Growth
from
Knowledge

Data Science League

Timo Buchner
Isabella Eigner
Lisa Maag
Valery Manokhin
Nataliia Plotnikova

21.02.2022

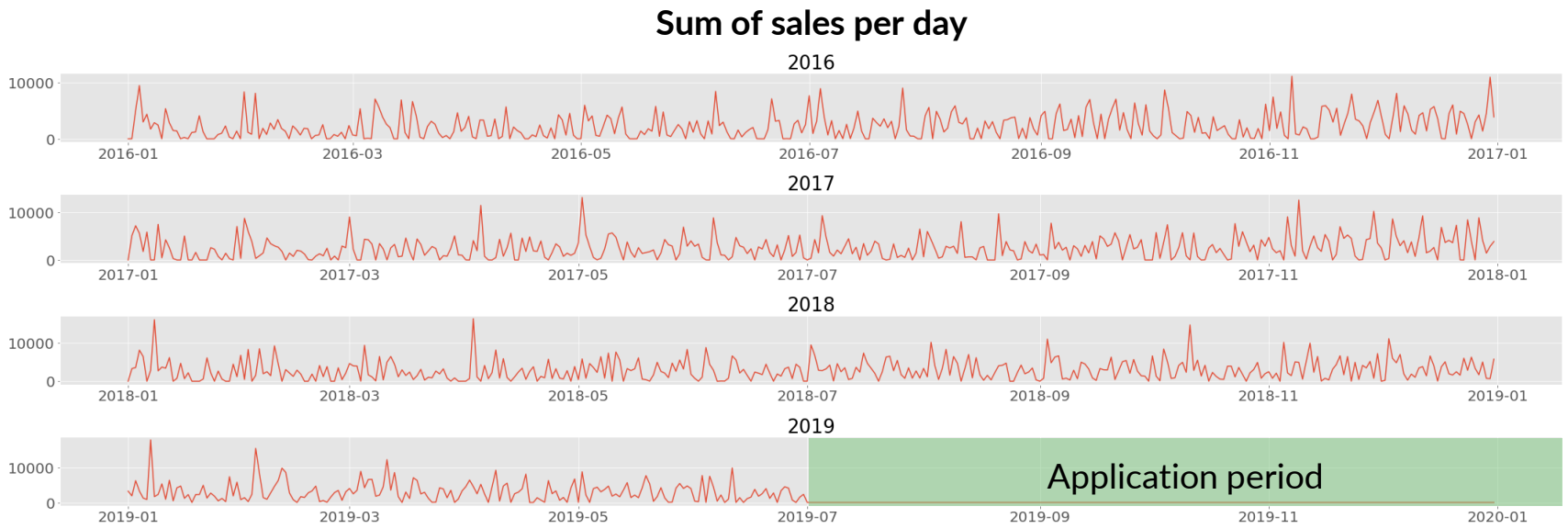
Agenda

- Task overview
- Data exploration
- Approaches
- Validation
- Final results

Task overview

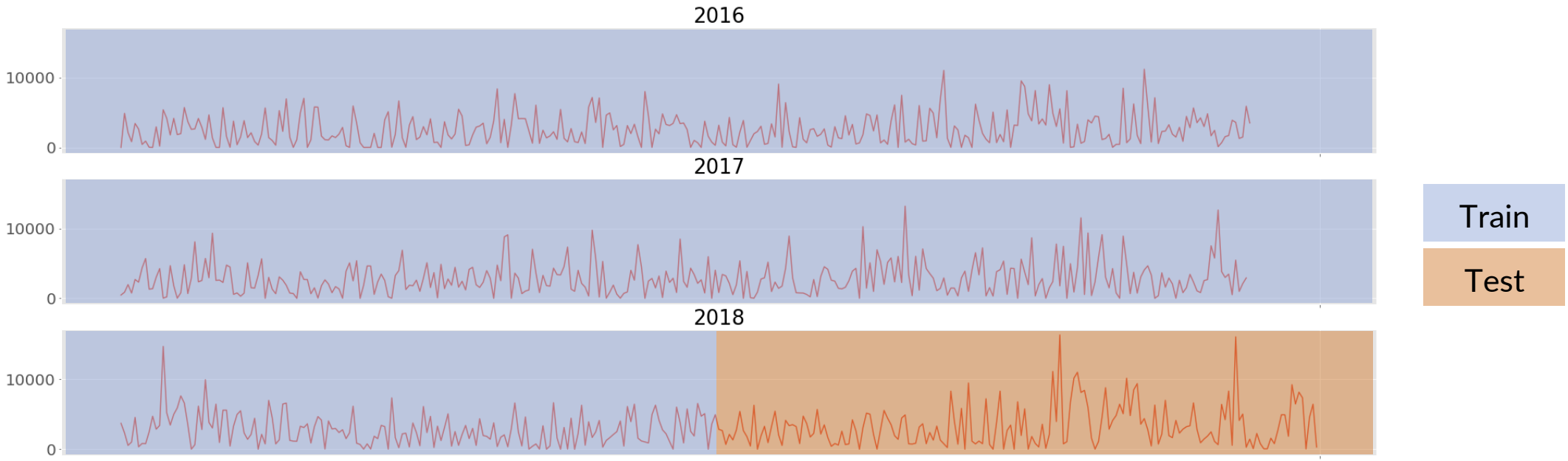
Goal

- Predict **sales** for **26 weeks** based on B2B transactions from **one company**
