Big Data Programming-2 Project

Adding new functionality to Python Package: Scikit-Learn

HOCHSCHULE
SRH HEIDELBERG
Intelligence in Learning

Presented on 01.04.2020 by:

Shekhar Singh (11011694)
Sanika Medankar (1011861)
Rishabh Garg (11011875)
Rohit Keshav Bewoor (11011831)

Students of Big Data and Business Analytics 2018-20 batch
SRH Hochschule Heidelberg

Content

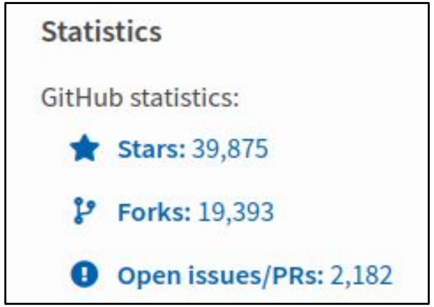
STAATLICH
ANERKANNTE
HOCHSCHULE

- Introduction and Motivation
- Logic
  - Explanation
  - Modules impacted
- Issues and Workaround
- Pull Request
- Environment setup using package uploaded to Test Pypi
- Demo
- Q&A

Introduction

- Popular Machine Learning packages: SCIKIT-LEARN and STATSMODELS
- About Scikit-learn package
  - Source Code (PyPI): https://pypi.org/project/scikit-learn/#files
  - For Machine Learning, built on top of Scipy
  - Website: http://scikit-learn.org
- In Big Data 1 project, we choose to dissect Scikit-learn's ***train_test_split*** function
- Big Data 2 project:
  - Currently Functionality:
    - train_test_split function always splits each input array into a "Train" and "Test" subset.
  - New functionality:
    - Allow splitting each input array into three subsets: "Train", "Test" and ***an additional "Validation".***
- **Issue created on Github of Scikit-learn:** https://github.com/scikit-learn/scikit-learn/issues/13990

Source: https://pypi.org/project/scikit-learn/



Statistics

GitHub statistics:

★ **Stars:** 39,875

Forks: 19,393

Open issues/PRs: 2,182

Motivation

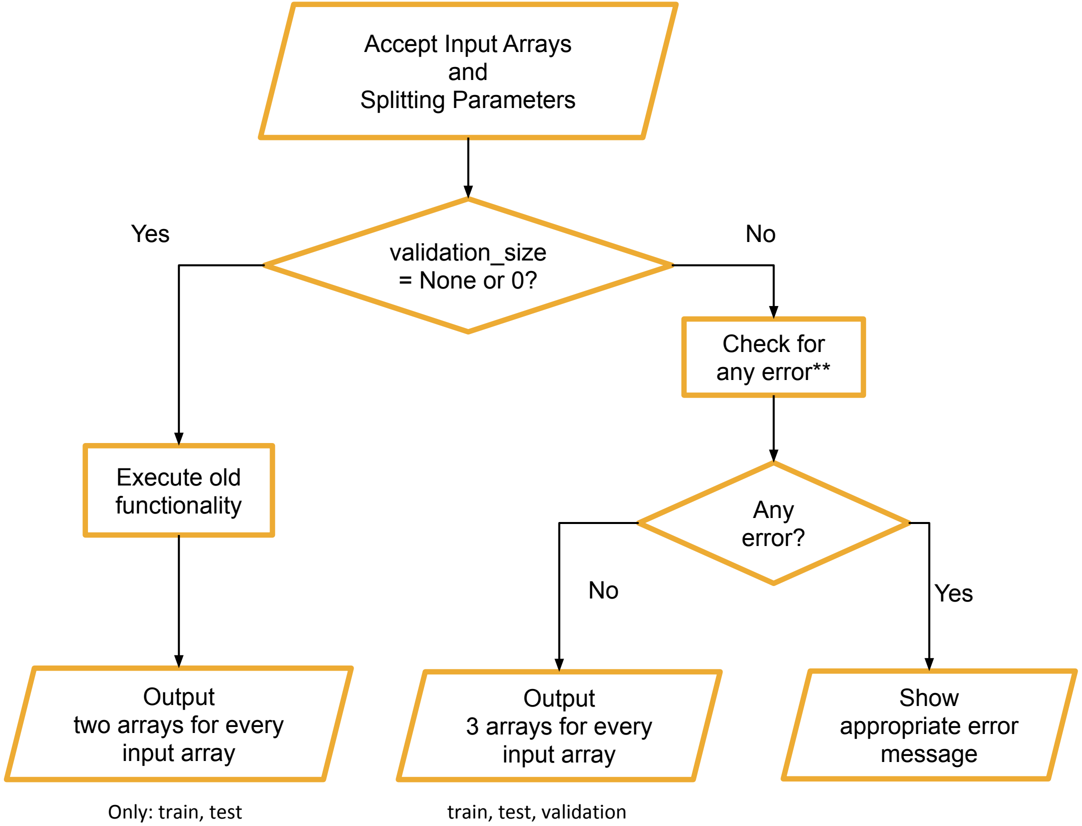
- Why do we need three subsets?
  - Traditionally, ML models are trained on “Train” set. Then, the “Test” set was used for model accuracy.
  - Nowadays, as part of training process a “Validation” set (aka. “Dev” set) is used at end of each epoch to track model accuracy.
  - *Only after training,* Test set (unseen data by model) used for final model accuracy.

