CHAPTER 7

Predictive Analytics and Size Recommendations

> *I am more than my measurements.*
>
> —Ashley Graham, model

Fit is a vaguely defined, complexly intertwined technical and emotional topic. Each individual has a different definition of how they want their clothing to fit. The way that we use words to describe fit even varies from one person to another. “Baggy” to one person may look like something different to someone else.

The Fit Problem

Sizing systems are an alphanumeric organization of garment dimensions created to help individuals find garments that fit. Unfortunately, not all brands assign the same measurements to the same sizes. This means for customers, it can be difficult to figure out which size to buy.

L. Luce, *Artificial Intelligence for Fashion*,
https://doi.org/10.1007/978-1-4842-3931-5_7

According to a Body Labs 2016 retail survey, $62.4 billion worth of global apparel and footwear are returned annually because of incorrect sizing or fit. In these cases, the problems the customer experienced with the fit of the product meant that they were unsatisfied with the purchase and returned it. This was especially prevalent in e-commerce transactions.

All bodies are different. For brands, it is well-known that trying to create products that will fit anyone is fairly unrealistic, maybe impossible. While getting the perfect fit would require tailoring and hands-on personal attention, matching a customer with the best possible garment using historical data can still help reach higher levels of customer satisfaction and decrease returns. This process of fit matching using predictive analytics has shown promise in returning higher levels of customer satisfaction.

What Are Predictive Analytics?

The term **predictive analytics** encompasses a pool of techniques, from statistics to machine learning. The key characteristics are the use of historical data to predict future events.

Humans are really good at recognizing patterns. Having a "hunch" or "intuition" for something is generally more than just a dark art. Predictive analytics are modeled after the idea that by recognizing patterns in events that happened in the past, we can develop a framework, or model, by which we can predict events happening in the future.

A **model** in predictive analytics, as in other machine learning techniques, is a predictive algorithm used to simulate real-world activity. A predictive algorithm is a statistical method that uses historical data to infer what the data might be in the future. The use of a predictive model has two distinct stages: training and predicting, as shown in Figure 7-1.

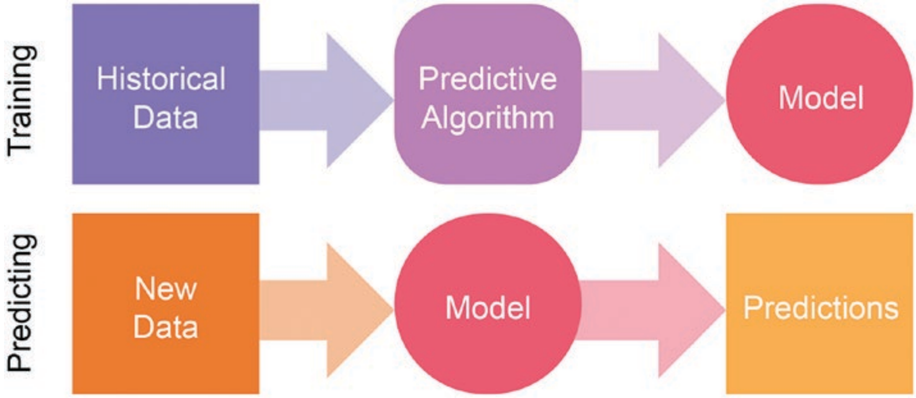


Figure 7-1. *The process of training a model and then using it to make predictions about future events*

