BLOG

Demand Forecasting 1: Econometric Models

By Semantive August 6, 2018 No Comments

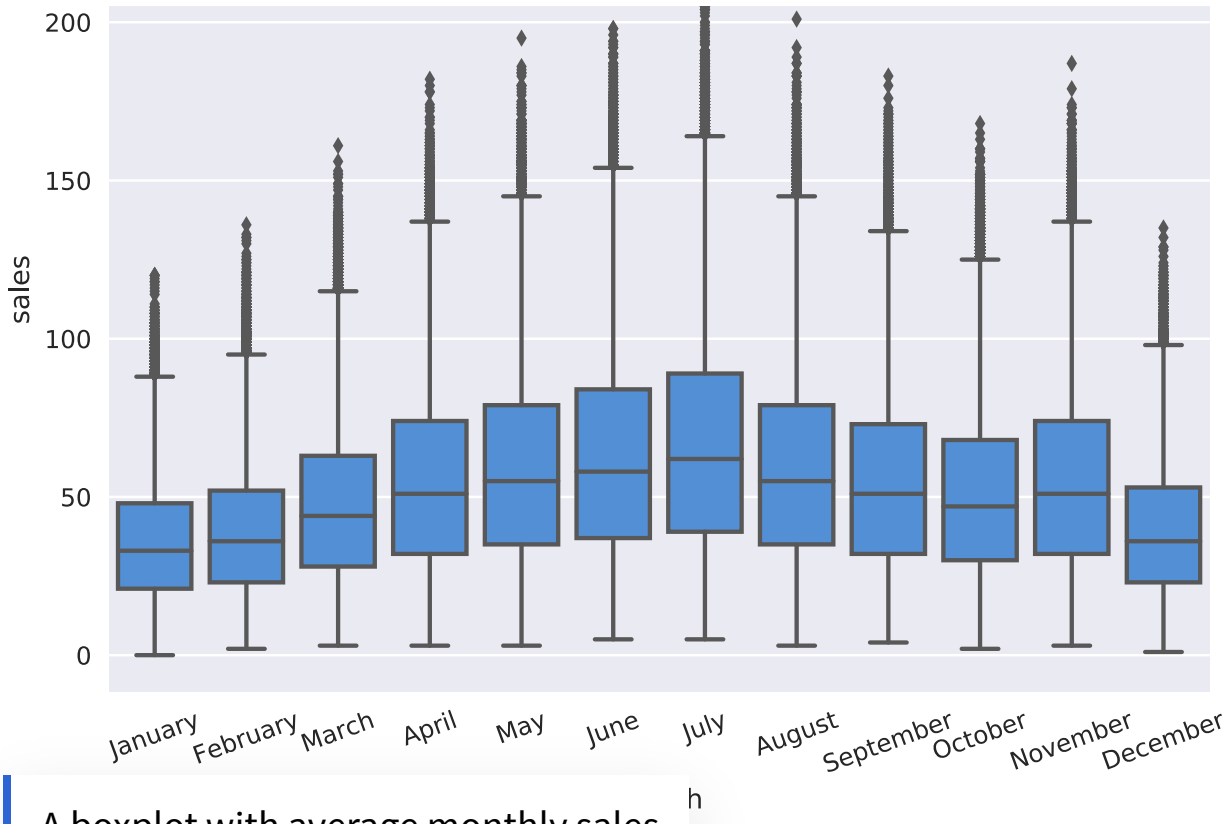
This post is a part of our series exploring different options for long-term demand forecasting. To better understand our journey, you might want to check out our **introductory blog post:** Long-Term Demand Forecasting

If you are interested in using historical data to make time series forecasts, undoubtedly, a good starting point for your analysis are statistical models used in econometric analysis for years. Such models, usually have strong theoretical foundations and often have strict assumptions about their dependent and independent variables, which enables inferencing and interpreting the results as well

