Dataloop

The Top 5 Data Challenges in Retail: Solved!

dataloop.ai

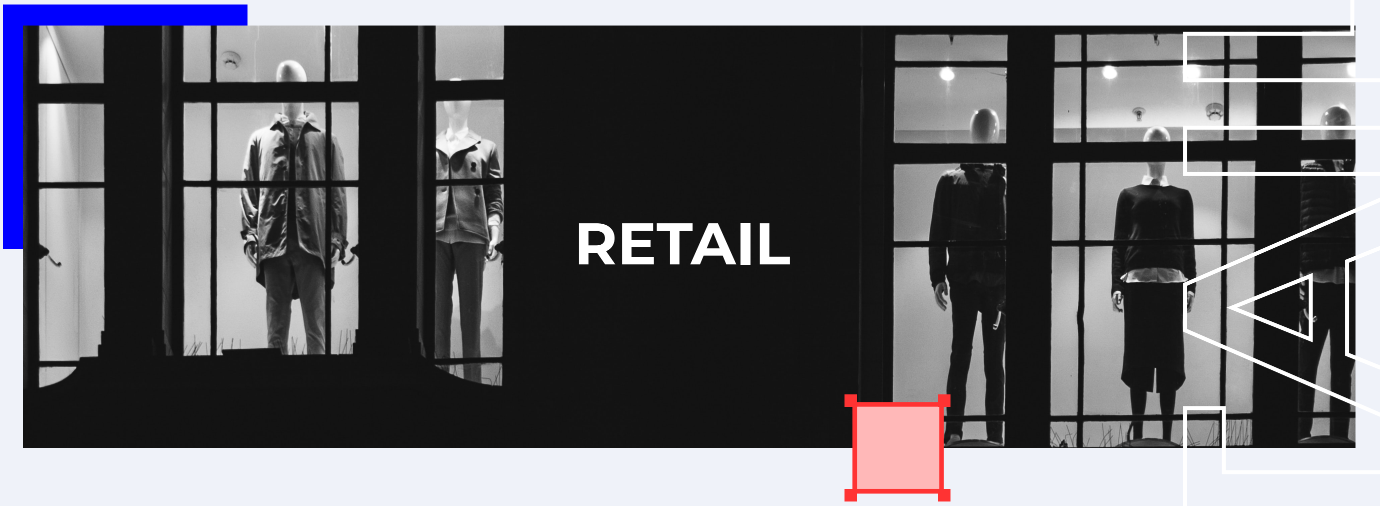
AI-driven applications are rising in importance in every industry, and retail is no exception. Retailers have been paying careful attention to the progress that AI is making, and they can see the opportunities it holds. It's no surprise that AI for the retail market is predicted to reach $23,426.3 million by 2026, with a CAGR of 33.7%.

AI is already having a dramatic impact on numerous retail use cases, from improving logistics and demand forecasting to increasing the accuracy of personalized offers and promotions. A recent study by IBM found that retail executives who have already implemented intelligent automation are reporting benefits across the value chain, with 83% seeing an improvement in their decision-making, 79% enjoying increased operational efficiency, 70% experiencing a rise in revenue growth, and 66% succeeding in improving customer experience.

EXECUTIVES USING INTELLIGENT AUTOMATION TODAY RANKED IMPACTS THEY ARE EXPERIENCING

83%	Improve quality and speed of decisions
81%	Increase operational agility
79%	Increase operational efficiency
71%	Extend and expand capabilities
70%	Increase revenue growth
67%	Enable insights through enhanced analytics
66%	Improve customer experience
65%	Reduce risk by improving visibility and processes
64%	Enhanced insights from integrated data sources
64%	Improve employee experience
63%	Reduce costs

