Data Science for Supply Chain Forecast

Nicolas Vandeput

1st edition – 2018

Contents

Preface

Tomorrow's supply chain is expected to provide many improved benefits for all stakeholders, and this across much more complex and interconnected networks than the current supply chain.

Today, the practice of supply chain science is striving for excellence: innovative and integrated solutions are based on new ideas, new perspectives and new collaborations, thus enhancing the power offered by data science.

This opens up tremendous opportunities to design new strategies, tactics and operations to achieve greater anticipation, a better final customer experience and an overall enhanced supply chain.

As supply chain generally account between 60% and 90% of all company costs (excluding financial services), any drive toward excellence will undoubtedly be equally impactful on a company's performance as well as on its final consumer satisfaction.

This book, written by Nicolas Vandeput, is a carefully developed work emphasizing how and where data science (with its systems integration and knowledge) can effectively lift the supply chain process higher up the excellence ladder.

This is a gap-bridging book from both the research and the practitioner's perspective, it is a great source of information and value.

Firmly grounded in scientific research principles, this book deploys a comprehensive set of approaches particularly useful in tackling the critical challenges that practitioners and researchers face in today and tomorrow's (supply chain) business environment.

Prof. Dr. Ir. Alassane B. NDIAYE
Professor of Logistics & Transport Systems
Universite Libre de Bruxelles, Belgium

Data Science for Supply Chain Forecast

> *Artificial intelligence is the new electricity*
> Andrew Ng[1]

In the same way electricity revolutionized the second half of the XIXth century, allowing industries to produce more with less, AI will drastically impact the following decades. While some companies already use this new electricity to cast new light upon their business, others are definitely still using old oil lamps; or even candles, using manpower to manually change these candles every hour of the day in order to keep the business running.

As you will discover in this book, artificial intelligence (AI) & machine learning (ML) are not just a question of coding skills. Using data science to solve a problem will require a scientific mindset more than coding skills. We will discuss many different models and algorithms in the latter chapters. But as you will see, you do not need to be an IT wizard to apply these models. There is another more important story behind these: a story of experimentations, observation and questioning everything; a true scientific method applied to supply chain. In the field of data science as well as for supply chain, simple questions do not come with simple answers. To answer these questions, you will need to both be a scientist and use the right tools. In this book, we will discuss both.

Supply Chain Forecast Within all supply chains lies the question of planning. The better we evaluate the future, the better we can prepare ourselves. The question of future uncertainty, how to reduce it or how to protect yourself against this unknown has always been crucial for every supply chain. From negotiating contract volumes with suppliers to setting safety stock targets, everything relates to the ultimate question:

[1] Andrew Ng is the co-founder of Coursera the leading online-classes platform

What is tomorrow going to be like?

Yesterday, big companies provided forecast softwares that allowed businesses to use a statistical forecast as the backbone of their S&OP[1] process. These statistical forecast models were proposed 60 years ago by Holt & Winters[2] and didn't change much in the last 50 years: at the core of any statistical forecast tool, you still find exponential smoothing. Software companies sell the idea that they can add a bit of extra intelligence into it, or some less-known statistical model, but in the end it all goes back to exponential smoothing, which we will discuss in the first part of this book. Yesterday, one analyst on his own personal computer couldn't compete with these models.

Today, things have changed. Thanks to the increase in computing power, the in-flow of data and the availability of free tools, one can make a difference. **You** can make a difference. With a bit of coding skills and an appetite for experimentation, powered by machine learning models, you will be able to bring to any business more value than any off-the-shelf forecast software can deliver.

We often hear that the recent rise of artificial intelligence (or machine learning) is due to an increasing amount of data available as well as cheaper computing power. This is not entirely true. Two other effects explain the recent interest in machine learning. In the previous years, many machine learning models were improved, giving better results. As these models were becoming better & faster, the tools to use them became more user-friendly. It is much easier today to use powerful machine learning models than it was 10 years ago.

Can I do this? Is this book for me? This book has been written for supply chain practitioners, forecasters and analysts who are looking to go the extra mile[3]. You do not need technical IT skills to start using the models of this book today. You do not need a dedicated server or expensive software licences: only your own computer. You do not need a PhD in mathematics: we will only use mathematics when they are directly useful to tweak and understand the models. More often than not – especially for machine learning – a deep understanding of the mathematical inner workings of a model will not be necessary to optimize it and understand its limitations.

