Demand Forecasting Data Scientist Challenge

Carlos Espeleta

Agenda

Dataset

Orders: how to clean outliers

Temperature: how to deal with missing values

Marketing spend: how to model marketing campaigns

Seasonality analysis

Cross validation setup

Prediction intervals

Model performance and cross validation predictions

Predictions on the test set

Exploratory Data Analysis

Dataset

Contains 4 variables and 790 observations

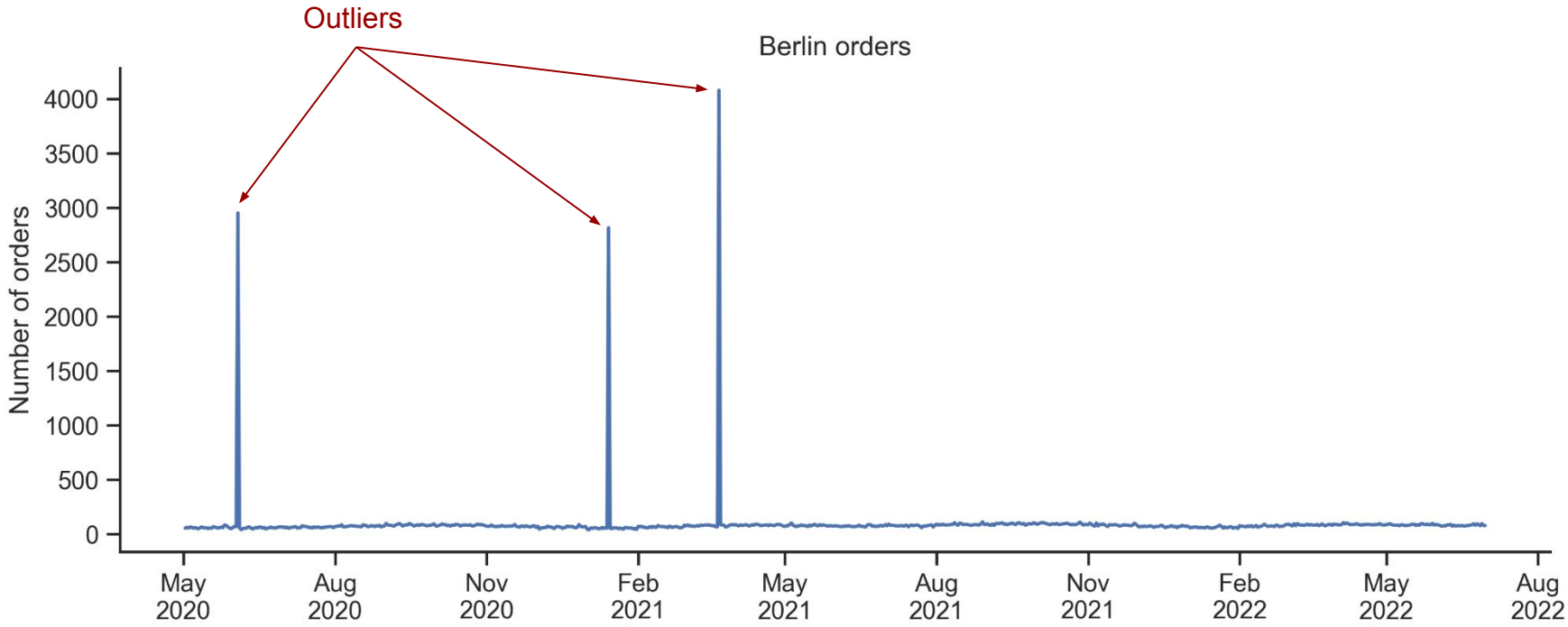
First observations

- **Date**: daily frequency, from May 2020 to June 2022
- **Orders**: positive values, with extreme values
- **Temperature**: there are negative and positive values. Contains missing values.
- **Marketing**: this variable contains many zeros

	date	orders	temperature	media_spend
count	790	790.00	769.00	790.00
mean	2021-05-31 12:00:00	89.06	17.36	0.50
min	2020-05-02 00:00:00	40.00	-9.98	0.00
25%	2020-11-15 06:00:00	67.25	11.00	0.00
50%	2021-05-31 12:00:00	78.00	18.29	0.00
75%	2021-12-14 18:00:00	86.00	24.02	0.00
max	2022-06-30 00:00:00	4080.00	37.95	14.99
std	nan	200.77	9.88	2.67

