The Journal of Open Source Software

Scikit-Longitudinal: A Machine Learning Library for Longitudinal Classification in Python

Simon Provost [1] **and Alex A. Freitas** [1]

1 School of Computing, University of Kent, Canterbury, United Kingdom

DOI: 10.21105/joss.08481

Software

- Review
- Repository
- Archive

Editor: Evan Spotte-Smith

Reviewers:

- @TahiriNadia
- @blengerich

Submitted: 09 May 2025
Published: 01 August 2025

Introduction

Longitudinal data, characterised by repeated measurements of variables over time, presents unique challenges and opportunities in machine learning. This paper introduces Scikit-Longitudinal, a Python library designed to address these challenges by providing a comprehensive set of tools for longitudinal data classification. Built to integrate with the popular Scikit-learn library, Scikit-Longitudinal offers a robust solution for researchers and practitioners working with longitudinal datasets.

Summary

Scikit-Longitudinal, also abbreviated Sklong, is an open-source Python library that enhances machine learning for longitudinal data classification and integrates with the Scikit-learn environment (Pedregosa et al., 2011).

Longitudinal data, which consists of repeated measurements of variables across time points (referred to as waves (Ribeiro & Freitas, 2019)), is extensively utilised in fields such as medicine and social sciences. Unlike standard tabular datasets, longitudinal data contains temporal relationships that necessitate specialised processing (Kelloway & Francis, 2012).

Sklong addresses this with a novel library that includes (Provost & Freitas, 2024):

- **Data Preparation**: Utilities such as LongitudinalDataset for loading and structuring data, defining temporal feature groups, and other techniques.
- **Data Transformation**: Methods to treat the temporal aspect of tabular data, by either (1) flattening the data into a static representation (i.e., ignoring time indices) for standard machine learning to be performed (e.g., MarWavTimeMinus, or SepWav), or (2) keeping the temporal structure (e.g., MerWavTimePlus), yet saving it for later use in longitudinal-data-aware preprocessing or estimators steps (Ribeiro & Freitas, 2019).
- **Preprocessing**: Longitudinal-data-aware feature selection primitives, such as CFS-Per-Group (Pomsuwan & Freitas, 2017), utilising the temporal information in the data to proceed with feature selection techniques (see feature selection review in (Theng & Bhoyar, 2024)).
- **Estimators**: Longitudinal-data-aware classifiers (Kotsiantis et al., 2007; Ribeiro & Freitas, 2019), such as LexicoRandomForestClassifier (Ribeiro & Freitas, 2024), LexicoGradientBoostingClassifier, and NestedTreesClassifier (Ovchinnik et al., 2022), which leverage the temporal structure of the data to ideally enhance classification performance.

In total, the library implements 1 data preparation method, 4 data transformation methods, 1 preprocessing method, and 6 estimators, 2 of which have been published as standalone methods in the literature (the above named LexicoRandomForestClassifier and NestedTreesClassifier methods).

Sklong emphasises highly-typed, Pythonic code, with substantial test coverage (over 88%) and comprehensive documentation (over 72%).

Finally, Sklong is available on PyPI. Feel free to explore the official documentation for various installation methods.

Longitudinal Classification

Longitudinal classification is a variant of the standard classification task where the data includes features taking values at multiple time points / "waves'' (Ribeiro & Freitas, 2019), e.g., cholesterol values measured at different waves. Longitudinal classification is particularly relevant in biomedical applications, since biomedical data about patients is often collected across long time periods.

The challenge is to learn a model that predicts the class label (Y) for an instance while accounting for the evolution of features' values over time, i.e., to learn a predictive model (classifier function) of the form:

$$Y \leftarrow f(X_{1,1}, X_{1,2}, \ldots, X_{1,T}, \ldots, X_{K,1}, X_{K,2}, \ldots, X_{K,T})$$

where $X_{i,j}$, for $i = 1, \ldots, K$ and $j = 1, \ldots, T$, is the value of the i-th feature at the j-th wave (time point), K is the number of features, and T is the number of waves. The classifier function $f(\cdot)$ can either operate on a transformed version of the data where temporal information is "flattened", allowing the application of standard machine learning algorithms, or handle the temporal dependencies between a feature's values across time and between different features directly. Note that, in the type of longitudinal classification task for which Sklong was designed, the features are longitudinal, but the class variable is not; i.e., the goal is to predict the class label of an instance at a single time point (usually the last wave).

There are two broad approaches for coping with longitudinal data (Ribeiro & Freitas, 2019): (1) **Data Transformation**: this approach involves preprocessing methods that convert longitudinal data into a standard, "flattened" non-longitudinal format, enabling the use of any standard, non-longitudinal classification algorithm on the data but potentially losing relevant information regarding how a feature's values change over time. (2) **Algorithm Adaptation**: this approach entails modifying classification algorithms to directly handle temporal dependencies inherent in longitudinal datasets, preserving the temporal dynamics of the data but requiring more complex tooling.