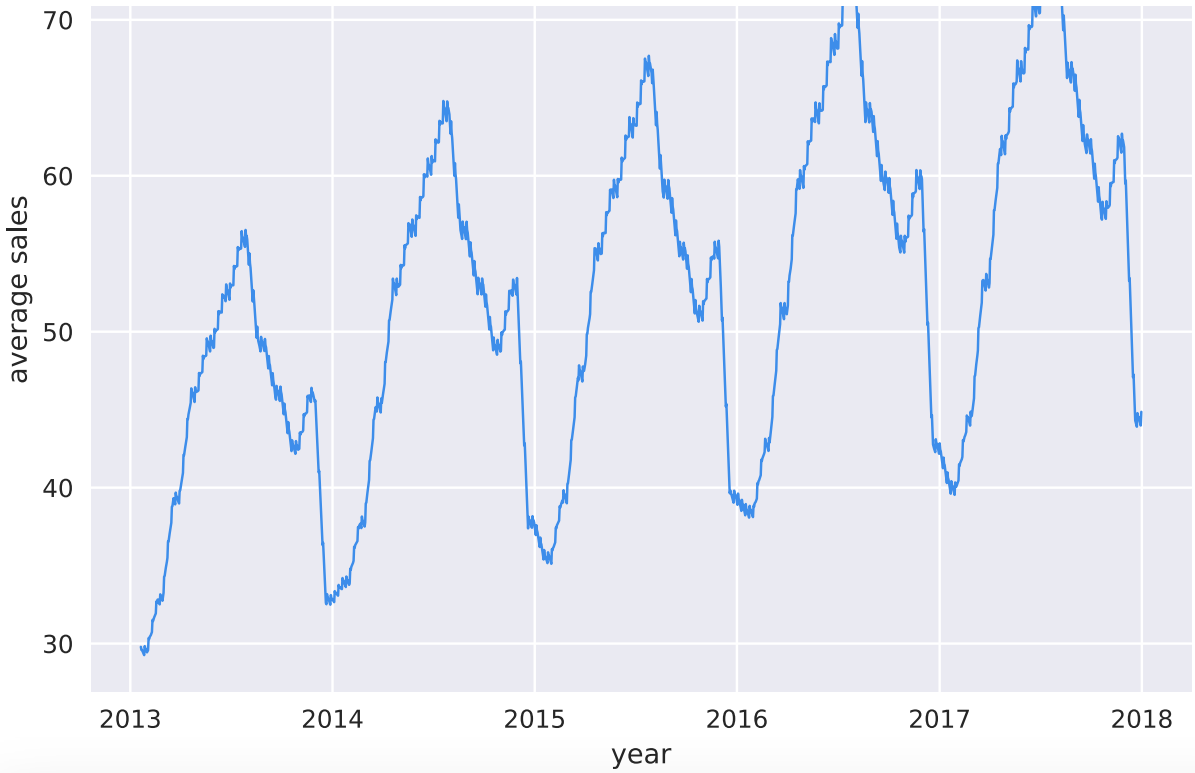
models provide acceptable results.

Dealing with seasonality

The time series we were to forecast had strong weekly and yearly seasonal components. Weekend sales substantially exceeded sales during the workweek. Average sales tend to increase during the summer, have a short spike in November and sink in December to start a steady growth in January. Many models we used could deal with simple seasonality, but often wouldn’t handle the complex seasonality consisting of multiple seasons.



A boxplot with average monthly sales



> A smoothed, average daily sales plot, note the regularity of the time series

We tried differencing to deal with seasonality, but sticked with averaging – calculating average values for particular time intervals of each season and subtracting them from the original ones. To get rid of weekly seasonality, we used the training data to calculate the average sales for every day of the week, and for the yearly component we used averages for each week of year. This allowed us to remove complex seasonality from the series.

For more information please check our deseasonalization utils on GitHub.