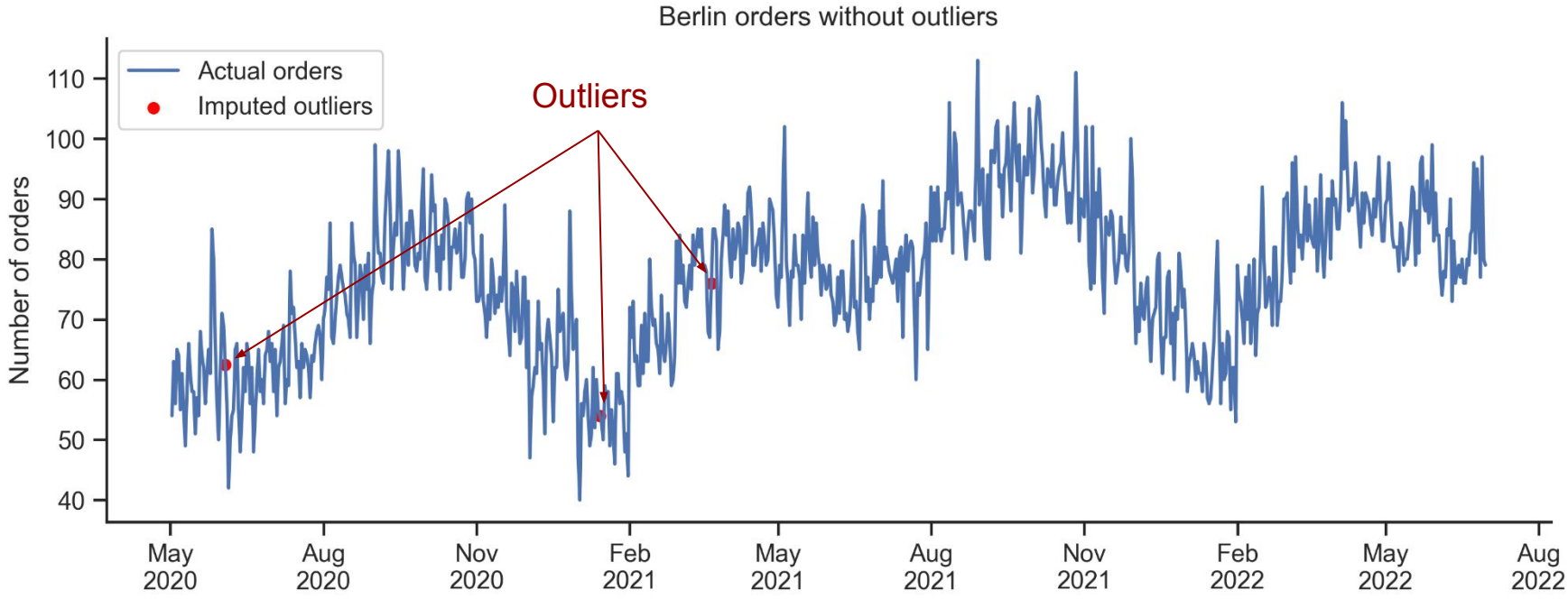
Orders

Contain 3 extreme points



Orders

After removing outliers



Outliers identification and correction

This is a two step approach, first identify outliers and then treat them

Understand why there are outliers:

- Is it because of a problem in our ETL?
- Is it because of a marketing campaign? In this case the impact seems to be to high.
- Is it because of an special event?
- Is it because a direct competitor of Wolt had problems in Berlin and Wolt received more orders than expected?

The approach followed in this exercise:

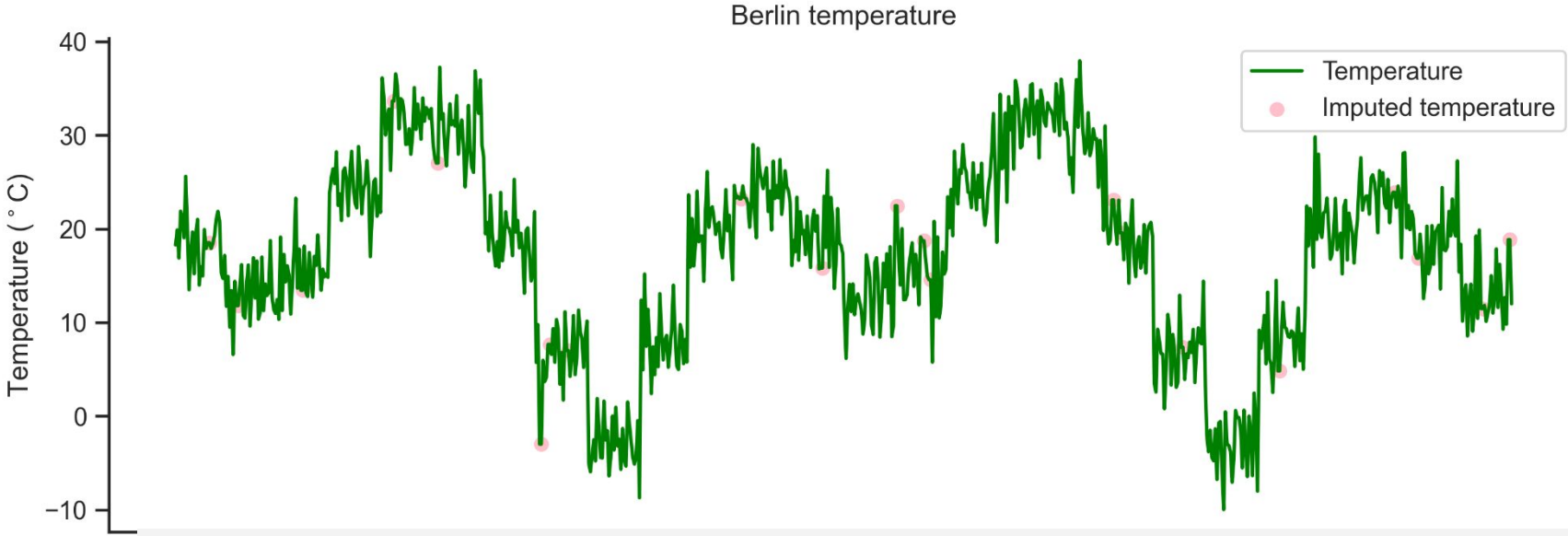
- Manual identification of the outliers
- Correct outliers with a linear interpolation

