Annual progress report: 7 November 2024 - 6 November 2025

Contents

Aims of your award

Items

Item 1

Title of the aim

Extend the capabilities of scikit-learn's visualization displays

Description of the aim

Scikit-learn provides visualization tools for analyzing predictive modeling results. However, these tools were limited to a single trained predictor. Best practices in experimentation recommend using cross-validation to assess the variability and uncertainty in predictive modeling. We therefore aim to extend the scikit-learn API to be compatible with scikit-learn's existing cross-validation framework.

What is the progress against this aim?

In Progress

Comment on progress

This work is ongoing, and we achieved a major milestone by merging the pull request that adds this capability to ROC curves (https://github.com/scikit-learn/scikit-learn/pull/30399). This establishes the standard we will follow for upcoming visualization methods. This work was highlighted as a release feature in the latest scikit-learn release: https://scikit-learn.org/dev/auto_examples/release_highlights/plot_release_highlights_1_7_0.html#plotting-roc-curves-from-cross-validation-results.
We are developing additional visualization methods, including precision-recall curves (https://github.com/scikit-learn/scikit-learn/pull/30508) and DET curves (https://github.com/scikit-learn/scikit-learn/pull/32235). Additionally, we are implementing two new visualization displays that will benefit from this framework (https://github.com/scikit-learn/scikit-learn/pull/32732 and https://github.com/scikit-learn/scikit-learn/pull/28972).

Item 2

Title of the aim

Improve the visual representation of models in interactive environments

Description of the aim

Scikit-learn provides basic visual representations of machine learning pipelines. However, these diagrams have limitations and should be extended. Specifically, we want to provide visualizations that display model parameters and track feature name propagation through the pipeline.

What is the progress against this aim?

In Progress

Comment on progress

This work is ongoing, but we have already achieved important milestones. The parameter list, documentation, and links to the web documentation are now available during interactive development. These features were highlighted as a release feature in the current scikit-learn release: https://scikit-learn.org/dev/auto_examples/release_highlights/plot_release_highlights_1_7_0.html#improved-estimator-s-html-representation. We continue developing in three key areas: enabling users to check the number of features at different stages of a pipeline (https://github.com/scikit-learn/scikit-learn/pull/31937), displaying information about fitted attributes in a pipeline (https://github.com/scikit-learn/scikit-learn/pull/31442), and showing available methods for a pipeline (https://github.com/scikit-learn/scikit-learn/pull/31698). Additionally, we have pull requests for maintenance and minor improvements related to these features.

Item 3

Title of the aim

Implementation of a callback API

Description of the aim

We recently prototyped the foundations of this framework and
we intend to (i) finalize this framework, (ii) further develop different types of callbacks, and (iii) extend all scikit-learn
estimators so that they use these callbacks. Such a framework should, of course, be made available to third-party
libraries compatible with scikit-learn.

What is the progress against this aim?

In Progress

Comment on progress

We have recently started this project and recruited an engineer to lead the work.
We are close to finalizing the framework and will soon begin implementing the first callback mechanism to monitor the training process of predictive models using a progress bar.

Item 4

Title of the aim

Improve model explainability techniques

Description of the aim

We aim to establish foundational work in model explainability. Scikit-learn provides several explainability techniques; however, we want to develop a unified API to simplify how users interact with these tools. We therefore intend to pursue the work outlined in the following scikit-learn enhancement proposal.

What is the progress against this aim?

In Progress

Comment on progress

We are working on this topic, but we have slightly modified our goal. Rather than developing a new API, we are laying foundational research to understand the theoretical aspects of statistics and improve the library. This work has allowed us to identify limitations in one method currently implemented in scikit-learn and explore alternative approaches. We are currently discussing how to make these findings available to our users (https://github.com/scikit-learn/scikit-learn/pull/31279).

Item 5

Title of the aim

Inclusive community of developers

Description of the aim

To deliver this technical work while improving the repository's inclusiveness, we plan to scale our scikit-learn open-source mentoring program.

What is the progress against this aim?

In Progress

Comment on progress

To date and for the coming year, we have onboarded three new developers through a grant to contribute to these initiatives.
During this process, we also promoted two developers to core contributors in scikit-learn and added one of our mentees to the maintainer team.

Human participants

Does your award involve human participants or human biological material?

Yes - only involves human participants

If your award involves any human participants, does it include a clinical trial?

No

Lived Experience

Have you engaged with people with lived experience or from affected communities as part of your award?

No

Location

Are there any changes to the location(s) where you will be conducting research/grant activities or redirecting funds?

No

Team

Staff directly funded by the award

Yes

List all roles whose salaries or stipends are currently and directly funded through this award.

Role	Number of staff	How many of these roles are clinically active?
Other	4	0

Are any of the staff listed here enrolled for a PhD with fees paid for by this award?

No

Indicate the next destination of any staff/personnel who were funded through this award and who have now left the award

Next destination sector	Number
Other	0

Are you experiencing any challenges with staff recruitment?

No

Lived experience

No

Collaborations

No

Publications

Is your award directly contributing to the generation of any publications?

No

Health Interventions

Is your award directly contributing to the generation of any Health Interventions?

No

Methodological advances

Is your award directly contributing to the generation of methodological advances?

No

Physical tools

Is your award directly contributing to the generation of Physical tools?

No

Computational tools or software

Is your award directly contributing to the generation of any Computational tools or Software?

Yes

Items

Item 1

Title of the Computational tool or software

scikit-learn

Provide a short description of this Computational tool or software

scikit-learn is a free and open-source machine learning library for the Python programming language (https://en.wikipedia.org/wiki/Scikit-learn).

It is widely used by the research and industry in many fields.

Who are the intended users or audience of the Computational tool or software?

The audience of scikit-learn is wide since it offers a library implementing applied mathematical algorithm used for predictive modeling. Therefore, it can be used by anyone interested in creating statistical predictive model and not specialized to a specific domain.

This library is the most used framework in machine learning.

Provide a link to the relevant repository about this output as per the guidance:

Here is the GitHub repository where all the work of the grant is contributed openly: https://github.com/scikit-learn/scikit-learn

Users will always look at the documentation website when it comes to use the library: https://scikit-learn.org

Provide link(s) to any publication(s) describing this output where applicable. Where possible include a Digital Object Identifier (DOI) linked to the publication, if not then provide a URL.

Databases or datasets

Is your award directly contributing to the generation of any Databases or datasets?

No

Influence on policy and practice

Is your award directly contributing to the generation of any influence on policy and practice?

No

Other research outputs

Is your award directly contributing to the generation of any artistic or creative products?

No

Is your award directly contributing to the generation of any educational resources?

No