5 Reasons Most Machine Learning Initiatives Fail and How to Fix Them

pandio

Machine learning and artificial intelligence (AI): They're more than just buzz words. Today, companies that successfully adopt machine learning and AI have opportunities to scale their businesses at unprecedented levels, offering next-level, customized, intelligent services for clients.

Unfortunately, many companies shy away from the technologies and their benefits.

Why?

Because few can do it well and even fewer can do it in a scalable manner where they can actually generate profits off their investments.

Pull stat:

95% of businesses surveyed say they need to manage unstructured data.
Forbes, "Big Data Goes Big"

So why is adopting machine learning and AI so challenging? Because it's a relatively new concept for most companies and many don't have the skills, knowledge, or resources to successfully implement the technology or overcome the obstacles that can stop forward momentum.

But you don't have to miss out on all the benefits machine learning and AI can bring to your business.

With proper planning and strategies to overcome the most common obstacles—paired with a distributed messaging solution that can handle AI and machine-learning demands—you can soon be on your way to scaling your business in ways you previously just dreamed about.

VISIT **PANDIO**.COM

Here are
5 Reasons Most Machine Learning Initiatives Fail and How to Fix Them

Obstacle 1
Accuracy Issues

When we talk about machine learning and AI, big data means BIG data—and lots of data that must be ingested, processed, and distributed quickly. The more your company scales with AI and machine learning, the bigger your data gets.

When companies first adopt machine learning, many don't see the big picture or future scalability issues so they start with processes that may handle their short-term needs, but, as they grow over time, they begin to hodge-podge processes together.

Think of it Like This: It's a game of telephone. You set your data up with a simple line and you can send and receive messaging clearly and quickly. But as you grow and scale—and process more messages and data—that single line becomes many with a variety of connections, and before you know it, by the time your message gets to its endpoint, it has filtered through a variety of those processes and loses accuracy and data integrity.

When this happens, companies begin to question their data value.

- ✓ Where did it come from?
- ✓ Where does it traverse on its journey?
- ✓ Was it modified on its journey?
- ✓ If so, who modified it?
- ✓ Is the data traceable?

Overcoming the Obstacle: The key to overcoming this big data hurdle is within provenance—being able to pinpoint the earliest known history, or in this case, origin of where that data came from. You also need to be able to determine when and how the data changed over time and how it was aggregated into your systems.

The Pandio Solution: Pandio, built on Apache Pulsar, delivers more reliability and durability with less data loss than other distributed messaging solutions such as Apache Kafka. In one recent study, for example, a tester ran similar data-loss tests on Pulsar and Kafka. Pulsar consistently successfully published messages with no disruptions, data loss, or data accuracy issues in more than seven different testing scenarios. The same cannot be said about Apache Kafka.

Pandio can store changes to your data indefinitely to always ensure provenance. If you encounter a data issue, you can easily discover when and where the data change occurred so you can quickly and effectively resolve accuracy issues, track data changes and transfers over time, and ensure future messaging issues don't arise from data loss or data accuracy concerns.

Obstacle 2
Accessing Your Data

Data silos are common in business. In smaller enterprises, that could be files stored on someone's desktop or laptop, or in larger enterprises, across a variety of servers or other storage devices. Not only do you generally have issues with accessibility, but that data also is often not uniform because depending on user or location, the data could be stored in a variety of formats or through disparate applications or systems that don't communicate with one another.
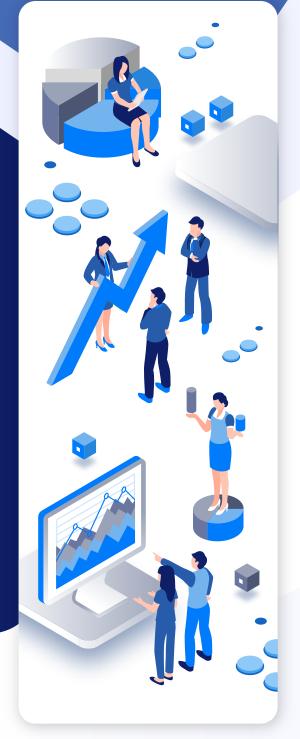
What happens when you need to quickly access all that data and make important, timely business decisions? With older distributed messaging solutions, it can be challenging to not just access that data, but enable it to communicate when and how you need it.

