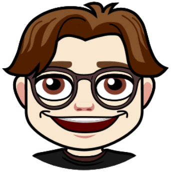
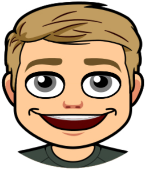
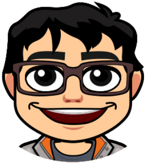
Ensemble methods for PLAsTiCC Astronomical Classification

Youssef Berrada [1], Georgy Guryev [1], Sheng Yao[1]

[1]Massachusetts Institute of Technology
6.867 Machine Learning, Fall 2018

December 11, 2018

Outline

Context and Motivation

The project involves using time series light-curve observations and object specific information to classify different astronomical objects observed by the Large Synoptic Survey Telescope (LSST).

Outline

Data Preparation

Time Series Data

- Object_id: Unique Object ID
- mjd: Modified Julian date from 01/01/2022 to 12/31/2024
- passbands: passband integer
- flux: Simulated brightness
- flux_err: uncertainty on the measurement of the flux
- detected: boolean to detect if measure different from template.

Meta Data

- Object_id: Unique Object ID
- ra,decl,gal_l,gal_b: Position indicators.
- hostgal_specz: spectroscopic hostgal_photoz:photometric
- distmod: distance
- MWEBV: extinction of light
- target: Label

Data Exploration

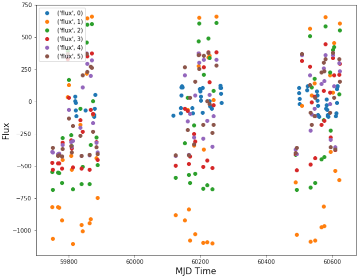


Figure: Flux Measure from object id 615

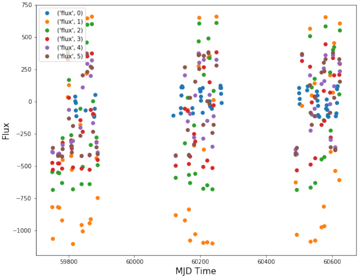


Figure: Flux Measure from object id 130

The measurements of passbands are all observed on a periodic basis, which is related to how frequent an object is observed. The negative measurements are due to logarithm scaling used by the data-provider.

Feature Engineering

- Data Split for class purpose
  - **Training Set**: 7848 labelled objects $\rightarrow$ 70/30 for the class
  - **Testing set**: 3.49 M not labelled objects
- Statistical Aggregation (`tsfresh` library)
  - Aggregate on statistical measures such as $AR(p)$ coefficients
  - Less affected by noise but also compress information.
  - Features are of similar magnitudes.
- Time-series Encoding
  - Recurrent Neural Network
    - Preserve most information in data.
    - Suffer from outliers: adopt logarithm scaling as a remedy.

Outline

Loss Function

To achieve accurate predictions for all classes, we adopted loss function from the competition. The w_j parameter governs the weight assigned to loss for each category. The table below shows the weights we assigned.

$$L = -\frac{\sum_{j=1}^{M} w_j \sum_{i=1}^{N} \frac{1}{N_j} \tau_{i,j} \ln P_{i,j}}{\sum_{j=1}^{M} w_j}$$

class	6	15	16	42	52	53	62	64	65	67	88	90	92	95
weights	1	2	1	1	1	1	1	2	1	1	1	1	1	1

Survey of Machine Learning Algorithms

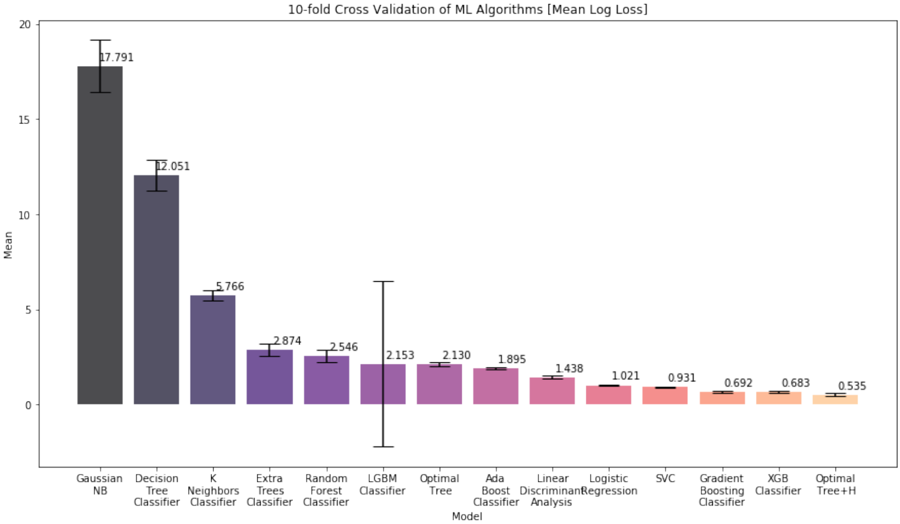


Figure: Survey of Machine Learning performance with 10-fold cross validation

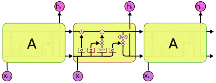
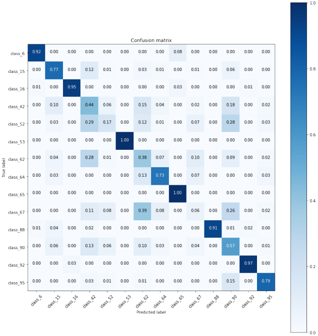


