# **Tree**-based predictions for **w**eather and **e**nergy - *treewe*

Introducing an open-source Python library

# I’m Sebastian Haglund

- Energy engineer by training (**KTH**)
- Worked at **Fortum** with energy trading
- Co-founder at **rebase.energy**
- Part of **Stockholm.AI** community
- Love working with **Python** ❤️

Connect with me on

LinkedIn

# 6 Whys

- Why Energy?
- Why Weather?
- Why Trees?
- Why Python?
- Why Open-Source?
- Why *treewe*?

# Why energy predictions?

# Common prediction problems in the energy sector

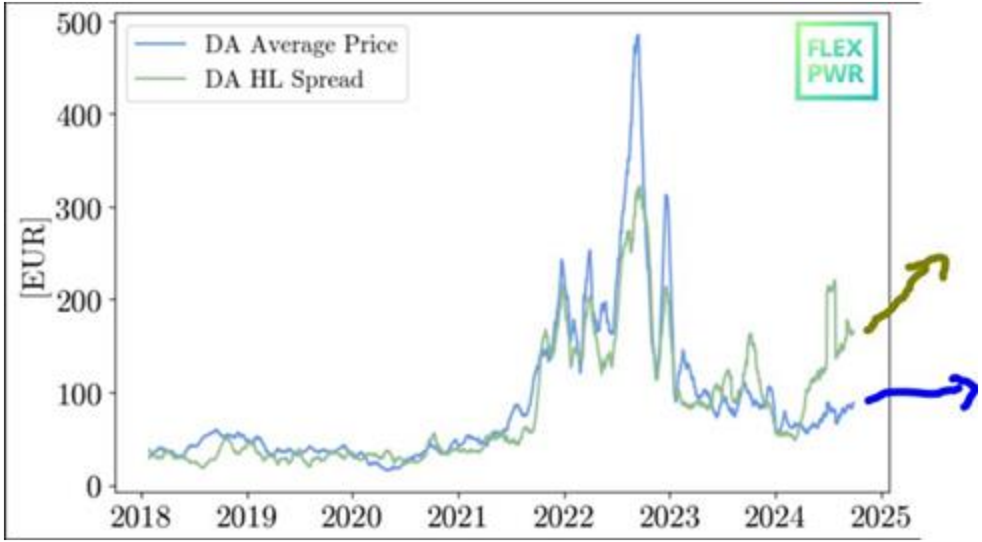
- Power forecasting
- Icing forecasting
- Wind speed estimation
- Anomaly detection

- Power forecasting
- Anomaly detection

- Demand forecasting
- Demand estimation
- Peak prediction
- Anomaly detection

# Why weather predictions?

# Weather intermittency will drive volatility in energy markets




Source: [flex-power.energy](flex-power.energy)

Decoupling of DA average price and DA HL Spread is largely driven by **increased weather dependency in the energy market**

# Why tree-based models for energy predictions?

# Why does deep learning struggle with tabular datasets?

**Why do tree-based models still outperform deep learning on tabular data?**


**Léo Grinsztajn**
Soda, Inria Saclay
leo.grinsztajn@inria.fr

**Edo**
ISIR, CNRS


**Abst**

While deep learning has enabled tremend
its superiority on tabular data is not clear.
standard and novel deep learning method
XGBoost and Random Forests, across a lar
ter combinations. We define a standard set
clear characteristics of tabular data and a l
for both fitting models and finding good h
based models remain state-of-the-art on m

TABULAR DATA: DEEP LEARNING IS NOT ALL YOU NEED


**Ravid Shwartz-Ziv**
ravid.ziv@intel.com
IT AI Group, Intel

**Amitai Armon**
amitai.armon@intel.com
IT AI Group, Intel


November 24, 2021

ABSTRACT

A key element in solving real-life data science problems is selecting the types of models
Tree ensemble models (such as XGBoost) are usually recommended for classification and regı
problems with tabular data. However, several deep learning models for tabular data have recentl
proposed, claiming to outperform XGBoost for some use cases. This paper explores whethe

