Lehre & Studium

ANALYSIS OF THE FACTORS AFFECTING THE STABILITY OF LIME FOR UNIVARIATE TIME-SERIES CLASSIFICATION

GABRIEL WITTE
(BACHELORSTUDIUM APPLIED ARTIFICIAL INTELLIGENCE)

Betreuer: Priv.-Doz. Dr. Sven-Joachim Kimmerle, Prof. Dr. Marcel Tilly

Kurzfassung

KI-Systeme treffen zunehmend wichtige Entscheidungen, erklären ihre Schlüsse aber oft nicht. Ein verbreitetes Erklärwerkzeug namens LIME liefert solche Erklärungen, doch diese sind nicht immer verlässlich: Schon kleinste Änderungen an den Eingabedaten können zu völlig anderen Erklärungen führen. Diese Arbeit untersucht in über 350.000 Experimenten, welche Faktoren diese Verlässlichkeit (Stabilität) beeinflussen, und zeigt praktische Wege auf, wie sich stabilere Erklärungen für Zeitreihenmodelle erzeugen lassen.

The problem

Artificial intelligence now helps make decisions that matter, such as approving a loan, flagging a possible heart condition, or predicting financial markets. The catch is that the most accurate AI systems are black boxes: they give you an answer, but not a reason. When the stakes are high, and increasingly under laws such as the European Union's AI Act and the GDPR (General Data Protection Regulation), that is not good enough. We need these systems to explain themselves. This is the goal of a field called Explainable AI.

How the explanations are made

A popular tool for this is LIME (Local Interpretable Model-agnostic Explanations). It operates much like a translator for complex algorithms. LIME takes a single AI decision and asks the "black box" thousands of small "what if" questions—what if this person earned a little less? What if they were a few years older? By analyzing the answers, it calculates which inputs mattered most, resulting in a simple, readable summary of the AI's reasoning.

In Figure 1, LIME is applied to a flower classification model. For this specific data point, the model classified the flower as the Versicolor species. To understand why, LIME perturbed the data by asking multiple hypothetical questions, such as "what if the sepal width was larger than 0.59cm?" or "what if the petal width was smaller than 0.2cm?"

By aggregating all of these responses, LIME generates a quick, intuitive summary of the decision:

- Supporting Evidence: The petal length, petal width, and sepal length all strongly pushed the model toward classifying it as Versicolor.
- Contradictory Evidence: A valuable feature of LIME is that it also highlights inputs that argued against the final decision. In this case, LIME shows that the flower's sepal width (less than 0.59cm) is unusually short for a Versicolor, which slightly pushed the model away from that classification.

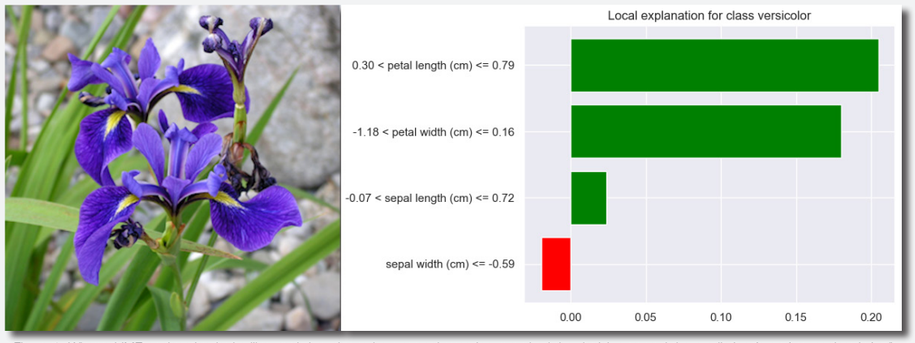


Figure 1. What a LIME explanation looks like: each bar shows how strongly one input pushed the decision toward the prediction (green) or against it (red). Author's own illustration.

The catch: how it was tested

During my research I ran into a troubling issue. If you ask LIME to explain the same decision twice, changing the input by only a hair, it can hand back a completely different reason. An explanation that keeps changing its story cannot really be trusted. Researchers call this reliability stability. So my thesis asked one question: what makes these explanations wobble, and how do we make them hold still?

LIME builds its explanations in four distinct steps: Segmentation, Perturbation, Weighting, and Model Fitting. I put state-of-the-art techniques for each of these phases to the test across a broad spectrum of AI models—ranging from basic decision trees to complex neural networks. I applied them to five vastly different real-world time-series datasets: human heartbeats, distant starlight, spoken words, household electricity use, and food scans. In total, this comprehensive analysis required 352,800 experimental measurements.

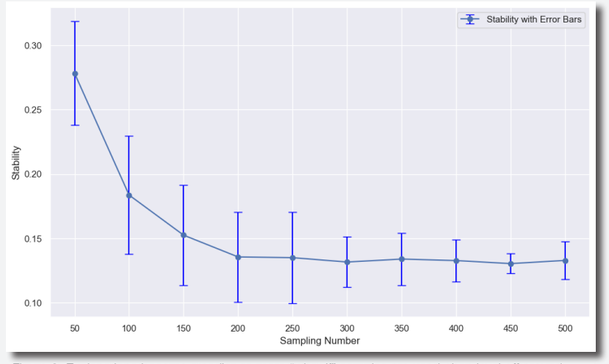
What I found

Two steps turned out to be decisive: how the data is divided into pieces, and how the “what if” questions are generated. Getting these right makes explanations far steadier. A third factor, how many “what if” questions you ask, has a clear sweet spot: reliability improves quickly up to about 150 questions, then flattens out, so asking more simply wastes computing power (Figure 2). Just as usefully, two factors that many experts assumed were important turned out to make almost no difference at all.

Figure 2. Explanations become steadier as more “what if” samples are used, then level off around 150, the practical sweet spot (lower values mean steadier, more reliable explanations). Author's own work.

Why it matters

Taken together, these results form a practical recipe for making AI explanations consistent enough to rely on. It is a small but real step toward being able to use AI safely in the high-stakes areas such as healthcare and finance, where trust is not optional.



References

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. Proceedings of KDD 2016.

Schlegel, U., Lam, D. V., Keim, D. A., & Seebacher, D. (2021). TS-MULE: Local interpretable model-agnostic explanations for time-series forecast models. arXiv:2109.08438.

Sivill, T., & Flach, P. (2022). LIMESegment: Meaningful, realistic time-series explanations. Proceedings of AISTATS 2022.

Zhou, Z., Hooker, G., & Wang, F. (2021). S-LIME: Stabilized-LIME for model explanation. Proceedings of KDD 2021. arXiv:2106.07875.

Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. arXiv:1806.08049.

European Commission (2024). Regulation laying down harmonised rules on artificial intelligence (EU AI Act).