- Currently, Scikit-learn allows only two-way splitting. ***User needs to call function twice*** to:
  - Split Full data into Train and Intermediate sets
  - Split Intermediate set into Validation and Test sets
  - User needs to track the number of samples (or percentage of split) manually and this is prone to errors

- New approach:
  - ***Added New function to split three ways = train_test_val_split()***
  - No chance of manual calculation errors
  - Maintained Backward compatibility to allow two-way split also

Logic - Flowchart

STAATLICH
ANERKANNTE
HOCHSCHULE



**Error Checks:

1. validation_size + train_size + test_size <= n_samples (or <= 1.0 if proportion)
2. validation_size must be an Integer or Float
3. If integer; ensure
   0 <= validation_size < n_samples
4. If float; ensure
   0.0 <= validation_size < 1.0

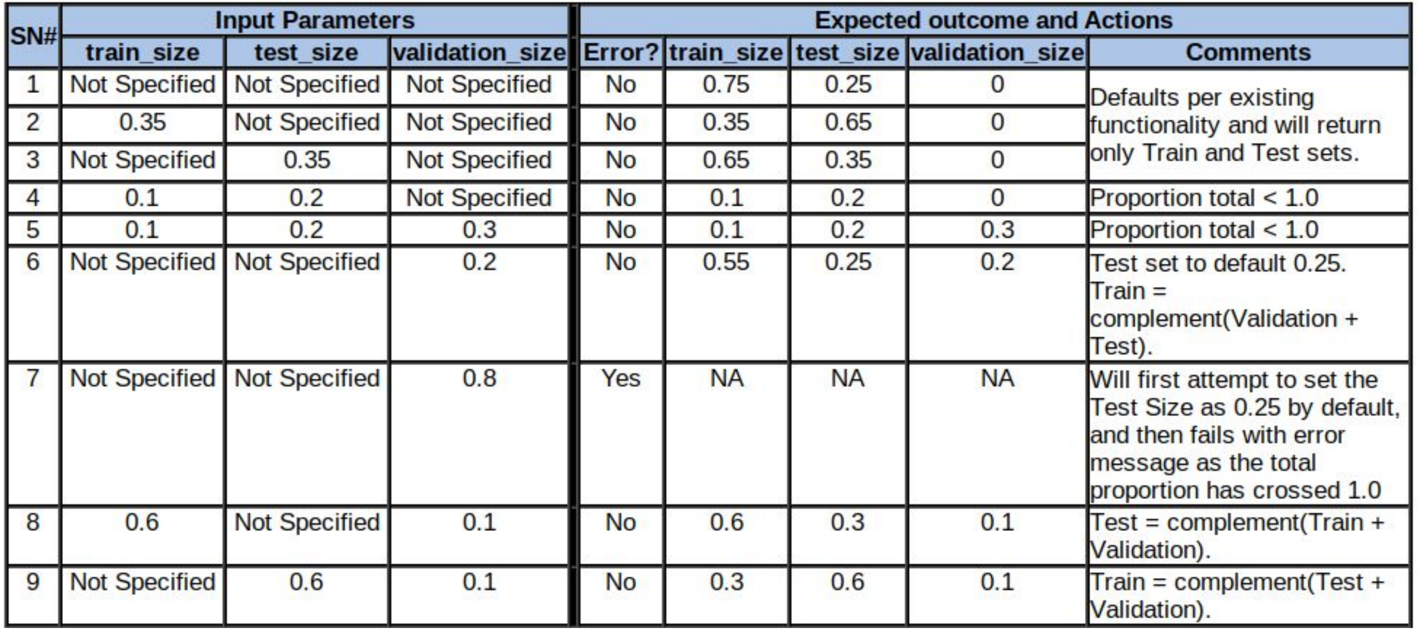
Default Value of validation_size= 0.0

- Allows existing functionality for two-way split

Logic - Truth Table

Optional parameters during call to NEW function train_test_val_split():
train_size, test_size, validation_size, shuffle, stratify

Acceptance for processing and the actions taken in each scenario.



SN#	Input Parameters			Expected outcome and Actions				
	train_size	test_size	validation_size	Error?	train_size	test_size	validation_size	Comments