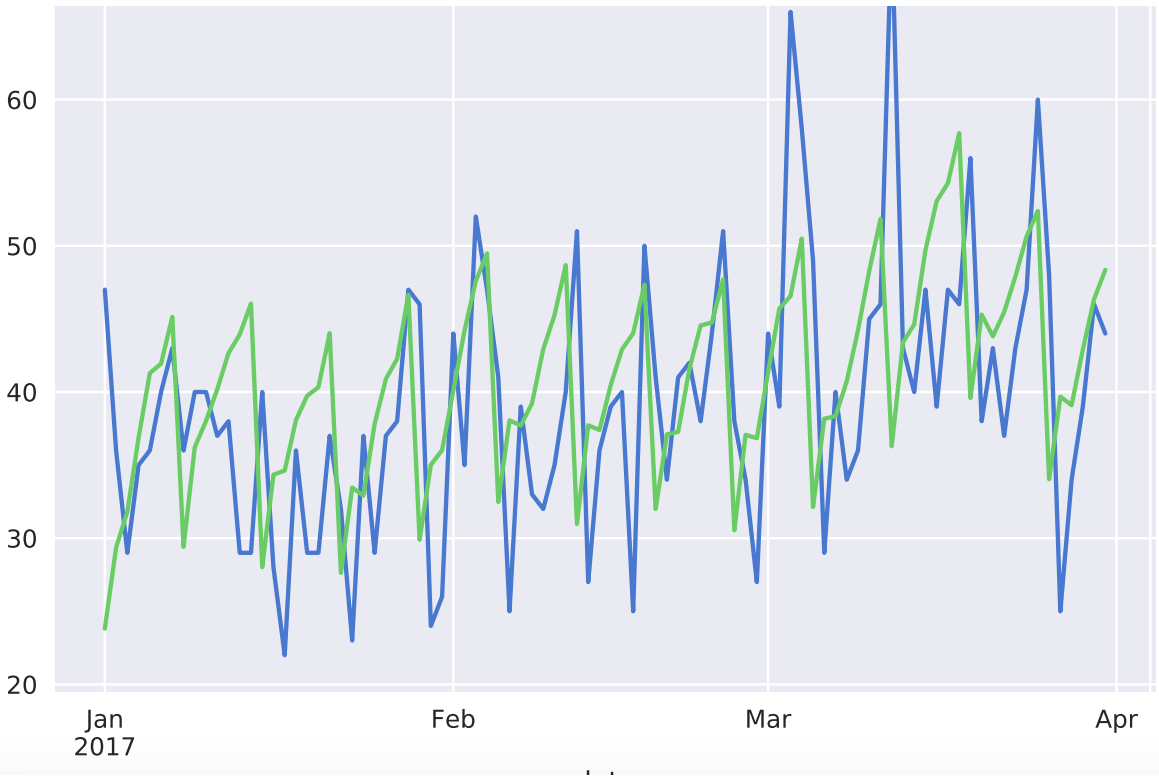
Holt-Winters exponential smoothing

seasonality of the data we decided to use triple exponential smoothing also known as Holt-Winters model. The model works fine as a baseline because of its simplicity, but it is still made for the univariate time series analysis, so if you have to deal with hundreds time series at once, it might take a lot of time to compute all the instances of the model. On the other hand, it has an advantage of not requiring data to be stationary.

Here is how the model's code looks like:

```
hw_trainer = lambda df: smt.ExponentialSmoothing(endog=df.loc[-365:,'normed'],
```

The values of the parameters were the same for each model: we recognized trend and seasonal component as 'additive' and we set the seasonal period to 7. Then we trained 500 models and computed the forecast, as well as combined accuracy results using SMAPE. Eventually, our first model achieved a SMAPE score 17.198 on Kaggle test set.



A sample Holt-Winters 90 days forecast for the single store (no. 6) and the single item (no. 14)