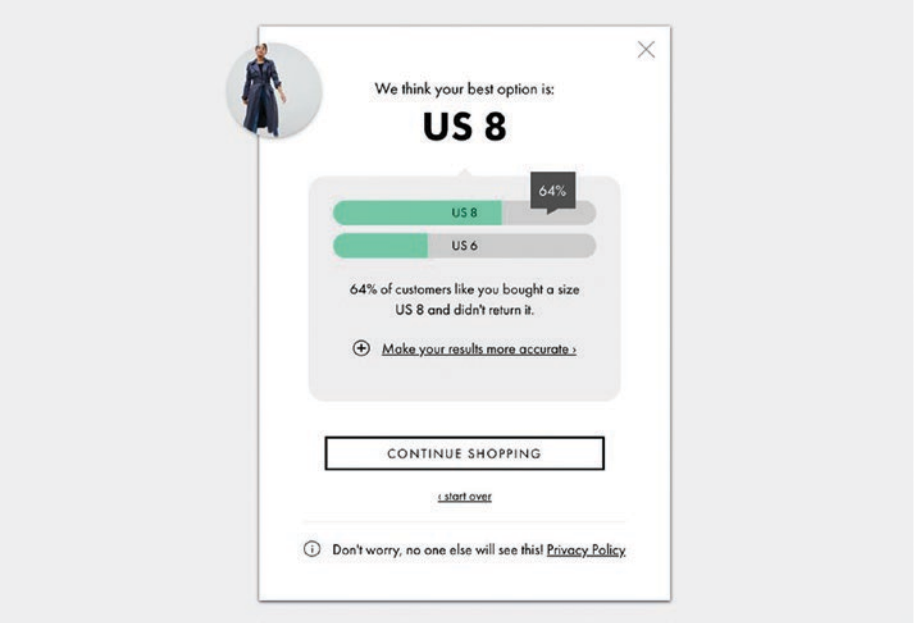
Learning Fit

Traditionally, size recommendations have been made on e-commerce platforms through a static sizing chart with dimensions such as waist, hip, and bust measurements. These charts, especially on retail platforms that sell multiple brands using different measurements, have lacked the accuracy that customers need to purchase garments confidently.

Size recommendations apply predictive analytics to match a customer with the size that will fit them for a specific garment.

The way predictive analytics can be applied is by creating a recommendation engine, like those explained in Chapter 6, that provides recommendations based on fit. Recommendation engines are an application of predictive analytics.

A customer might experience size recommendations in an online store by going through these steps: they open an e-commerce site and choose a garment to purchase. If the customer doesn't know their size, they can use the interface to gain insights about what other customers with similar body sizes purchased, as shown the screenshot in Figure 7-2.



We think your best option is:

US 8

64% of customers like you bought a size US 8 and didn't return it.

Don't worry, no one else will see this! Privacy Policy

Figure 7-2. *The Fit Analytics Fit Finder on the ASOS web site*

Fit Analytics is one company that provides fit matching and other insights as a service. It powers over 250 million size recommendations every month. Its Fit Finder interface can be found on the e-commerce sites of brands like The North Face, ASOS, and Tommy Hilfiger. After the customer inputs information about themselves (height, weight, age, and fit preference), Fit Finder will return a best-fit recommendation like the one in Figure 7-2. While this concept was new just a few years ago, it has quickly become the new normal on fashion e-commerce sites.

Other Applications for Predictive Analytics

For many companies, using predictive analytics is not a new idea. It can be used across many departments within a business, from sales to marketing to design. Here are a few other scenarios:

- Predicting which customers will continue making purchases from your business and which are likely to leave the platform
- Targeting marketing campaigns to those who are most likely to make purchases
- Identifying suspicious transactions and detecting fraud

Implementing Predictive Analytics Systems

Depending on the business and strategy, a fashion brand can implement predictive analytics by outsourcing and using third-party services like Fit Analytics, described previously. They can also start their own predictive analytics initiatives internally.

For brands that do want to implement their own predictive analytics practices, there are basic steps to take to start a project from scratch. Implementing predictive analytics is done in a series of well-defined stages: define the problem, collect data, create a predictive model, train the model, and then use the model to create predictions. We'll explore each of these steps in the next section.

The most important thing to do is to identify the business value in investing internal time and resources in a predictive analytics project. Without a clear goal, there are too many directions to go and not enough information to learn. To turn to a team and say, "Add predictive analytics to our web site" would be like saying. "Add thread to our clothes." It's too vague and poorly defined to be actionable. Immediately, they will turn to you and ask a bunch of questions about the thread, the business value, and the application: What kind of thread? Who's the vendor? Is it cheaper? Is it stronger? Can the factory use that thread? Which product line? Which category? Which color? Do you want to add that thread to all of our products?

If your role is management at one of these brands and you're looking to start a predictive analytics project, take a step back and evaluate your biggest pain points before diving in.

BETTY & RUTH WORKING WITH E-COMMERCE CLOTHING REVIEWS

At Betty & Ruth, we want to know whether garment size is correlated with customer satisfaction for that garment. In this problem, the input variable is the person's size, and the output is a rating of how likely that person is to be satisfied with the garment.

