DATA SCIENCE SUMMIT

sponsor diamentowy: accenture

SAMSUNG

sponsorzy srebrni: lingaro Nethone VERTICA

XGBOOST AS A TIME-SERIES FORECASTING TOOL

Filip Wójcik
Objectivity Digital Transformation Specialists
PhD Student on a Wroclaw University of Economics
filip.wojcik@outlook.com

Agenda

01 What is XGboost

02 Rossmann – legendary competition

03 Forecasting – training challenges

04 Data preparation

05 Training method

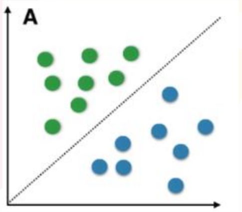
06 Comparison to other models

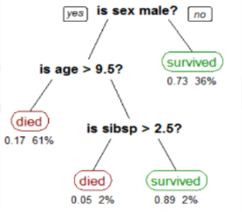
07 Results analysis

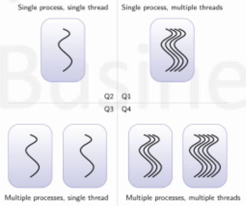
08 Other data sources

09 Summary

What is XGBoost







Classification & regression algorithm

- Primarily intended for classification and regression
- Optimizes many different scoring and error functions
- Supports target functions:
  - Softmax
  - Logit
  - Linear
  - Poisson
  - Gamma

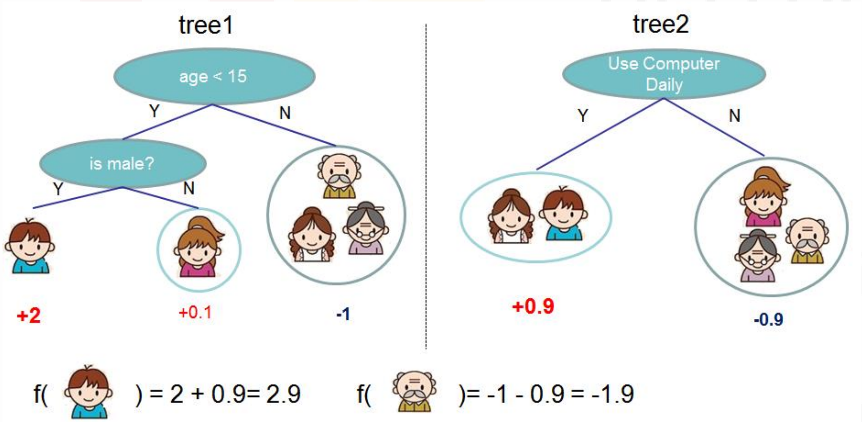
Based on trees

- Base estimators are decision trees
- Composite algorithm – an ensemble
- Booting type algorithm – increasing weight of harder examples
- Each tree improves previous one

Good parallelization

- Trees are dependent on each other
- Parallelization occurs at the level of a single tree – for building successive nodes
- XGBoost uses compressed data format to save memory

What is XGBoost



$l(x_1, x_2)$ – cost function

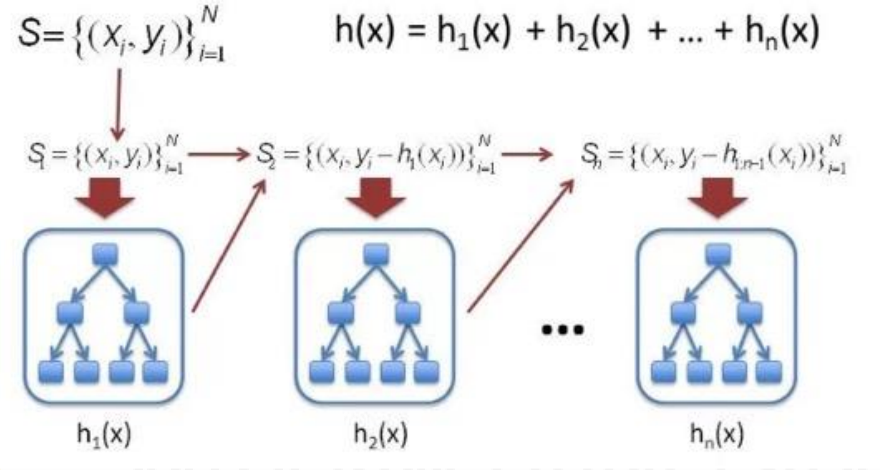
$f_i(x)$ – i – th tree builing function

$$\begin{aligned}
\hat{y}_i^{(0)} &= 0 \\
\hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\
\hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\
&\dots \\
\hat{y}_i^{(t)} &= \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)
\end{aligned}$$

$$\begin{aligned}
\text{obj}^{(t)} &= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i) \\
&= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + constant
\end{aligned}$$

Source: http://xgboost.readthedocs.io/en/latest/model.html#tree-ensemble

What is XGBoost?



