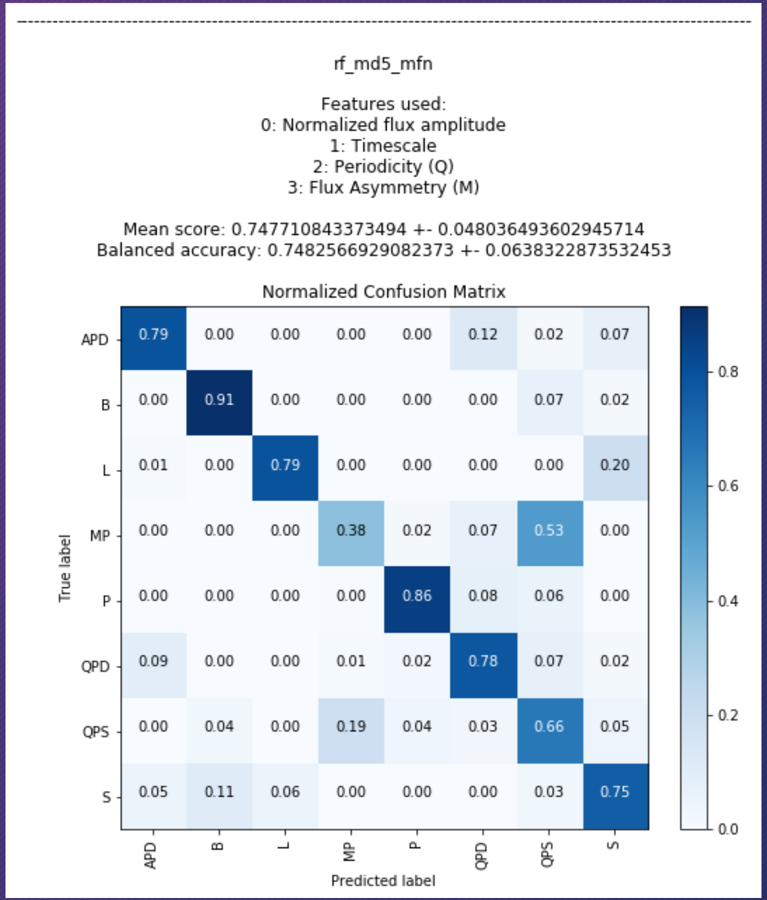
Feature Number	Value
0	Normalized Flux Amplitude
1	Timescale
2	Quasi-periodicity
3	Flux Asymmetry

