SURVEY

The Industry Is Ready for **Machine Learning Observability At Scale**

Delays, painful processes, and team misalignment all stand to benefit from a modern approach to ML monitoring and observability according to a survey of over 900 ML teams and technical executives

arize

Table of contents

Introduction

The AI revolution is here. Machine learning (ML) and artificial intelligence are relied on in nearly every industry today to increase profitability, productivity, and even save lives. In all, *IDC forecasts* that global enterprise spending on AI will eclipse $204 billion by 2025.

Unfortunately, investments in needed infrastructure may not be keeping pace. Many enterprises are shipping AI blind or relying on outdated model monitoring approaches to catch issues with models in production.

In order to understand the scope of the problem and provide insights on potential solutions, Arize AI – an early pioneer and leader in ML observability – conducted this survey of data scientists, engineers and executives.

The results speak to a distinct need for ML observability to quickly visualize where and why problems are emerging and enable faster root cause analysis when models fail.

Highlights

26.2%

of data scientists and ML engineers say it takes their team a week or more to detect and fix an issue with a model in production

48.6%

of data scientists say their jobs are more difficult in the wake of COVID-19 due to elevated drift, data quality and performance issues

50.8%

of data scientists and ML engineers want deeper capabilities to monitor model drift

87.6%

of teams report that their business executives cannot quantify the ROI of ML initiatives at least some of the time

Key Findings

Today's status quo is frequently painful and slow for ML teams
84.3% of data scientists and ML engineers say the time it takes to detect and diagnose problems with a model is an issue for their teams at least some of the time – likely exacerbated by a post-pandemic environment where drift and performance issues are elevated.

ML teams and enterprise executives need ML observability at scale
Over half of data scientists, ML engineers and technical executives say their teams would benefit from deeper capabilities around drift monitoring and troubleshooting as well as in explainability. Over four in ten want better model performance monitoring.

A chasm exists between executives and AI teams
Despite the fact that ML models are more critical to business results in the wake of COVID-19 for nearly one-third of those surveyed, 54.0% report issues with business executives not being able to quantify AI ROI often or somewhat often. Meanwhile, 19.3% of technical executives say that they often encounter issues around sharing data.

The industry has a way to go on AI ethics
79.7% of ML teams report that they "lack access to protected data needed to root out bias or ethics issues" at least some of the time, and nearly half (42.1%) say this is an issue at least somewhat often.

Explainability is important, but it isn't everything
Although technical executives and data scientists place high importance on explainability, ML teams in healthcare and ML engineers – who are generally the ones getting models into production and maintaining them once there – place higher importance on monitoring and troubleshooting drift and data quality issues.

Recommendations

Evaluate and implement an ML observability platform that helps expose and eliminate AI blindspots

Virtually all ML teams monitor known knowns – model metrics such as accuracy, F1, etc. Most also try to address black box AI — known unknowns — through explainability. However, often what's still missing are solutions that expose the issues teams are not actively looking for: blindspots, or unknown unknowns. True ML observability can help eliminate blindspots by automatically surfacing hidden problems before they impact business results. Rather than writing seemingly interminable queries to get to the bottom of performance degradation, teams using a modern ML observability platform can quickly visualize the full array of potential problems and perform root cause analysis in a few clicks. See *The Definitive ML Observability Checklist* for a list of product and technical requirements to consider when assessing an ML monitoring and observability platform.

Grow internal visibility, increase ML literacy, and quantify AI ROI

As AI becomes increasingly important to business-critical operations, the technical teams deploying ML models and their executive counterparts need to be in total alignment. Unfortunately, this is proving elusive according to the survey. The way out is straightforward: business executives need increased education, access to tools and digestible, relevant KPIs – including, most importantly, a way to quantify AI ROI. By tying ML model performance metrics to key business metrics and giving executives access to tools that visualize these metrics over time, ML teams can ensure broader buy-in for their efforts.

Implement a modernized data policy that grants AI practitioners access to protected data where needed

To act on AI bias or fairness issues, teams need to reliably quantify them. Modernizing policies around access to data and in some cases expanding permissions for data scientists is a worthwhile exercise. While many ML teams historically lacked access to protected class data for legal liability reasons, that may need to change precisely because such data across the full ML lifecycle is critical to delivering accountability and ensuring a model's outputs are not biased or discriminatory.