[1] The sales and operations planning (S&OP) process focuses on aligning mid and long-term demand and supply.

[2] See page 32 for more information about Holt & Winters.

[3] Even though we will focus on supply chain demand forecast, the principles & models explained here can be applied to any forecast problem.

The Data Scientist's Mindset

As the business world discovers data science, many supply chain practitioners still rely on rules of thumbs and simple approximations to run their businesses. Most often, most of the work is done directly in Excel. A paradigm shift will be needed to go from manual approximations done in Excel towards automated powerful models in Python. We need to leave oil lamps behind and move to electricity. This is what we will do – step by step – in this book. Before discussing our supply chain data-scientist tools, let's discuss what our data scientist mindset should be.

Data is gold If artificial intelligence is the new electricity – allowing us to achieve more in a smarter way – data is the modern gold. Gold, unfortunately, does not grow on trees; it comes from gold ore that needs to be extracted and cleaned. Just like data: it needs to be mined, extracted and cleaned. We even have to think where to mine to get the data. As supply chain data scientists, we are both goldsmiths and miners. Even though this book does not cover the specific topic of data cleaning nor the question of data governance, it will show you how to magnify gold in jewelry. Always treat data as if it were gold ore: it is precious but needs to be cleaned and refined in order to become usable.

Start small, iterate It is easy to lose yourself in details as you try to model the real business world. To avoid this, we will always start tackling broad supply chain questions with simple models. And then, we will iterate on these models, adding complexity layers one by one. It is an illusion to think that one could grasp all the complexity of a supply chain at once in one model. As your understanding of a supply chains grows, so does the complexity of your model.

Experiment! There is no definitive answer nor model to each supply chain question. We are not in a world of one-size-fits-all. A model that worked for another company might not work for you. This book will propose you many models & ideas and will give you the tools to play with them and to experiment. Which one to choose in the end is up to you! I can only encourage you to experiment small variations on them – and then bigger ones! – until you find the one that suits your business.

Unfortunately, many people forget that experimenting means trial & error. Which means that you will face the risk of failing. Experimenting with new ideas and models is not a linear task: days or weeks can be invested in dead ends. On the other hand, a single stroke of genius

can drastically improve a model. What is important is to fail fast and start a new cycle rapidly. Don't get discouraged by a couple of days without improvement.

Automation is the key to fast experimentation As you will see, we will need to run tens, hundreds or even thousands of experiments on some datasets to find the best model. In order to do so, only automation can help us out. It is tremendously important to keep our data workflow fully automated in order to be able to run these experiments without hurdle. Only automation will allow you to scale your work.

Automation is the key to reliability As your model will grow in complexity and your datasets in size, you will need to be able to reliably populate results (and act upon them). Only an automated data workflow together with an automated model will give you reliable results over and over again. Manual work will slow you down and create random mistakes, which will result *in fine* in frustration.

Don't get mislead by overfitting and luck As we will see in chapter 8, overfitting (i.e. your model will work extremely well on your current dataset but fail to perform well on new data) is the #1 curse for data scientists. Do not get fooled by luck or overfitting. You should always treat astonishing results with suspicion and ask yourself the question: *"Can I replicate these results on new unseen data?"*.

Sharing is caring Science needs openness. You will be able to create better models if you take the time to share their inner workings with your team. Openly sharing results (good and bad) will also create trust among the team. Many people are afraid to share bad results, but it is worth doing so. Sharing bad results will allow you to trigger a debate among your team to build a new and better model. Maybe someone external will bring a brand-new idea that will allow you to improve your model.

Simplicity over complexity As a model grows bigger, there is always a temptation to add more and more specific rules and exceptions. Do not go down this road. As more special rules add up in a model, the model will loose its ability to perform reliably well on new data. And soon, you will loose the understanding of all the different interdependences. You should always prefer a structural fix to a specific new rule (also known as *quick fix*). As the pile of quick fixes grows bigger, the potential amount of interdependences will exponentially increase and you will not be able to identify why your model works in a specific way.

Focus on the point As a last piece of advice, you should always focus on the point. Clarity comes from simplicity. When communicating your results, always ask yourself some questions:

Who am I communicating to?
What are they interested in?
Do I show them everything they are interested in?
Do I show them only what they are interested in?

In a world of big data, it is easy to drown someone in numbers and graphs. Just keep your communication simple and straight to the point. Remember that our brain easily analyzes and extrapolates graphs and curves. So prefer a simple graph to a table of data when communicating your results.