Alternative approaches:

- Rolling standard deviation, quantiles
- Time series decomposition

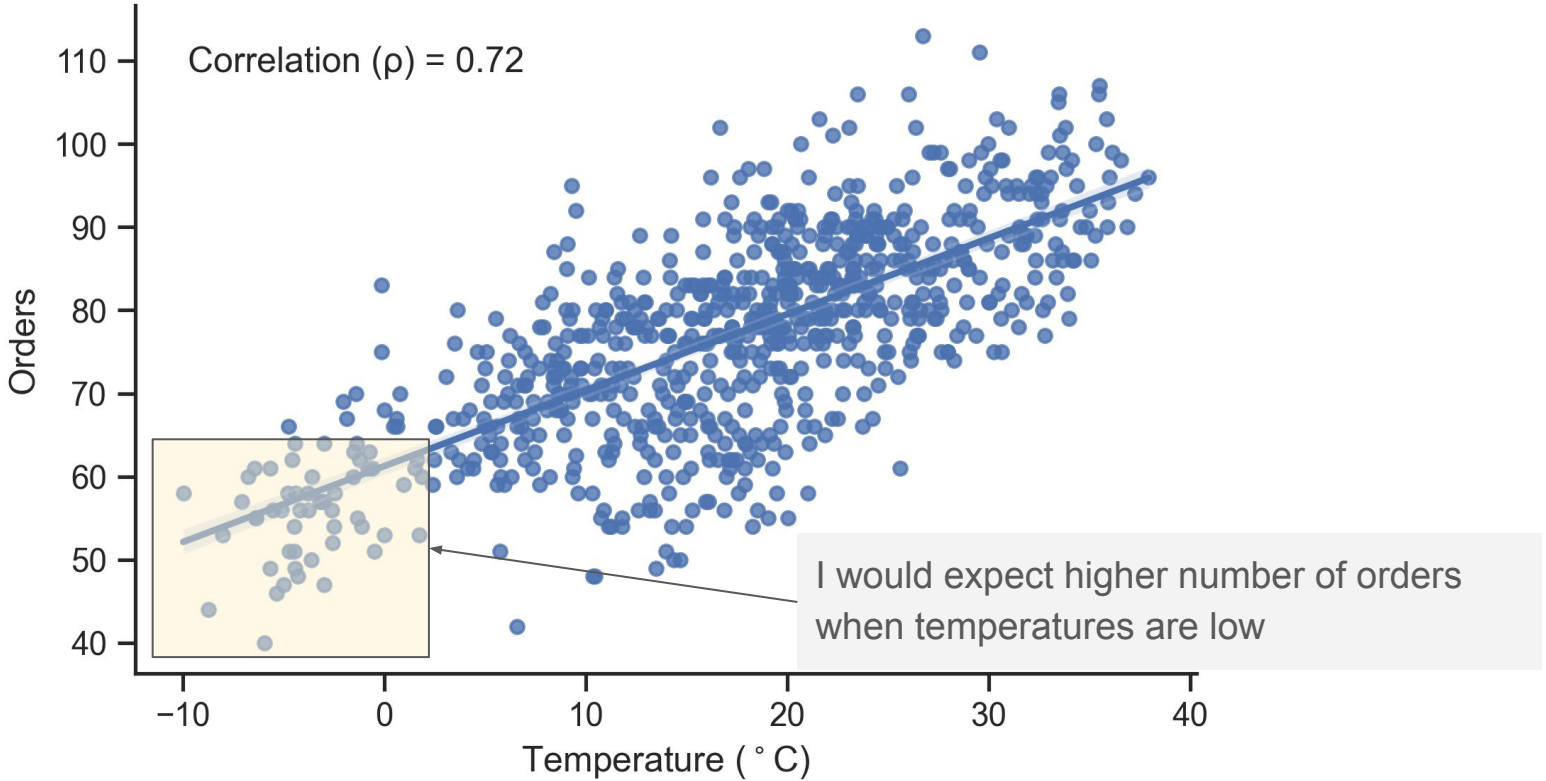
Temperature with missing values imputed

Missing values have been imputed using fill-forward method



The average temperature of a city has a gradual transition, that is, it does not change abruptly from one day to the next. For that reason, missing values have been imputed using previous day's value.

