Modern Practices in Business Forecasting:

Insights From the Success Story of Gradient Boosting Decision Trees (GBDTs)

Filotas Theodosiou

Contact Details:
- filotas.theodosiou@vives.be
- www.linkedin.com/in/filotas-theodosiou

- Agenda

1) A Brief Overview of Business Forecasting
2) Introduction to Gradient Boosting Decision Trees (GBDTs)
3) Real World Implementations & Practical Tips.
4) Deployment in Production: Suggestions & Potential Pitfalls

Section 1.

A Brief Overview of Business Forecasting

- What is the Point of Forecasting?
- Why Consider Forecasting Systems in your Business?
- What to Actual Forecast?
- The Missing Key to the Puzzle.

What is the Point of Forecasting

“A wall of wood alone shall be uncaptured, a boon to you and your children”, 480 BC

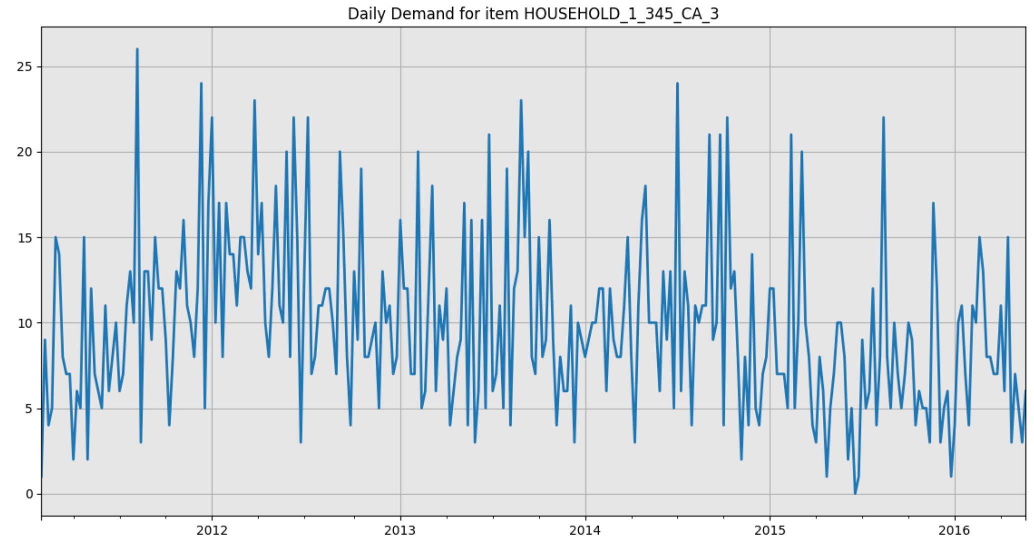
- Herodotus, The Persian Wars, Volume 1

King Algeus in front of Pythia at the Oracle of Delphi, 440-430 BC

Why do we Consider Forecasting Systems?

Quick answer: Assist in our Decision Making.

Example: Inventory Replenishment



But…

What about our managers?

What about our years of experience?

Why do we Consider Forecasting Systems?

Quick answer: Assist in our Decision Making.

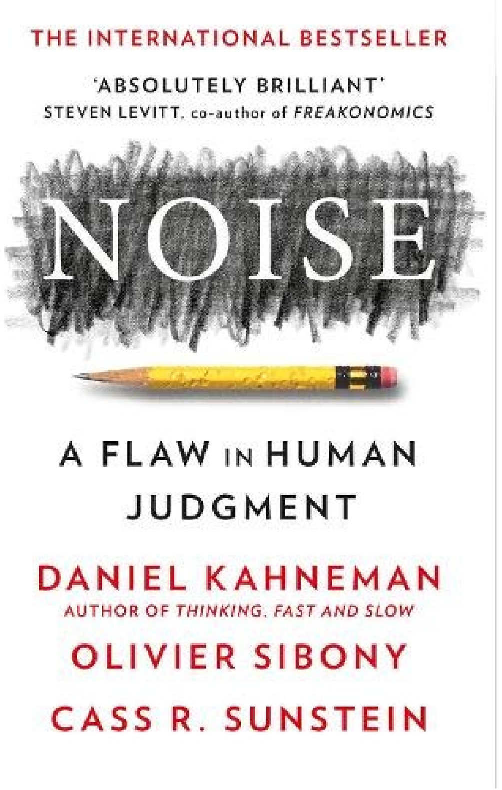
Example: Inventory Replenishment

But…

What about our managers?

What about our years of experience?

1. Biases & Noises*



Kahneman, Daniel, Olivier Sibony, and Cass R. Sunstein. *Noise: a flaw in human judgment*. Hachette UK, 2021.

Why do we Consider Forecasting Systems?

Quick answer: Assist in our Decision Making.