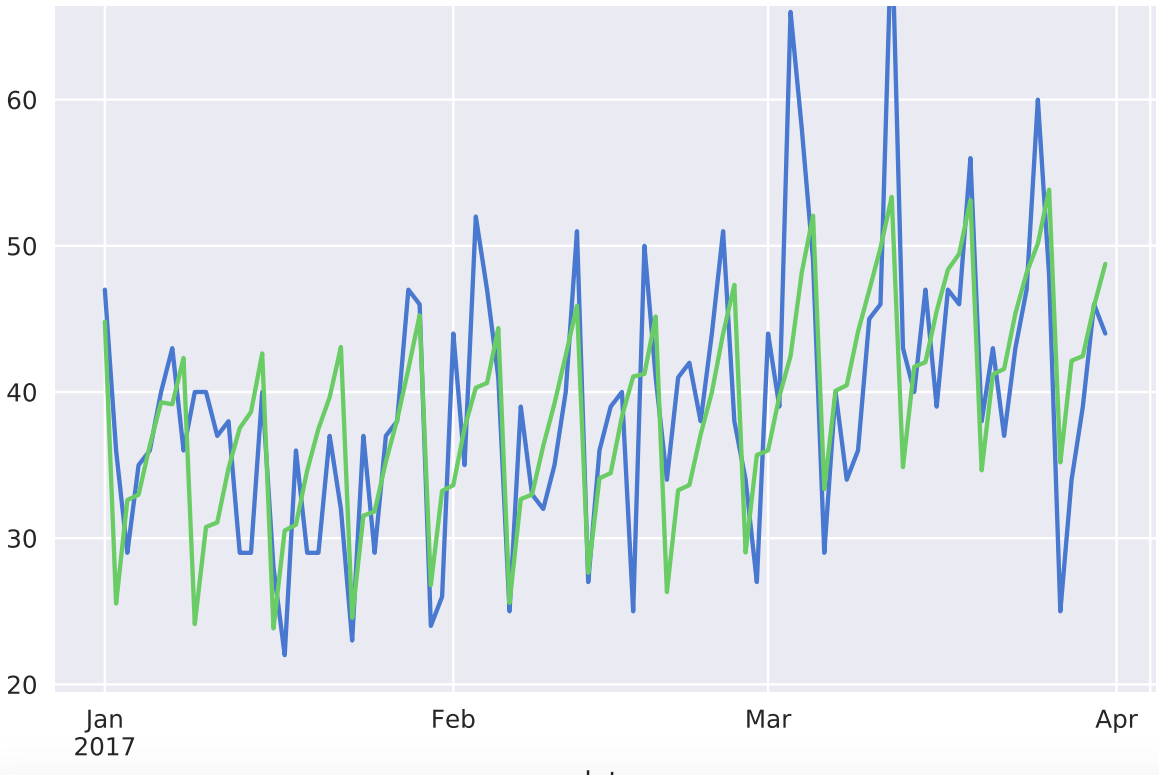
SARIMA

Seasonal Autoregressive Integrated Moving Average model which extends two popular models, AR and MA, while also being able to handle basic seasonality and non-stationary series. Unfortunately, this model is a generalization, so it has a lot of parameters and finding their correct values might be a little bit tricky and time consuming. It also requires theoretical knowledge and analysis of the autocorrelation

```
from pyramid.arima import auto_arima

arima_trainer = lambda df: auto_arima(df.loc[-365:,'sales'], m=7, n_jobs=1, max_p=7, max_q=7,
                                      max_P=7, max_Q=7, max_ord
                                      trace=False, error_action
```

SARIMA can only handle univariate time series, so we needed to find the best parameters separately for all 500 time series. Doing it manually would be long and tedious, so we decided to run a time-consuming parameters search to do it. What is more, SARIMA is not effective for the complex seasonal component, so it was necessary to remove the yearly seasonality separately. To train and tune the models we used pyramid.arima Python package which copies the functionality of auto.arima from the R language, used to automatically find the right SARIMA parameters. Eventually, the whole training process took almost 2 days on a quite decent PC, but the mean SMAPE decreased to 14.897.