Recommendations (cont'd)

Do not rely on explainability alone; take a proactive approach to model performance management

Focusing on explainability in the pre-production phases of the model lifecycle – training a model and validating it before deployment – can be useful. However, continuing to expend the bulk of resources on explainability once a model is in production is of limited utility since it creates a passive feedback loop. In other words: while explainability is useful for sorting when troubleshooting model performance in production, it does not help you surface blindspots the same way that data quality monitoring helps proactively catch potential issues before a substantial shift in inference distributions occurs, for example. By setting up automated performance monitors across your model, ML teams can have a first line of defense — especially if able to A/B compare datasets and perform data quality checks. Drift monitoring across environments or prior periods of production can also be an early signal that model outputs are shifting. Leveraging these and other techniques, teams stay a step ahead.

Methodology

This survey was in the field between November 10, 2021 and December 31, 2021. In order to reach its primary target of AI practitioners (ML engineers, data scientists, etc.), the survey was promoted in relevant technical industry publications, newsletters, and community channels. Participation was incentivized through drawings for gift cards and other promotional items, such as a guided *"ML Observability Moment of Zen"* meditation session with MLOps.Community founder Demetrios Brinkmann.

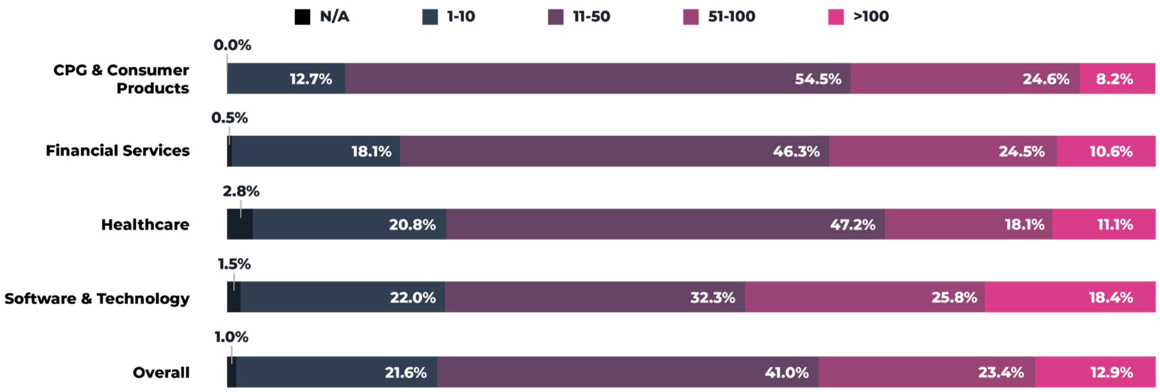
The survey's 945 respondents are comprised primarily of machine learning operations (MLOps) and other engineering professionals, including: data scientists (25.1% of the total), software engineers (24.1%), technical executives (14.4%), ML engineers (13.7%), and IT/devops/SRE (7.8%). Except where otherwise specified, the results in this paper are exclusively from data scientists and ML engineers.

It is worth acknowledging that given the nature and format of this survey, the findings may not be perfectly representative of ML teams in every industry; both selection bias and observational bias may be present in the data. That said, hopefully this report – the only one to focus specifically on ML observability for the teams that build, validate, deploy, and monitor ML models – is useful in providing a signpost on where the industry is headed.