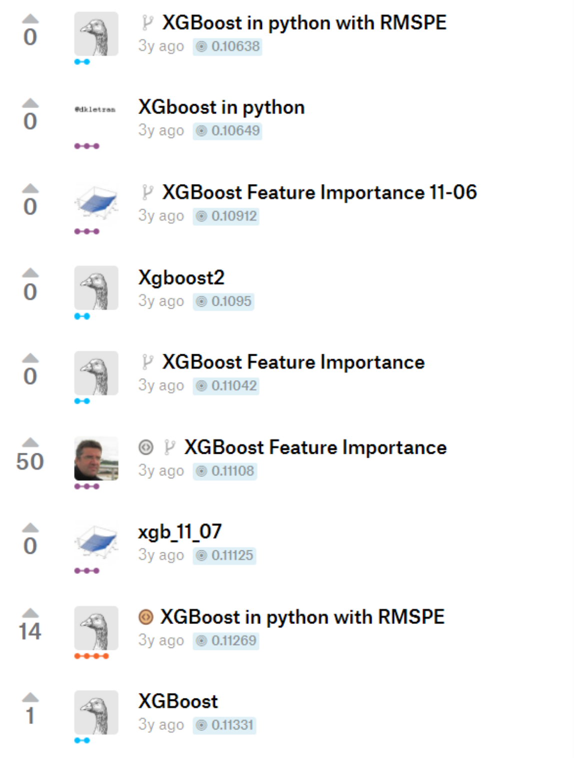
Source: http://xgboost.readthedocs.io/en/latest/model.html#tree-ensemble

Rossmann – legendary competition

- Kaggle.com contest started in 2016r – the goal was to forecast turnover based on historical values and data from macrosurroundings of stores
- 3303 teams took part
- A significant majority of kernels implemented the Xgboost algorithm
- Evaluation Metric – RMSPE Root mean squared percentage error

$$\mathrm{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2}$$

- Best score – approx. 10% error
- Typical problem of forecasting – not classification or regression

