CHAPTER 10

Deep Learning and Demand Forecasting

> *Machine learning drives our algorithms for demand forecasting, product search ranking, product and deals recommendations, merchandising placements, fraud detection, translations, and much more.*
>
> —Jeff Bezos, CEO, Amazon

Demand forecasting is a branch of predictive analytics that focuses on gaining an understanding of consumer demand for goods and services. If demand can be understood, brands can control their inventory to avoid overstocking and understocking products. While there is no perfect forecasting model, using demand forecasting as a tool can help fashion businesses better prepare for upcoming seasons.

For fashion brands, estimating the amount of inventory to manufacture can be a complex game of combining historical data, applying intuition, and forecasting fashion trends. Placing too large an order of a particular garment style can lead to inefficiencies for the brand, from decreasing margins when items are discounted on the retail floor, to complete loss if the products don't move. For these predictions, investing in demand forecasting strategies could be worth billions of dollars every year for the fashion industry. The waste created by overproduction is both a financial and an environmental disaster.

L. Luce, *Artificial Intelligence for Fashion*,
https://doi.org/10.1007/978-1-4842-3931-5_10

While demand forecasting remains a challenge, developments in machine learning have provided dramatic improvements. In this chapter, we'll explore deep learning in demand forecasting for fashion.

What Is Demand Forecasting?

Forecasting as a general topic in fashion can address needs such as capacity planning, stock keeping, and pricing strategy. In simplest terms, demand forecasting estimates future sales. Forecasting can encompass long-term planning (for example, how much revenue will we generate this year, and how should we grow our workforce next year) as well as short-term planning (for example, how many shirts of this style will be purchased this year). In this chapter, we will discuss the latter, though many of these techniques can be applied to both.

Capacity planning and stock keeping usually fluctuate over time. The forecasting techniques to predict these future demands rely on **time-series data**, data that is successive over a period of time. This type of prediction is referred to as **time-series forecasting**.

Forecasting Methods

Demand forecasting is an entire field on its own. There are many methods for forecasting demand, ranging from more qualitative to quantitative approaches. Examples of methods used are averages, time-series analysis, surveys, as well as machine learning approaches including deep neural networks. Depending on the data, some methods work better than others.

Fashion's Challenges in Forecasting

Across retail sectors, demand forecasting has become a critical tool. For grocers whose products may expire over a few days, predicting demand can radically change profit margins. For fashion, the idea is not much different, though in some ways it is more complex. There is a lack of historical data to rely on, because new items are being constantly introduced, and seasonal trends add an extra layer of unpredictability. The challenges to overcome in fashion include overproduction, expiration, fast and short seasons, and unpredictable consumer behavior.

Overproduction

As the pain point of overproduction has become increasingly detrimental to apparel brands, the entire industry has begun to shift to accommodate. One of these changes is a movement toward vertical integration, which has led to faster response times across an organization. Some experts have suggested that the best solution is to be building flexible manufacturing infrastructure that allows the industry to respond to changing demands in real time. While flexible manufacturing does provide a solution to the same problem, it's not necessarily a substitute. Improved forecasting can provide shorter-term relief and reduce waste and spending.

Fast and Short Seasons

Understanding demand is so critical in the fashion industry partly because the product life cycle is short, the selling season is short, and the replenishment lead times are long. For major retailers like Zara, there are around 20 seasons in a year, leaving only a couple of weeks to sell the styles in a season.

Although fashion goods do not expire physically, the desire for those goods does. The timelines have shortened as the movement of goods in fashion has increasingly taken on the patterns of a consumable.

In some ways, this is quenching a thirst for creativity and self-expression. In other ways, it may be illuminating a toxic aspect of our culture. The more we consume in fashion, the more waste we generate.

Consumer Behavior

Consumer behavior in this space is emotionally driven and often highly impulsive. Purchasing patterns are also highly impacted by external factors. They're volatile in that even the weather and the news can affect sales dramatically. Aside from all of these challenges, the product itself has a wide range of variety in style, color, surface treatment, print, and so many other variables.