The State of AI Teams By Industry, Role

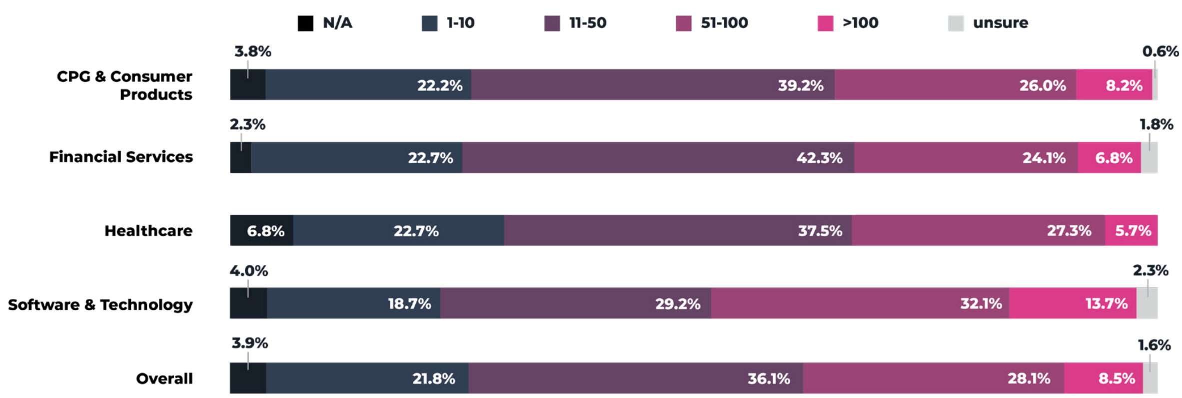
Overall, the most common size for ML teams is between 11 and 50 employees. Unsurprisingly, the software and technology sector has the largest data science teams – with 18.4% topping 100 employees – followed by the financial services and consumer packaged goods industries.

How big is your ML and data science team?



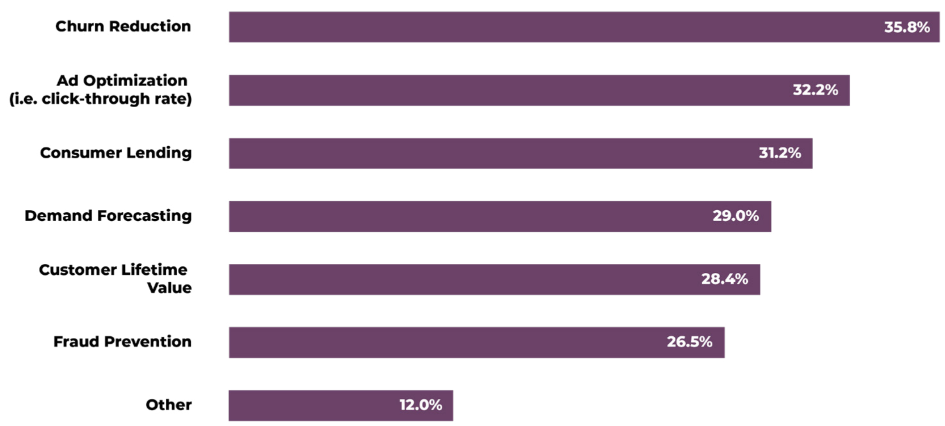
The software and technology sector also reports the most models in production, with nearly half (45.8%) of respondents reporting that they have over 50 models in production and over one in ten (13.7%) reporting over 100 deployed models.

How many models do you have in production?



The most common use case among those surveyed is churn reduction, with 35.8% of data scientists and machine learning engineers reporting that is the primary purpose of their ML models in production. The most cited "other" in the responses is recommendation systems.

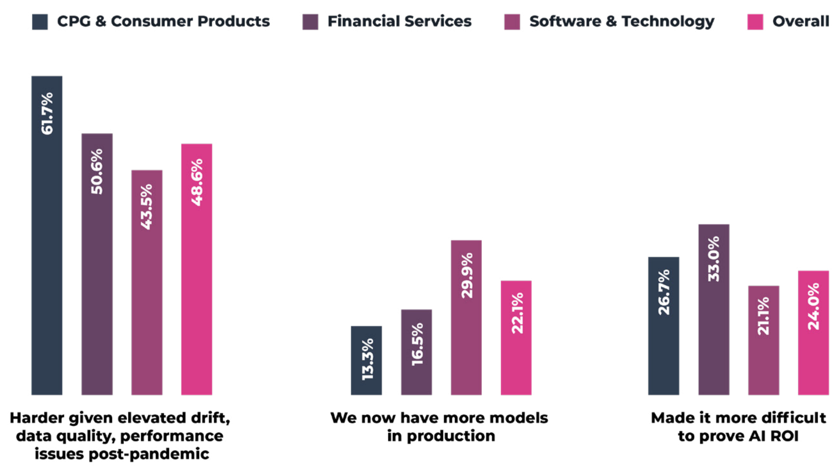
What is the primary purpose of your supervised ML models?