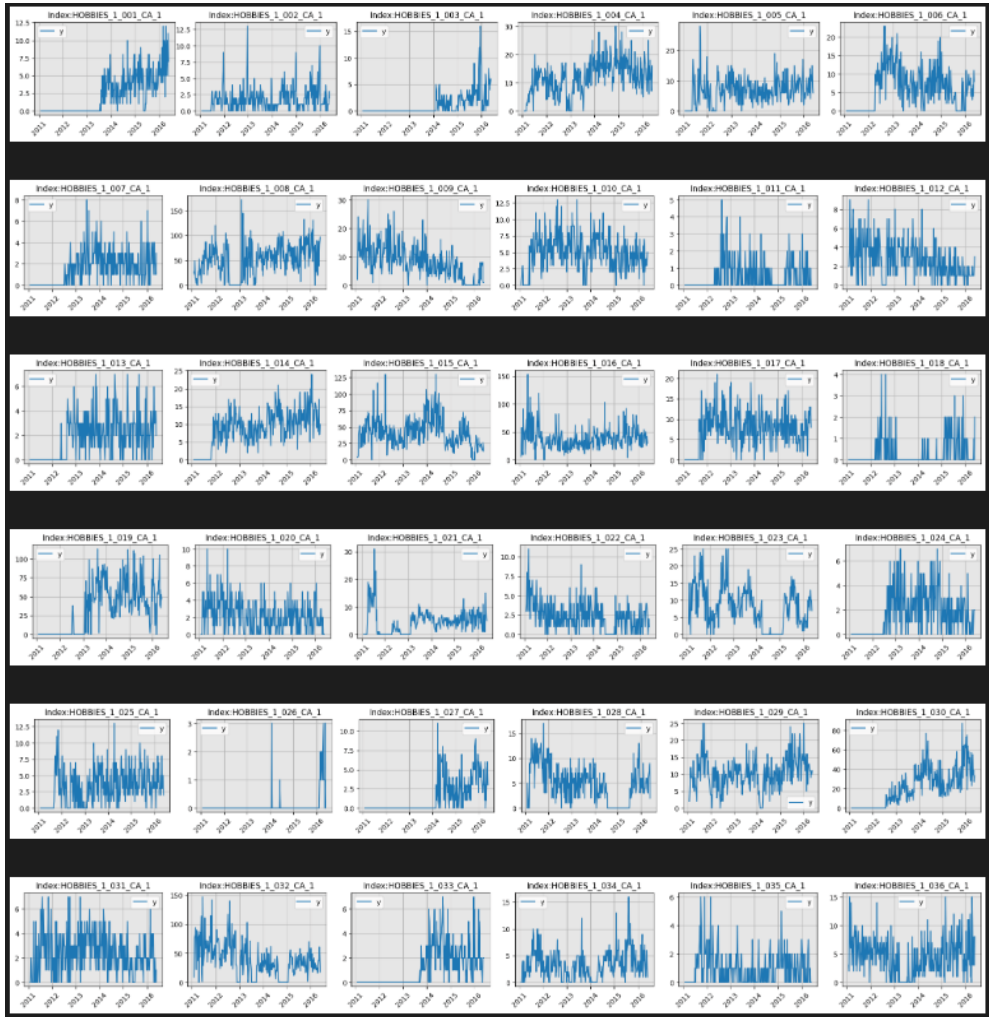
Example: Inventory Replenishment

But…

What about our managers?

What about our years of experience?

1. Biases & Noises*
2. Limited brain capacity

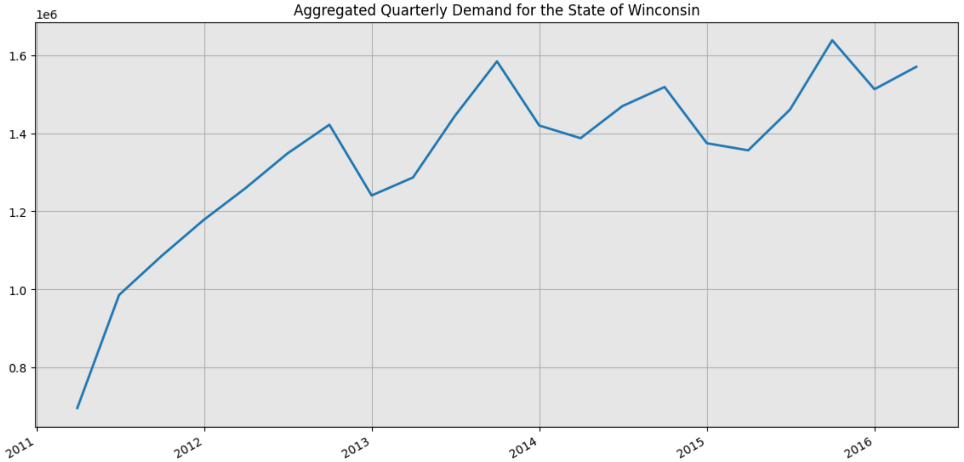


Kahneman, Daniel, Olivier Sibony, and Cass R. Sunstein. *Noise: a flaw in human judgment*. Hachette UK, 2021.

What to Actually Forecast

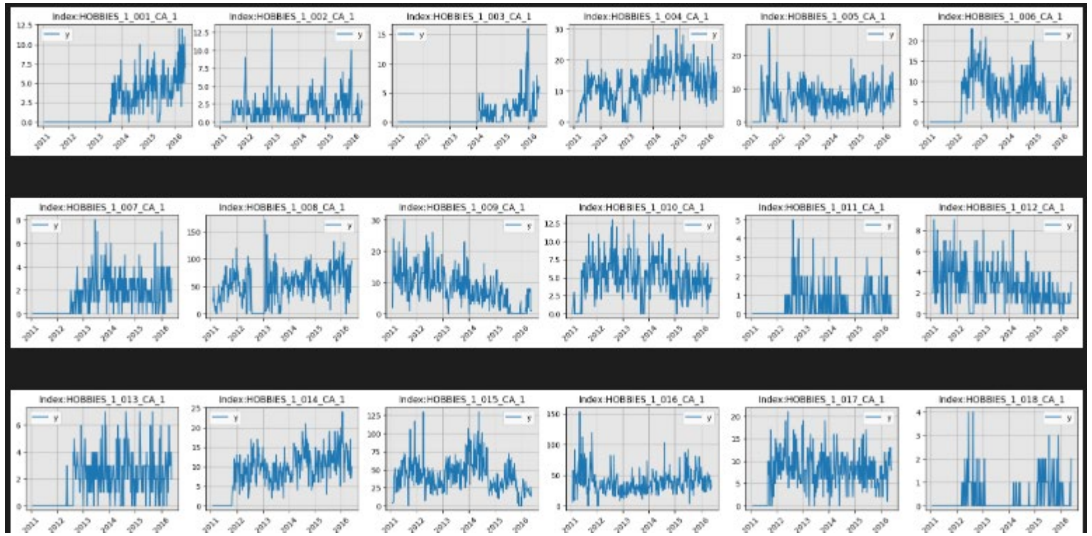
Different types of forecasts for different organizational decisions