The potential is enormous. AI and machine learning promise to impact the accuracy of retailers' promotions and targeted offers; reduce money wasted on unsold products; optimize pricing decisions; streamline and speed up payments through automated checkouts and contactless payments; cut down on missed sales due to a lag in restocking; improve in-store and online customer service; and more. **Over 75% of retail executives expect to see AI impact supply chain planning, demand forecasting, customer intelligence, and targeted marketing.**



85%
Supply Chain Planing

85%
Demand forecasting

79%
Customer intelligence

75%
Marketing, advertising and campaign management

73%
Store Operations

73%
Pricing and promotion

But it’s not easy to reach this AI nirvana. At Dataloop, we have assisted many retail AI technology vendors to tap into the power of AI and ML, and we’ve seen them repeatedly face the challenge of **efficient data labeling and data management at scale.**

Successful ML projects depend above all on the data that they receive. You need accurately-labeled data to train ML models correctly, so that they can accurately predict the pattern recognition that you rely on to guide your projects. But converting your raw data into labeled data means applying attributes, complex ontologies, and various annotation types to the heap of data which is constantly arriving.

Retailers have no shortage of data — in fact, the problem is that they have too much of it. Data points flood in from customer loyalty programs, IoT devices that track inventory levels, delivery people and couriers, marketing campaigns, PoS systems, cameras, social media monitoring tools, and more. The real difficulty is to process this mound of raw data into usable, labeled data. 19% of businesses adopting AI report that lack of data and poor data quality are their biggest obstacles to success, while employees involved in AI projects say that processing, cleaning, and labeling data for ML models takes up 80% of the total project time.

Data labeling is the foundation of ML success; when it’s not carried out effectively, the entire project is doomed. In our own experience, we’ve seen how project after project has advanced once their data labels improved in quality, accuracy or both.

Because data labeling plays such an important role in AI development, and specifically as there are retailers that plan to tap into the benefits of AI, we took a closer look at the underlying factors which hold enterprises back from efficient data labeling. **We found 5 main issues that hamper successful data labeling:**

→ Workforce management

→ Dataset quality

→ Financial obstacles

→ Data privacy

→ Smart tooling

Once we understand the roots of these data labeling challenges, we can develop ways to solve them and increase success rates for AI and ML retail projects.

01 HUMAN WORKFORCE MANAGEMENT FOR DATA LABELING IN THE RETAIL INDUSTRY

Workforce management is one of the biggest data labeling challenges for **two reasons:**

> - **You need enough workers to process the massive volume of unstructured data**
> - **You need to keep quality high and consistent across a large and varied workforce**

Smart retail solutions depend upon ML models that can identify products quickly and accurately so that they can predict demand with a low margin of error. It's relatively easy to track stock levels in a warehouse or storeroom, but in a busy store it's far harder to identify items on a shelf.

You want your system to see instantly whether items are missing, or placed in the wrong area of the store, so you need enough data for your model to learn to recognize the type and category of each item just through a camera image or video. That in turn depends on a workforce which is familiar with all the items you carry, so that they are able to recognize and label each data point correctly.

When it comes to data labeling, speed and quality are equally important. Enterprises need to walk a fine line between expanding their labeling workforce fast enough to keep up with their flood of data, and ensuring that a large, disparate group of workers all receive the training and oversight they need. We've seen successful startup teams, and even enterprises, begin by managing data labeling and other data processing needs in-house. This works, but only as long as datasets remain a manageable size.

> "
>
> **When it comes to data labeling, speed and quality are equally important.**

As companies grow, their labeling workload grows too. It makes sense for a growing retail solution to scale up their external data labeling workforces, but a larger workforce brings **new issues:**

- **Training many labelers for their assigned tasks.**
- **Distributing work seamlessly across large, varied teams and dividing tasks up into individual assignments.**
- **Tracking individual progress without losing track of the project as a whole.**
- **Ensuring smooth communication and collaboration between labelers and data scientist(s) to maintain quality control, validate data, and resolve workforce issues.**
- **Overcoming language, geographic, and cultural barriers between labelers who might fail to annotate data correctly because they'll miss certain cultural cues.**

02 MANAGING CONSISTENT QUALITY ACROSS RETAIL DATASETS

Visual product recognition is in high demand, but it depends on high quality data in order to produce reliable and accurate results. A rising number of retailers are building complementary apps that help consumers to find the products they want, compare colors, styles, and sizes, and check stock levels in their nearest store. Others are using in-store product recognition apps to track stock levels, manage in-store product placement, and support smart checkouts.