Deep Neural Networks and Tabular Data: A Survey


Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk and Gjergji Kasneci


*Abstract*—Heterogeneous tabular data are the most commonly used form of data and are essential for numerous critical and computationally demanding applications. On homogeneous data sets, deep neural networks have repeatedly shown excellent performance and have therefore been widely adopted. However, their adaptation to tabular data for inference or data generation tasks remains highly challenging. To facilitate further progress in the field, this work provides an overview of state-of-the-art deep learning methods for tabular data. We categorize these methods into three groups: data transformations, specialized architectures, and regularization models. For each of these groups, our work offers a comprehensive overview of the main approaches. Moreover, we discuss deep learning approaches for generating tabular data, and we also provide an overview over strategies for explaining deep models on tabular data. Thus, our first contribution is to address the main research streams and existing methodologies in the mentioned areas, while highlighting relevant challenges and open research questions. Our second contribution is to provide an empirical comparison of traditional machine learning methods with eleven deep learning approaches [illegible] and with different learning objectives. Our results, which we have made publicly available as competitive benchmarks, indicate that algorithms based on gradient-boosted tree ensembles still mostly outperform deep learning models on supervised learning tasks, suggesting that the research progress on competitive deep learning models for tabular data is stagnating. To the best of our [illegible] approaches for tabular data; as such, this work can serve as a valuable starting point to guide researchers and practitioners interested in deep learning with tabular data.

contrast to image or language data – are heterogeneous, leading to dense numerical and sparse categorical features. Furthermore, the correlation among the features is weaker than the one introduced through spatial or semantic relationships in image or speech data. Hence, it is necessary to discover and exploit relations without relying on spatial information [9]. Therefore, Kadra et al. called tabular data sets the last "*unconquered castle*" for deep neural network models [10].

Heterogeneous data are the most commonly used form of data [7], and it is ubiquitous in many crucial applications, such as medical diagnosis based on patient history [11]–[13], predictive analytics for financial applications (e.g., risk analysis, estimation of creditworthiness, the recommendation of investment strategies, and portfolio management) [14], click-through rate (CTR) prediction [15], user recommendation systems [16], customer churn prediction [17], [18], cybersecurity [19], fraud detection [20], identity protection [21], psychology [22], delay estimations [23], anomaly detection [24], and so forth. In all these applications, a boost in predictive performance and robustness may have considerable benefits for both end users and companies that provide such solutions. Simultaneously, this requires handling many data-related pitfalls, such as noise, impreciseness, different attribute types and value ranges, or the missing value problem and privacy issues.

Meanwhile, deep neural networks offer multiple advantages over traditional machine learning methods. First, these methods

Sources: arxiv.org/abs/2207.08815, arxiv.org/abs/2106.03253
and arxiv.org/abs/2110.01889

# Tree-based methods win in energy prediction competitions


Global Energy Forecasting Competition (2014)


EEM20 Wind Power Forecasting Competition (2020)


Hybrid Energy Forecasting and Trading Competition (2024)

| Tree-based models | Tabular data | Deep learning |
| --- | --- | --- |
| Effectively handles different data types | Mixed data? | Requires careful feature standardisation |
| Supports sharp decision boundaries | Data smoothness? | Assumes smoothness |
| Good at modelling fairly direct relationships | Data complexity? | Good at extracting meaning from low-level abstractions |
| Performances well on smaller ~10k samples | Data sizes? | Performances better with +100k samples |
| High interpretability | Interpretability? | Low interpretability |
| Efficiently filters out irrelevant data | Irrelevant data? | Sensitive to irrelevant data input |
| Robust to missing data | Missing data? | Sensitive to missing data |