When it comes to adding machine learning and AI to your processes, you need immediate access to your data and you need to ensure that the data accessed is the most current and accurate data within your systems—and that all your data is speaking a common language.

Think of it Like This: Message storage is similar to selecting cardboard boxes for moving. If you buy a variety of box types and sizes, it can be challenging to properly load your moving truck when you're ready to go. You have to play a game of Tetris, pulling one box from one spot and stacking it in another. Eventually, it takes significantly longer to get all of your boxes organized and set up so you can move on. Whereas, if all your boxes were made from the same materials, were the same shape and the same size, you could quickly load your truck and be on your way.

Overcoming the Obstacle: Older messaging solutions spread your data across many servers and many locations, sometimes using a variety of formats and processes, and often limiting connections to that data. To overcome data access issues, you need a distributed messaging solution that makes access to your data simple and uniform.

The Pandio Pulsar Solution: Pandio Pulsar has hundreds of connectors that make it easy to access your data. And, if there is a connector you need that's not already built, the platform empowers users with the right tools and frameworks to quickly build new connections, breaking down barriers to ensure data uniformity.

Obstacle 3
Piecemealing Processes for Success

Some companies, when they stand on the end of the board ready to dive into machine learning and AI, try to build their current and future distributed messaging services with solutions that can't hold water.

Apache Kafka, for example, only streams messages. It can't natively queue messages. If you want to use message queueing, you may need to use an additional piece of technology, like RabbitMQ. Sure, that might work initially, but as you scale and process more data more quickly, you may discover those piecemealed processes start to feel a little bit like a house of cards, not a foundation for your business.

Think of it Like This: Let's say you have five critical operational systems for your company. It's all very linear. For processes to perform effectively, System 1 communicates to System 2. System 2 speaks a perfect language to System 3. System 3 then communicates to System 4. And System 4 perfectly delivers its data to System 5.

But what happens if you need System 1 to communicate to System 5 before data goes to System 2? Or what happens if one department can only access Systems 3 and 4, but they need the data from the endpoints of System 5?

That linear system no longer works for your company. In these instances, you'll need some pretty smart people on your team to craft workarounds to deliver those core functions. The more those individual needs change over time, the more work and resources you have to put into hodge-podging things together. Every time you need a change in your processes, it may feel like reinventing the wheel, which can add an additional layer of challenges to your business.

Overcoming the Obstacle: Instead of patching together old processes or piecemealing additional technologies to make an incomplete or outdated platform work (temporarily) for you, use a distributed messaging solution that can handle all your needs intuitively from within a single platform.

The Pandio Solution: When it comes to your distributed messaging solution, you'll want all the pieces of your system to communicate in a standardized way. By building your messaging pipeline within Pandio, you can eliminate the choke points encountered with other solutions.

If you need a straight pipeline from System 1 to System 5, you can create that, but then, whenever you need to change it or readjust your flow, you don't have to start all over. You can simply swap components in and out with consistent message deliverability and accuracy.

Obstacle 4
Labor Issues

Across the industry, technology is growing and changing so rapidly, professionals are having a hard time keeping up. There's a general shortage of skilled professionals and that's creating a number of gaps. With machine learning and AI, specifically, there's a growing communication trench between machine learning data scientists and operational professionals, known as MLOps.

The MLOps gap is creating cascading issues for machine learning and AI adoption.

PULL STAT:

Only 1 out of 10 data science projects make it into production.
VentureBeat, "Why Do 87% of Data Science Projects Never Make it Into Production?"