We created a hypothesis before beginning. We notice a pattern that correlated certain sizes and styles. We believe that we can use customer review data to determine customer satisfaction for a given size per style and make recommendations accordingly. For example, based on reviews for size Medium that say that the style fits small, we can make predictions about how other size Medium customers will respond and recommend sizing up through our web site.

Another benefit to these findings could be that the technical design team learns more about the customers they're developing product for. The information from this experiment might influence their grade rules and technical specs during the product development process.

Data Collection

Once there is a clear goal to the project, collecting data is the next step. No matter the company, you're probably collecting data about your customers and products all the time. **Data collection** is simply about getting the information you need to apply a predictive model.

How to Get Data

For experimentation, Betty & Ruth uses **Kaggle**, which can be a valuable resource for finding datasets. In many cases, these datasets are already prepared for use. Kaggle is a web site where people share and compete on data science challenges. There's a specific focus on predictive modeling and analytics. On this platform, companies and users will often upload clean datasets and challenges. This allows the data scientists on the platform to skip the time-consuming part of collecting and cleaning data and jump straight to writing code to analyze that data.

When we're working with our own data, there are often ways to export information such as customer reviews as a CSV. CSV stands for comma-separated values, and this type of document stores information in a table format by giving each record its own line. If that record contains more than one field, it is separated by a comma. CSVs are a very common file format for importing data.

Some e-commerce platforms, like Shopify and Magento, have a feature built into their Product Review apps for exporting reviews as a CSV file. Even for platforms that don't offer an easy way to do this, there are other resources you can leverage to get it. For example, several web sites and browser extensions will allow you to save reviews from certain e-commerce sites as a CSV.

Cleaning Data

For datasets that is aren't already cleaned and prepared, cleaning data is a vital step to getting quality results. **Cleaning data** refers to the process of removing outliers, spikes, missing data, and anomalies. Data with a lot of outliers is often described as **noisy** because the abundance of irrelevant data makes it harder to interpret the core information. The noise can significantly and unnecessarily change the output of your predictive model.

It might sound inconceivable that data would need to be “cleaned” to process, but it’s rare that you would have access to a perfect dataset related to your product and customers. Most experts in this space will tell you that it’s better to have clean data than a lot of data. This is true for all applications of machine learning. By some reports, data scientists can expect to spend up to 80% of their time cleaning and preparing data, depending on the project and its data sources.

When data is brought together from multiple platforms (such as Amazon and Shopify, for example), the structures might be different and the fields may not match up perfectly. This is another common reason for needing to clean data. Multiple sources can lead to discrepancies.

Missing data can also be the result of changes over time. On the Betty & Ruth e-commerce web site, we recently started listing fiber contents on every new product. All of the products before that change did not contain the fiber content data. Extracting statistics like “25% of our garments are made out of cotton” without first removing the garments that did not list fiber content, would mess up our statistics. That statement would not be an accurate representation of our product offering. If missing data is ignored or overlooked, which is easy to do (especially in large datasets), the output predictions will be wrong.

Another reason that a dataset needs to be cleaned before being used is because of the presence of **outliers**. Maybe a person who is otherwise a size 0 orders a sweatshirt that is a size XL and thinks that the fit is very snug. This is the kind of outlier that should be removed from your data because it will probably not be predictive of the behavior of the majority of customers. Outliers can be created through human error or machine error and sometimes won’t even make sense, given the variables.

Without a high-quality dataset, it is not possible to get high-quality results.

Data Visualization

There are a lot of existing ways for developers and data scientists to visualize data even before jumping in with machine learning and other predictive analytics. By using visualization tools, they can uncover trends in the data early on. Some example programs for visualizing data are Tableau, Chartio, Plotly, Infogram, and Google Charts. Creating these **data visualizations** is helpful for understanding it by giving a visual index of the data.

In a screenshot from Kaggle, you can see how representing a variable in a simple bar graph gives more insight about the dataset than looking at a list view of the same information. Figure 7-3 shows the age distribution of this particular example.