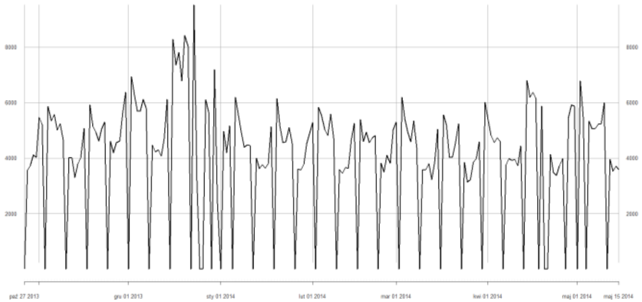


Forecasting – training challenges

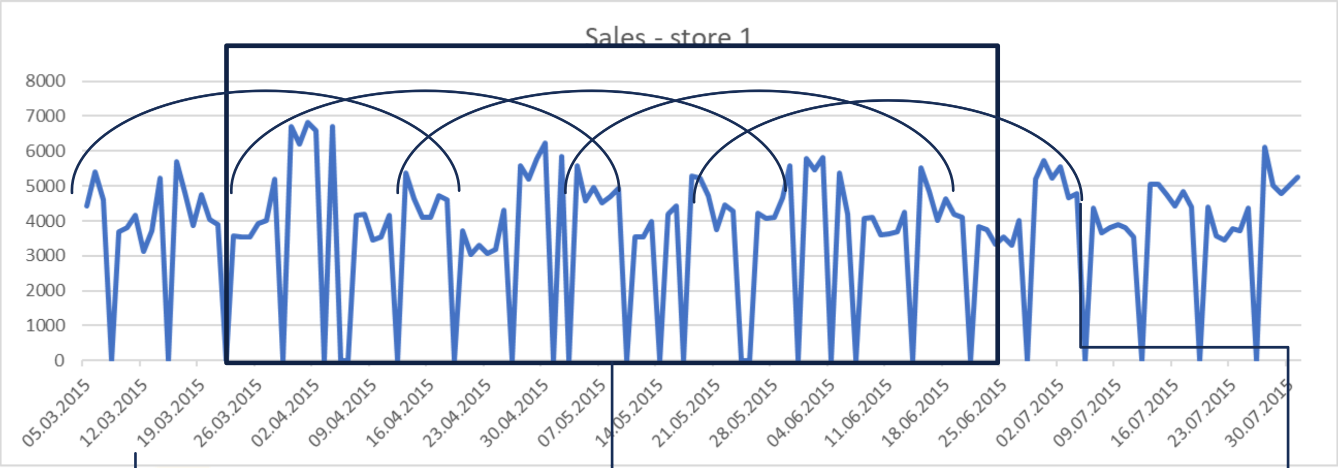
- Two types of variables
  - **Static** – describing each store characteristics
  - **Time series**– turnover and customer count
- Challenge in training time <u>series forecasting model</u>
  - Standard cross-validation does not work
  - Random selection ruins chronology and order
  - *OOB error* is not the best estimator anymore
- Huge prediction variance between subsequent runs
- <u>Maintaining order and chronology is crucial to teach the model</u>
  - Seasonal trends
  - Autocorrelation

Store	DayOfWeek	Date	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday
1	5	2015-07-31	5263	555	1	1	0	1
2	5	2015-07-31	6064	625	1	1	0	1
3	5	2015-07-31	8314	821	1	1	0	1
4	5	2015-07-31	13995	1498	1	1	0	1
5	5	2015-07-31	4822	559	1	1	0	1
6	5	2015-07-31	5651	589	1	1	0	1
7	5	2015-07-31	15344	1414	1	1	0	1
8	5	2015-07-31	8492	833	1	1	0	1
9	5	2015-07-31	8565	687	1	1	0	1
10	5	2015-07-31	7185	681	1	1	0	1
11	5	2015-07-31	10457	1236	1	1	0	1
12	5	2015-07-31	8959	962	1	1	0	1
13	5	2015-07-31	8821	568	1	1	0	0

$\neq$



Data preparation



Day: 19, month: 3, year: 2015, Q: 1

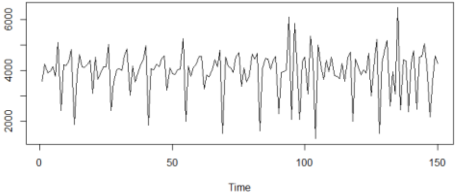
Numerical time indicators

- Number representing Day-Of-week
- Month Number
- Quarter number

$$\mu_{q2} = 3835, \sigma_{q2} = 1805$$

Seasonal indices

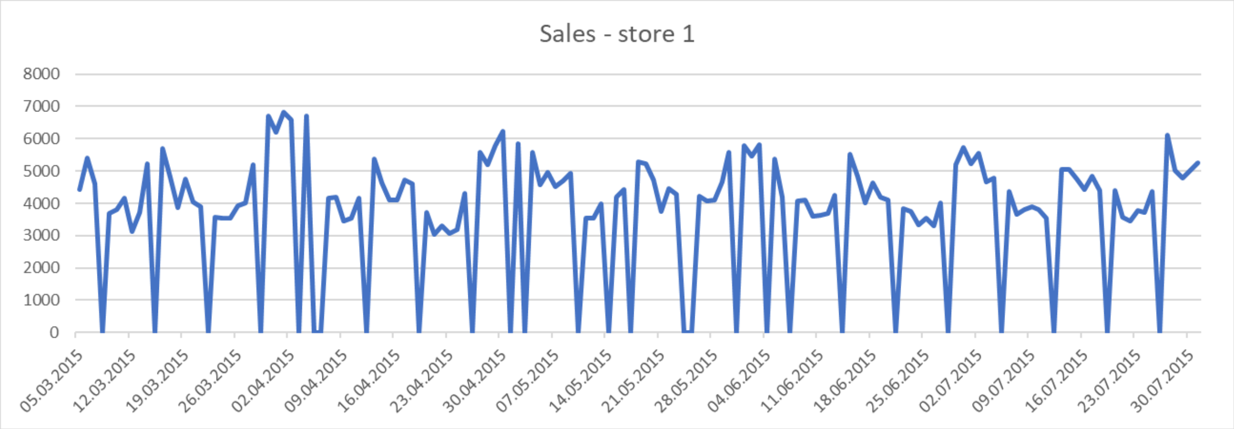
- Seasonal means and standard deviations
- Calculated per day-of-week/week/month/quarter



Additional seasonal indices:

- Moving averages
- Different orders – last day/5-days/2 weeks/etc.