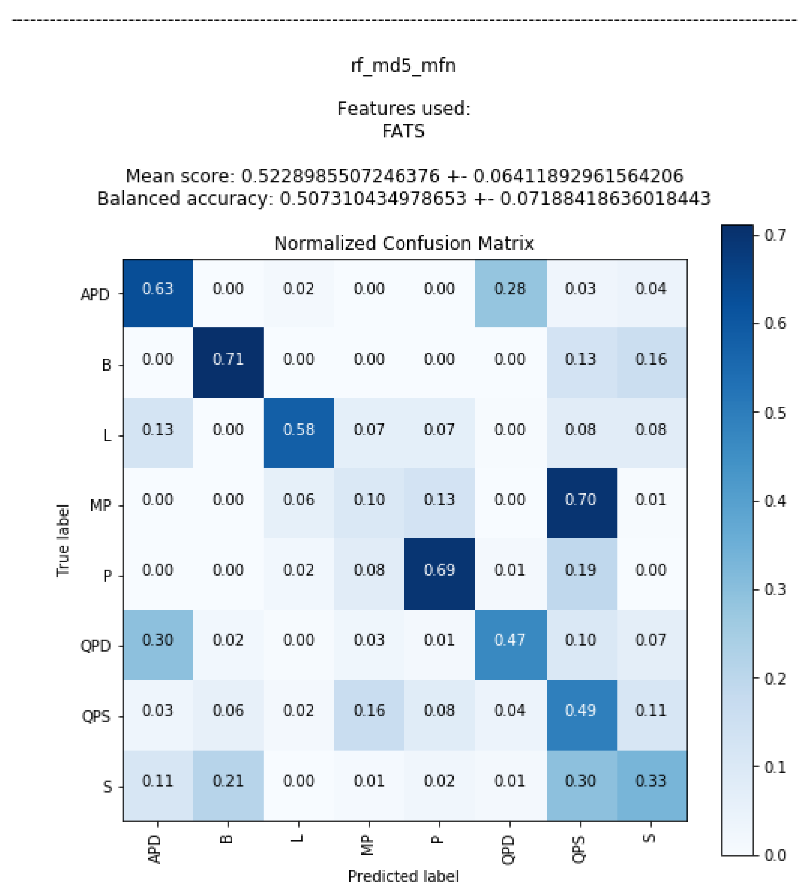
Results: Feature Importances

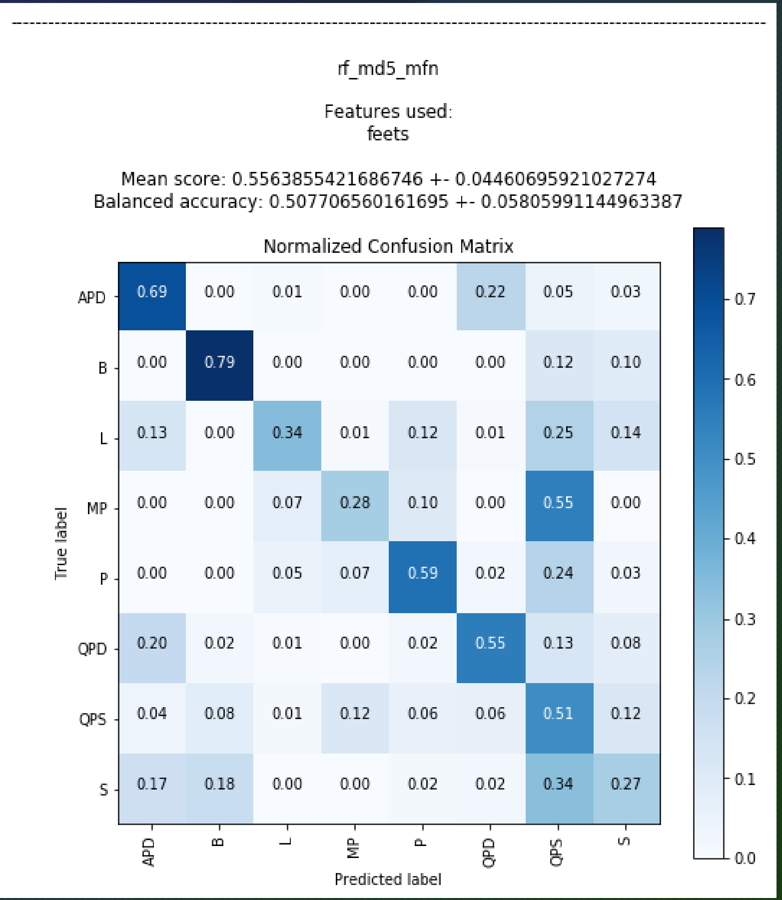


Feature Number	Value
2	Quasi-periodicity
3	Flux Asymmetry
26	smoothed light curve polynomial fit rms error
9	Stetson k index
10	half-magnitude amplitude ratio