# Difference in smoothness between RF and MLP

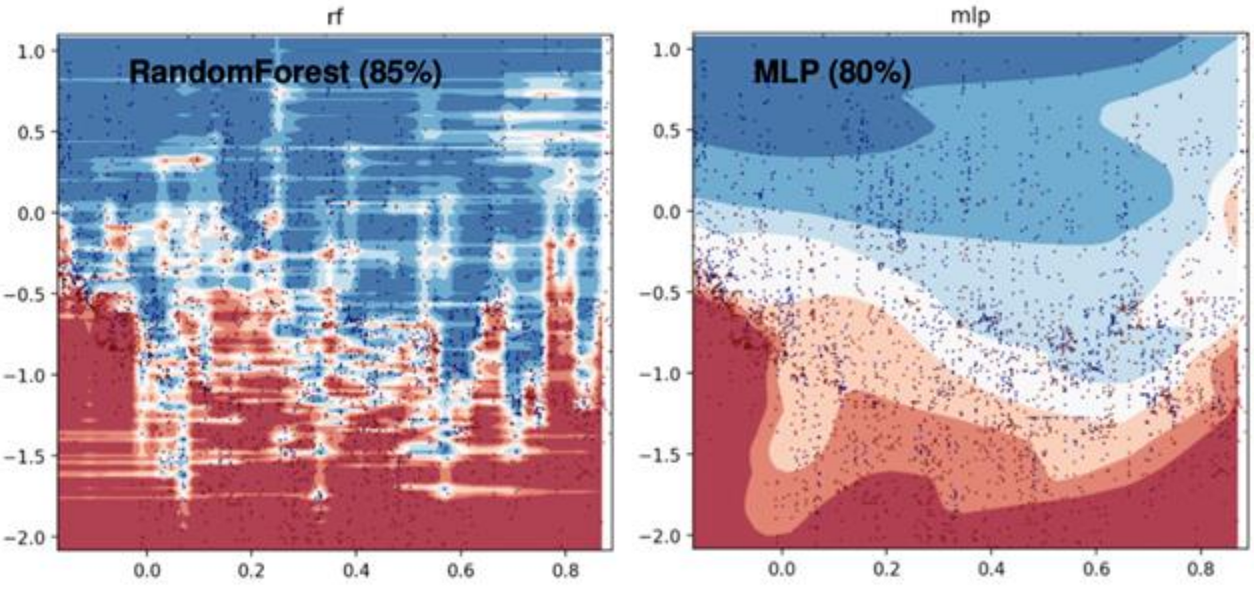
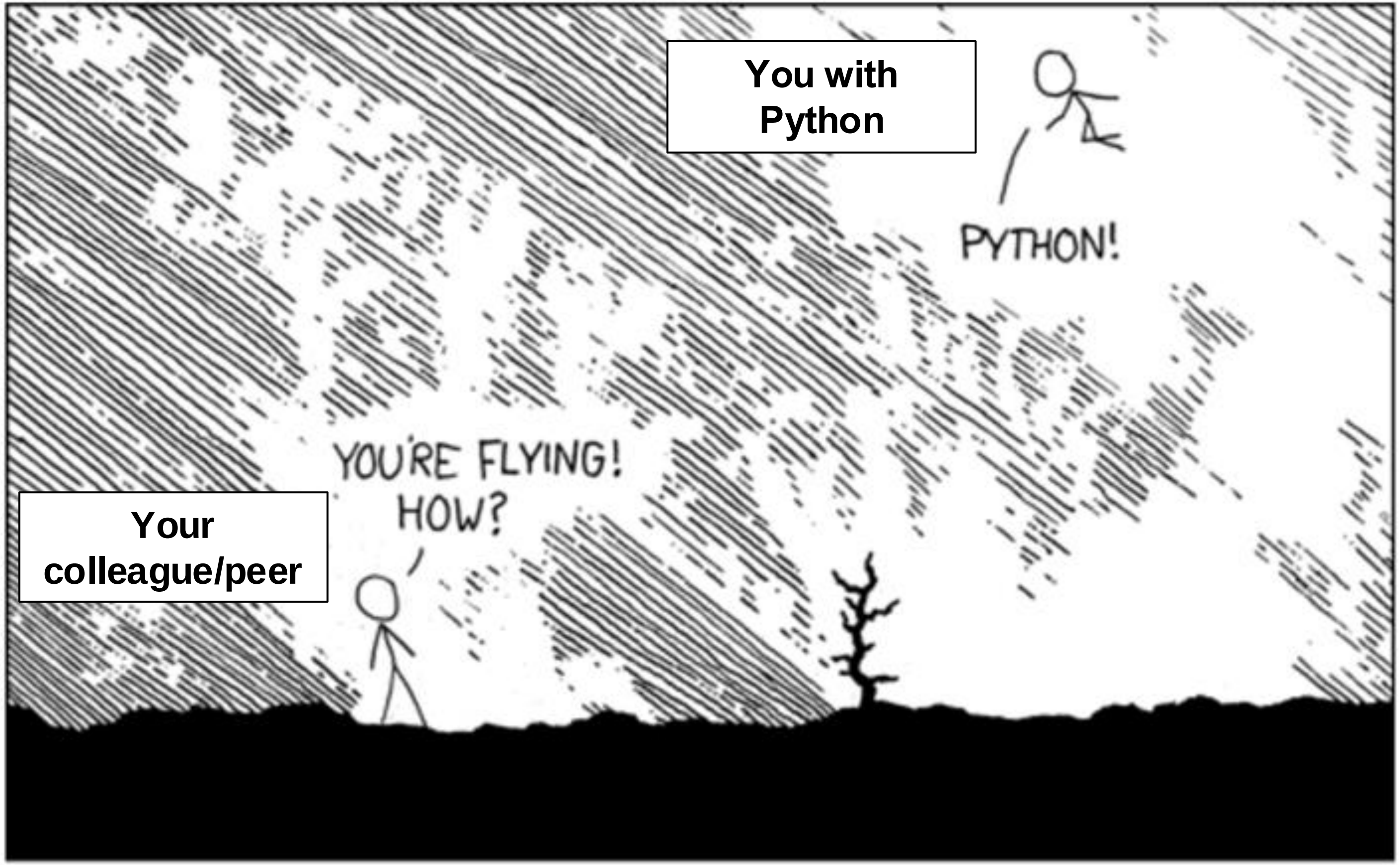