Nonforecasting Solutions

Without the ability to accurately forecast demand of individual fashion items, brands and retailers have turned to measures to create demand. One of the most straightforward ways to create or control demand is through price.

Price Prediction

One common variable to optimize in sell-through of a product is price. If you're selling through a product every time without optimizing the price, there might be an opportunity to charge more for that item. On the flip side, if an item isn't selling, decreasing the price might be a simple change to help increases sales of that item. Markdowns and promotions are ever-popular techniques for addressing inventory management in the fashion industry.

Historically, pricing strategy has relied on a mix of markup cost, competition pricing, and the merchant's judgement or intuition about what a customer will pay for the item. This strategy has its shortcomings, as it relies on methods that are often immeasurable.

We're mentioning price here as a nonforecasting strategy to manipulate demand, but the relationship can also go the other way around. Forecasting can be used to determine the optimal price for a product by interpreting related information including competitor data, historical data, or sales of similar items.

Deep Learning

> *The analogy to deep learning is that the rocket engine is the deep learning model, and the fuel is the huge amounts of data we can feed to these algorithms.*
>
> —Andrew Ng, general partner, AI Fund

How can deep learning help resolve some of the complex challenges in demand forecasting for the fashion industry? Deep learning has been used to handle complex problems with many variables that are difficult to model manually. These techniques work best in a data-rich environment, or in use cases in which there is a lot of data.

Deep neural networks have the potential to generate more-accurate predictions than a linear model because of compatibility when it comes to the complexity of these problems. One of the major downsides to a deep learning approach is that it's difficult to interpret how the algorithm arrived at its results, which are crucial for business decisions.

What Is Deep Learning?

Deep learning is a specialized subset of machine learning. Deep learning usually refers to large neural networks trained on large datasets. Deep learning is being used for a vast array of applications, not only inside the fashion industry and retail, but in medicine, self-driving cars, and beyond. While deep learning is often recognized for successes in image generation, like the GANs described in Chapter 8, it can also be used for other tasks.

In fact, several of the techniques discussed in this book can be considered under the umbrella of deep learning, such as deep neural networks, convolutional neural networks, and recurrent neural networks.

The distinction between basic machine learning and deep learning is difficult to define. One explanation is that basic machine learning tends to be more task specific, and deep learning more learning based. Deep learning encompasses neural networks that have more hidden layers.

Traditional methods often rely on **linear regression** models for demand forecasting and have had less success in fashion because of the chaotic nature of the industry. While these might give some high-level estimation to demand, they don't give enough resolution to make predictions about what will happen in the future.

In order to get a high-resolution forecast, the predictions need to fit the curve that demand does. Much of machine learning is focused on **curve fitting**, with the goal to more accurately track reality as represented by a graph. Figure 10-1 shows how the real world, a linear model, and a machine learning model might look.

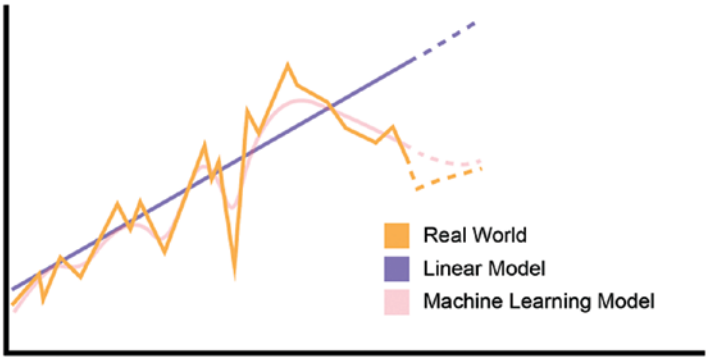

Figure 10-1. *In a machine learning model, complex curve fitting can track closer to the real world. In a linear regression model, the average turns out to be an oversimplification.*

Deep Learning for Demand Forecasting

Neural networks provide greater flexibility in demand forecasting because they are nonlinear models that can take in a lot of variables and then output simple predictions. However, their benefits are greatly affected by the training data. Without a lot of high-quality data, it can be difficult for these models to perform. This is actually a major inefficiency of using deep neural networks for the application of demand forecasting; they are not very data efficient.