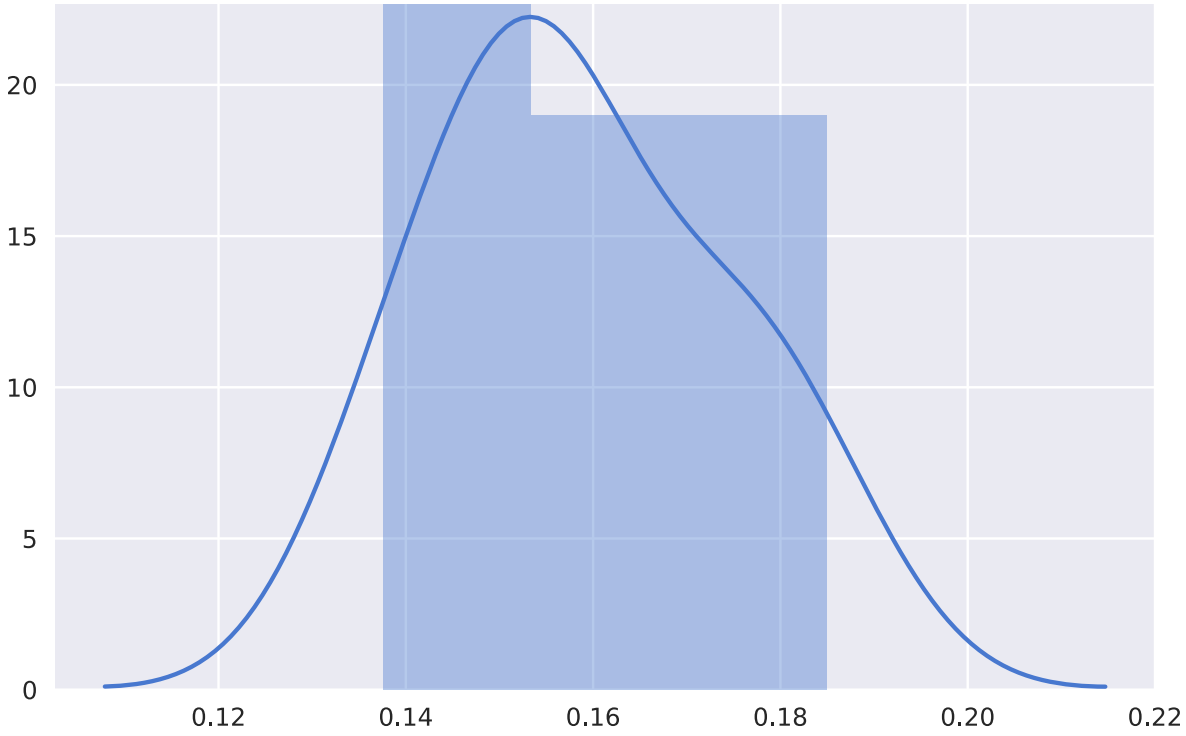
A sample SARIMA 90 days forecast for the single store (no. 6) and the single item (no. 14)