Figure 20: Decision boundaries of a default MLP and RandomForest for the 2 most important features of the *electricity* dataset

Source: arxiv.org/abs/2207.08815

# Why Python for data work in energy?



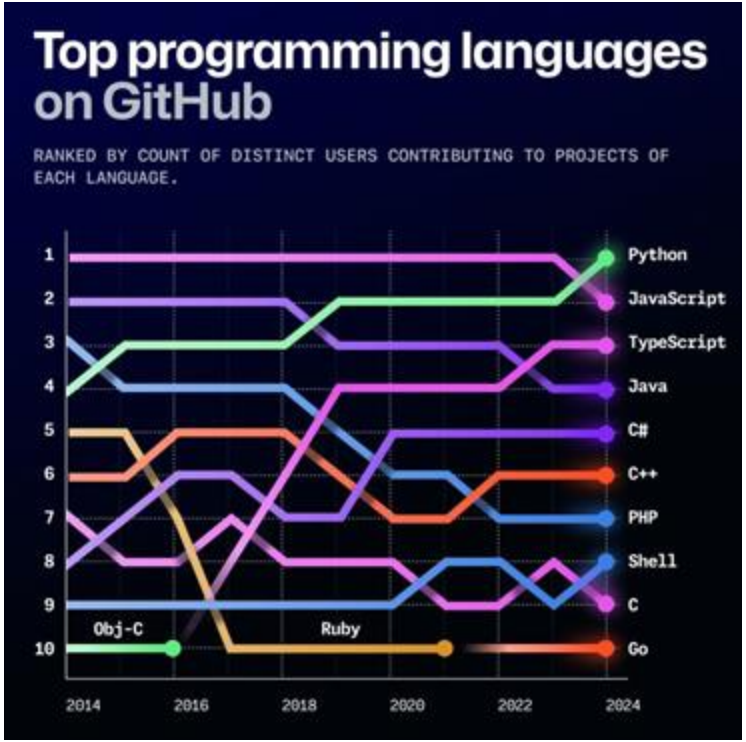

Hint: `import antigravity`

# Python is the lingua franca of data work

python™

As of 2024, Python is the most commonly used programming language on Github!