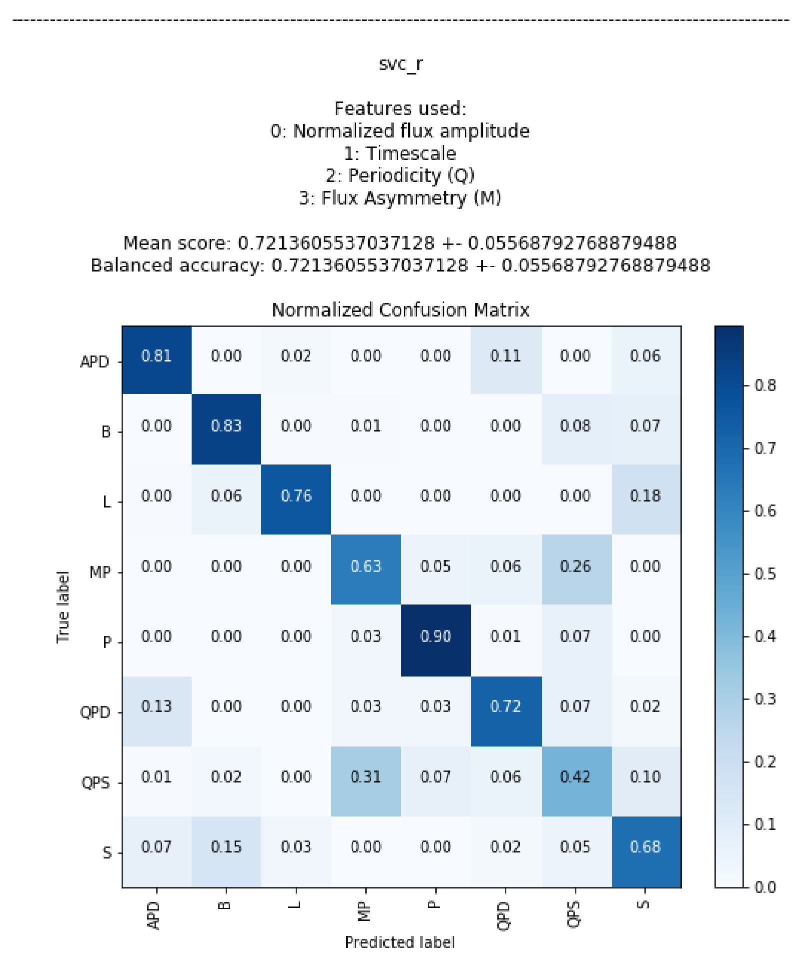
Results: Specialized vs. FATS vs. feets Features