Correlation between orders and temperature

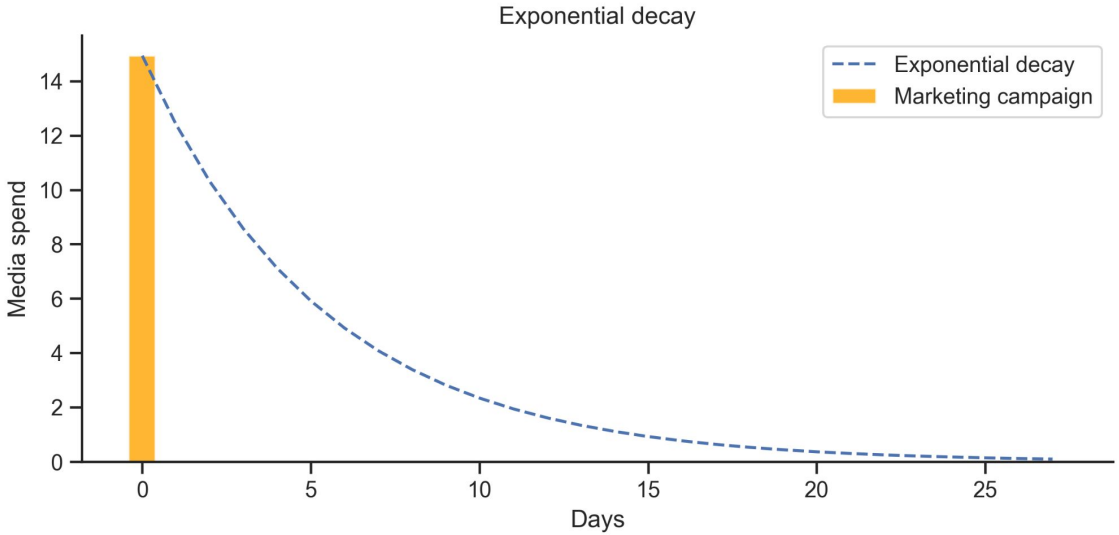


Marketing

Modeling marketing events as exponential decay

Things to consider when modeling a marketing campaign:

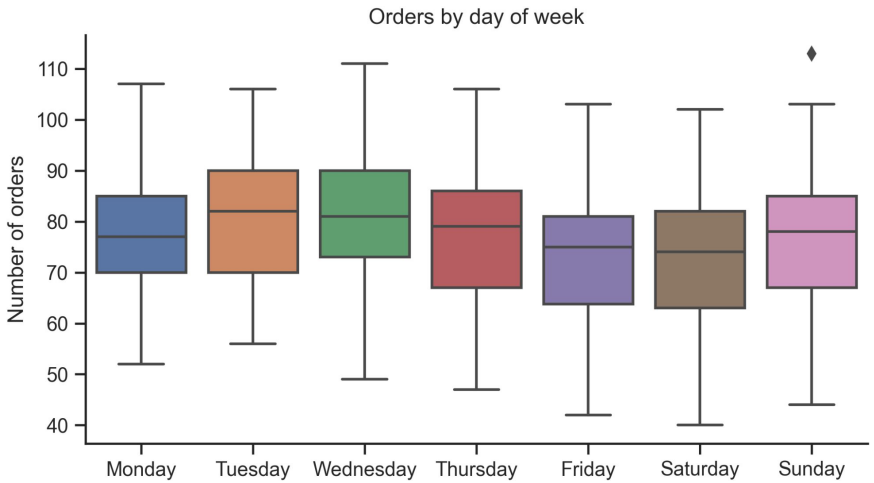
- There is a time lag between the campaign and its effect.
- The effect of the campaign lasts over time



