Accuracy comparison between Sparse Autoregressive and XGBoost models for high-dimensional product sales forecasting

A master thesis by Blai Ras in collaboration with Accenture

UNIVERSITAT DE BARCELONA

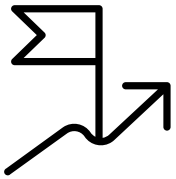
accenture

Index

Project

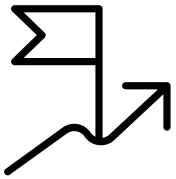

Internship with Accenture
Start of the project
End of the internship
End of the project
November 2020
March 2021
July 2021
September 2021

Goals



1. Proving that inter-product relationships may affect a forecast accuracy.

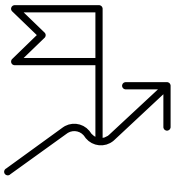
Goals



1. Proving that inter-product relationships may affect a forecast accuracy.

2. Background research of time series analysis.

Goals

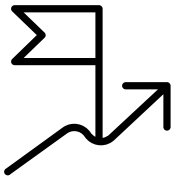


1. Proving that inter-product relationships may affect a forecast accuracy.

2. Background research of time series analysis.

3. Understanding of Vector Autoregressive models.

Goals



1. Proving that inter-product relationships may affect a forecast accuracy.

2. Background research of time series analysis.

3. Understanding of Vector Autoregressive models.

4. Design a framework of model building, training and comparison following ML principles.

Sales forecasting 

What?

Predicting future sales using historical data.

Why?

Resource allocation.

Budgeting.

Informed decision-making.

Dataset

- Weekly sales records from a home improvement, gardening, and workshop retailer.
- Observed during 2 years and 11 months (**153 weeks**).
- **4370 different products**, divided in **23 segments**. On average, a segment has **113 products**.
- We have information about applied promotions for each product.
- Overall, ≈2 million data points.

Date	Sales_322471	Sales_322472	...	Discount_322471	Discount_322472	...	Duration_322471	Duration_322472	...
24/8/2015	86	63	...	20	0	...	7	0	...
31/8/2015	105	65	...	0	0	...	0	0	...
7/9/2015	89	64	...	0	20	...	0	7	...
...	...	...	...	...	...	...	...	...	...

Pre-processings

- **Missing values management**

  2 types of NaN:

  - At the beginning of the time serie.
  - In the middle.

Pre-processings

- **Missing values management**

  2 types of NaN:

  - At the beginning of the time serie
  - In the middle

- **Stationarity**

Pre-processings

- **Missing values management**
  - 2 types of NaN:
    - At the beginning of the time serie
    - In the middle
- **Stationarity**
- **Standardization**

Vector Autoregressive models (VAR)

• Statistical model able to capture relationships between variables as they change over time.

• Able to predict sales accordingly:

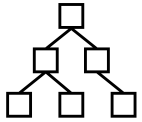
$$Y_t = \sum_{i=1}^{p} \Phi_i y_{t-i} + \sum_{j=1}^{s} B_j x_{t-j} + \varepsilon_t \,, \quad t = 1, \ldots, T \,,$$

Dimensionality problem

- Common statistical models can’t handle inputs of ≈10 or more variables.
- Not every product is related to each other!

Group-Lasso Regression

- Variable selection by forcing some coefficients to be zero.
- Easier to interpret.
- Less parameters, less bias.

XGBoost 

- Leading Machine Learning algorithm when working with tabular data.

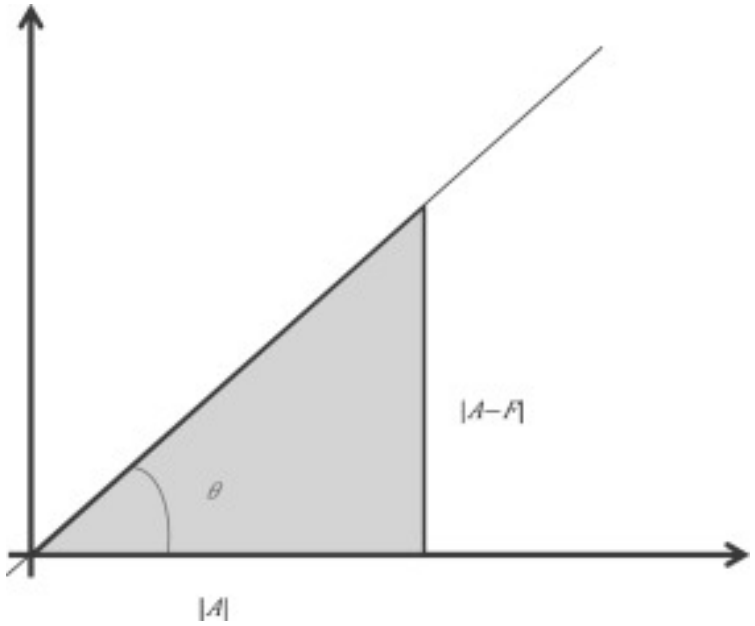
- Offers a scalable and distributed gradient boosting tree implementation.

- Is not specialized on detecting this inter-product associations.

Evaluation metrics

$$\text{RMSE} = \sqrt{\frac{\sum_{i=0}^{N} (a_i - f_i)^2}{N}}$$

$$\text{MAAPE} = \frac{100}{N}\sum_{t=1}^{N}\arctan\left(\left|\frac{a_i - f_i}{a_i}\right|\right)\%$$



Slope can be measured as a ratio of |A-F|, which can go from zero to infinity, or as an angle, ranging from 0 to 90°

Experiments

- Train & Test split.
- 1-month ahead sales prediction (4 weeks).
- 2 evaluations: standardized and at business scale.
- 2 error types: weekly and monthly.