Retailers need to consistently identify and tag large datasets of product images, so that these tasks can be carried out successfully, but the datasets are not just large, they are also constantly changing. **30% of products are changed in some form or other every quarter, creating a mix of old and new products and generating an enormous number of different product versions and properties for accurate labeling.**

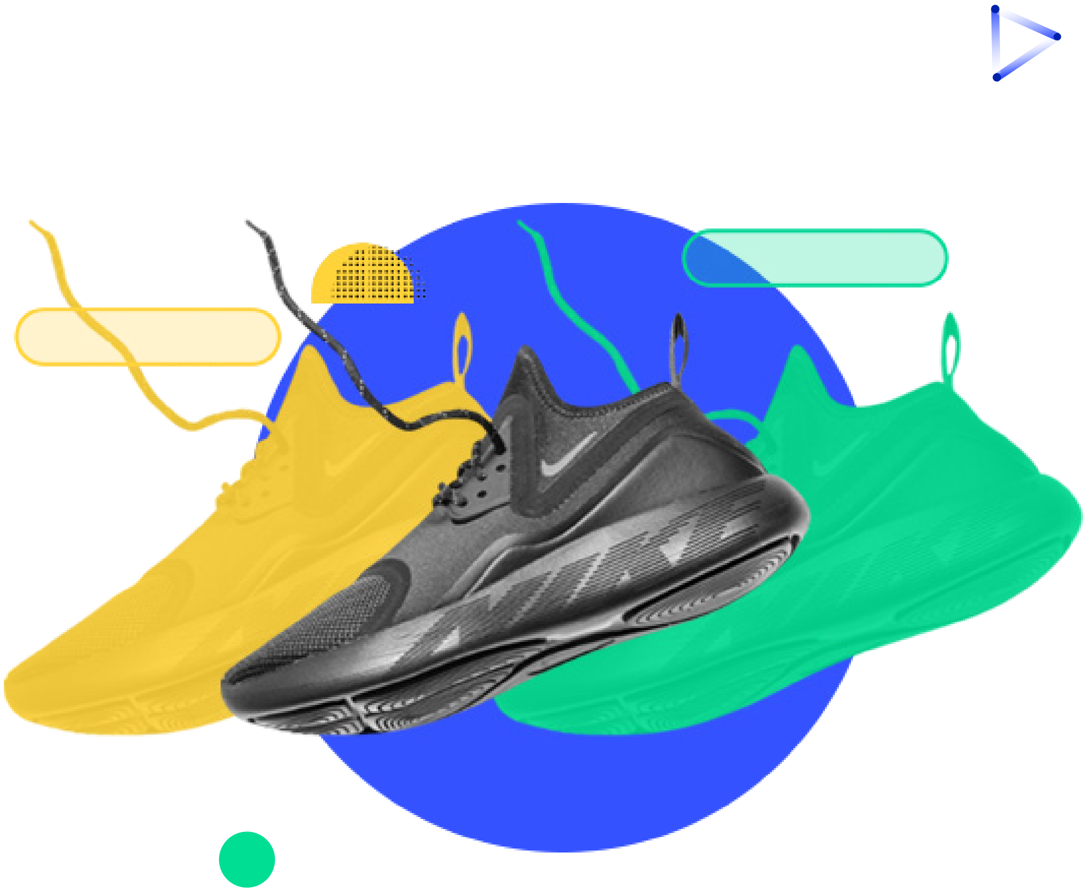
For retailers to tap into good, reliable data, they need high dataset quality across two types of data: **subjective data,** and **objective data.** Subjective data includes labels that don't have a single definitive "truth," while objective data labels do have a measurably correct answer.