1	Not Specified	Not Specified	Not Specified	No	0.75	0.25	0	Defaults per existing functionality and will return only Train and Test sets.
2	0.35	Not Specified	Not Specified	No	0.35	0.65	0	
3	Not Specified	0.35	Not Specified	No	0.65	0.35	0	
4	0.1	0.2	Not Specified	No	0.1	0.2	0	Proportion total < 1.0
5	0.1	0.2	0.3	No	0.1	0.2	0.3	Proportion total < 1.0
6	Not Specified	Not Specified	0.2	No	0.55	0.25	0.2	Test set to default 0.25. Train = complement(Validation + Test).
7	Not Specified	Not Specified	0.8	Yes	NA	NA	NA	Will first attempt to set the Test Size as 0.25 by default, and then fails with error message as the total proportion has crossed 1.0
8	0.6	Not Specified	0.1	No	0.6	0.3	0.1	Test = complement(Train + Validation).
9	Not Specified	0.6	0.1	No	0.3	0.6	0.1	Train = complement(Test + Validation).

Default values: Test = 0.25, Validation = 0.0, Train = complement(Test + Validation).

=> If no values specified: then original functionality where train_size = 0.75 and test_size = 0.25 and no validation set is created.

Same results can be achieved with or without stratify

Impacted modules

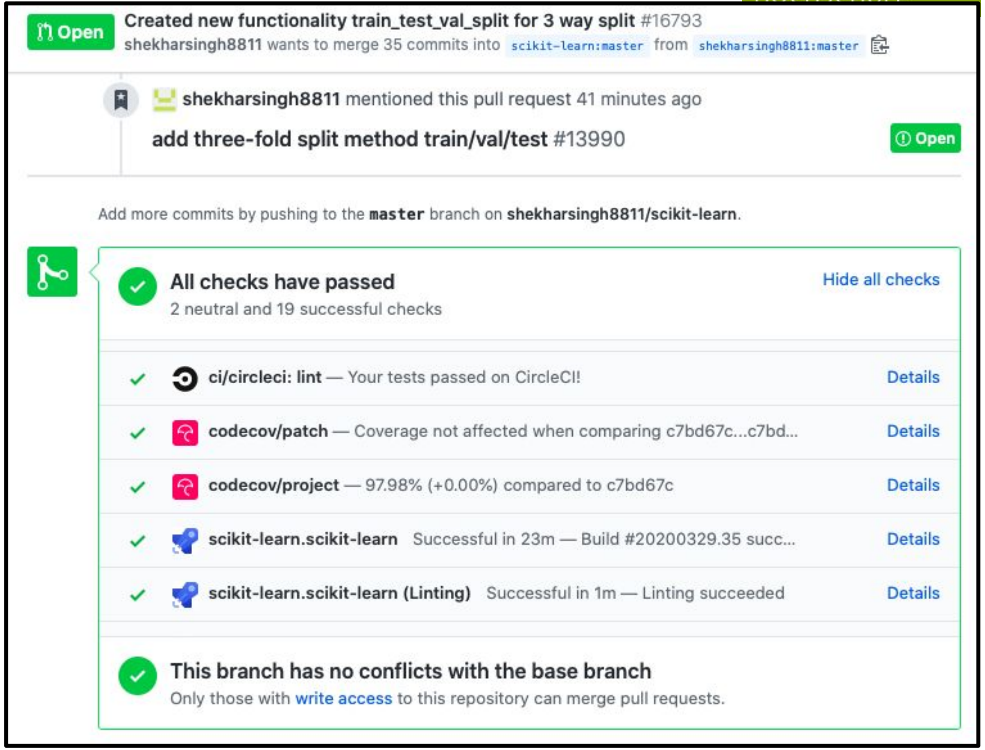
- The following modules are changed for this project:
    - sklearn / model_selection / _split.py
      - Logic changes here
    - sklearn / model_selection / __init__.py
      - Exposed new function train_test_val_split()
    - sklearn / model_selection / tests / test_split.py
      - Changes to test new functionality