Strategic Forecasts



- Small number of time series
- Long Forecasting Horizon
- Often no need for forecasting systems
- E.g. Possible Expansion

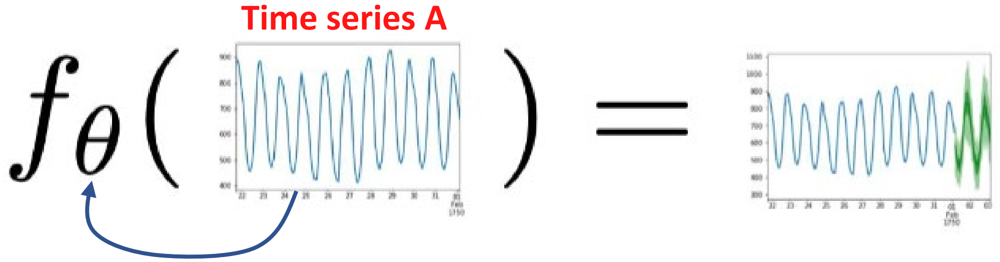
Operational Forecasts



- Large number of time series
- Short forecasting horizons
- Forecasting systems & high degree of automation is needed
- E.g. Inventory Replenishment

Januschowski, Tim and Stephan Kolassa "A classification of business forecasting problems". Foresight the International Journal of Applied Forecasting. International Institute of Forecasters (2019). Issue 52

Modern Operational Forecasts: The missing key to the puzzle



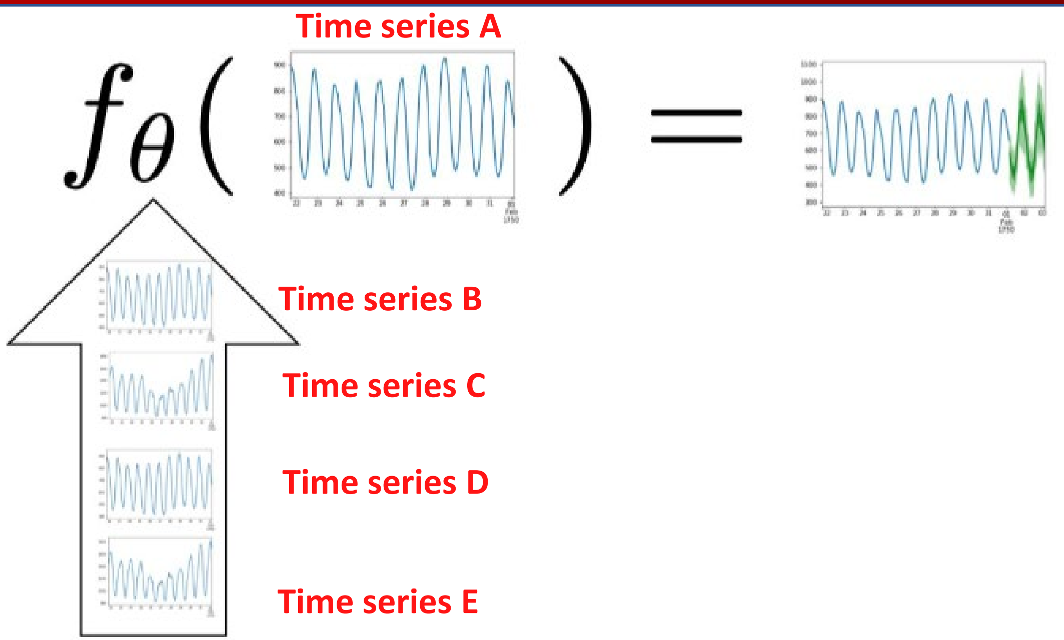
Local Learning: A model per time series

Most “traditional” approaches.

- Computational issues
  - 10000 models for 10000 time series?
- Models of limited capacity.
  - Many time series have limited historical values to properly fit a mod
  - Limited data ⇔ Limited number of parameters ⇔ “Simpler” Models.

Montero-Manso, Pablo, and Rob J. Hyndman. "Principles and algorithms for forecasting groups of time series: Locality and globality." *International Journal of Forecasting* 37.4 (2021): 1632-1653.
Gasthaus. Jan & Januschowski , Tim. “Modern Forecasting in Practise, Session 1”, Sphere Online Course. URL: https://www.getsphere.com/cohorts/modern-forecasting-in-practice

Modern Operational Forecasts: The missing key to the puzzle



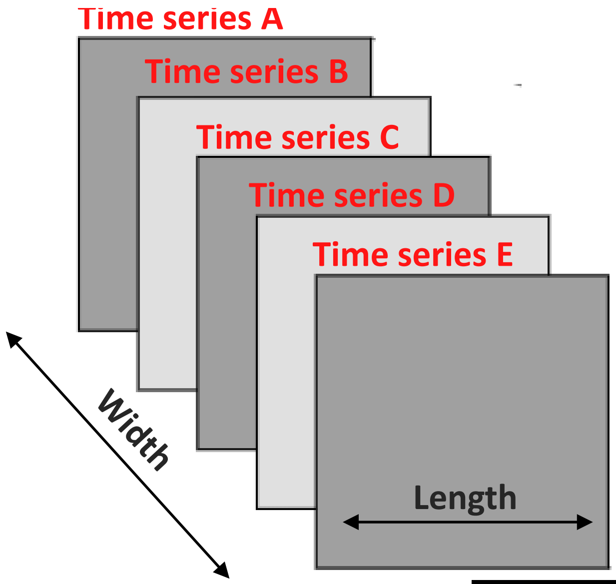
Local Learning: A model per time series

Most "traditional" approaches.

- Computational issues
    - 10000 models for 10000 time series?
- Models of limited capacity.
    - Many time series have limited historical values to properly fit a model
    - Limited data ⇔ Limited number of parameters ⇔ "Simpler" Models.

Global Learning: A paradigm shift.