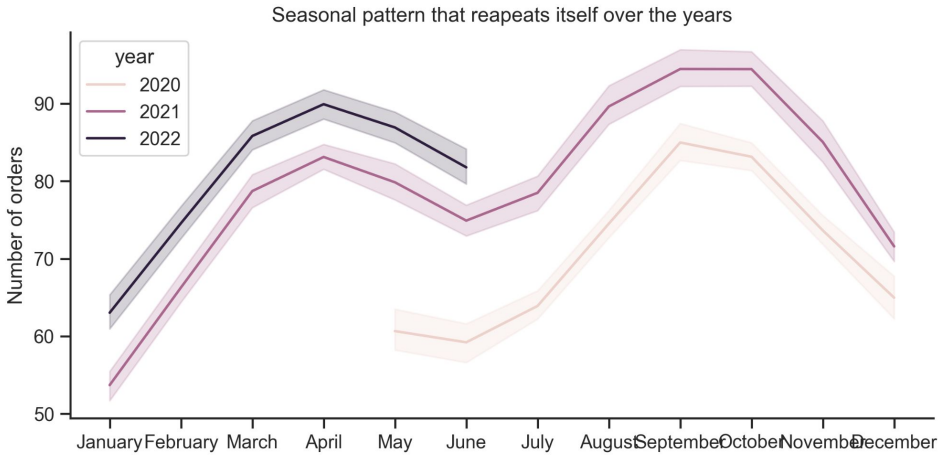
Seasonality

Patterns that are repeated over time

Weekly seasonality



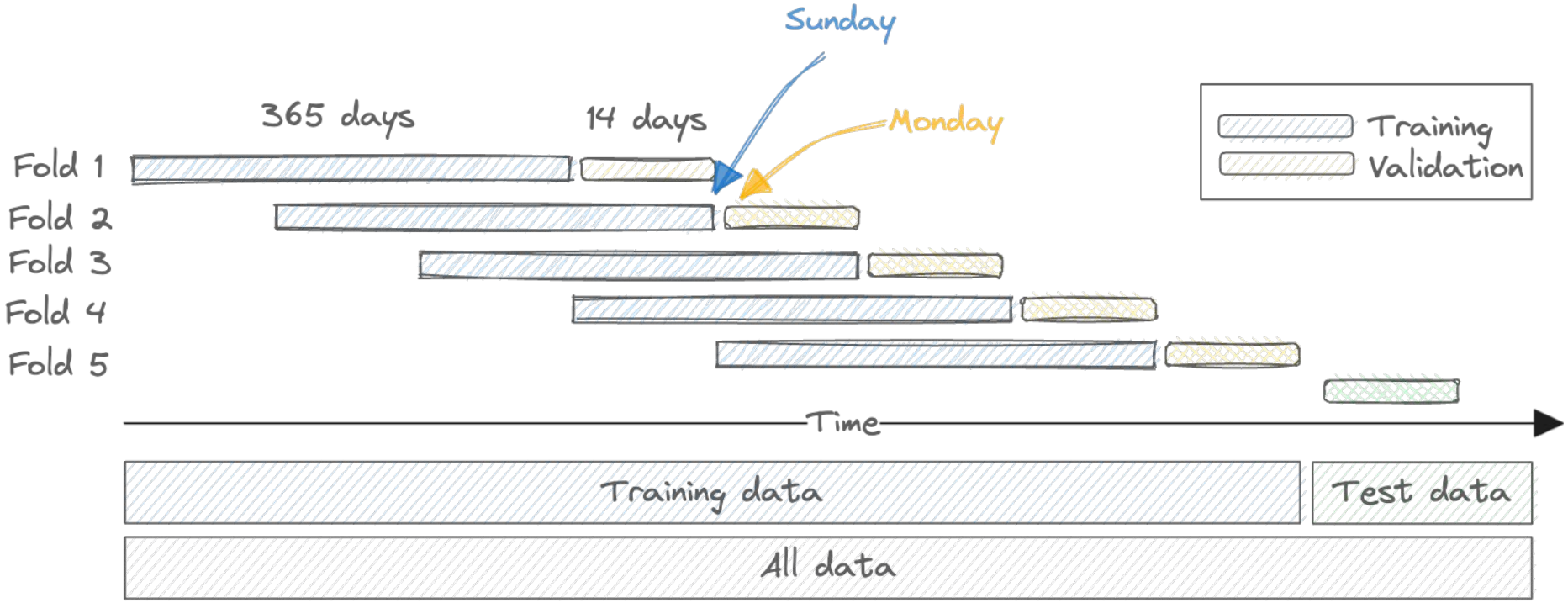
Monthly seasonality



Modeling

Time series split and cross validation setup

Data is splitted according to time



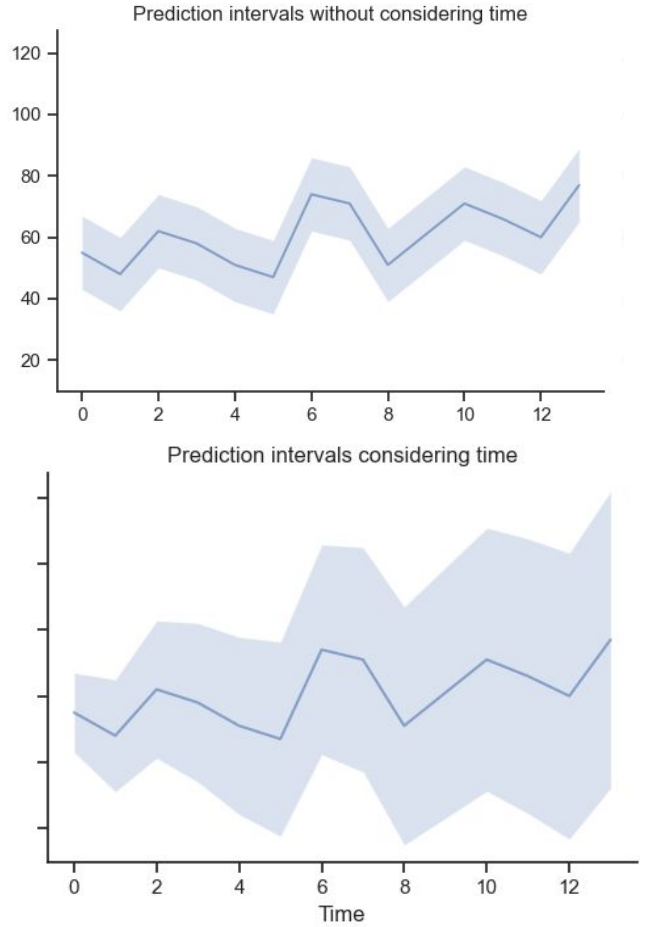
Prediction intervals