Data preparation



Time - indicators

Static variables

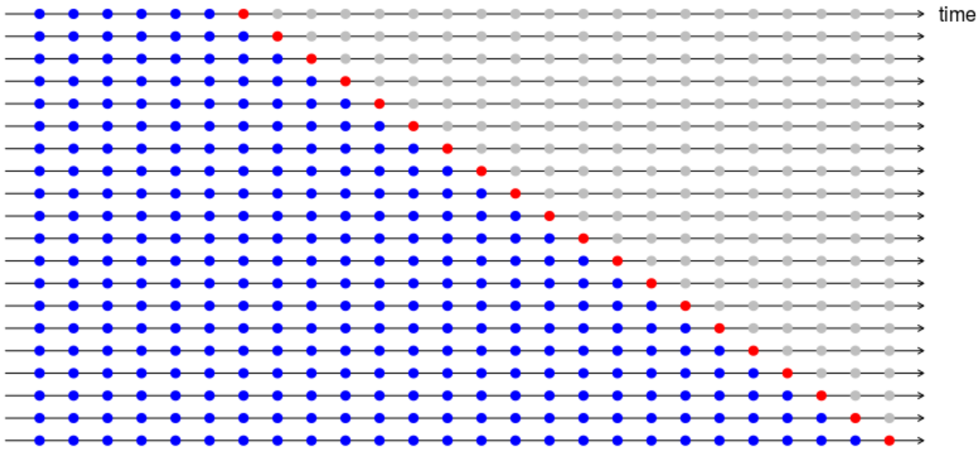
y

Seasonal indices

Shop	Date	Day of week	Month	Quarter	Shop type	Promo	Assortment	Sales	$\mu_{sales}^{day\ of\ week}$	μ_{sales}^{month}	$\mu_{sales}^{quarter}$
1	10-01-2016	7	1	1	A	True	B	1536	954	1100	950
2	15-10-2016	6	10	4	C	False	A	764	1005	1256	1954

Training method

- Classic **cross-validation does not work** due to variance changing in time
- **Methodology characteristic for forecasting models** (like ARIMA) was used:
    - Gradually move prediction window and training data
    - Keep order
    - Move one-time-chunk at a time
- Model was trained on larger and larger data, and predicting **one-step ahead**
- Additionally – using **classic XGB metrics** like OOB score



Source: https://robjhyndman.com/hyndsight/tscv/

Comparison to other models

The goal of study and assumptions

Research question

Comparative study of different forecasting methods using exogenous (static) variables