- Increase the width and not the length
- Allow for higher capacity models (NNs, GBDTs, etc..)
- Theoretical evidence for improved performance



Montero-Manso, Pablo, and Rob J. Hyndman. "Principles and algorithms for forecasting groups of time series: Locality and globality." *International Journal of Forecasting* 37.4 (2021): 1632-1653.
Gasthaus. Jan & Januschowski , Tim. "Modern Forecasting in Practise, Session 1", Sphere Online Course. URL: https://www.getsphere.com/cohorts/modern-forecasting-in-practice

Section 2.

Introduction to Gradient Boosting Decision Trees (GBDTs)

- What the Media do not Tell you.
- Facts About GBDTs in Forecasting.
- The Theory Behind GBDTs.
- How to Make GBDTs Work for You.

What the Media Does not Tell you.

Generative Pre-Tr

Will ChatGPT Take Your Job?

Is chatGPT a real threat to Humanity

ChatGP

medical

moment as 'milestone for artificial intelligence'

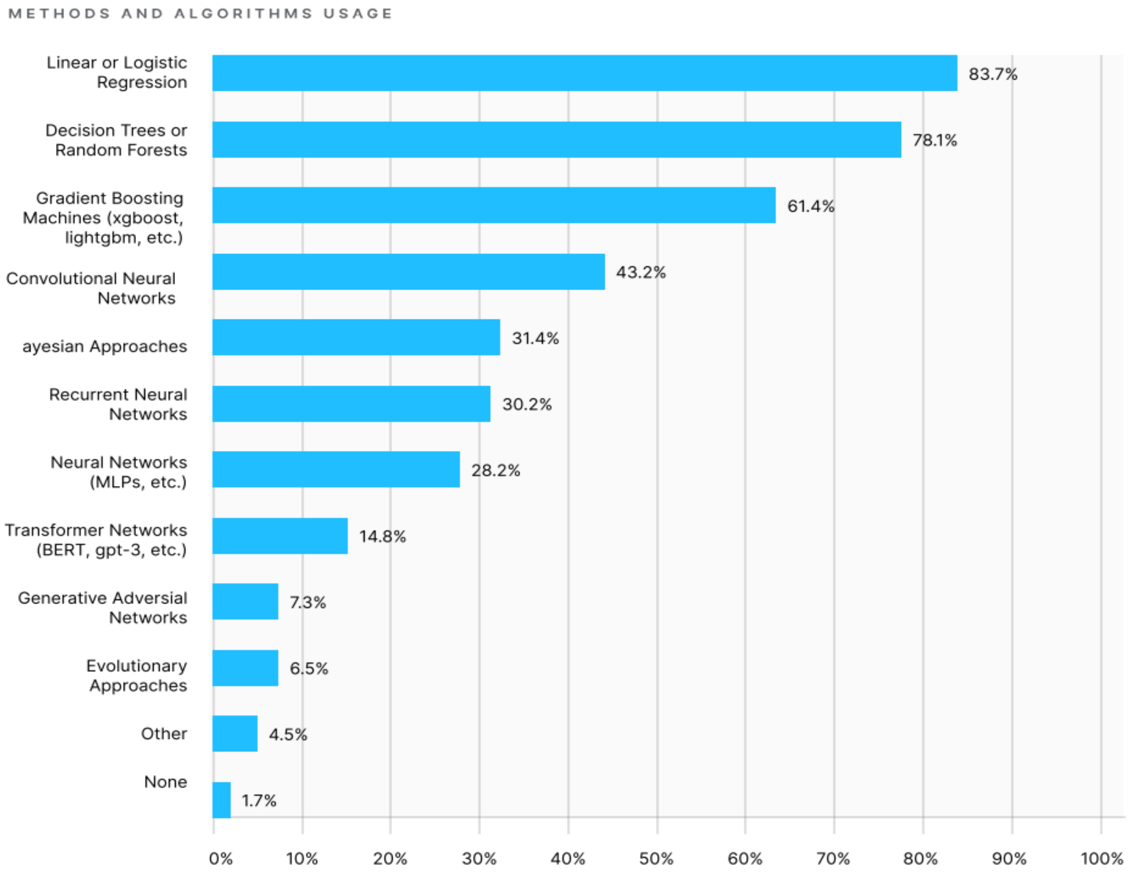
ChatGPT Is a Plague Upon Education

exams

despite 'mediocre' performance

What the Media Does not Tell you.

Kaggle Annual Report 2019



Generative Pre-Trained Transformers (GPTs) are all over the news.

And for a good reason. They dominate NLP and Computer Vision tasks

But:

- **Industry deals mostly with structured (Tabular) Data.**
- **GBDT models have been dominating applied ML competitions**
- **GBDT models are widely applied in practice.**

“ The most glaring difference between what is used on Kaggle (real-world datasets) and what is fashionable in academia”
– Antony Goldbloom, CEO of Kaggle when asked about the prominence of GBDTs.

Kaggle is the biggest Data Science Platform with over 10million active users. It is used by educators, practioneres and by big companies (e.g. Walmart, Netflix)

Facts About GBDTs in Forecasting.

2010s Situation: Despite some evidence of performance benefits

- Statistical methods were mostly preferred both in academia and practice
- Local Neural Networks (and other ML models) as an alternative

Competition	Place
Walmart	3rd
Rossman	1st, 2nd, 3rd
Wikipedia	2nd
Corporacion Favorita	1st, 3rd
Recruit Restaurant	1st
M4	2nd
M5	1st

Submissions that utilized GBDT models

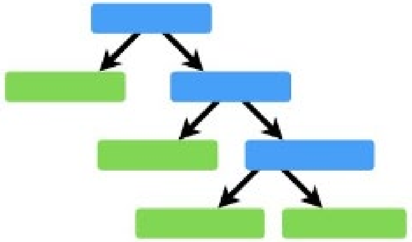
2020 - A critical moment: M5 Forecasting Competion

- The biggest forecasting competition (over 6000 participating teams)
- Walmart Retail Forecasting (42.840 daily time series)
- GBDTs (and Global Learning) dominated the competition.
  - Heavily outperformed statistical bencmarks
  - Accuracy improvement over Deep Learning Models
  - Socking moment: 59th place with 0 model optimization

Bojer, Casper Solheim, and Jens Peder Meldgaard. "Kaggle forecasting competitions: An overlooked learning opportunity." *International Journal of Forecasting* 37.2 (2021): 587-603.