- **Training data**
  - **3.5 years**
  - 2016 / 2017 / 2018 + 2019 (Jan – June)
- **Application data:**
  - **0.5 year**
  - 2019 (July – Dec)



Validation Setup

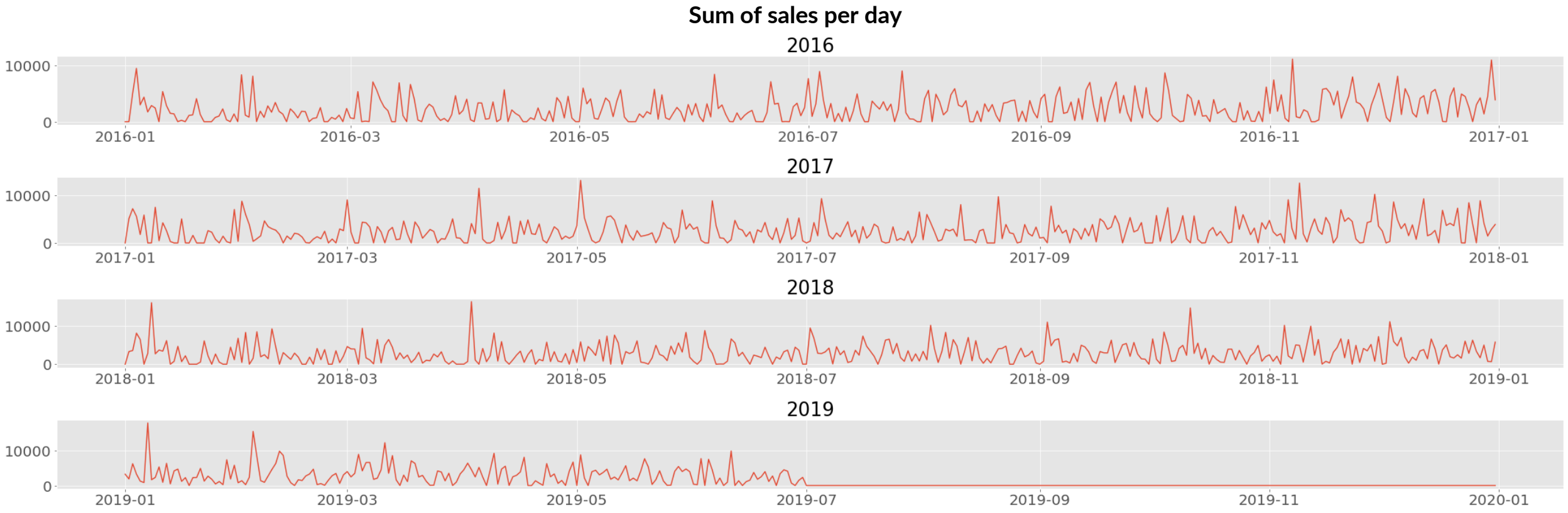
- **Validation approach:**
  - Train on 2,5 years: 2016 + 2017 + 2018 (Jan – June)
  - Test on 0,5 year: 2018 (July-Dec)