Long short-term memory (**LSTM**) is one of the most talked about deep learning models used for time-series forecasting. LSTMs are a type of recurrent neural network with unique memory-like behavior that allows them to learn patterns better.

Techniques for Smaller Datasets

While data is an important aspect of all machine learning, companies with small amounts of data can still apply deep learning through **transfer learning** and other forecasting models that are specialized for smaller datasets.

Transfer Learning

In transfer learning, an approach specific to deep learning problems, pre-trained models are used to bootstrap another more specific model. This significantly reduces the computation and time it takes to carry out tasks. It doesn't apply to just forecasting, but to other deep learning applications as well.

One way to think of transfer learning is as a design methodology. It doesn't refer to a difference in the model itself, but in the way it is trained and run. Rather than starting from nothing, you can start from a model that is trained on data that was used for a similar task. For example, if I trained an image classifier to identify whether an image contained a sneaker, I could use that pretrained model to recognize other items (such as boots, maybe). The differences in design methodology are expressed in Figure 10-2.

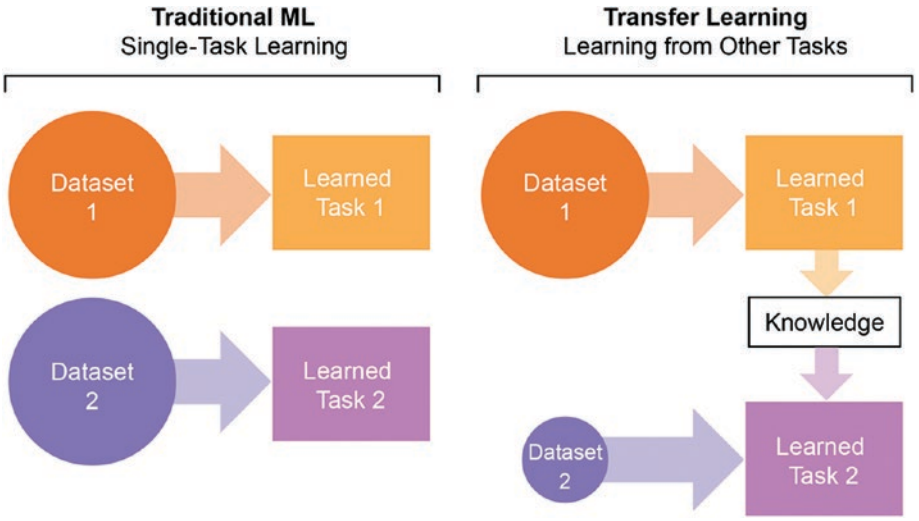

Figure 10-2. *Comparing transfer learning to traditional methods of machine learning*

Transfer learning isn't just a solution for small datasets, but can be used as a way to make the most of older data that might have less relevance in making predictions. What a brand was selling in 1990 might be irrelevant to them today, but the data can still be useful in training a forecasting model.

Other Forecasting Models

There are conflicting experiments and opinions about which models work best for forecasting applications. While deep learning methods such as LSTM appear to work well for applications with a lot of data, other models could outperform LSTM in certain use cases.

This chapter focuses on deep learning as a demand forecasting method in order to explain deep learning. However, there are other models for forecasting, including **autoregressive integrated moving average** (**ARIMA**) and **Prophet**.

ARIMA

Each of the components of ARIMA explains what the model does. Autoregression (AR) shows a changing variable, integrated (I) refers to taking the difference in the equations between current and previous values, and **moving average** (MA) incorporates a lagged average, such as a three-day average in which each point on a graph contains the average of the previous three days. Until recently, this was the state-of-the-art method for forecasting.