Widely used for machine learning, engineering, statistics, automations…



# Popularity of Python stems from its ease-of-use


First released in 1991 by Guido van Rossum


“The Zen of Python”
Emphasises versatility, readability and ease-of-use

# Why open-source software?

# What is the value of open-source software?

- **Transparency** → I can read and understand the source code
- **Flexibility** → I can modify and adopt the source code to my needs/use case
- **Collaboration** → I can increase development speed and share investment costs

# Why do we need *treewe*?

# Limitations of standard tree-based prediction libraries

- Existing libraries not focused on time series
- Existing libraries not handling trends and extrapolation well
- Existing libraries have different naming conventions
- Existing libraries are not focused on energy use cases

CatBoost

DMLC XGBoost

scikit learn

LightGBM

pvlib

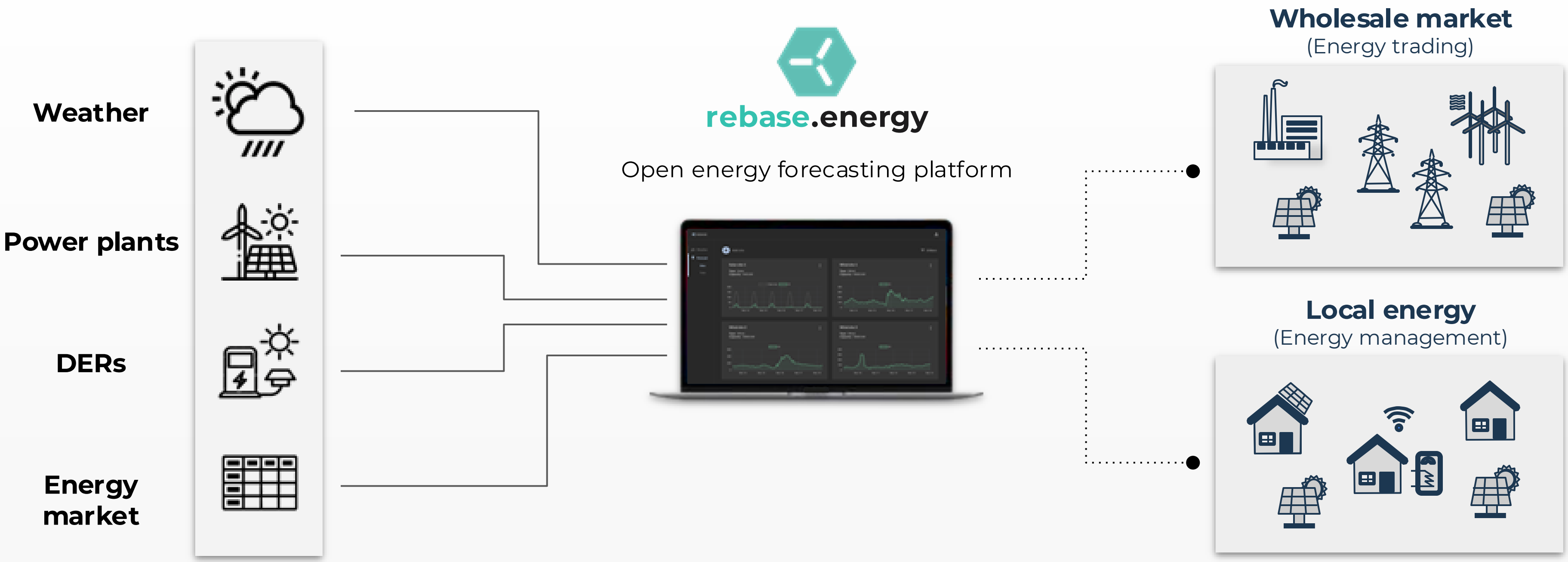
windpowerlib

# Live coding

Open source is not everything…

## Our Platform

# **Python-first and open** energy forecasting platform



# Connect with us!

**Follow us on Linkedin!**

**Join us on Slack!**