Data exploration

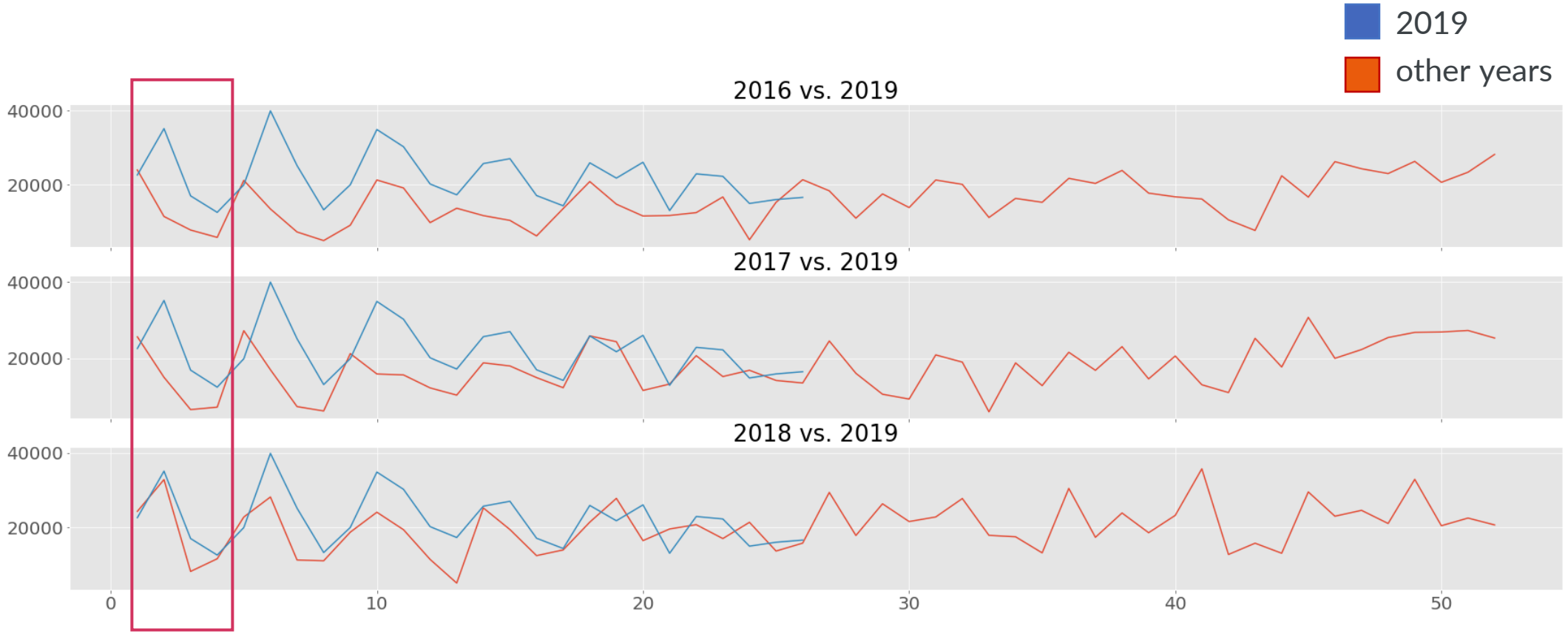
Data exploration

Sum of sales by date



Data exploration

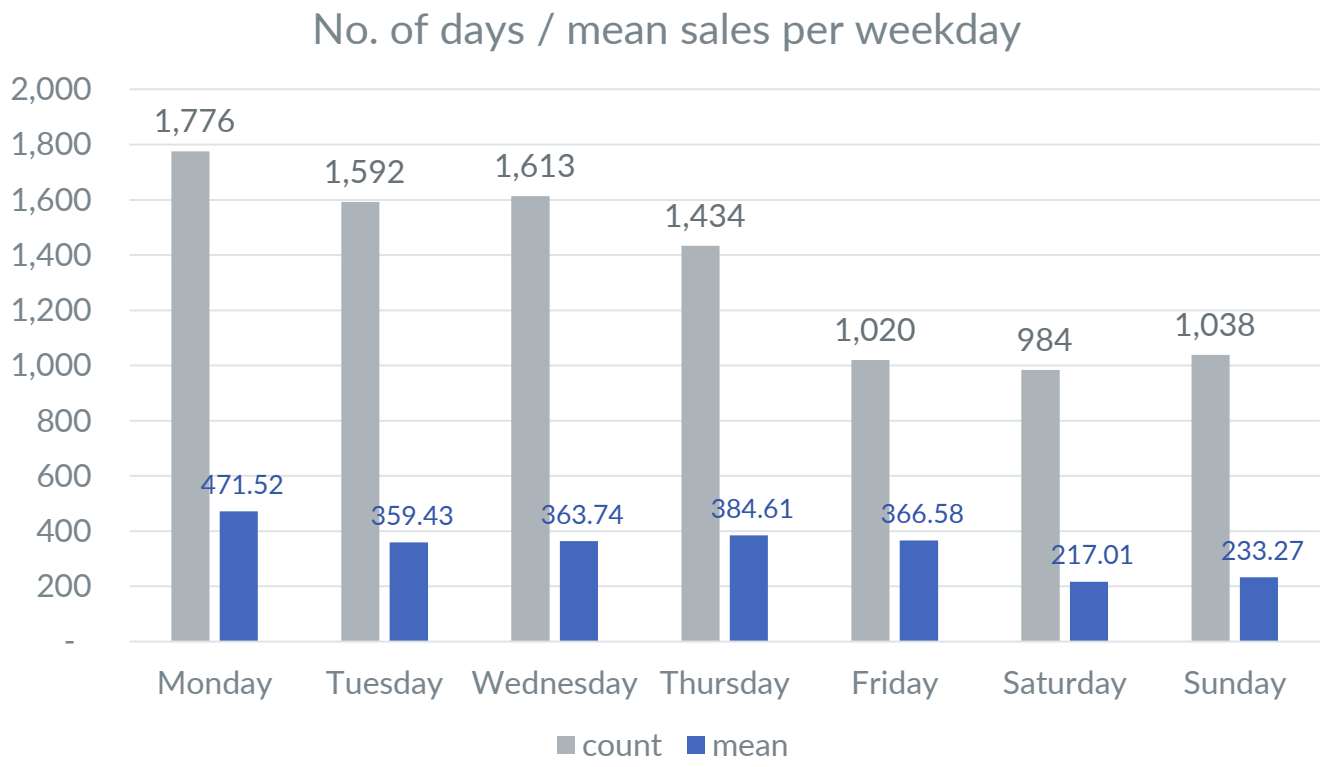
Compare previous years to 2019 (Sales by week)



Data exploration

Missing days

- No missing values
- Missing 123 days in the time period – no pattern in days



Preprocessing steps

Potential features