Data scientists and ML engineers report a wide variety of challenges in the wake of COVID-19. The most pronounced, according to the survey: elevated drift, data quality, and model performance issues post-pandemic. In all, 61.7% of ML practitioners in the consumer packaged goods (CPG) industry cite this as a challenge – more than any other industry.

Overall, 22.1% of teams have more models in production compared to before the pandemic, and 24.0% report that today's environment makes proving AI ROI more difficult.

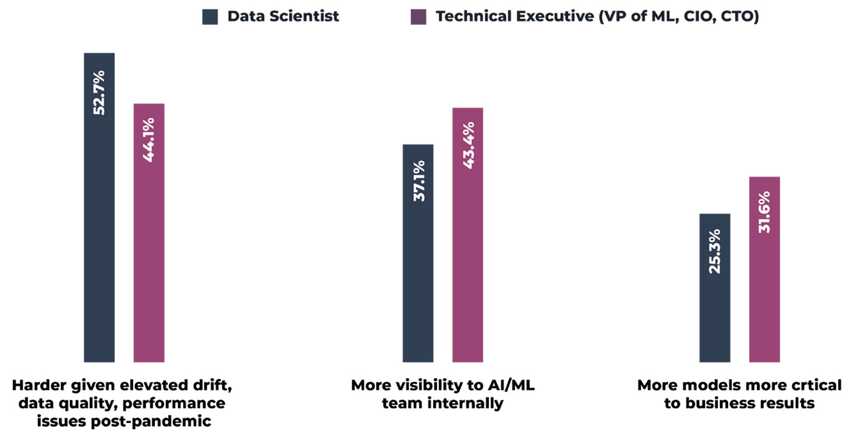
Which of the following best describes how your role has changed since the onset of COVID-19?



Despite the increased stress and stakes for many ML teams in the wake of the pandemic, one positive corollary is increased visibility and headcount. Over one-third (36.9%) of data scientists and ML engineers report more internal visibility for the AI/ML team – especially in the financial services industry – and nearly as many (29.8%) say their budgets increased.

Breaking results out by role, some interesting differences emerge between technical executives and data scientists. Specifically, CTOs and other technical executives are more likely than data scientists to believe that models are more important to business results; however, executives are also likely to underestimate how a post-pandemic environment makes the lives of data scientists harder.

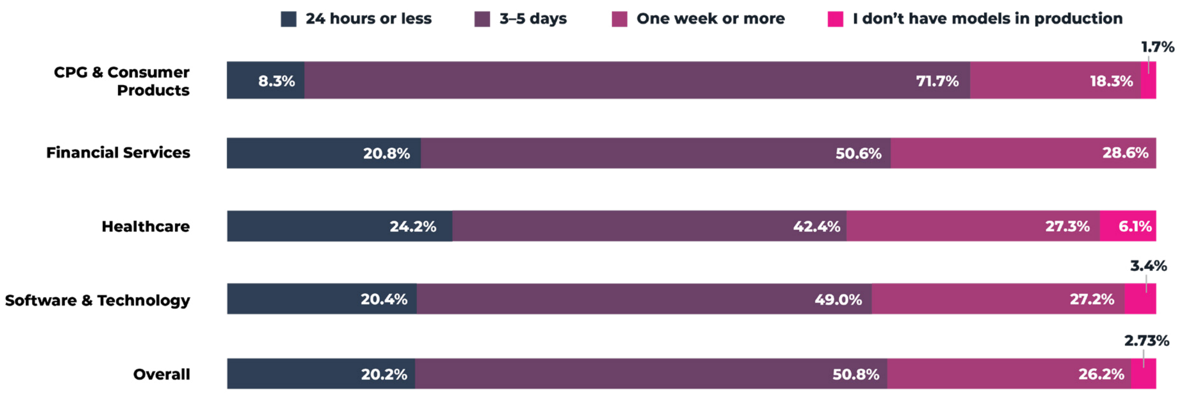
Which of the following best describes how your role has changed since the onset of COVID-19?