Statistical comparison of prediction quality

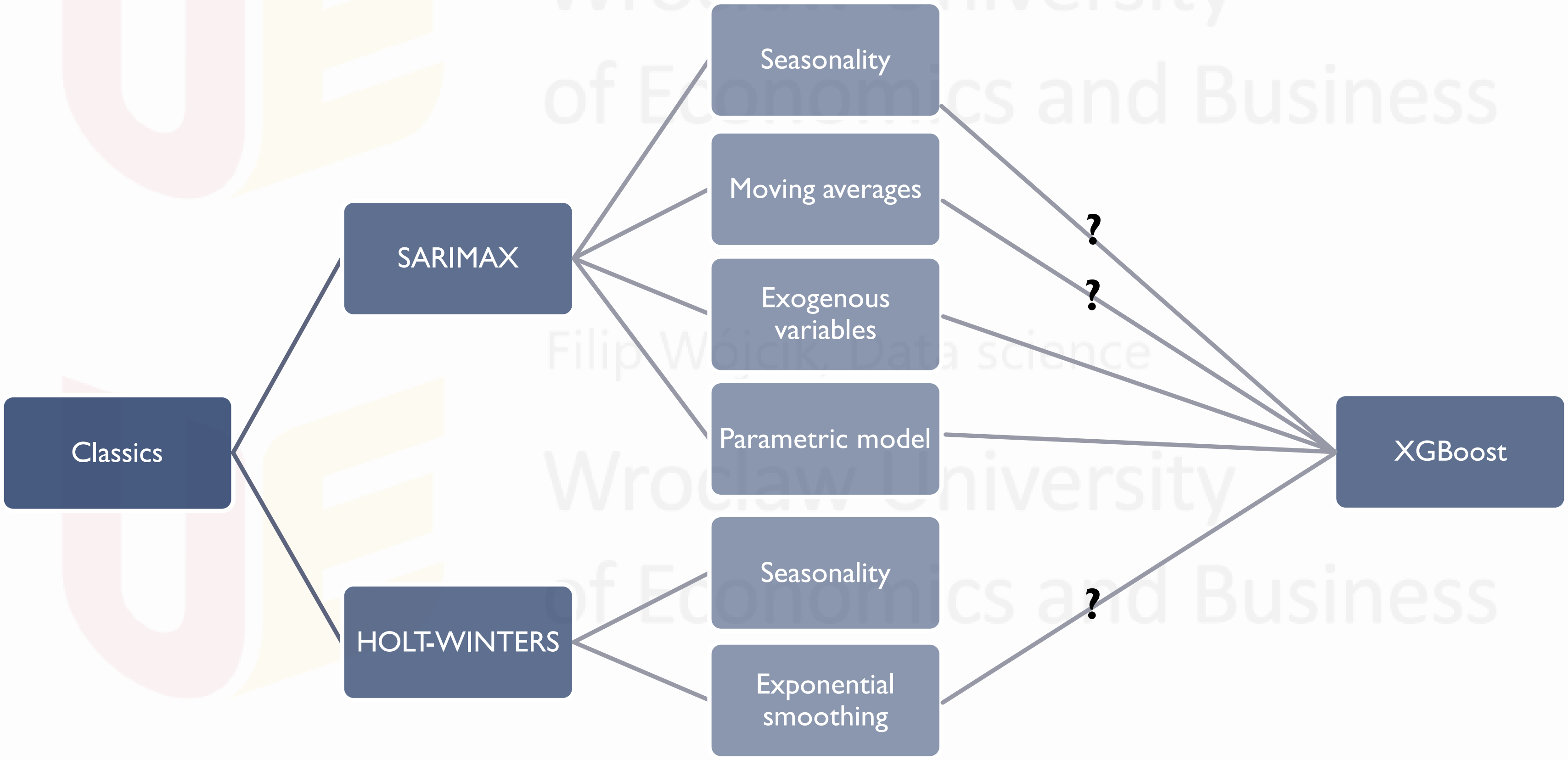
- Systemathic comparison of forecasts
- Is there a statistically SIGNIFICANT difference?

Coefficient importance check

- In case of „classic" models – parameters interpretaion
- In case of XGBoost – feature importance calculation

Comparison to other models

SELECTION OF CLASSIC MODELS



Comparison to other models

TRAINING METHODS

CLASSIC MODELS

One model per store

For each store – separate model was trained

Automatic params tuning

Params for each model were selected automatically using optimization techniques (AIC, BIC, RMSPE). Random sample was manually cross-checked

Missing values interpolation

In case of missing values – polynomial interpolation was used

XGBOOST

One model for full dataset

Experimental study indicated that seasonal indices + exogenous variables are enough for model to generalize. One model for full dataset is enough