Perfection is achieved, not when there is nothing more to add, but when there is nothing left to take away.
Antoine de Saint-Exupery (1940-1944)

The Data Scientist's Tools

We will use two tools to build our models, experiment and share our results.

Excel Excel is the data analyst's Swiss knife. It will allow you to easily perform simple calculations and plot data. The big advantage of Excel compared to any programming language is that we can **see** the data. It is much easier to debug a model or to test a new one if you see how the data is transformed by each step of computation. Therefore, Excel can be a first go-to to experiment with new models or data.

Excel also has many limitations. It won't perform well on big datasets and will hardly allow you to automate difficult tasks.

Python Python is a programming language initially published in 1991 by Guido van Rossum, a Dutch computer scientist. If Excel is a Swiss knife, Python is a full army of construction machines awaiting instructions from any data scientist. Python will allow you to perform computations on huge datasets in an automated and fast way. Python also comes with many libraries dedicated to data analysis (`pandas`), scientific computations (`NumPy` & `SciPy`) or machine learning (`scikit-learn`). These will soon be your best friends.

Why Python? We chose to use Python over other programming languages as it is both user-friendly (it is easy to read & understand) and one of the most used programming languages in the world. In 2018, it was the programming language that was the most googled and it is the most commonly used for machine learning.

Should you start learning Python? Yes, you should.

Excel will be perfect to visualize results and the different data transformation steps you perform, but it won't allow you to scale your models to bigger datasets nor to easily automate any data cleaning. Excel is also unable to run any machine learning algorithm.

Many practitioners are afraid to learn a coding language. Everyone knows a colleague who uses some macros/VBA in Excel – maybe you are this colleague – and the complexity of these macros might be frightening to the profane. **Python is much simpler than Excel macros**. It is also **much more powerful**. As you will see by yourself in the following chapters, even the most advanced machine learning models won't require so many lines of code or complex functions. It means that you do not have to be an IT genius to use machine learning on your own computer. You can do it yourself, today. Python will give you a definitive edge compared to anyone using Excel. Today is a great day to start learning Python. Many resources are available: videos, blogs, articles, books... You can, for example, look for Python courses on the following online platforms:

www.edx.org
www.coursera.org
www.udemy.com
www.datacamp.com

I do personally recommend the MIT class "*Introduction to Computer Science and Programming Using Python*" available on EdX [8]. This will teach you everything you need to know about Python to start using the models presented in this book.

We will also briefly introduce the most useful concepts in chapter 1 to help you out with the first code extracts.

How to Read This Book

Data Science for Supply Chain Forecast is written the way I wish someone had explained me how to forecast supply chain products some years ago. It

is divided into two parts: we will first discuss old-school statistical models and then move to machine learning models.

Old-school statistics and Machine Learning One could think that these statistical models are already outdated and useless as machine learning models will take over. But this is wrong. These old-school models will allow us to *understand* and *see* the demand patterns in our supply chain. Machine learning models, unfortunately, won't provide us any explanation nor understanding of the different patterns. Machine learning is only focused on one thing: getting the right answer. The *how* does not matter. This is why both the statistical models and the machine learning models will be helpful for you.

Concepts & Models The book is divided into many chapters: each of them is either a new model or a new concept. This will allow us to build our understanding of the field of data science & forecast step by step. Each new model or concept will allow to overcome a limitation or to go one step further in terms of forecast accuracy.
On the other hand, obviously, not all forecast models are explained here. We will focus on the models that have proven their value in the world of supply chain forecast.

Do It Yourself We also take the decision not to use any prebuilt forecast function from Python or Excel. The objective of this book is not to teach you how to use a software. It is twofold. First, it is for you to be able to experiment with different models on your own datasets. This means that you will have to tweak the models and experiment with different variations. You will only be able to do this if you take the time to implement these models yourself. Second, it is for you to acquire an in-depth knowledge on how the different models work as well as their strengths and limitations. Implementing the different models yourself will allow you to learn by doing as you test them along the way.
At the end of each chapter, you will find a *Do-It-Yourself* section that will show you a step-by-step implementation of the different models. I can only advise you to start testing these models on your own datasets ASAP.

Other resources

You can download the Python codes shown in this book as well as the Excel templates on `www.supchains.com/book-ressources` (password: `SupChains`).

There is also a glossary at the end of the book where you can find a short description of all the specific terms we will use. Do not hesitate to consult it if you are unsure about a term or acronym.