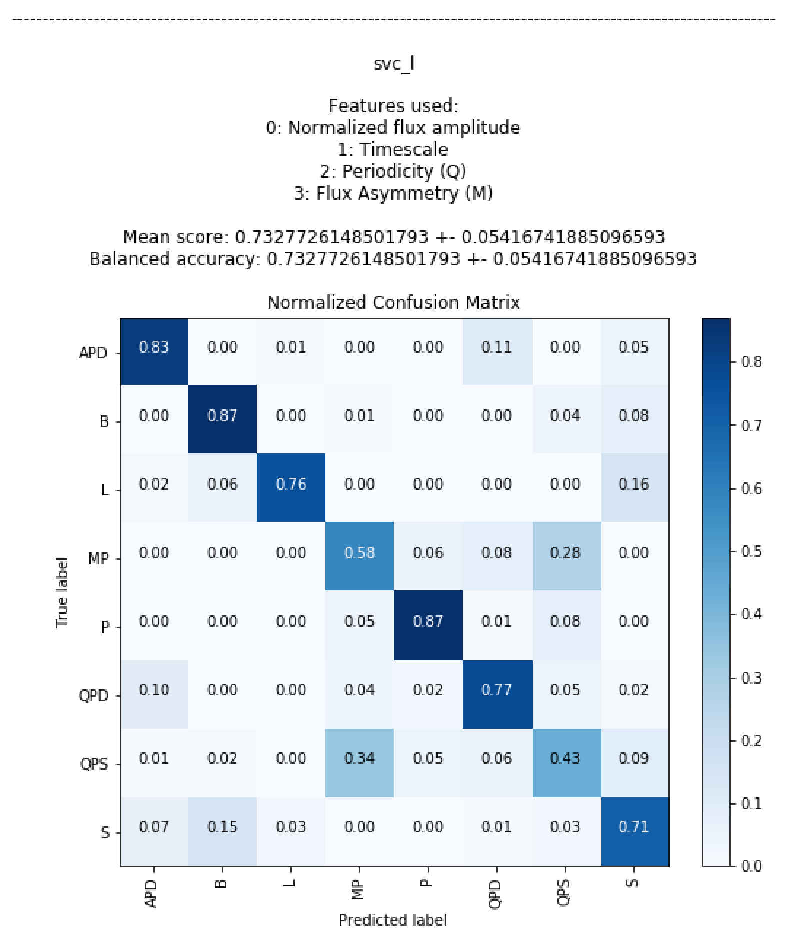




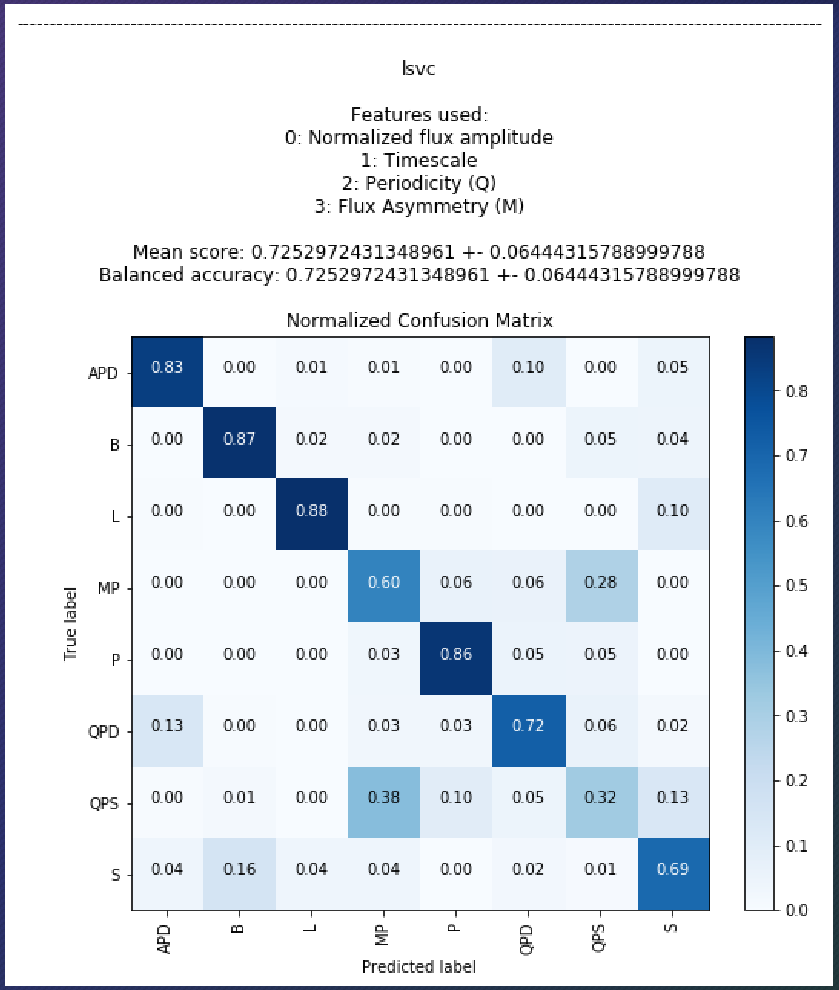


Results: Classification Performance Highlights (SVC)





Results: Classification Performance Highlights (LinearSVC)