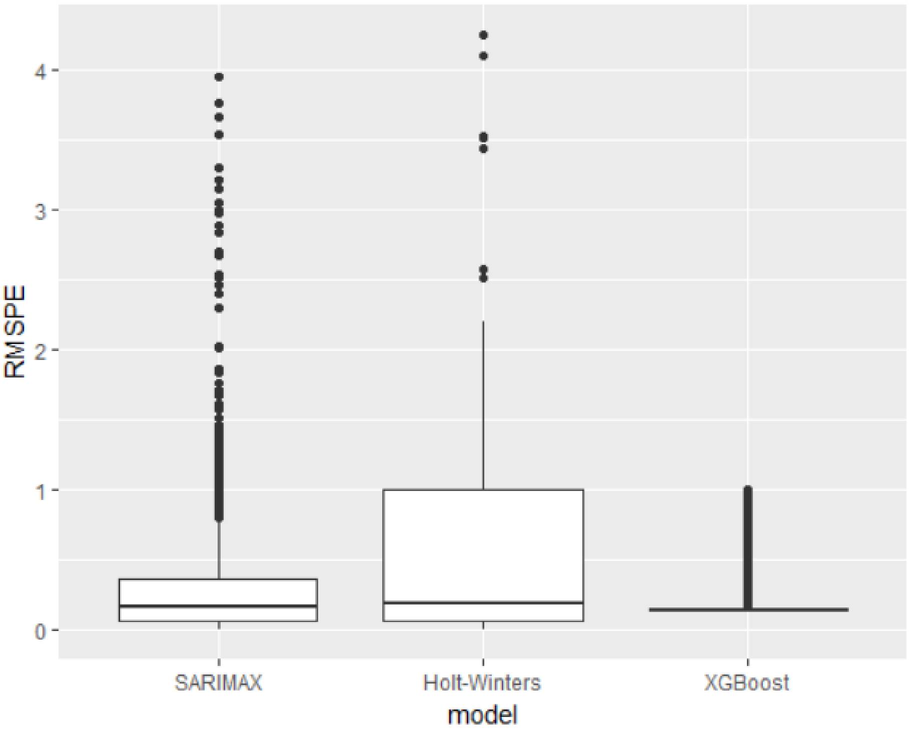
Time-series validation

One-step Ahead validation technique was used, enriched with 1000 last observations from ordered dataset