Industry Pain Points & Potential Solutions

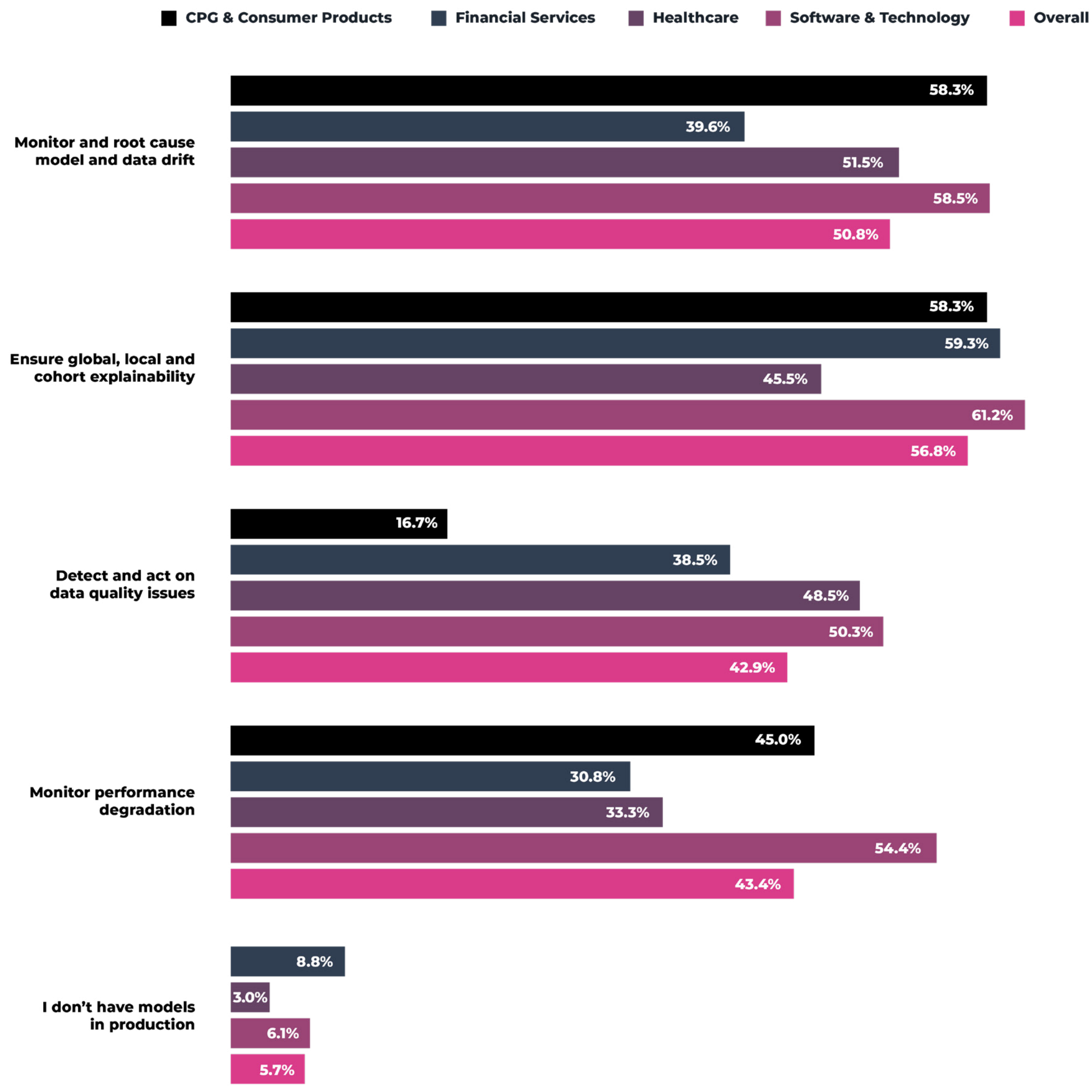
Over one in four (26.2%) ML practitioners admit that it takes them one week or more to detect and fix an issue with a model (i.e. retraining a model in production after detecting concept drift). Delays of one week or more are most common in financial services, followed closely by healthcare and technology. Among respondents who say that their models "prevent physical harm or bad health outcomes," 24.5% report delays of a week or more (note: this does not mean that models are live in production during this time, as many may fall back on a prior model version or no model while retraining occurs).

How long does it typically take your team to detect and fix an issue with a model (i.e performance, drift, data quality)?