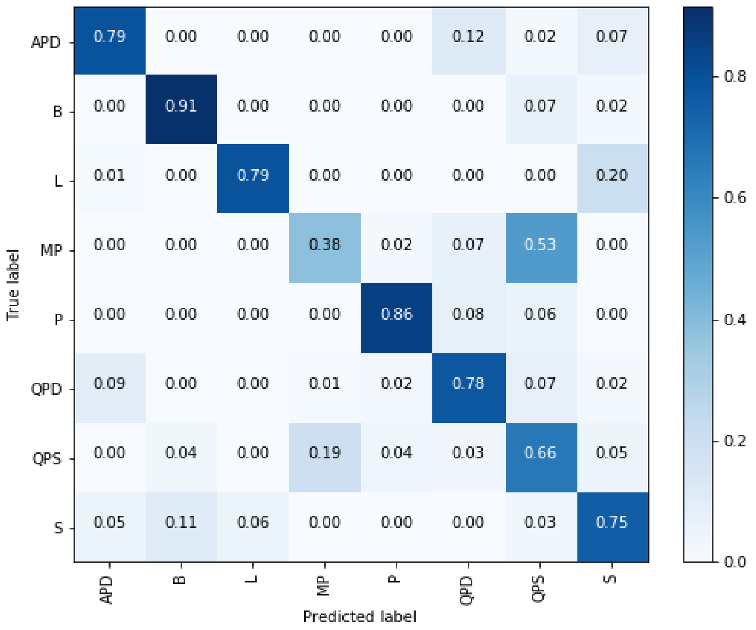
Results: Classification Performance Highlights (Random Forest)

rf_md5_mfn

Features used:
0: Normalized flux amplitude
1: Timescale
2: Periodicity (Q)
3: Flux Asymmetry (M)

Mean score: 0.747710843373494 +- 0.048036493602945714
Balanced accuracy: 0.7482566929082373 +- 0.0638322873532453

Normalized Confusion Matrix



Conclusions

Conclusions

- Feature Importances of specialized features
  - Periodicity and flux asymmetry dominate specialized features
  - Next best features were smoothed light curve polynomial fit rms error, Stetson k index, and half-magnitude amplitude ratio

Conclusions

- Feature Importances of specialized features
  - Periodicity and flux asymmetry dominate specialized features
  - Next best features were smoothed light curve polynomial fit rms error, Stetson k index, and half-magnitude amplitude ratio
- Specialized vs. FATS vs. feets features
  - Specialized features work best

Conclusions

- Feature Importances of specialized features
  - Periodicity and flux asymmetry dominate specialized features
  - Next best features were smoothed light curve polynomial fit rms error, Stetson k index, and half-magnitude amplitude ratio
- Specialized vs. FATS vs. feets features
  - Specialized features work best
- Classification Results Highlights
  - Best classification results capped at ~75% weighted, balanced accuracies
  - Random Forest leads slightly as best classifier with weighted accuracy of 75 ± 5 % and balanced accuracy of 75 ± 6 %
  - Classifiers struggle with MP vs. QPS, APD vs. QPD, S in general

Conclusions

- Feature Importances of specialized features
  - Periodicity and flux asymmetry dominate specialized features
  - Next best features were smoothed light curve polynomial fit rms error, Stetson k index, and half-magnitude amplitude ratio
- Specialized vs. FATS vs. feets features
  - Specialized features work best
- Classification Results Highlights
  - Best classification results capped at ~75% weighted, balanced accuracies
  - Random Forest leads slightly as best classifier with weighted accuracy of 75 ± 5 % and balanced accuracy of 75 ± 6 %
  - Classifiers struggle with MP vs. QPS, APD vs. QPD, S in general

- Supervised learning has great potential for automated variable young star classification!

Future Work

Future Work

Future Work

Future Work

- New features
  - targeted to differentiate variability types current classifiers confuse
    - MP vs. QPS, APD vs. QPD, S in general, etc.

Future Work

- New features
  - targeted to differentiate variability types current classifiers confuse
    - MP vs. QPS, APD vs. QPD, S in general, etc.
- Applying ML methods to other time-domain datasets

Future Work

- New features
  - targeted to differentiate variability types current classifiers confuse
    - MP vs. QPS, APD vs. QPD, S in general, etc.
- Applying ML methods to other time-domain datasets
- Increasing scope of data analysis pipeline
  - Processing (labelled) raw data -> processed data -> feature data-> machine learning methods