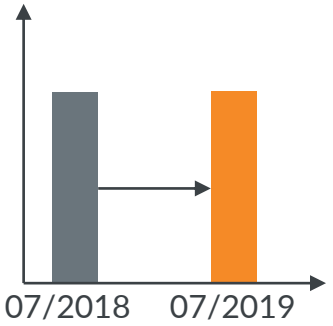
Variable	Aggregation level	Characteristic
Day of week	day	Mon – Sun
Holiday	day	0 / 1
Holiday	week	0 / 1
Time of month	week	1 – 4
Week of year	week	1 – 52
Month	week	1 – 12
Year	week	2016 – 2019

Approaches

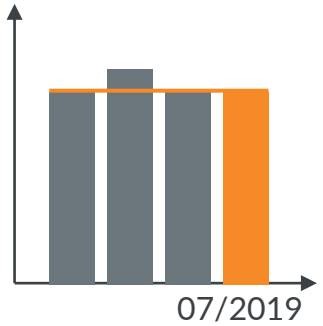
Approaches

Base models

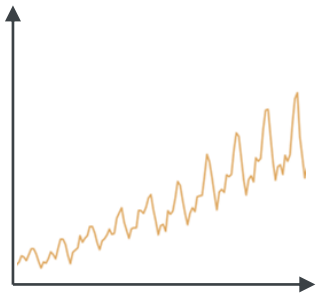
Naive approach
(copy last year)