SUBJECTIVE DATA

Subjective data labeling concerns how to define the label in cases where there's no single source of truth. We frequently see how the labeler's domain expertise, language, geography, and cultural associations can all influence the way that they interpret the data before them.

Imagine that a large shoe retailer wants to build a shoe shopping app that can automatically detect a running sneaker. The subjective data challenge is deciding whether a given sneaker is a running sneaker, a basketball sneaker, a hiking sneaker, etc. The retailer needs an enormous amount of data to successfully train the models so the app can distinguish between each type of shoe.

Since there's no single "correct" answer for subjective data, the data ops team needs to set clear instructions to guide how the workforce understands each data point.



OBJECTIVE DATA

Unlike subjective data, objective data does have a single correct answer, but it still presents challenges. For a start, there's a risk that the labeler might not have the domain expertise needed to answer the question correctly. Then there's also the fact that there can be multiple "correct answers," but only one is the one you're looking for.

To continue our example of a shoe retailer's running sneaker app, it would have to work out how to define the right attributes to classify a running sneaker. A sneaker could be appropriate for running and for hiking, or it could be a multi-purpose sneaker that is also suitable for runners. The decision is not always easy, and without good directions, the labelers wouldn't know how to correctly label each item.

Finally, it's impossible to entirely eradicate human error, no matter how good your dataset quality verification system.

This leaves data science teams to find ways to resolve both subjective and objective quality issues by setting up a closed-loop feedback process that checks for errors.

03 BALANCING THE FINANCIAL COST OF QUALITY LABELED RETAIL DATA

When asked why their AI projects are failing, 26% of enterprises blamed a lack of budget. Our own experience backs this up; we've frequently found businesses that struggle to manage a data labeling budget because there's no standard pricing or established metrics to use for comparison. We've often encountered this lack of transparency into exactly what enterprises are paying for in their data labeling projects, whether it's in-house or contracted out.

Organizations that outsource data labeling generally need to choose whether they'll pay for data labeling per hour or per task. Paying per task is more cost effective, but it incentivizes rushed work as labelers try to get more tasks done in a given timeframe. In our experience, most enterprises prefer to pay per hour.

The high price tag that accompanies quality data labeling is often off-putting to retailers. Other AI applications are handling a relatively narrow list of object types, but retail inventories in even small stores can run to thousands of product categories, each with thousands of items. You need very complex ontological structures and relationships to make it possible to easily find the right product name in a list of several thousand, and creating them pushes the price up even higher.

> "
>
> **The high price tag that accompanies quality data labeling is often off-putting to retailers.**

To add to the complexity (and thus the cost), self-service checkouts rely on static cameras that take images of the product from a number of angles at the same time. This makes the data pipelines even more complicated, since the labelers need to select the best images before they begin labeling the actual products.

However, the price of data labeling needs to be set against the cost to the business of not having sufficient data to run processes at maximum efficiency. This is especially crucial when it comes to retail logistics and supply chain management.