Regression trees

Base estimators were regression trees and RMSPE – error function

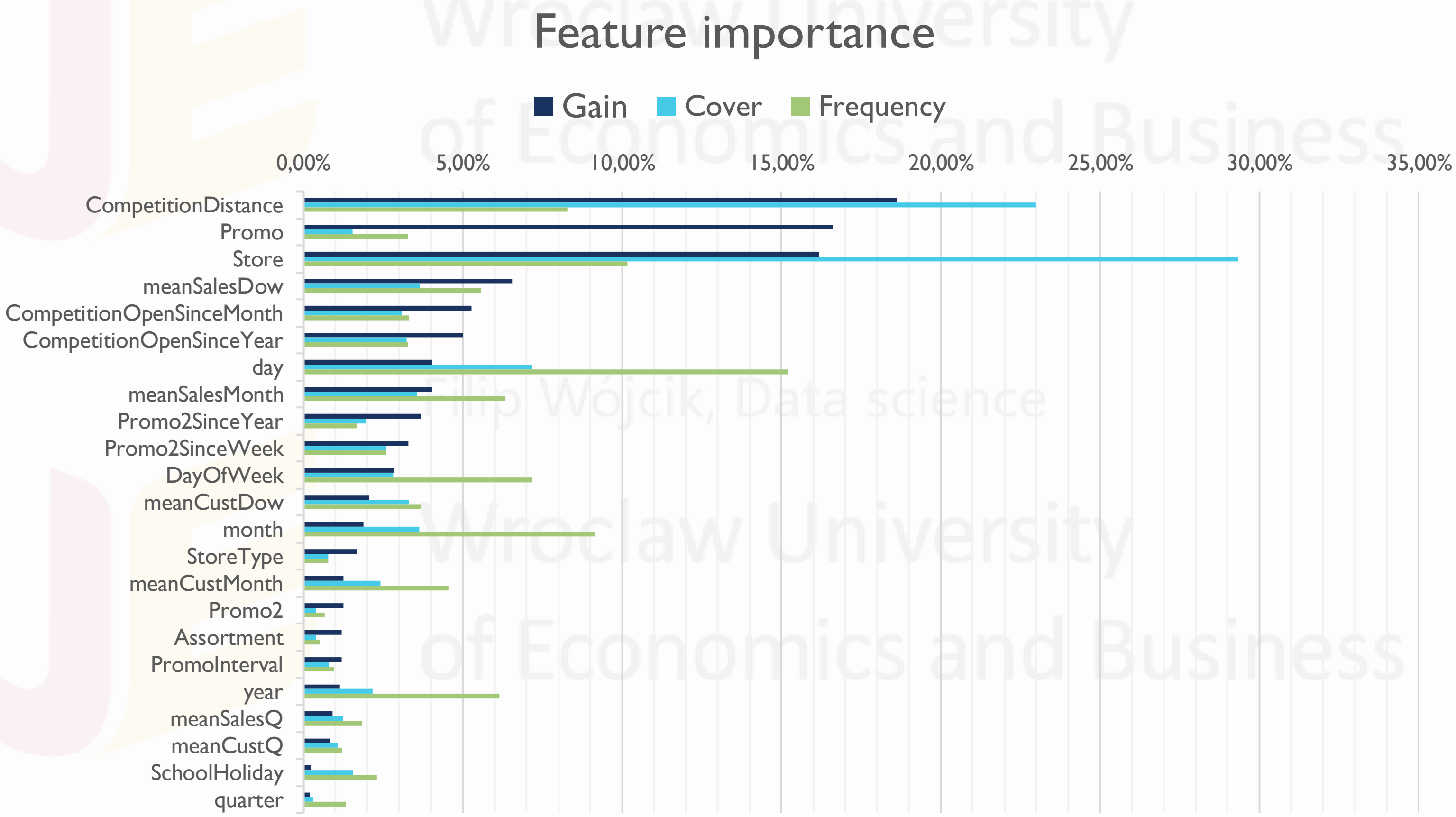
Results analysis



Metric (median)	SARIMAX	Holt-Winters	XGBoost
Theil's coefficient	0.061	0.059	0.1364
R^2	0.838	0.54	0.92
RMSPE (valid.)	0.17	0.18	0.13
RMSPE (leaderboard)	0.16	0.367	0.121

Models	RMSPE diff	Confidence from	Confidence to	P-val
XGBoost - SARIMAX	-0.126	-0.141	-0.111	<< 0.01
XGBoost - Holt-Winters	-0.218	-0.235	-0.200	<< 0.01

Results analysis



Results analysis

01 XGBoost – better results

Values for all metrics are better for the XGBoost algorithm.

02 Lower variance

The predictions of the XGBoost are more stable, compared to the rest of models, with much less variance

03 Lower training time

Training one model globally, for all stores, takes much less time than training 1115 SARIMAX or Holt-Winters models

04 Feature importance – time indicators

Among the first 15 most important attributes, there are time indicators– day, month, year. XGBoost was able to identify the impact of seasonality

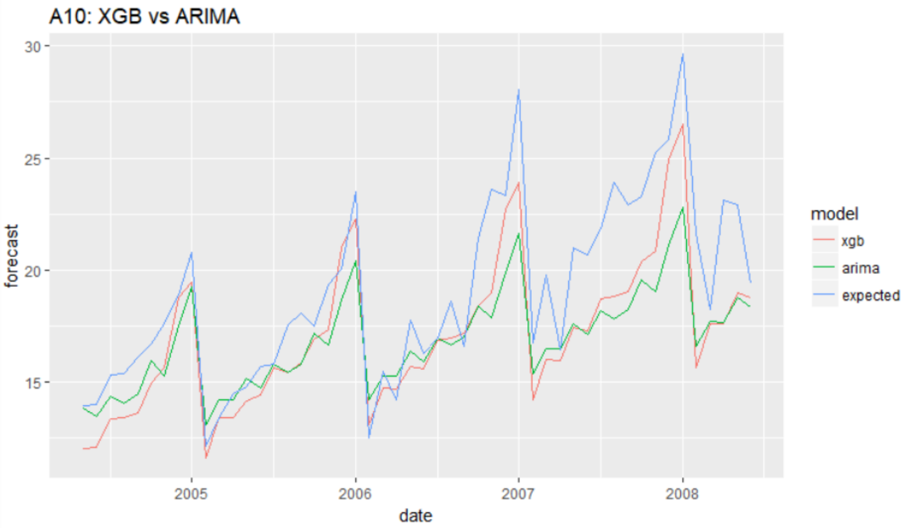
05 Feature importance – seasonal indices