Provide uncertainty of point estimates

Process to generate intervals:

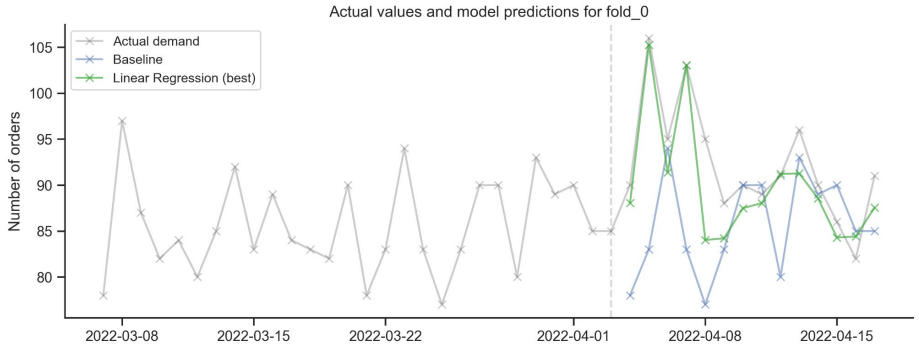
1. Model cross validation
2. Make out-of-fold predictions
3. Compute residuals
4. Bootstrap residuals with replacement
5. Calculate median of the samples
6. Calculate standard deviation of medians
7. Include time dependent component as:

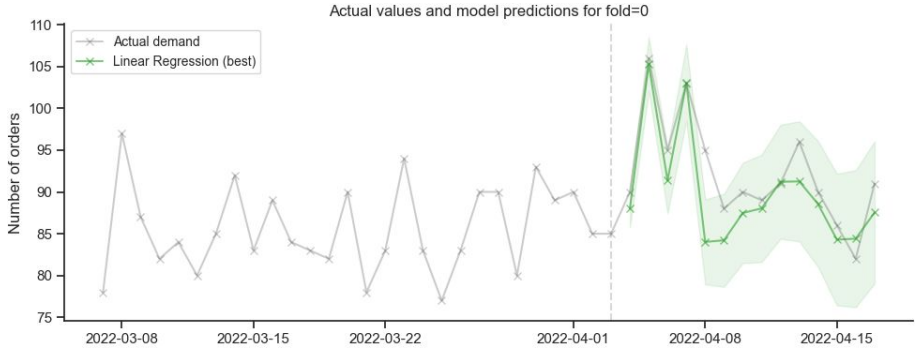
$$y\_pred +/- std * sqrt(t)$$

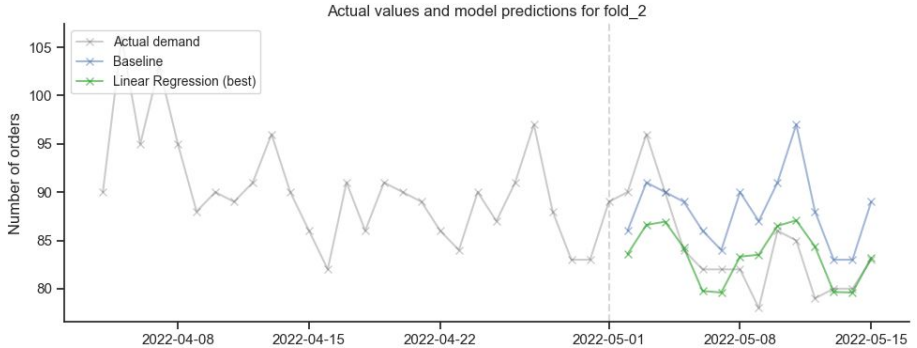


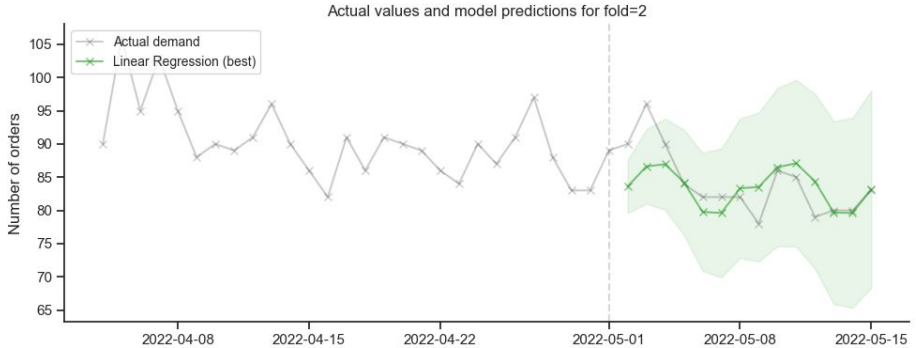
Linear regression forecast across different folds

Point forecast on the left and prediction intervals on the right



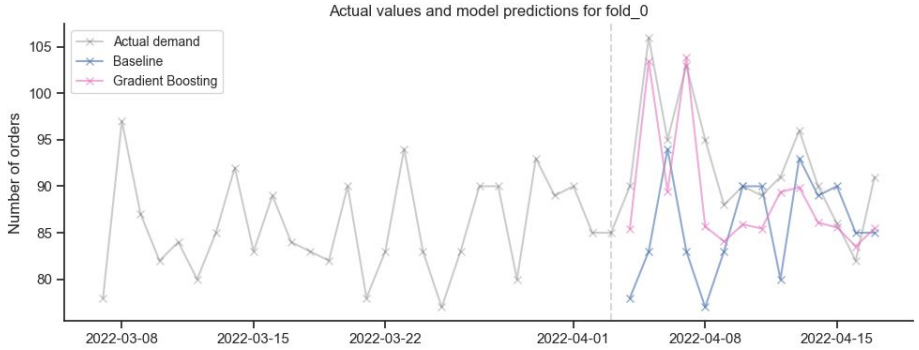


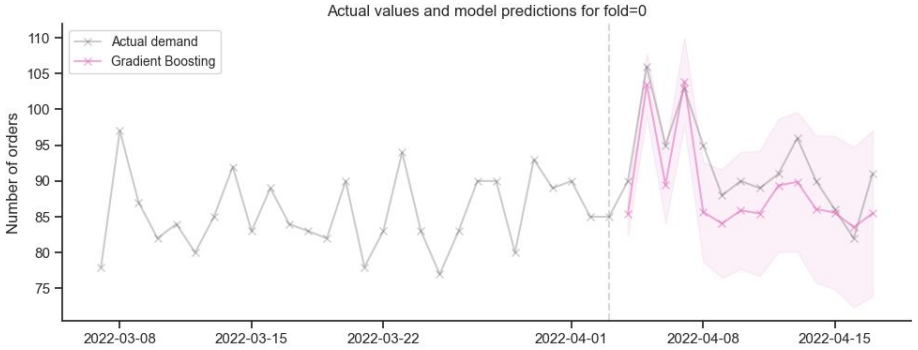


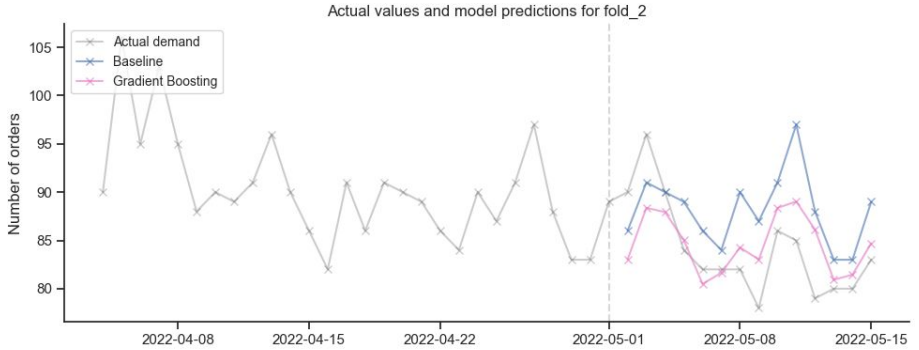


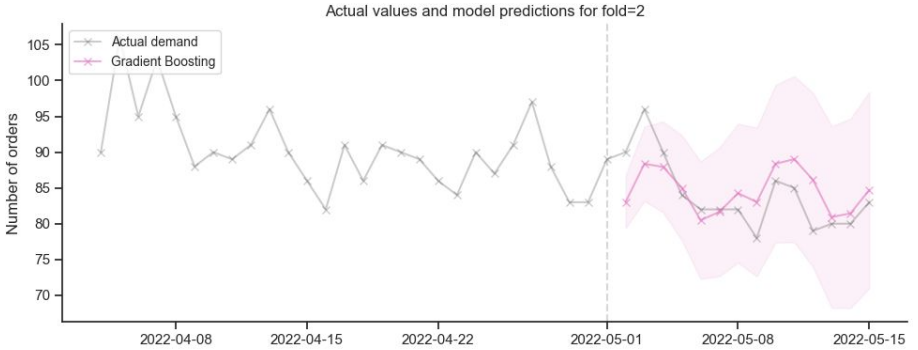
Gradient boosting forecast across different folds