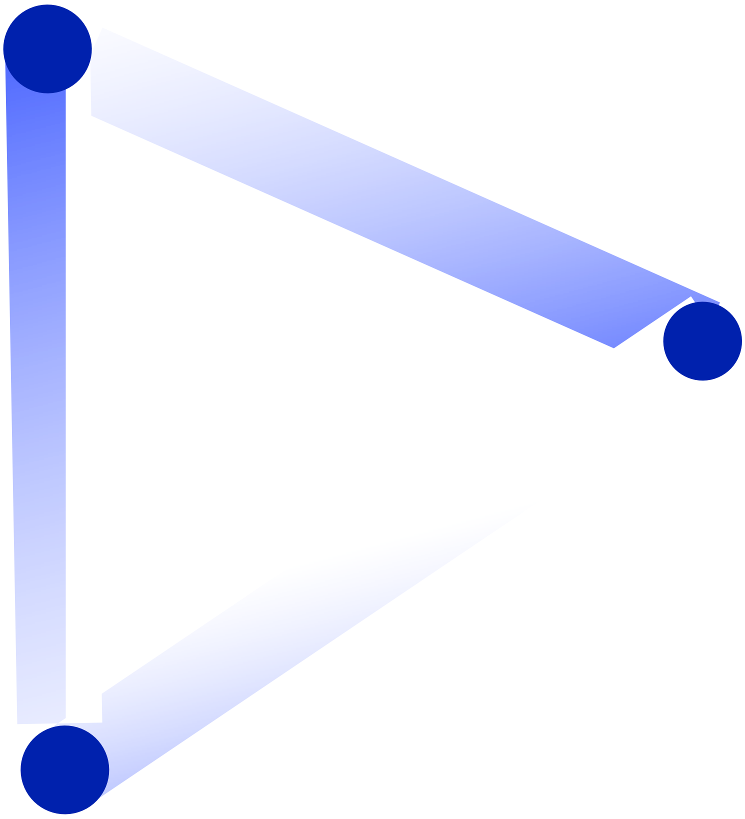
Advanced ML models that are trained effectively on accurate and adaptable data are able to track stock levels, customer demand, transport speeds around the world, last mile delivery (where relevant), and more, so that they can speed up delivery times, cut logistics costs, and make sure that they never run out of high-demand items or are left with unsold items.

Without sufficiently accurate and sophisticated models, the size of the datasets and required liability will force self-checkout retailers to place human employees, often remote cashier workers, to work in tandem with the AI models to back up the systems in real time and ensure billing is 100% accurate. **From our experience, this could drag self-checkout response time from 30 seconds to 20 minutes.**

Retailers also face higher data labeling costs because of the need for repeated long tail edge cases analysis. This is because AI-powered self-service checkouts are constantly being attacked by shoppers, trying to fool the system to get products for free. For example, a shopper might take 2 products but only pay for one of them. Training the system to recognize this fraud demands a very difficult annotation definition of shopper actions, as well as those needed for product annotation definitions. Because users will keep looking for new loopholes in the modeling environment, **retailers will keep having to refine definitions and properties for greater accuracy, driving labeling costs ever higher.**

Poor retail logistics causes $1.1 trillion of losses for retailers each year, but when done well, it can maximize profits and increase revenue. A significant part of Amazon's success rests on its "just in time" approach to restocking, which means that it doesn't have to pay for warehouse space for excess stock, nor turn away customers because it runs out unexpectedly. Morrisons' grocery stores reduced on-shelf gaps by 30% with better ML modeling.

Small, in-house manual data labeling teams are expensive, due to the time and training needed to reach true expertise. As data quantity grows exponentially, prices rise in tandem. We know that it's impossible to predict the final volume of data for processing, which only adds to the complexity of balancing the cost of labeled data against the cost of not having it.



04 COMPLYING WITH PRIVACY REQUIREMENTS FOR RETAIL CUSTOMER DATA

One of the biggest and most important use cases for ML in retail is to improve the customer experience and sharpen the accuracy of marketing. By analyzing data on customer behavior and preferences, retail marketers can segment their customers more effectively to personalize content and promotions for each tranche of their audience without guesswork.

However, using data to run ML models that improve smart retail can be a minefield of compliance issues. Data confidentiality and privacy regulations like GDPR, DPA and CCPA restrict the data that businesses can draw on. Just as companies are gathering even more data and tapping into its value, data confidentiality laws are multiplying worldwide. All customer data has to be anonymized, which slows down the data labeling process and increases the amount of work. When it comes to labeling unstructured data, this includes anonymizing or blurring faces, license plates, and any other identifying data that might appear in images.