Issues Faced and Workaround

- Test Pypi does not allow a simple linux wheel to be uploaded. Possibly existing issue:
  - Binary wheels for linux are not supported #120 (https://github.com/pypa/pypi-legacy/issues/120)
  - Recommend manylinux wheels in the Error 400 response for "linux" package uploads #6545 (https://github.com/pypa/warehouse/issues/6545)
- Supposedly, a many-linux wheel version can be uploaded. But we are unable to create it.
- **Therefore, only uploaded the source and not a wheel (i.e. output of sdist and not bdist_wheel).**
- **E**xtensive testing for all scenarios done using our own test script:
  - Script location: https://github.com/rbewoor/BigData2_Project_Bkup_Two_Functions/blob/master/Dedicated_Test_Cases/Test_Cases_SciKit-Learn_fromTestPypi_Two_Functions.ipynb

Pull Request

- First tried to make changes in existing function train_test_split()
    - Faced many issues with creating a pull request (indentation, circleCI checks, azure pipeline, etc). Fixed as much as possible but there were still unresolved errors
- Changed approach and created a new function train_test_val_split()



Open **Created new functionality train_test_val_split for 3 way split** #16793

shekharsingh8811 wants to merge 35 commits into `scikit-learn:master` from `shekharsingh8811:master`

shekharsingh8811 mentioned this pull request 41 minutes ago

add three-fold split method train/val/test #13990 Open

Add more commits by pushing to the **master** branch on **shekharsingh8811/scikit-learn**.

All checks have passed Hide all checks

2 neutral and 19 successful checks

✓	**ci/circleci: lint** — Your tests passed on CircleCI!	Details
✓	**codecov/patch** — Coverage not affected when comparing c7bd67c...c7bd...	Details
✓	**codecov/project** — 97.98% (+0.00%) compared to c7bd67c	Details
✓	**scikit-learn.scikit-learn** Successful in 23m — Build #20200329.35 succ...	Details
✓	**scikit-learn.scikit-learn (Linting)** Successful in 1m — Linting succeeded	Details

This branch has no conflicts with the base branch

Only those with write access to this repository can merge pull requests.

Pull request link: https://github.com/scikit-learn/scikit-learn/pull/16793

Environment Setup from Test Pypi

STAATLICH
ANERKANNTE
HOCHSCHULE

- Test Pypi details:
  - Project: Scikit-learn-VAL-TestPypi, Latest Version: 0.0.3
  - Link: https://test.pypi.org/project/scikit-learn-VAL-TestPypi/

- Automatic install of dependencies was not working, so manual install required. ***Resolved problem by using an extra index url in command***:
  - pip3 install --index-url https://test.pypi.org/simple/ --no-cache-dir **--extra-index-url https://pypi.org/simple/** scikit-learn-VAL-TestPypi

- Environment from Test Pypi - build with source option only

  - Automatic install of dependencies not working; so ***manual install required***:
    - Cython>=0.28.5, setuptools, wheel, numpy>=1.14.0, scipy>=1.1.0, joblib>=2.0.0, threadpoolctl>=2.0.0

  - Now install from Test Pypi:
    - pip3 install -i https://test.pypi.org/simple/ scikit-learn-VAL-TestPypi

  - Additional packages for our test script: https://github.com/rbewoor/BigData2_Project_Bkup_Two_Functions/blob/master/Dedicated_Test_Cases/Test_Cases_SciKit-Learn_fromTestPypi_Two_Functions.ipynb
    - Jupyter, pandas

Demo

- Various test cases for combinations of splitting variable values:
  - Expecting only Train+Test vs. Train+Test+Validation
  - With and without use Shuffle
  - With and without Stratification

Q&A

- Thank you. Open to questions!