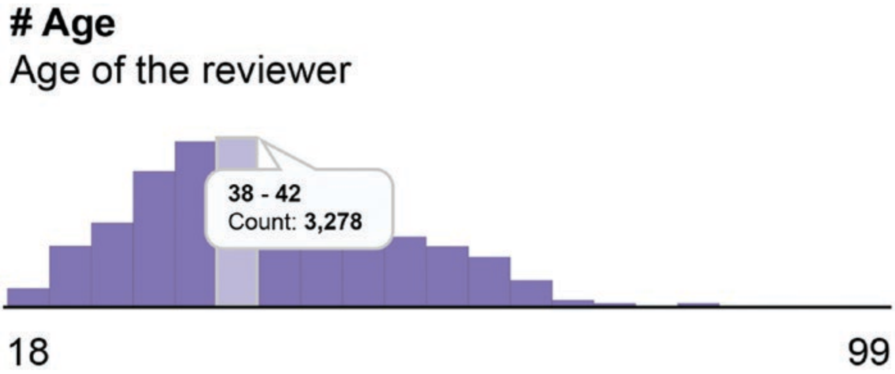


Figure 7-3. *A histogram based on the age variable of an example dataset*

It isn't always obvious at first what type of model you'll need to build. Some exploration and experimentation will have to be done to find useful insights about the dataset and establish more informed hypotheses before getting started on a machine learning project.

Models

In a few paragraphs, it is impossible to explain the challenges that could arise in the actual building and implementation of a predictive model. Some use cases will have easier outcomes to predict than others. Not all predictive analytics models require machine learning, but it is a useful tool, especially in dealing with large amounts of data.

Fortunately, we probably don't need to build a predictive algorithm ourselves to get started using predictive analytics. There are tools available that can be used without even needing to code. They give users access to out-of-the-box algorithms that work for a wide array of problems. Out-of-the-box techniques are useful for getting started, but in practice may not give you the best result.

There are some common algorithms that can be implemented out of the box. They usually fall into two high-level categories: classification and regression. You can learn more about the actual algorithms themselves on blogs such as towardsdatascience.com or in books about machine learning and data science.

Platforms such as MLJAR have these algorithms available and ready to use. The platform provides a straightforward graphical interface for uploading data and running experiments and predictions. While it does automate many of the processes required to get up and running, the tools are built for people with domain knowledge in machine learning as their audience. However, a growing number of tutorials are becoming available to those who are curious and want to learn more and experiment without learning to code.

Enterprise Tools

Companies have built graphical user interfaces to make experimentation with data science easier than ever to implement. These interfaces by no means take the work out of implementation, but they make it easier to get started quickly without needing to use **command-line interfaces** (**CLIs**), which can be daunting for nonprogrammers. Companies like MLJAR are making interfaces so that as long as you understand core concepts, you can get started. Your data science team might use these tools to streamline their process.

There are also companies, such as DataRobot and RapidMiner, that are offering artificial intelligence for enterprise. These tools usually have an even further developed GUI, making it simpler to use. While they might have less transparency into what's going on behind the scenes, they can help solve large and common problems that might be expensive and specialized to tackle independently—like fraud detection, for example. For more information about the growing number of AI enterprise tools, refer to Chapter 12.

Technology Blogs at Fashion Companies

Several methodologies can be used to match the measurements of a person and a garment. While this chapter does focus on predictive analytics as a methodology particularly for fit evaluation, it's important to mention that variations on this method have also been explored within the fashion industry.

Stitch Fix uses an interesting approach to solve for fit, which is described on its engineering blog. In one example, for each garment, there are three "answers": it's too small, it's too large, and it fits just right. If a customer describes themselves as a size 4 but say a size 4 garment is too large, that garment may be ranked as an in-between size, such as 2–4, and vice versa. Both the user and the garments are being ranked on a measurable scale. Both variables are moving targets, and the system is continually adapted based on information gathered about the user.

Software companies in the fashion space often share their techniques, approaches, and experiments publicly. You can read more about Stitch Fix's approach on its technology blog, `multithreaded.stitchfix.com`.

Stitch Fix isn't the only company writing online about how it is solving problems such as fit in fashion. With a small amount of effort, you can find information about the inner technology at a lot of large software-based fashion companies. A few more examples are Lyst (you can read about its tech at making.lyst.com) and Rent the Runway (whose engineering blog can be read at dresscode.renttherunway.com). These blogs can provide a source of inspiration and information for other companies looking to update their practices or simply gain insight to some of the mechanics behind these companies.