The Theory Behind GBDTs: Decision Trees -> Bagging->Boosting

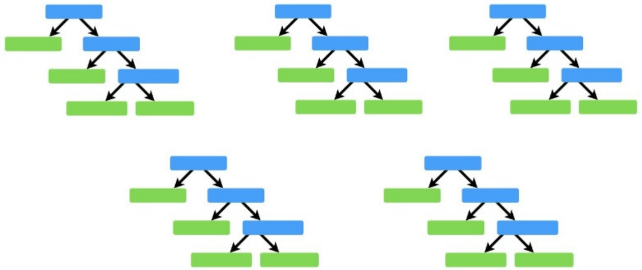
Decision Trees



Weak Base Learner

- Very simple model
- Splits features at nodes
- Takes the average

Bagging (Random Forest)

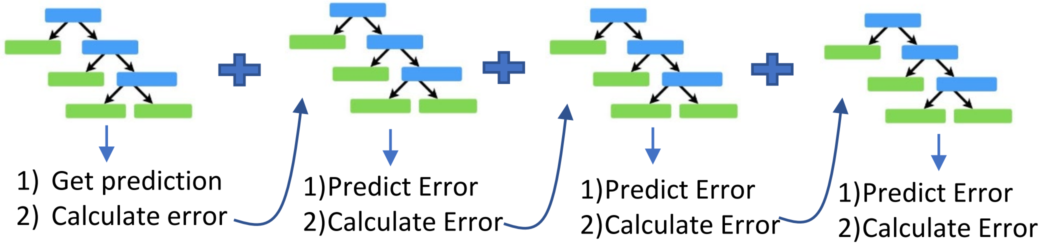


- Combines multiple weak learners
- Fit a unique tree to different data samples
- Combines independent predictions

(Partially) Deals with overfitting

- Cannot extrapolate.

Boosting: (GBDTs)



- Combines multiple weak learners
- Fits trees sequentially instead of parallel
- Iteratively corrects errors of the previous tree
- (Partially) Deals with overfitting
- Cannot extrapolate.

XGBoost LightGBM CatBoost

The Theory Behind GBDTs: Modelling the Feature Space

1. GBDTs: Decision Trees for Different Sets of Features:
   - Different weak learners for diverse features
   - Informative feature space ⇔ More holistic view for the task.
   - Require a detailed description of the problem
     (eg paper size, average number of dots, total folds etc)

2. Neural Networks: Transform the Feature Space
   - Multiple affine transformations
   - Find a meaningful representation of the feature space
   - Unfold the paper (given enough data) and count

Imaginary Problem:
Count dots inside the folded paper

Main Takeaway: GBDTs excel when given an accurate feature space

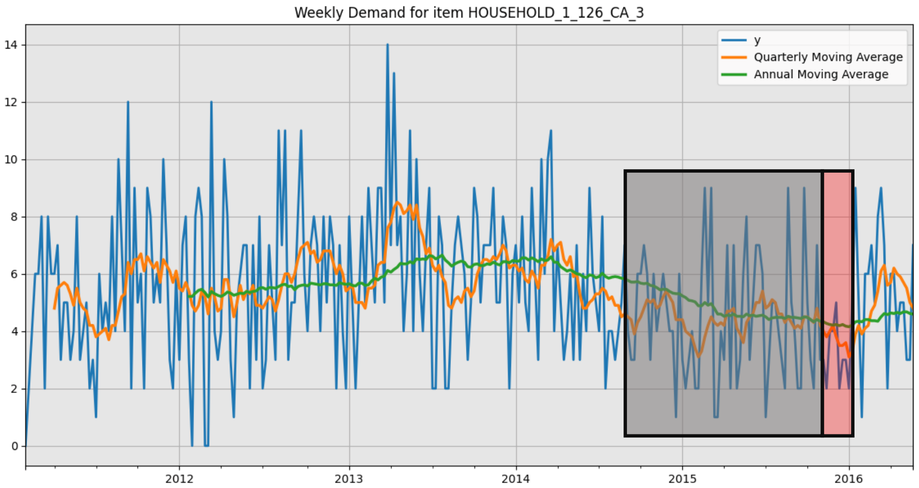
Original Example: Chollet, Francois. *Deep learning with Python*. Simon and Schuster, 2021.

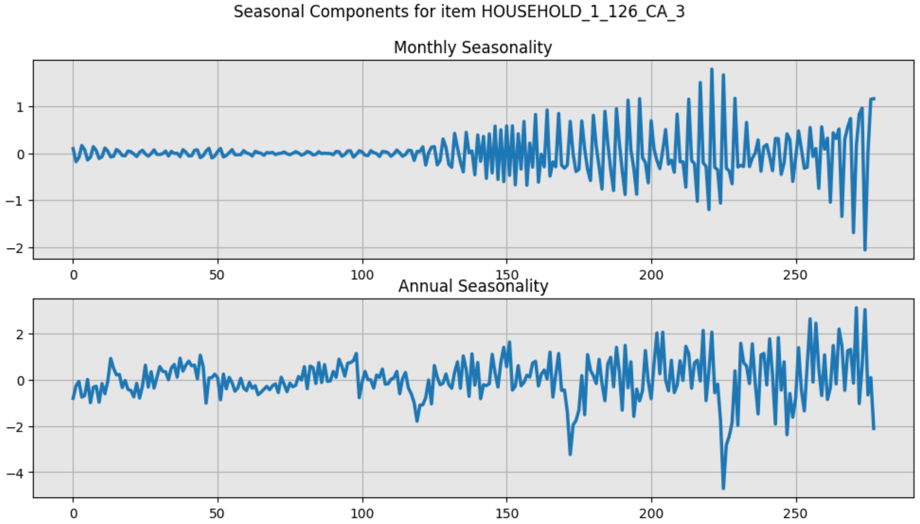
How to Make GBDTs Work:
The Importance of Feature Engineering

