Today we begin a short series of posts about the long-term forecasting using different machine learning techniques. Forecasting sales is a common activity that almost all businesses need, so we decided to dedicate our time to testing different approaches to this problem. We took part in a Kaggle competition to see how various models' predictions compare to the top results and came up with some interesting conclusions that we wanted to share. This article is just an introduction to a series in which we will describe different approaches in greater depth.

we had to predict sales volume for every day separately – not just the total sales accumulated over this period. To evaluate the accuracy of the models, the authors of the competition chose Symmetric Mean Absolute Percentage Error (SMAPE) metric described by this equation:

$$SMAPE = \frac{100\%}{n} \sum_{t=1}^{n} \frac{|F_t - A_t|}{(|A_t| + |F_t|)/2}$$

Detailed problem description and datasets are available on Kaggle.

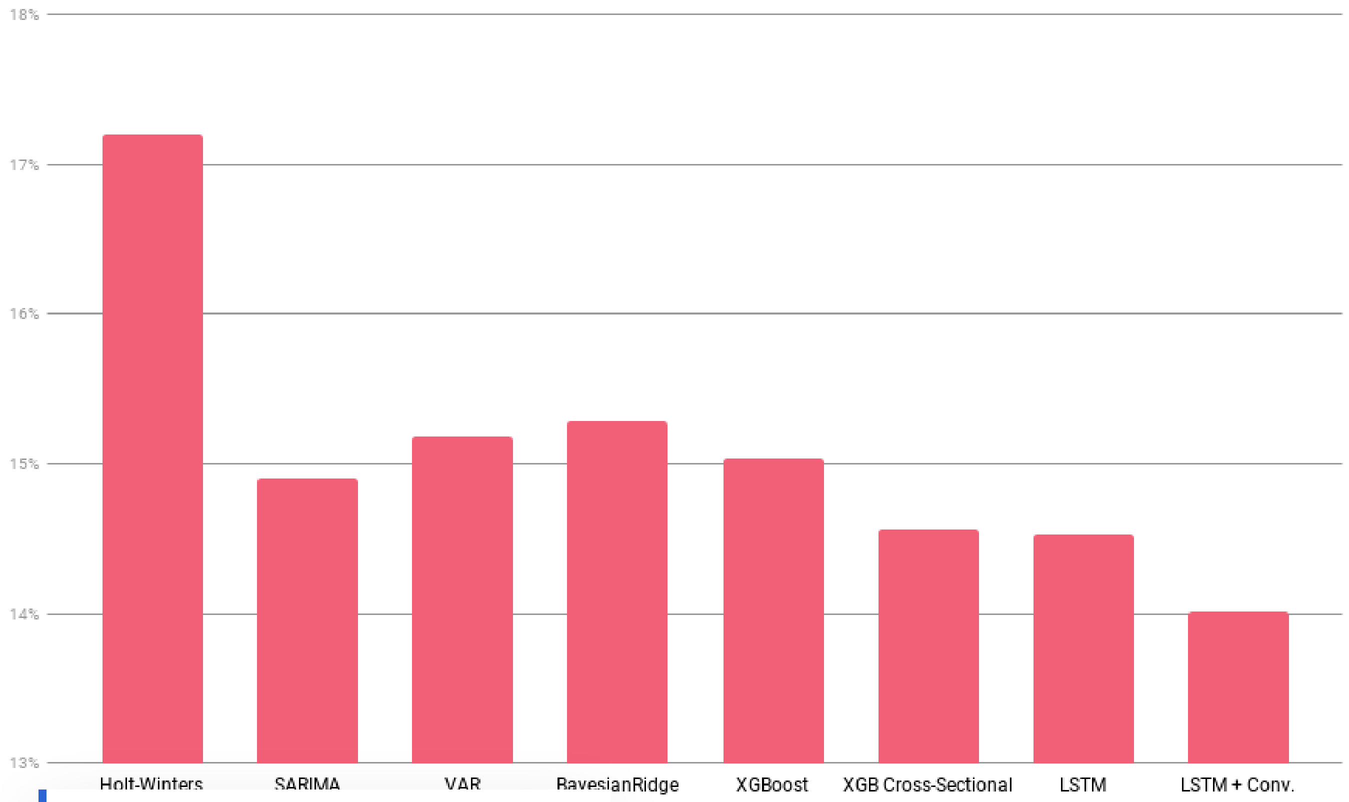
After initial exploratory analysis, it turned out that the sales of most items are seasonal and have a steady growing trend. This seemed like something that would allow for very accurate predictions, but in the end actually increased the need for pre-processing for some of the models. It also meant that some models would do surprisingly well and others unexpectedly bad.

Trying various approaches

For a start, we decided to check out some established econometric models that have strong theoretical foundations and are easily explainable. In particular, we tested Holt-Winters model, Vector Autoregression and SARIMA. We established that for this particular problem the last one was significantly more accurate. It has to be said, that it also required significant pre-processing and took a long time to tune its parameters, but this is a topic for another article.

After initial success with econometric methods, we decided to move on to classic machine learning. We have tried numerous regression models, starting from a basic linear regression and ending with XGboost. Most of our time was spent on engineering features as well as testing different approaches to training the model and

Finally, we decided to roll out the big guns and try deep recurrent neural networks, in particular Long Short-Term Memory Networks (LSTMs). We were afraid that the dataset is not large enough to train such models, but to our surprise it worked perfectly and scored the best results. We have experimented with different network architectures, all of which will be discussed in the future blog posts.



SMAPE Scores (lower is better)

Choosing the right model

Neural networks achieved the best results among our submissions. It is still important to remember about other types of models, as all of them have various

This was just a brief introduction to our exploration of forecasting methods and we want to share details of all of them with you – starting next Monday with econometric methods. We have also made all our code publicly available on GitHub.

See the next post: Demand Forecasting 1: Econometric Models

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