Think of It Like This: You want to tackle machine learning and AI with an internal team, so you know you'll need a great data scientist to help drive your program. First, because there is a shortage of skilled professionals in the industry, it's difficult to find a data scientist. And, if you can find one, and he or she is a great fit for your company, there's a decent chance that person doesn't have industry experience.
Instead of waiting longer to find an alternative, you hire the person without industry experience, but you have to set aside time for onboarding. That process could be three times longer than what it would take to onboard someone with industry expertise. The next hurdle is ensuring your new data scientist can access all the needed company data for the project. With siloed data and disparate systems in most companies, this can be a slow-moving and frustrating process for everyone. After a lot of delays, a new challenge pops up when your data scientist finally gets your model ready. **After handing it off to your team, do they understand what that model is? What it's designed to do? How to use it?**

The hiring challenges don't stop here.

You'll need an architect who can create a tight feedback loop to ensure your data scientist gets the data needed from the model. The data scientist will need to routinely analyze that data so things learned can be integrated into model updates. The goal is to make the model smarter and that requires a lot of strategy, patience, and often luck, when you're relying on people to do this for you.The reality is, this knowledge gap can be a death-knell for your machine learning and AI program. It's rare to have a successful MLOps hire and it may very well be the unicorn hire for business.

Overcoming the Obstacle: While internal teams bring a lot of value and business benefits to the table, sometimes you have to look beyond internal resources for success. Knowing there's a shortage of skilled labor professionals, your company may benefit from working closely with outside advisors who have created solutions that facilitate routine updates for your ML and AI models.

The Pandio Solution: Pandio, built on Apache Pulsar, is so intuitive it runs your distributed messaging for you. Its neural network looks at all the messaging within your system with a granular approach. The system reports on existing model effectiveness in real-time on a per-individual component basis. It then consumes that data and makes a prediction about what your system should do. It makes changes accordingly and then evaluates if that change was positive or negative. Over time, Pandio learns if it's making the right decisions and makes changes to constantly improve your system performance. In most cases this means that within 24 hours, Pandio delivers performance and reliability unmatched in the industry.

Obstacle 5
Big Data Costs Can Quickly Exceed Revenues

There's a general belief that when it comes to big data, the more data you gather, the more accurately you can produce results over time.

PULL STAT:
More than 150 zettabytes (150 trillion gigabytes of data) will need analysis by 2025.
Forbes, "Big Data Goes Big"

Think of it Like This: In year 1, your company successfully collects a terabyte's worth of data for your distributed messaging system. That's great! You've never had the ability to do that before. Since big data generally means more data for many companies, you set goals to double that in the second year, and then double the second year's data collection in the third. Unfortunately, a lot of that data is stored, at least partially, within disparate systems, for example, your content management system (CMS). The more data you retain within those systems, the quicker costs add up.

Overcoming the Obstacle: Big data that runs machine learning and AI doesn't necessarily equate to "all the data" or "retaining every piece of data." Just because by year three, for example, you have five times the data you had before you started doesn't mean all that data has value. You need a distributed messaging solution that can help you synthesize that data, analyze what you actually need to retain for your current initiatives, and then store what you need in one location while easily moonlighting the rest—which you can access again later if you need it.

The Pandio Solution: Pandio turns the old model of data storage used by messaging solutions like Apache Kafka on its head. With Kafka, for example, you have to purchase storage space for all the data you might use, quickly adding additional costs. And, worse yet, Kafka can't separate data storage from compute. So you either have to leave all your data within the Kafka system or you have to manage retention data on your own.

Pandio uses a tiered-storage option instead, meaning you can keep data within your Pulsar cluster for as long as your data retention policy says you need to, and then offload that data to a cheaper storage option. When you need that offloaded data, it's quickly and easily accessible within the solution. Not only does Pandio help you decrease data storage costs, the platform's other benefits mean you can keep scaling your data and your system without having to hire more people.

Overcome the Challenges for Machine Learning and AI Today

Adding artificial intelligence and machine learning into your business is easier than you think. Sure, from data accuracy and accessibility, and from brittle systems to added costs and labor concerns, you'll face some challenges. But, with the right direction and planning, you can quickly cut through the hype and put them to work for you.

Need help getting started?

VISIT **PANDIO.COM**

pandio