Feature Selection & Engineering is the most crucial step when forecasting with GBDTs

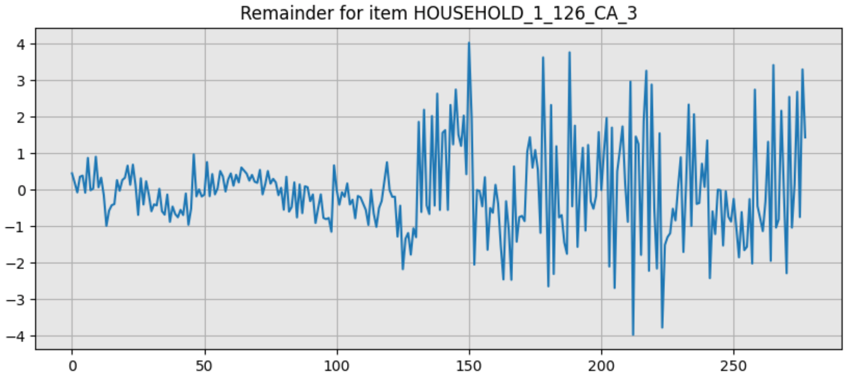
Provide A Detailed Description.

1) Recent History
2) Trend
3) Seasonality



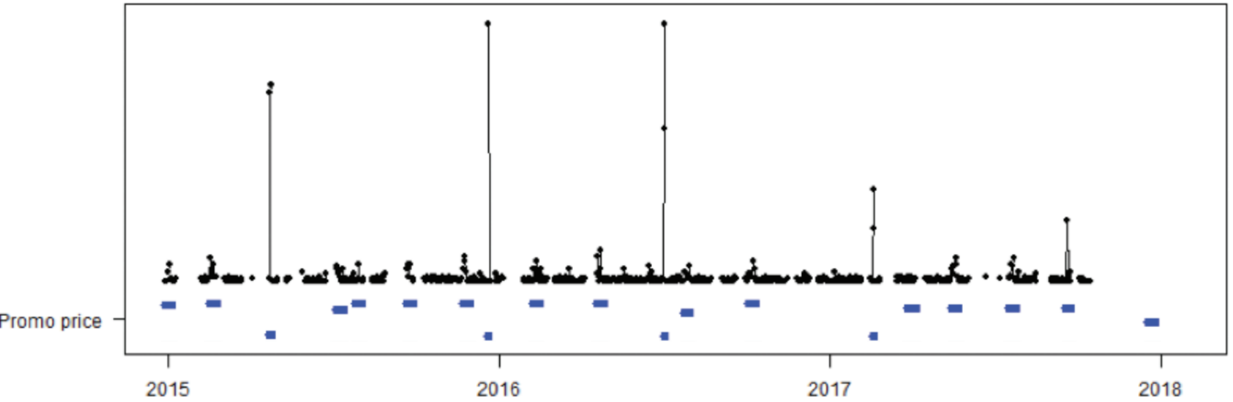


Question:
What about the reminder?
Is it just random fluctuations?



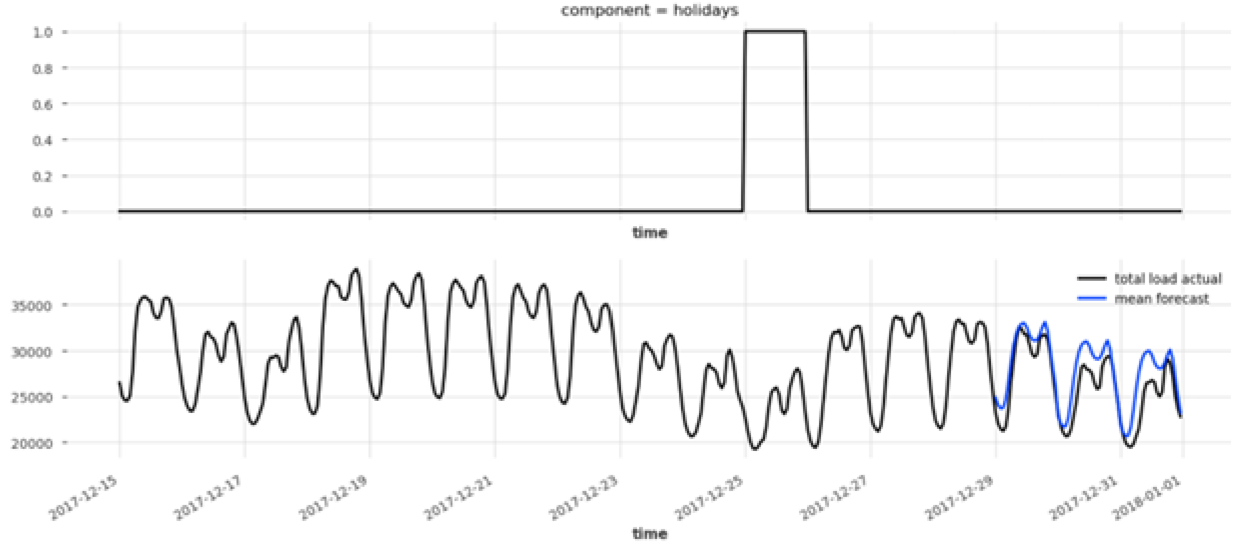
How to Make GBDTs Work:
The Importance of Context Expertise