Future Work

- New features
  - targeted to differentiate variability types current classifiers confuse
    - MP vs. QPS, APD vs. QPD, S in general, etc.
- Applying ML methods to other time-domain datasets
- Increasing scope of data analysis pipeline
  - Processing (labelled) raw data -> processed data -> feature data-> machine learning methods
- Using new classifiers
  - E.g. Decision Tree, Extremely Randomized Trees, AdaBoost, etc.
  - Neural Networks
    - Convolutional Neural Networks

Future Work

- New features
  - targeted to differentiate variability types current classifiers confuse
    - MP vs. QPS, APD vs. QPD, S in general, etc.
- Applying ML methods to other time-domain datasets
- Increasing scope of data analysis pipeline
  - Processing (labelled) raw data -> processed data -> feature data-> machine learning methods
- Using new classifiers
  - E.g. Decision Tree, Extremely Randomized Trees, AdaBoost, etc.
  - Neural Networks
    - Convolutional Neural Networks
- Unsupervised Learning
  - clustering algorithms
  - re-evaluating variable type classifications, focusing on different light curve properties

Acknowledgements

- Mentor: Dr. Lynne Hillenbrand
- ZTF Summer School
- Dr. Ann Marie Cody
- scikit-learn documentation
- Flintridge Foundation

References

- Armstrong, D. J., Kirk, J., Lam, K. W. F., et al. 2016. K2 variable catalogue II: Machine learning classification of variable stars and eclipsing binaries in K2 fields 0-4. *Monthly Notices of the Royal Astronomical Society, 456*, 2260-2272. doi:10.1093/mnras/stv2836
- Bloom, J. S. & Richards, J. W. (2012). *Data mining and machine learning in time-domain discovery and classification*. Boca Raton, FL: CRC Press.
- Cody, A. M., Hillenbrand, L. A. 2018. The many-faceted light curves of young disk-bearing stars in Upper Sco and ρ Oph observed by K2 Campaign 2. eprint arXiv:1802.06409
- Cody, A. M., Stauffer, J., Baglin, A., et al. 2014. CSI 2264: Simultaneous optical and infrared light curves of young disk-bearing stars in NGC 2264 with *Corot* and *Spitzer*- evidence for multiple origins of variability. *The Astronomical Journal, 147*, 47 pp. doi:10.1088/0004-6256/147/4/82
- Hedges, C., Hodgkin, S., Kennedy, G. 2018. Discovery of new dipper stars with K2: A window into the inner disk region of T Tauri stars. *Monthly Notices of the Royal Astronomical Society, Advance Access*. doi:10.1093/mnras/sty328
- Herbst, W. 2012. The variability of young stellar objects. *Journal of the American Association of Variable Star Observers, 40*, 448-455. Retrieved from https://www.aavso.org/ejaavso401448
- Hinners, T. A., et al. 2018. Machine Learning Techniques for Stellar Light Curve Classification. *The Astronomical Journal, 156*, 13 pp. doi:10.3847/1538-3881/aac16d
- Mackenzie, C., Pichara, K., Protopapas, P. 2016. Clustering based feature learning on variable stars. *The Astrophysical Journal, 820*, 15 pp. doi:10.3847/0004-637X/820/2/138
- Richards, J. W., Starr, D. L., Butler, N. R., et al. 2011. On machine-learned classification of variable stars with space and noisy time-series data. *The Astrophysical Journal, 733*, 20 pp. doi:10.1088/0004-637X/733/1/10
- Scikit-learn. (2018). Scikit-learn: Machine Learning in Python. Retrieved from http://scikit-learn.org/stable/
- Strobel, N. (2010, June 8). The Basic Scheme. Retrieved from http://www.astronomynotes.com/evolutn/s3.htm
- Strobel, N. (2007, June 2). Stage 3: T-Tauri. Retrieved from http://www.astronomynotes.com/evolutn/s4.htm
- Valenzuela, L., Pichara, K. 2017. Unsupervised classification of variable stars. *Monthly Notices of the Royal Astronomical Society, 474*, 3259-3272. doi:10.1093/mnras/stx2913