Prophet

In 2017, Facebook released its open source forecasting model, Prophet. In comparison to the data-hungry models presented by deep learning, Prophet is more data efficient. Prophet is also really useful for data that is seasonal.

Prophet represents an analyst-in-the-loop approach that can be thought of as human analysis augmented by machine tools. When you create a model, a trained analyst who understands the data can make it better. This is important because quality forecasting requires a highly specialized person, but there are very few of them. The Prophet model bridges the gap by getting the machine to do fairly good work, and having a less-trained analyst who can fill in the gaps. In this approach, the barrier to getting a good forecast has been lowered. This is also useful because it allows anomalies, such as holidays and other features that would be difficult to model, to be input by the analyst.

Each method has its trade-offs, Figure 10-3 shows four high-level approaches and their pros and cons.

Model	Pros	Cons
Linear Regression	· Easy to understand · Handles different components	· Sensitive to outliers · Strong assumptions
ARIMA	· Easy to understand · Fits historical data well · Forecasts unbiased	· Sensitive to outliers · Small forecast range
Prophet	· Easy to understand · Analyst in the loop · Data efficient · Fast	· Sensitive to compounding seasonality · Required data format
Deep Learning Neural Networks LSTM Transfer Learning	· Can take in many complex variables · Finds nonlinear patterns · Strong predictions · Easy to automate	· Difficult to understand · Requires a lot of data

Figure 10-3. *Pros and cons of the approaches discussed in this chapter*

While deep learning and other machine learning techniques are useful in the context of demand forecasting, they're just a tool. A highly skilled analyst in a particular industry has deep domain knowledge of a space that is often not captured in the data or that might go against the data at times. Rather than ignoring the knowledge of the analyst in these contexts and blindly following the machine's predictions, some models (for example, Prophet) keep the analyst in the loop**.**

Summary

Demand forecasting is a long-standing challenge, especially in fashion, which requires inventory and resource planning for the production of physical goods. Short seasons and irregular customer behavior make demand even more difficult to predict in this industry.

The use of deep learning models for demand forecasting is still a nascent field. Each year, new research is presented and improved upon. There is no perfect forecasting model, but these tools can be helpful predictors of demand even in volatile markets.

Demand forecasting is a complex and nuanced field. You can learn more from the following books and from references in the annotated bibliography at the end of this book:

- *Forecasting: Principles and Practice* by Rob J. Hyndman and George Athanasopoulos (OTexts, 2018)
- *Applied Predictive Modeling* by Max Kuhn and Kjell Johnson (Springer, 2016)

Thank you to Adam Bouhenguel, founder of Tesserai for providing expert advice in machine learning and demand forecasting as background for this chapter.

Terminology from This Chapter

Autoregressive integrated moving average (**ARIMA**)—A common approach to forecasting time-series data that does not involve machine learning.

Curve fitting—The method of finding an equation that most closely models the data. Compared to linear regression, curve fitting requires more-complicated formulas to represent the relationships between data points.

Deep learning—An area within machine learning usually referring to neural networks composed of many layers.

Demand forecasting—Part of predictive analytics that focuses on understanding and predicting future demand for goods and services using models trained on historical data.

Linear regression—A simple summarization tool in statistics in which a series of points on a map are averaged by plotting a line through them.

Long short-term memory (LSTM)—Neural networks composed of LSTM units. These units are made to include a mathematical representation of memory between units. Rather than a linear path through the network, an LSTM creates different paths based on input, output, and forget gates associated with the model. It gives the network a memory-like quality.

Moving average—Takes previous intervals and averages them in an attempt to more closely represent a trend over time.

Prophet—An open source forecasting tool released by Facebook in 2017. It's an additive regression model in which certain variables can be adjusted easily by an analyst to improve predictions.

Time-series data—Data collected successively over a set time interval.

Time-series forecasting—Predicting future events based on data that is collected over time.

Transfer learning—A methodology in which a neural network is trained on one dataset, and then the knowledge is stored and then applied to a different but similar problem.