Data Responsibility

This section might seem a little out of place for this chapter, but when dealing with personal information, companies have a responsibility to protect their users. Especially in recent times, awareness over privacy and security are becoming more prevalent among consumers and corporations alike.

Recent events in this area are driving companies to take security and privacy more seriously. Several major Internet security breaches have left millions of people exposed; for example, the Equifax hack in 2017 exposed the social security numbers of hundreds of millions of Americans. These situations pose a crisis for both financial theft and identity theft. There have also been efforts toward improving privacy and security for Internet users to combat them.

General Data Protection Regulation

In 2018, the European Union (EU) introduced a new policy called the **General Data Protection Regulation** (**GDPR**). You probably heard about it, because you would have received an e-mail from almost every account you had, letting you know about privacy policy updates. While it applies only to businesses that have European customers, it has forced a lot of large technology companies to make changes across their business. It covers a wide range of measures to protect user data by keeping it encrypted and private and to give users the ability to opt out of having an account and have their data permanently deleted. It is most strict about regulation around personally identifiable information (PII), which could put the physical safety and privacy of individuals at risk.

Data and Third-Party Vendors

“No one’s thinking of us; we don’t matter” is no longer a valid way of thinking about security and privacy for businesses on the Internet. That approach leaves businesses more vulnerable to attack because they are low-hanging fruit. It also increases the likelihood of vulnerabilities down the line, when the business does gain traction or publicity.

Using third-party services is common practice, especially for fashion brands whose core competency isn’t technology. What those vendors do with customer data is still the brands’ business. Consumers are becoming more vigilant and expect at least basic levels of privacy and security. The downside to ignoring the security practices of third-party vendors is that if they get hacked, it could reveal sensitive customer data and destroy customer trust in the brands partnering with them.

Legal

For businesses using recommendation systems, ad retargeting, and other marketing tracking strategies that use user data, their Terms of Service and Privacy Policy should reflect how the data will be used.

Summary

Predictive analytics can provide powerful tools for fashion businesses, and the applications are endless. The most important aspects of implementing an effective and accurate model are having high-quality data and a clear business goal.

With the introduction of new graphical user interfaces for data science and predictive analytics tools, it's becoming easier for businesses to implement powerful machine learning methods.

While implementing predictive analytics and collecting and analyzing data about users, it's a corporation's responsibility to be mindful of data privacy and security. It will prevent financial and personal harm for both the corporations and their customers.

Terminology in This Chapter

Cleaning data—The process of removing outliers, spikes, missing data, and anomalies. Without clean data, predictive analytics models run the risk of inaccuracy.

Command-line interfaces (**CLIs**)—On every computer, there's an interface, also called a *shell*. The CLI is not commonly used by a general computer user, because most users prefer a graphical user interface. However, for users who do a lot of programming it can provide a powerful interface for controlling programs and the operating system of the machine.

Data collection—The process of gathering information for a specific use. Often in the post-collection process, that data might be mined and cleaned. (See Chapter 9 for more information about data mining.)

Data visualizations—Charts and other graphics that help humans understand the information embedded within a set of data.

General Data Protection Regulation (**GDPR**)—A data privacy regulation passed in 2018 by the European Union to protect the rights and data of Internet users.

Kaggle—A platform for predictive modeling and analytics challenges that was acquired by Google in 2017. Corporations and individuals have hosted challenges on the platform in which they prepare the data and describe a problem; then data scientists can compete to prepare the best model.

Missing data—Information gaps in a dataset. Missing data is important to be aware of because it can impact the quality of the output in models applied to the dataset.

Model—A mathematical representation used to simulate real-world activity. A machine learning model refers to the resulting artifact of a trained machine learning algorithm.

Noisy data—Data with spikes, outliers, anomalies, and missing data. This is often the state of raw data before it has been cleaned.

Outlier—Exceptions in data. When something is far off from what is normal, it could be because of machine error, human error, or incorrect interpretation. If a dataset reports a human user to be 300 years old, but everyone else in the dataset is within a more feasible age of 25–55, we might assume that something went wrong with that data. It is an anomaly, and we can remove it from our dataset in order to prevent inaccuracies.

Predictive analytics—A wide range of topics and methods, from statistics to machine learning, that use historical data to predict an event in the future.