VAR

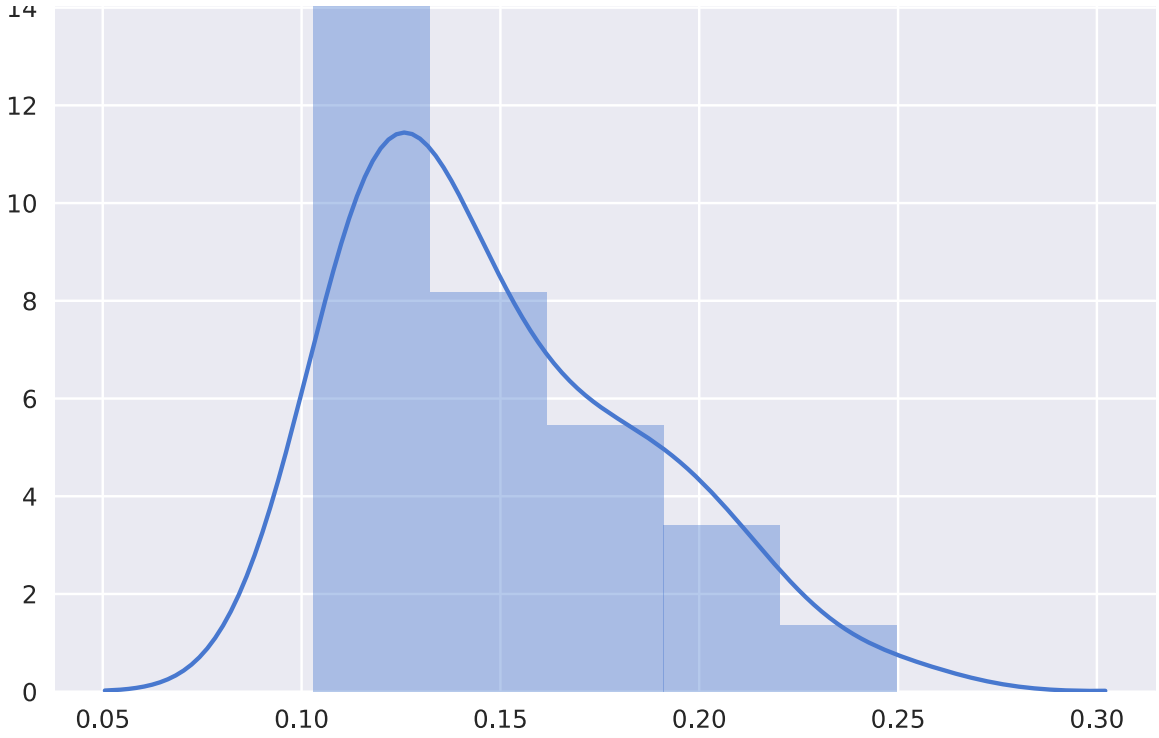
Vector Autoregression model is a generalization of the AR model from the last paragraph, that can be applied to multivariate time series and catch dependencies between individual variables. This model is a theory-free method, can only be used to make predictions about the future, and is not able to explain the dependencies

1. **deseasonalization of the training data**
2. **differencing the training data**
3. **modeling**
4. **forecasting**
5. **inverse differencing of the forecast**
6. **applying seasonality to the forecast**

We had only 1826 observations in each of the time series, which was not enough to run the VAR on all of the 500 series at once. Instead, we tried two different approaches: single VAR model for every store – 10 models total, and for every item – 50 models total. For each model, we had to remove seasonality and use target variable differencing to remove non-stationarity. Correspondingly, reverse operations had to be applied to the forecasts. On our validation set, 10 VAR models on data grouped by a store and 50 VAR models grouped by an item achieved 15.258 and 14.913 mean SMAPE respectively, so we chose the second model to submit to the Kaggle competition where it SMAPE score was 15.178.



> The SMAPE distribution on the validation set for the VAR models on data grouped by a store



> The SMAPE distribution on the validation set for the VAR models on data grouped by an item

Recommendations

The Holt-Winters model might be really useful as a baseline model, especially if you don't have a lot of time series to analyse. It is useful for the non-stationary, seasonal time series because it can handle them without any pre-processing. SARIMA, skillfully applied, may be an even better forecasting tool. It can handle non-stationary time series with seasonality and allows to interpret the results. Nonetheless, it has a lot of parameters and requires knowledge to tune the model correctly. VAR is a must-t technique for multivariate time series. It can't deal with non-stationarity but it's an

forecasting. Follow our blog if you want to learn more about our machine learning techniques, and in the meantime feel free to check out our code on GitHub.

See the next post: Demand Forecasting 2: Machine Learning Approach

Got a project idea? Let's schedule a quick, 15-minutes call to discuss how Big Data & Data Science professional services may give you the edge your business needs. Get in touch →

RECENT POSTS

4 modern AI solutions for manufacturing

Targi pracy IT 2019 w Warszawie. Co, gdzie, kiedy?

High-Performance computation in Python | NumPy

Text Summarization in Python

Data Science internship, and why Semantive program is worth it?

SEARCH

Search...

Leave a Reply

My comment is..

Name *

Email *

Save my name, email, and website in this browser for the next time I comment.

SUBMIT COMMENT

CONTACT US

+48 510 002 513 | contact@semantive.com

ul. Nowogrodzka 42/41, 00-695 Warsaw, Poland

SEMANTIVE

Big data | AI & Data Science | Cloud

Services that make your organization data informed