Promotions & Discounts



Scenario 1

Holidays & Special Events



Scenario 2

Main takeaway: Provide all information available to maximize the efficiency of the model

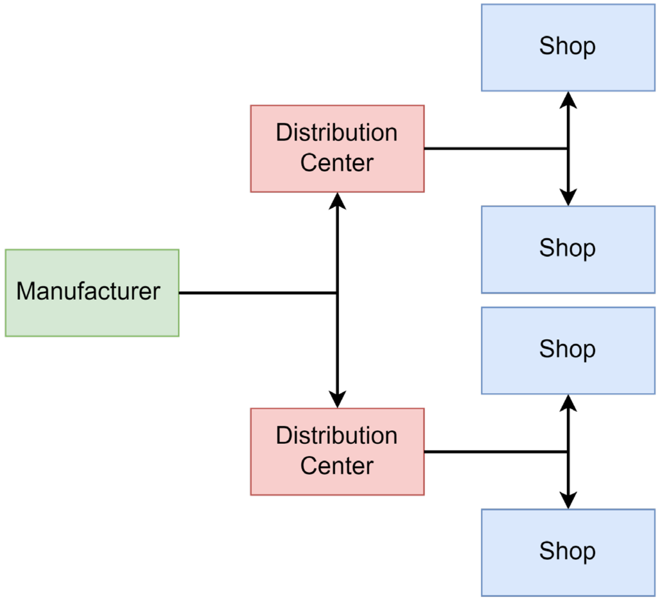
Januschowski, Tim and Stephan Kolassa "A classification of business forecasting problems". Foresight the International Journal of Applied Forecasting. International Institute of Forecasters (2019). Issue 52
Gasthaus. Jan & Januschowski , Tim. "Modern Forecasting in Practise, Session 2", Sphere Online Course. URL: https://www.getsphere.com/cohorts/modern-forecasting-in-practice

Section 3.

Real-World Implementations & Practical Tips

- PWO Example 1: Sales Forecasting
- PWO Example 2: Explainable Student Exam Grades Predictions
- Justified or Not?
  - The Case Against Statistical Models
  - The Bandwagon of Deep Learning

PWO Example 1: Sales Forecasting



Situation: Collaboration in a retail supply chain(add picture of supply chain)

1. Retailer orders weekly from a DC
2. DCs order bi-weekly from a Supplier
3. Supplier orders row products every month.

Aim: Reduce the overall inventory across the supply chain.

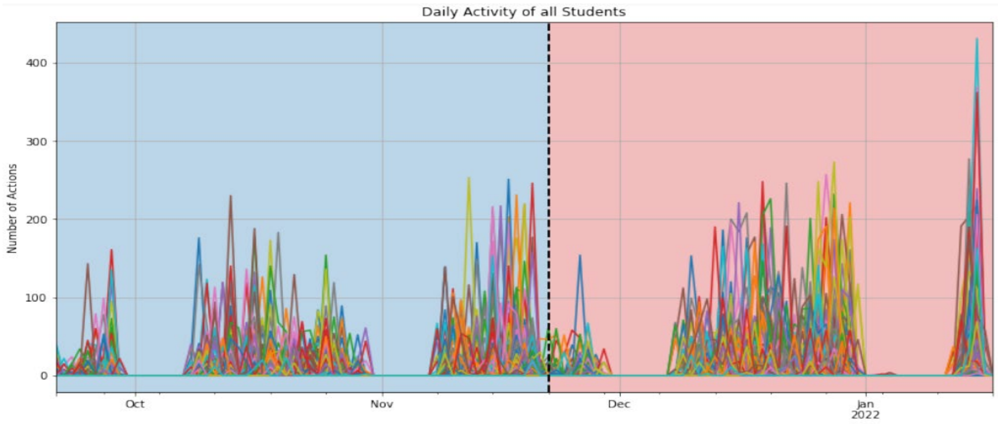
- Use LGBM to predict the demand for each level
- Feed the predictions to Neural Network to align them

	Level		
Model	Manufacturer	Warehouse	Supplier
LGBM	**2.3**	**3.06**	**6.2**
ETS	7.5	4.13	6.383
LASSO	3.13	7.61	
Stepwise Reg	2.32	10.65	

Table 1: Mean Interval Score of Demand Forecasts

Main Takeaway: Accuracy improvements regardless the hierarchical & the temporal level.

PWO Example 2: AI in Education



Situation:

Forecast students exam results based on their online activity

Aim:

1. Produce Early Predictions

   Allow enough time for adjustments

2. Produce Accurate Predictions

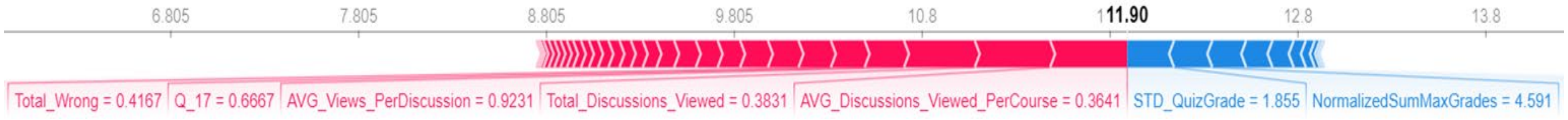
   Facilitate teacher's decion making

Model	MSE	Recall	Accuracy
XGBoost	10.83	0.68	0.53
CatBoost	**10.45**	**0.71**	**0.55**
LGBM	11.86	0.67	0.52
TabNet	14.58	0.65	0.42
Logistic Regression	-	0.38	0.4

Model	Accuracy
CNN *	0.68
CatBoost	**0.73**

Bonus point:

Explainability



Main Takeaway: GBDT models have high compatibility with explainability techniques

Justified or Not?
The Case Against Statistical Models

Notable Limitations

1. By definition local models
2. Limited performance in a production environment
   - Missing values
   - Cold start
   - Small history
   - Intermittency

On the other hand:

- They are less complex
- We know how they work
- They are battle-tested
- They are very hard to beat at some tasks

Main takeway: Always evaluate your fancy model with a good old Exponential Smoothing

Justified or Not?
The Bandwagon of Deep Learning

What about the fancy model on paper X?