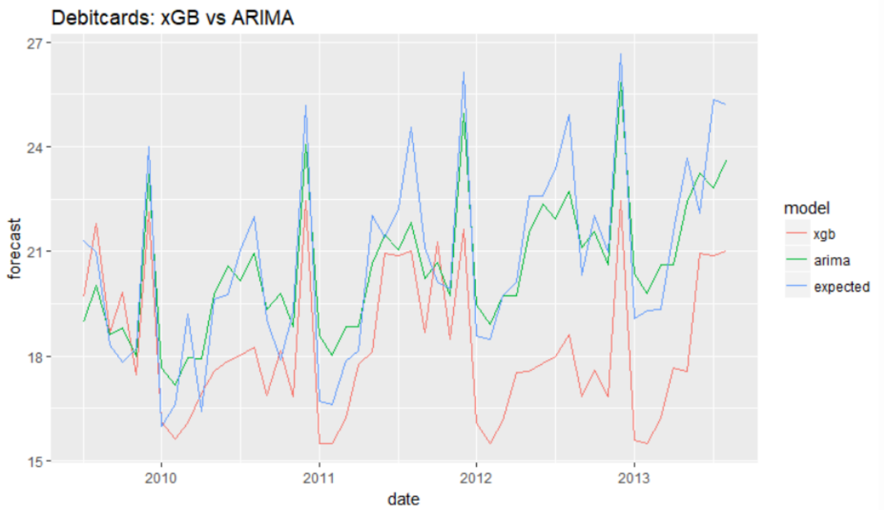
Among the first 15 key attributes, seasonal indices, such as average sales on the day of the week or month, have been identified as important.

06 Feature importance – static variables

Among the first 15 most important attributes, static variables/exogenous describing the characteristics of shops were correctly identified

Other datasets



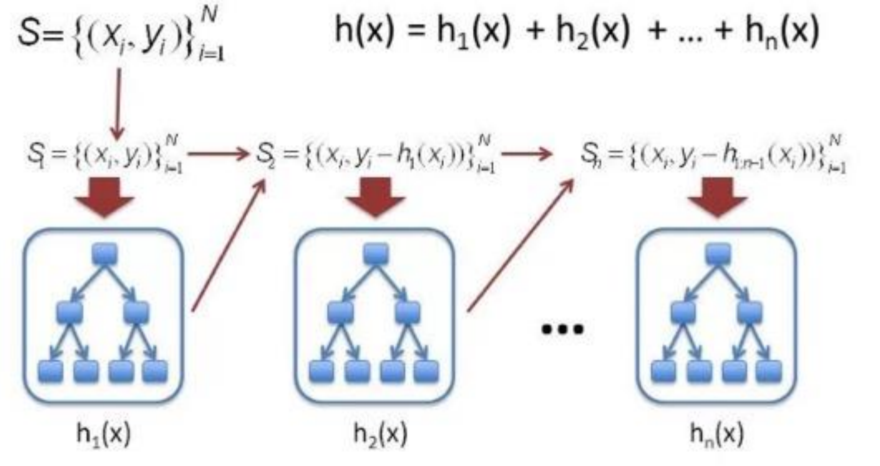


	RMSE	MAE	RMSPE
XGB	0.146	0.123	1.321
ARIMA	0.330	0.206	2.654

	RMSE	MAE	RMSPE
XGB	0.002	0.002	0.01
ARIMA	0.386	0.318	0.391

Summary

The initial results of the study seem to indicate that XGBoost is well suited as a tool for forecasting, both in typical time series and in mixed-character data.



Low variance

On all data sets tested, XGBoost predictions have low variance and are stable.

Trend and seasonality identification

The Model is able to recognize trends and seasonal fluctuations, and the significance of these attributes is confirmed by manual analysis.

Exogenous variables handling

The Model can simultaneously handle variables of the nature of time indexes and static exogenous variables

Papers

Chen, T. & Guestrin, C. (2016) „**XGBoost: A scalable tree boosting system**", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. New York, USA: ACM Press, p. 785–794

Ghosh, R. & Purkayastha, P. (2017) „**Forecasting profitability in equity trades using random forest, support vector machine and xgboost**", in *10th International Conference on Recent Trades in Engineering Science and Management*, p. 473–486.

Gumus, M. & Kiran, M. S. (2017) „**Crude oil price forecasting using XGBoost**", in *2017 International Conference on Computer Science and Engineering (UBMK)*. IEEE, p. 1100–1103.

Gurnani, M. *et al*. (2017) „**Forecasting of sales by using fusion of machine learning techniques**", in *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*. IEEE, p. 93–101

THANK YOU FOR ATTENTION