Empower ECL-ML: How to make ECL-ML library make easier to use?

Vivek Nair (vnair2@ncsu.edu)
Mentor: Arjuna Chala
Advisor: Tim Menzies

I. Problem Statement

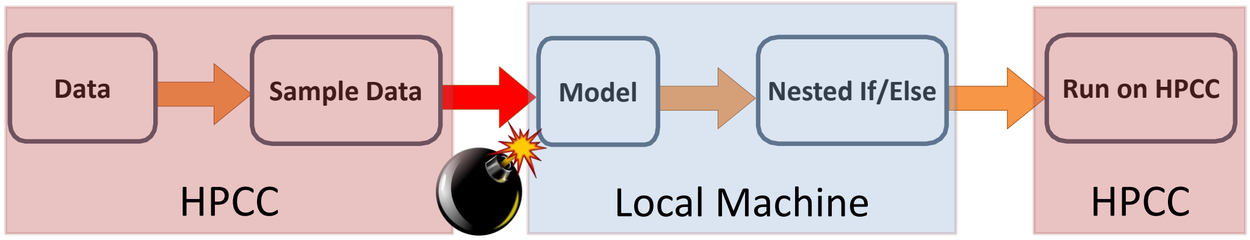
Create (i) Benchmarking Suite and (ii) Visual plugins for 'empowering' ECL-ML to make development faster and more reliable.

II. Why do we need these tools?

a) ECL is easy, but not so easy!

- ECL has a steep learning curve
- ECL-ML requires end-user to understand the details of the library, which needs extensive documentation

b) Current Workflow can be tedious!



c) Benchmark before Deploy!

- ECL-ML algorithms have never been benchmarked previously!
- Validation study is necessary for using ECL-ML in production

III. Contributions

a) Benchmarking Suite

- Compares performance of ECL-ML library with *scikit-learn*
- Supports Software Regression Testing of ML algorithms
- Extensible framework supports test cases for new ML algorithms

Tasks	Algorithms
Classification	Random Forest
	Decision Trees
	Logistic Regression
Regression	Linear Regression
Clustering	KMeans
Time Series	ARIMA

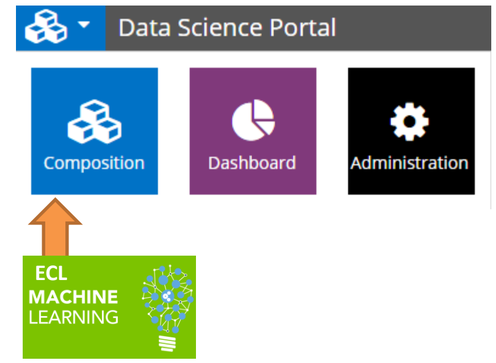
##	dataset_id	dataset_name	ecl_performance	scikit_learn_performance	status
1	1	continious_ecoliDS	0.8722222222222222	0.818382927	PASS
2	2	continious_glassDS	0.7093023255813954	0.703538248	PASS
3	3	continious_ionosphereDS	0.895	0.926703202	PASS
4	4	continious_ringnormDataDS	0.7307146237576905	0.95361631	FAIL
5	5	continious_segmentationDS	0.8883928571428571	0.967285828	PASS
6	6	continious_waveformDS	0.696892294593444	0.723008983	PASS

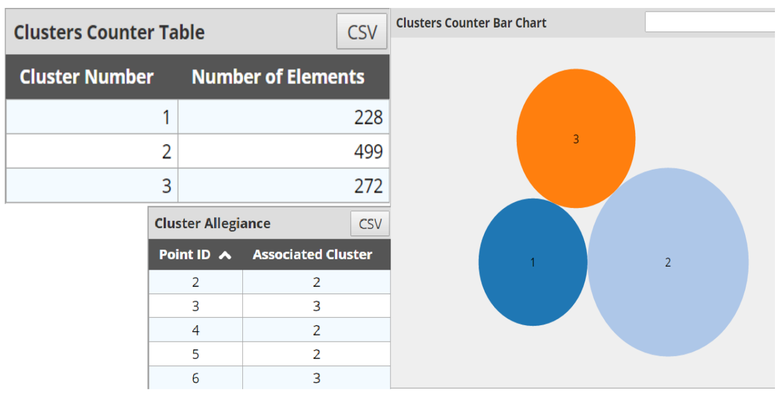
Exceptions | ECL Watch | Graphs | Classification_RandomForestC | Classification_RandomForestD | Classification

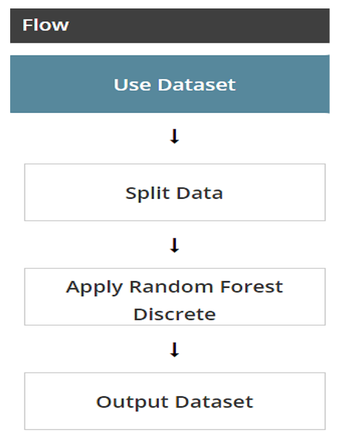
- Testing Suite found regression error in Logistic Regression
- Time required to run the suite is ~4 hours

b) ECL-ML Plugins

- Integrated with Data Science Portal (DSP)
- Access to distributed ML Algorithms
  - Classification
  - Regression
  - Clustering
- Supports
  - Quick data visualization support
  - Workflow based development







IV. Future Work

- Integrate the ML-Regression Suite to Platform Regression Suite
- Add more algorithms and make it available for production