Standardized results

RMSE

Algorithm	Period	Average	Median
VAR	Weekly	11.24	4.44
	Monthly	16.26	6.67
XGBoost	Weekly	31.93	10.68
	Monthly	75.79	24.23

MAAPE

Algorithm	Period	Average	Median
VAR	Weekly	78.95%	79.3%
	Monthly	77.95%	78.64%
XGBoost	Weekly	96.84%	95.7%
	Monthly	121.19%	122.95%

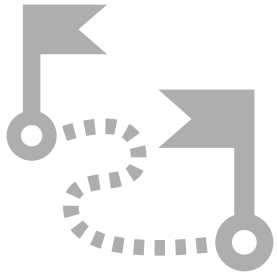
Re-escalated results

RMSE

Algorithm	Period	Average	Median
VAR	Weekly	126.00	42.01
	Monthly	333.61	109.54
XGBoost	Weekly	5257.62	626.36
	Monthly	17252.58	2276.32

MAAPE

Algorithm	Period	Average	Median
VAR	Weekly	38.69%	39.04%
	Monthly	32.00%	31.66%
XGBoost	Weekly	87.65%	85.05%
	Monthly	94.83%	92.42%



Conclusions

Product relationships affect a forecast accuracy.

Modified VAR models are able to detect these associations and predict accordingly.

Update

There’s a huge growing interest for sales forecasting taking in consideration product relationships.

Accenture is developing cannibalization assessment software partly based on my work.

Future work

- Assess different kinds of pre-processings, such as logarithmic transformations or polynomial interpolations.

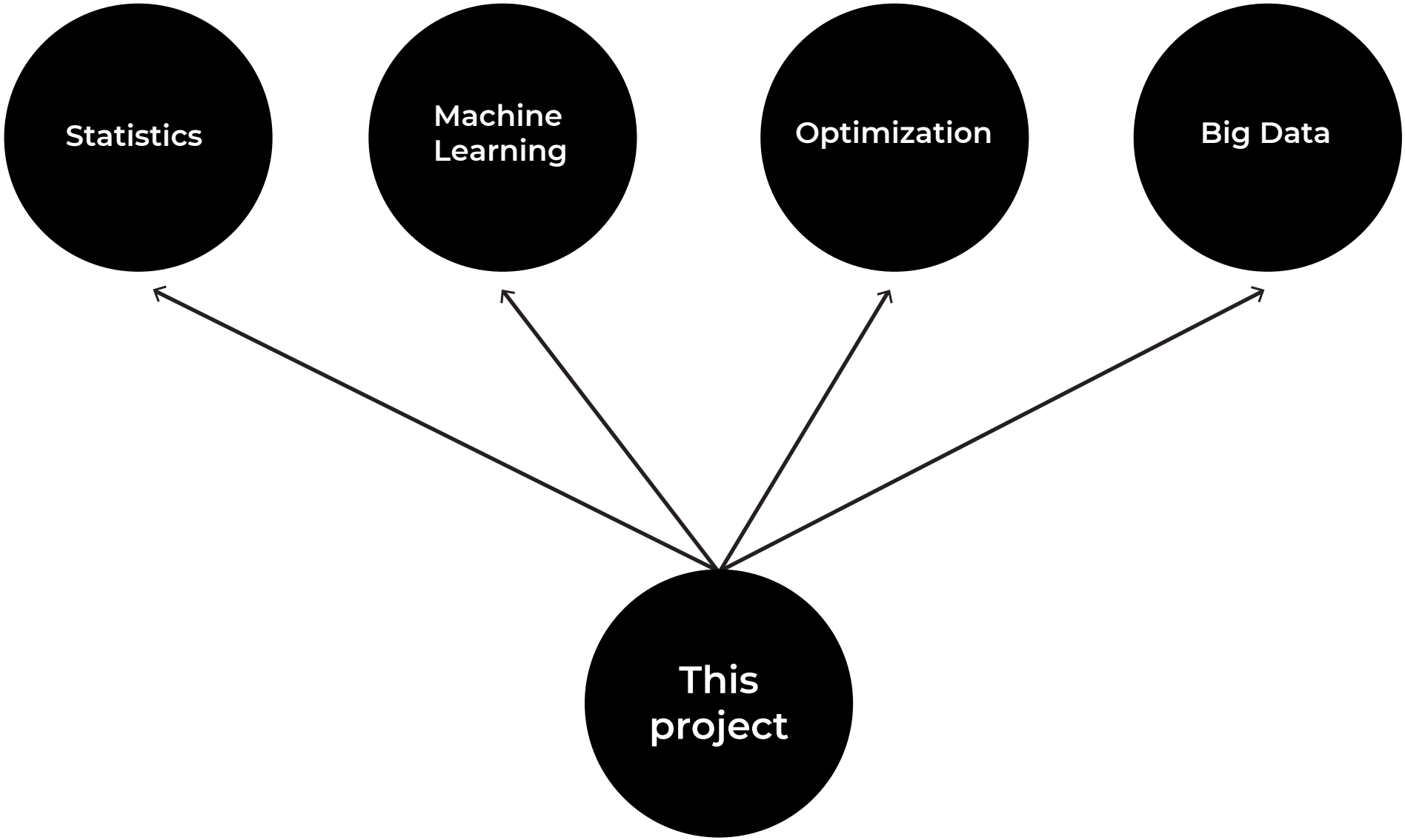
Future work

- Assess different kinds of pre-processings, such as logarithmic transformations or polynomial interpolations.
- Change the way we tell the algorithm that one product is new.

Future work

- Assess different kinds of pre-processings, such as logarithmic transformations or polynomial interpolations.
- Change the way we tell the algorithm that one product is new.
- Measuring product relationships.

Tools


Statistics
Machine Learning
Optimization
Big Data
This project

Thank you!

Blai Ras