Figure: The Long-Short Term Memory RNN

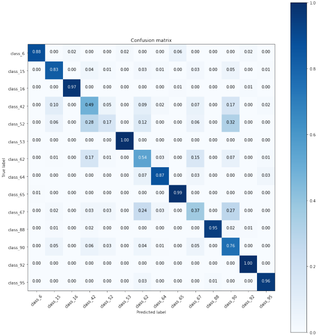


- Unlike traditional NN the RNN model facilitates processing of sequential data
- LSTM (RNN) model capture/memorize to learn long-term dependencies

Metrics	Score
RNN Accuracy	0.6702
RNN Precision	0.6430
RNN Recall	0.6906
RNN F-1 Score	0.6560

Gradient Boosting Tree

- Gradient boosting combines weak "learners" into a single strong learner in an iterative fashion
- At each iteration Gradient boosting generates the best possible estimator from a given functional space that minimizes the residual from the previous step



Metrics	Score
Light GBM Accuracy	0.7633
Light GBM Precision	0.7188
Light GBM Recall	0.768
Light GBM F-1 Score	0.7365

Optimal Classification Tree

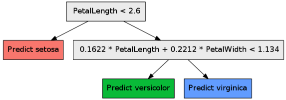
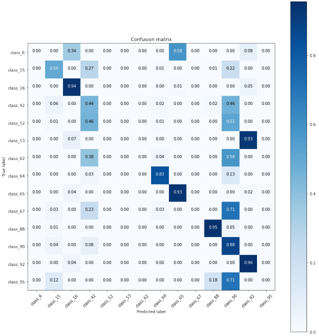


Figure: The Optimal tree representation with hyperplane split

- Standard decision tree models, that rely on top-down approach have a fundamentally greedy nature
- The Optimal Classification tree builds the entire and then performs the mixed-integer optimization to obtain an `Optimal` decision tree

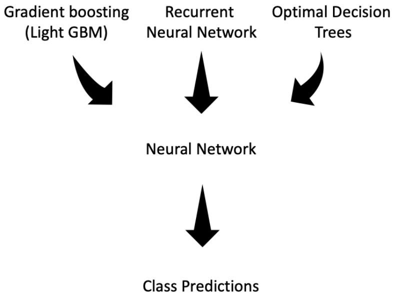


Metrics	Score
OPT_H Accuracy	0.67104
OPT_H Precision	0.390091
OPT_H Recall	0.45980
OPT_H F-1 Score	0.418617

Outline

Ensemble Learning

Model Description



- Ensemble design provides a solution to combine models built on different specification of features.
- To 'learn' the ensemble weights, we input the predictions of each model as features into feed-forward network with softmax activation output.

Ensemble Learning
Results

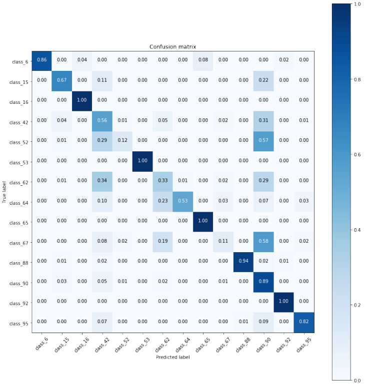


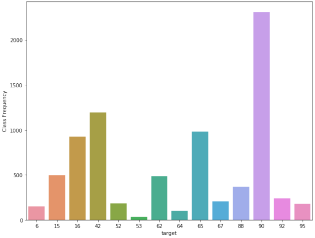
Figure: Confusion Matrix for the Ensemble Method

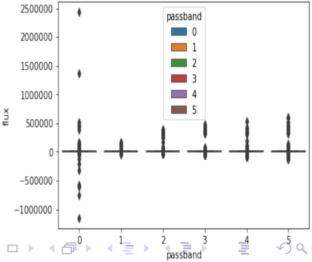
Metrics	Score	RNN	GBM	Opt
Accuracy	0.7814	+16%	+3%	+16%
Precision	0.7828	+21%	+9%	+100%
Recall	0.7029	+1%	-8%	+53%
F-1 Score	0.7281	+11%	-2%	+74%

Outline

Open Challenges and Experimented Design

- Unbalanced Data
  - Pseudo data set with balanced classes
    - Oversampling: risk overfitting
    - Undersampling: risk losing informative samples
  - Custom loss functions
    - Optimal error weighting matrix hard to estimate
- Time-series Encoding with Outliers
  - Statistical Aggregation + Ensemble design
- Class 99: Not in the training data
  - predict all class 99 and imply a weight and then fixed max probability.





Conclusion

- All three models under consideration provide a reasonably accurate classification for various astronomical objects
- The ensemble learning allows to improve an overall prediction accuracy and outperforms the most accurate individual classifier
- Future investigation is required to explore efficient balancing techniques for non-representative class samples