Median
(median of previous 40 weeks)



SARIMA



Gradient Boosting

CatBoost

LightGBM

NeuralProphet

Neural Prophet

Ensemble

Naive approach + Median

Naive * 0.6

+

Median * 0.4

SARIMA + Median

SARIMA * 0.5

+

Median * 0.5

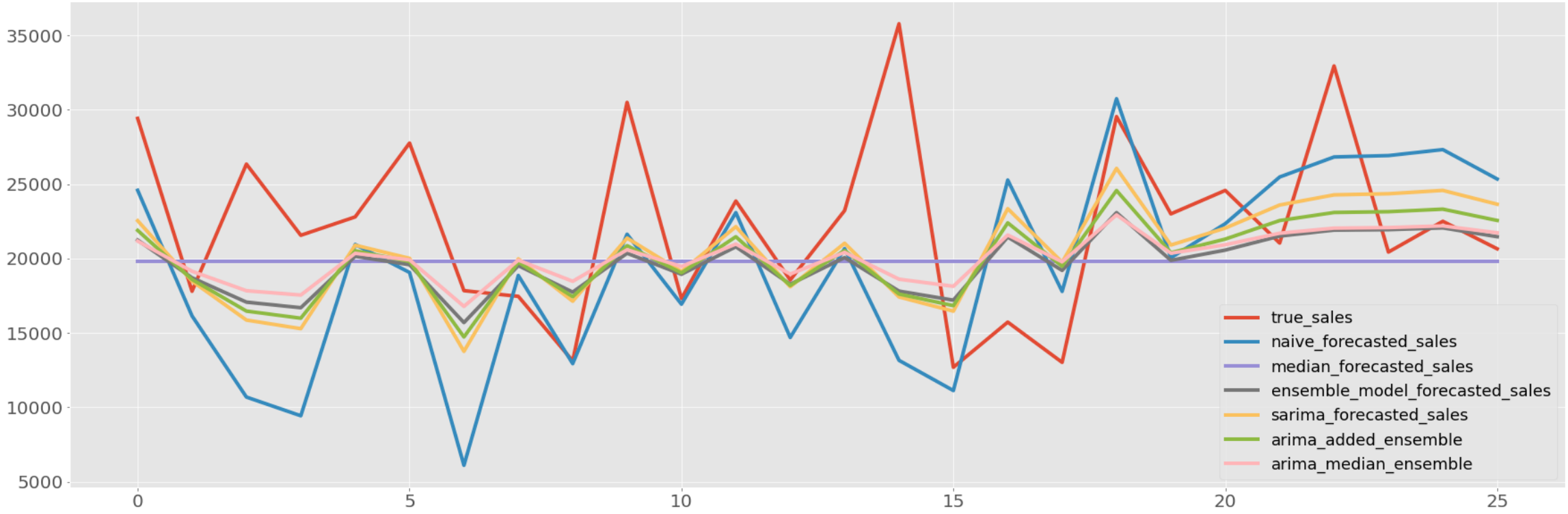
Naive + Median + SARIMA

Ensemble * 0.5

+

SARIMA * 0.5

Approaches

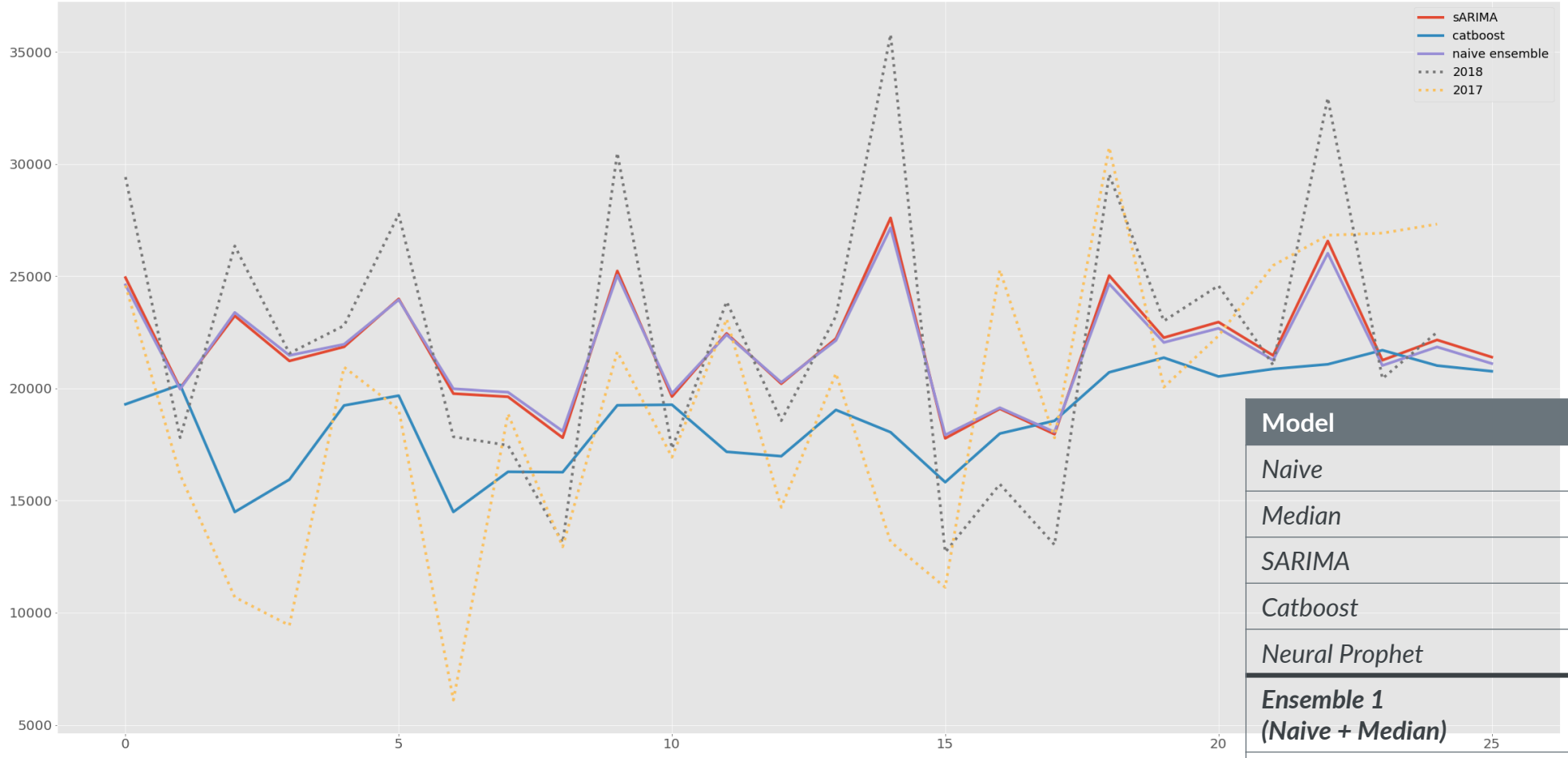


Approaches

Model	Advantages	Disadvantages
Naive	• Simple and straightforward • Easy to understand • Quick computation	• Doesn't consider trends
Median		• Very simplistic
SARIMA	• Considers seasonality	
Catboost	• Can consider additional features	• Complex hyperparamter settings
Neural Prophet	• Considers seasonalities, holidays, and trends	• More data needed • Complex hyperparamter settings

Final results

Final results



Model	MAPE
Naive	*0.245*
Median	*0.228*
SARIMA	*0.212*
Catboost	*0.21-0.22*
Neural Prophet	*0.25*
Ensemble 1 (Naive + Median)	***0.204***
Ensemble 2 (SARIMA + Median)	*0.206*
Ensemble 3 (Naive + SARIMA + Median)	*0.208*

Summary and Outlook

Lessons learned

- **Simple models** already show very good performance compared to more complex approaches
- **More data** required for more advanced methods such as Neural Prophet

Outlook

- Try out more **ensemble combinations**
- Optimize **hyperparameters** and **lags**
- More information on **missing days** required – is it really missing (= no sales) or should it be imputed?
- If more data is available, more **advanced methods** (e.g., NeuralProphet) could lead to better results
- Which **external data** could be included that may influence forecast? (stock market, Covid, competitors, etc.)

Growth
from
Knowledge

Thank you!