Statement of Need

To the best of our knowledge, no package in the Scikit-learn ecosystem provides an easy solution for longitudinal classification. Standard Python libraries, such as Scikit-learn itself, lack support for longitudinal data, leading to inefficient and inaccurate analysis. R includes statistical packages for longitudinal data (e.g., nlme (Pinheiro & Bates, 2000), GPBoost (Sigrist, 2022)). However, they often are not suitable for machine learning workflows often created in Python. On the other hand, systems like Auto-Prognosis (Alaa & Schaar, 2018) concentrate on longitudinal classification but do not have Scikit-learn's ease of use. Auto-Prognosis encompasses more than just longitudinal machine learning, making it difficult to identify and investigate specific problems. Furthermore, it focuses on algorithm adaptation for prognosis rather than providing both data transformation and algorithm adaptation paths like Sklong, which limits user flexibility.

Given the lack of Python libraries, lack of integration with the popular Scikit-learn API, and the absence of out-of-the-box solutions for longitudinal classification, there is a clear need for a library that provides tools for longitudinal data preparation, transformation, preprocessing, and estimation (model learning).

Limitations and Future Work

At present, Sklong primarily focuses on the classification task and does not yet include support for regression or neural networks. Future development could expand the library in these directions.

Acknowledgements

We thank the authors and contributors of Scikit-learn for their pioneering work in machine learning. We thank the NeuroData team for their contributions to Scikit-Tree (Li et al., 2024), which enables modification of Scikit-learn's Cython trees for optimising performance. We are grateful to the researchers who contributed to the design of many primitives within Sklong, including Dr. Tossapol Pomsuwan (Pomsuwan, 2018; Pomsuwan & Freitas, 2017), Dr. Sergey Ovchinnik & Dr. Fernando Otero (Ovchinnik et al., 2022), and Dr. Caio Ribeiro (Ribeiro, 2022; Ribeiro & Freitas, 2019, 2024).

References

Alaa, A. M., & Schaar, M. van der. (2018). *AutoPrognosis: Automated Clinical Prognostic Modeling via Bayesian Optimization with Structured Kernel Learning*. https://doi.org/10.48550/ARXIV.1802.07207

Kelloway, E. K., & Francis, L. (2012). Longitudinal research and data analysis. In *Research methods in occupational health psychology* (pp. 374–394). Routledge.

Kotsiantis, S. B., Zaharakis, I., Pintelas, P., & others. (2007). Supervised machine learning: A review of classification techniques. *Emerging Artificial Intelligence Applications in Computer Engineering*, *160*(1), 3–24.

Li, A., Panda, S., Xu, H., & Ogihara, I. (2024). *Treeple: Modern decision-trees compatible with scikit-learn in Python.* (Version v0.10.0). Zenodo. https://doi.org/10.5281/zenodo.14509519

Ovchinnik, S., Otero, F., & Freitas, A. A. (2022). Nested trees for longitudinal classification. *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, 441–444. https://doi.org/10.1145/3477314.3507240

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & others. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Pinheiro, J., & Bates, D. (2000). *Mixed-effects models in s and s-PLUS*. Springer science & business media. https://doi.org/10.1007/0-387-22747-4_8

Pomsuwan, T. (2018). *Feature selection for the classification of longitudinal human ageing data* [Master by Research thesis]. School of Computing, University of Kent, UK.

Pomsuwan, T., & Freitas, A. A. (2017). Feature selection for the classification of longitudinal human ageing data. *Proceedings of the 2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, 739–746. https://doi.org/10.1109/icdmw.2017.102

Provost, S., & Freitas, A. A. (2024). Auto-sklong: A new AutoML system for longitudinal classification. *Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 3645–3650. https://doi.org/10.1109/BIBM62325.2024.10821737

Ribeiro, C. (2022). *New longitudinal classification approaches and applications to age-related disease data* [Ph.D. Thesis]. School of Computing, University of Kent, UK.

Ribeiro, C., & Freitas, A. A. (2019). A mini-survey of supervised machine learning approaches for coping with ageing-related longitudinal datasets. *Proceedings of the 3rd Workshop on AI for Aging, Rehabilitation and Independent Assisted Living (ARIAL), Held as Part of IJCAI-2019, 5 Pages*.

Ribeiro, C., & Freitas, A. A. (2024). A lexicographic optimisation approach to promote more recent features on longitudinal decision-tree-based classifiers: Applications to the english longitudinal study of ageing. *Artificial Intelligence Review*, *57*, Article 84, 29 pages. https://doi.org/10.1007/s10462-024-10718-1

Sigrist, F. (2022). Gaussian process boosting. *Journal of Machine Learning Research*, *23*(232), 1–46.

Theng, D., & Bhoyar, K. K. (2024). Feature selection techniques for machine learning: A survey of more than two decades of research. *Knowledge and Information Systems*, *66*(3), 1575–1637. https://doi.org/10.1007/s10115-023-02010-5