Quiz Time!

1) How many parameters are included in GPT-3 (backbone of chatGPT)?

Tip:
ResNet-200 –used as backbone for many Vision Tasks- has 66M.

Justified or Not?
The Bandwagon of Deep Learning

What about the fancy model on paper X?

Quiz Time!

2) How much time would it take to train GPT-3 on my Laptop (Single mediocre GPU)

Tip:
OpenAI paid approximately $5M to train the model.

BELGIUM FORECASTING CENTRE

Justified or Not?
The Bandwagon of Deep Learning

What about the fancy model on paper X?

Quiz Time!

3) In your opinion, how many research labs in the world can build ChatGPT-like models?

Justified or Not? The Bandwagon of Deep Learning

What about the fancy model on paper X?

Be hesitant to jump on the bandwagon

- Are these models properly evaluated?
- Can they deployed in production?
- Reported evidence of their limitations.

A very complex Transformer NN is outperformed by a very simple model with 0 trainable parameters.

Dataset	Model	Autoformer		AverageTile	
	Metric	MSE	MAE	MSE	MAE
ETTm2	96	.255	.339	.263	.301*
	192	.281	.340	.321	.337*
	336	.339	.372	.376	.370*
	720	.422	.419	.471	.422
Electricity	96	.201	.317	.229	.290*
	192	.222	.334	.228	.294*
	336	.231	.338	.240	.307*
	720	.254	.361	.281	.338*
Exchange	96	.197	.323	.139*	.269*
	192	.300	.369	.235*	.352*
	336	.509	.524	.383*	.454*
	720	1.447	.941	.931*	.735*

Sun, Fan-Keng, and Duane S. Boning. "Fredo: frequency domain-based long-term time series forecasting." *arXiv preprint arXiv:2205.12301* (2022).

But under the right conditions they show remarkable performance gains

“Building Deep Learning Architectures is more an art than a science” - François Chollet, creator of Keras.

Notable Examples: DeepAR, MQ-CNN, TFT, NBEATS, etc…

Main Takeaway:

Unless you have teams of successful Data Scientists, Data and ML Engineers, using NNs in productions might backfire. Use established NNs instead

Grinsztajn, Léo, Edouard Oyallon, and Gaël Varoquaux. "Why do tree-based models still outperform deep learning on tabular data?." *arXiv preprint arXiv:2207.08815* (2022).
Shwartz-Ziv, Ravid, and Amitai Armon. "Tabular data: Deep learning is not all you need." *Information Fusion* 81 (2022): 84-90.

Section 4.

Deployment in Practice.
Suggestions & Potential Pitfalls

- Evaluate, Re-Evaluate and Evaluate Properly
- Improve the Business Process and not the Model

Evaluate, Re-Evaluate and Evaluate Properly

On November 2021, Zillow announced that they were shutting Zillow offers.

Nearly $245M losses

People blamed the Data Science Forecasting team for using Prophet a model developed by Facebook.

Did they properly evaluate??

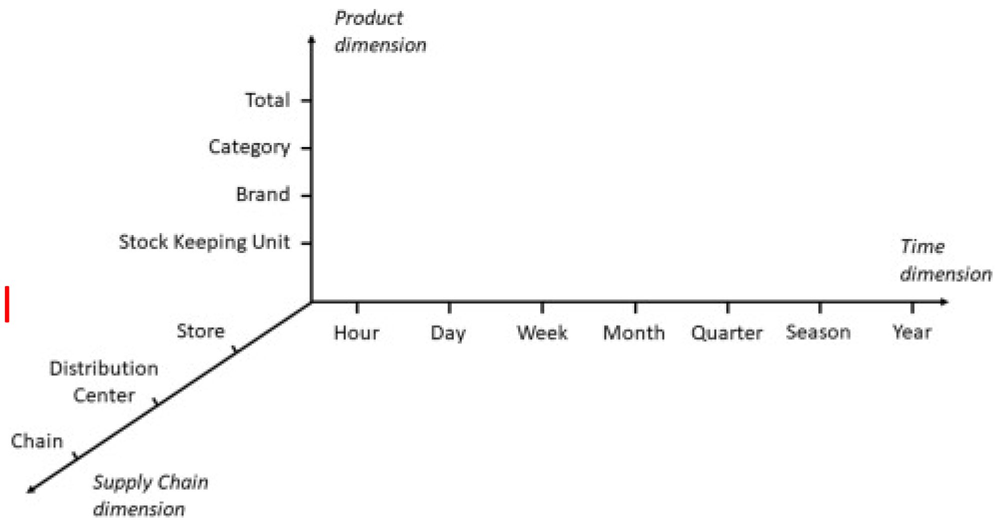
Tips about evaluations:

1. People assume accuracy metrics are directly connected with business Metrics: WRONG
   - Use business metrics (if possible)
   - Pick 2 or more (and possibly uncorrelated) evaluation metrics
   - Use different metrics for stakeholders and different for evaluating the models
2. Do not evaluate on a single forecasting horizon, use cross-validation.
3. Test the model under different scenarios

Improve the Business Process not the Model

- Have clear understanding on what you want to forecast and why

1. What is the business problem you are trying to solve
2. Is forecasting the most effective solution?
3. Is the problem forecastable?



- Design an Effective Forecasting Strategy before picking a model

- Build a Forecasting System

- Models are just a small piece to the forecasting puzzle
  - Data storage, Data pipelines, Decision Making Systems
- Build vs Buy: Minimizes expenses by drawing cost-efficient deployment plans.
  - Cloud vs Local infrastructures

The end ☺

Questions?

Filotas Theodosiou

Contact Details:
- filotas.Theodosiou@vives.be
- www.linkedin.com/in/filotas-theodosiou