Enterprises are obligated to comply with principles like "processing data lawfully, fairly and in a transparent manner in relation to the subject matter." They need to ensure that their data is secure, which means preventing data labeling workers from accessing it from an insecure device, downloading and transferring it to an unknown storage location, or working on data in a public location where it could be viewed by someone without security clearance.

> "
>
> **They need to ensure that their data is secure, which means preventing data labeling workers from accessing it from an insecure device...**

We've seen that in practice, this usually means that data has to be managed and stored on premises and accessed only from approved devices. We realize that this makes it challenging for organizations dealing with data to outsource tasks to third party data labeling providers while still complying with regulations. It limits where employees can work, and adds yet another layer of complexity to workforce management. It's essential that you and your data labeling service know which laws you need to follow and how to ensure compliance, whether your data is stored on-premise or in the cloud.

05 MAINTAINING SMART RETAIL DATA MANAGEMENT ML TOOLING AT SCALE

High quality data relies on a combination of well-trained workers and smart tooling, such as AI-assisted data annotation, automation, data management and data pipelines. We see that as AI reaches more domains and is expected to understand more human tasks, tooling requirements keep rising.

In our experience, organizations that begin with tools built in-house often discover that their annotation needs keep growing, so they have to work harder than expected to keep up. For example, a smart retail solution might want to develop in-house models that amass large amounts of data to teach models to classify and cluster different fruits and vegetables into the right categories.

The cost to develop this tool might be negligible, but what about when the amount of data grows?

Environmental changes must also be considered. Take the COVID-19 pandemic for example. The dramatic changes in consumer trends had a knock-on effect on warehouse management and customer behavior analysis costs. Models that had been ticking over smoothly suddenly had to consume a lot more data at a much faster rate in order to spot patterns and make predictions in the uncharted environment, raising costs for businesses. By using external platforms that support dynamic workflows, you can draw on existing tools and integrate pre-established functions into your processes, thereby cutting the cost of training new models.

In-house data labeling and data management tooling demands a great deal of both money and time to keep on supporting and extending the software, and over time, this level of investment in a tool that isn't core to the business distorts your focus as a company.

When you face the question of whether to build or buy, you need to consider the full impact of your decision. As stated above, building an internal data tool means risking paying over the odds in terms of time, cost of going to market, and continual maintenance so you don't fall behind. Your proprietary tools might cost too much to adapt to changing circumstances. However, before you buy you need to consider whether the tools you select provide all the services that you're seeking.

That's why **it's critical you find a platform that is robust enough to evolve with your projects, but also mature enough to ensure stability.**

> **The dramatic changes in consumer trends had a knock-on effect on warehouse management and customer behavior analysis costs.**

THE RIGHT DATA PLATFORM HELPS RETAILERS OVERCOME DATA LABELING CHALLENGES

As we've repeatedly discovered, data labeling and data management underpins success in every AI retail project. Finding the right platform to overcome these 5 most common data labeling challenges is vital so that you can efficiently create the workforce and establish the data labeling rules you need, keep track of costs and data confidentiality compliance, and tap into the necessary tooling for each data processing task.

Dataloop steps up to the plate to provide solutions for each issue. We grasp the critical role that effective data labeling plays for your organization, and have a deep understanding of the obstacles that lie in your way.

Our authentication and permission controls, military-grade encryption, and single point of access keep your data confidential and secure to underpin advanced customer management and smart checkouts, while our auto-scaling infrastructure is powerful and versatile enough to keep up with your growing data operations as your business scales and you add more consumer-facing shopping apps and internal supply chain solutions.

"

We reduce the challenge of workforce management and lower the costs of data labeling, bringing solutions like retail logistics within your price range.

Our proprietary platform integrates human and machine intelligence with cross-functional collaboration, closed loop feedback, and high data standards to ensure quality datasets for your product recognition and shelf management needs. We reduce the challenge of workforce management and lower the costs of data labeling, bringing solutions like retail logistics within your price range.