Point forecast on the left and prediction intervals on the right









Model performance

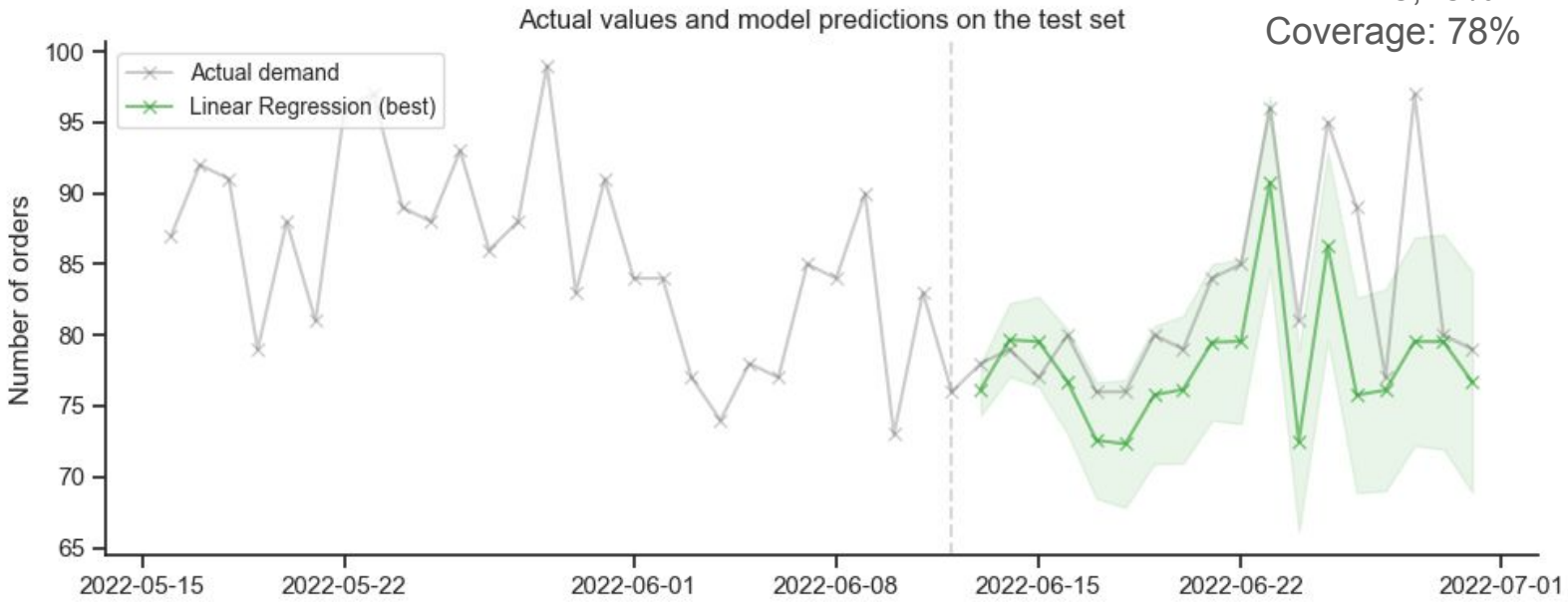
Summary performance of the cross validation results

	RMSE		MAPE		Bias		Coverage	
Model	**Train**	**Valid**	**Train**	**Valid**	**Train**	**Valid**	**Train**	**Valid**
Benchmark	9.66	8.77	9.30%	7.78%	-0.81	0.07	-	0.93
LR - Integers	7,70	6,44	7,57%	5,93%	0,78	-0,09	-	0,91
LR - Ohe	5,70	5,45	5,22%	4,71%	0,44	-0,16	-	0,91
LR - Temp + Marketing	4,82	**4,62**	4,62%	**3,96%**	0,35	**-0,14**	-	**0,94**
Gradient Boosting	4,41	4,68	4,35%	4,19%	0,45	-0,15	-	0,93

Test set evaluation

Predictions on the test set and performance metrics

RMSE: 6,6
MAPE: 5,73%
Coverage: 78%



Next steps and future work

- Include in the EDA analysis of the ACF
- Decompose the serie into trend, seasonality and residuals
  - Does any of the components correlate with temperature or marketing?
- Explore other methods to correct outliers
- Model seasonality with trigonometric features
- Model marketing as suggested in the analysis
- Include interactions in the linear regression model
- Test other models
- Fine tune parameters of the gradient boosting model
- Analyze coefficients of the model to gain interpretability
- Etc.

Thank you

Questions?