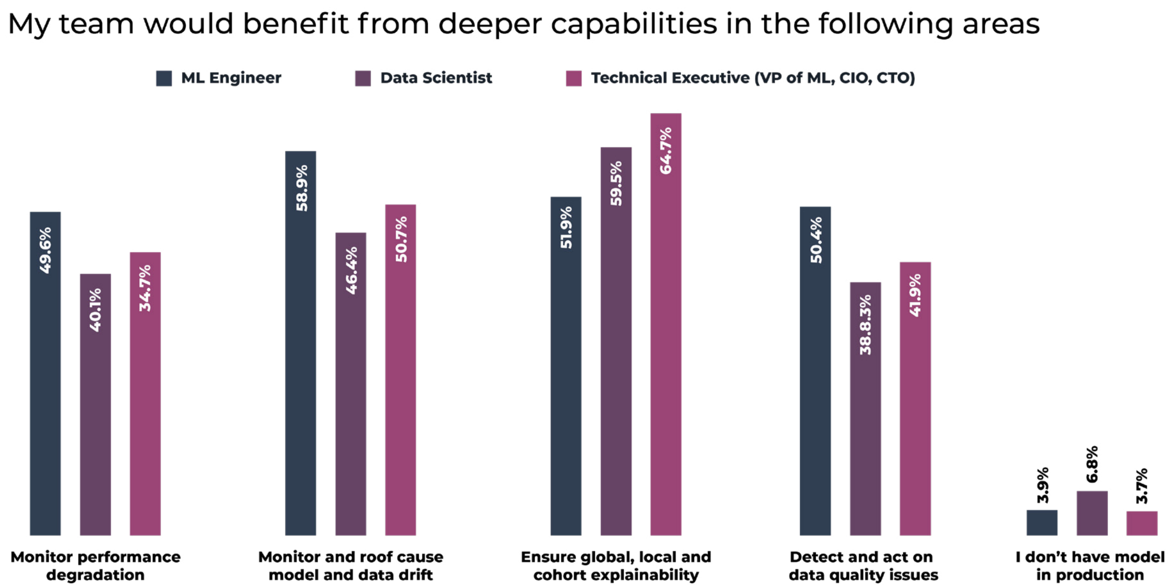


Across industries, ML teams are fairly consistent in wanting deeper capabilities in ML monitoring and observability – particularly for explainability and monitoring and troubleshooting drift and data quality issues. While explainability is the top priority among those listed overall, there are some interesting differences by industry. In healthcare, for instance, monitoring and root-causing drift and data quality issues are cited as higher priorities than using explainability.

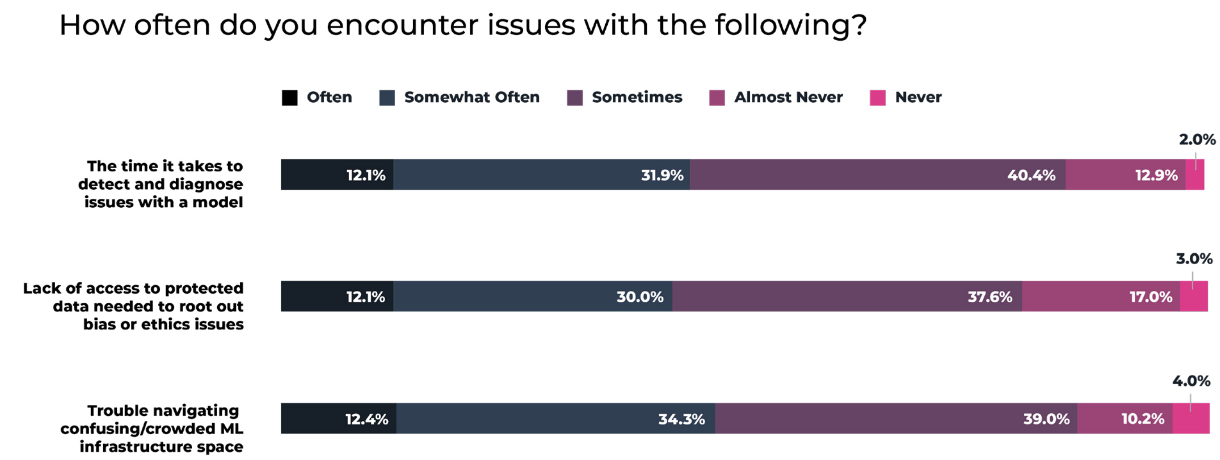
My team would benefit from deeper capabilities in the following areas (Select all that apply)



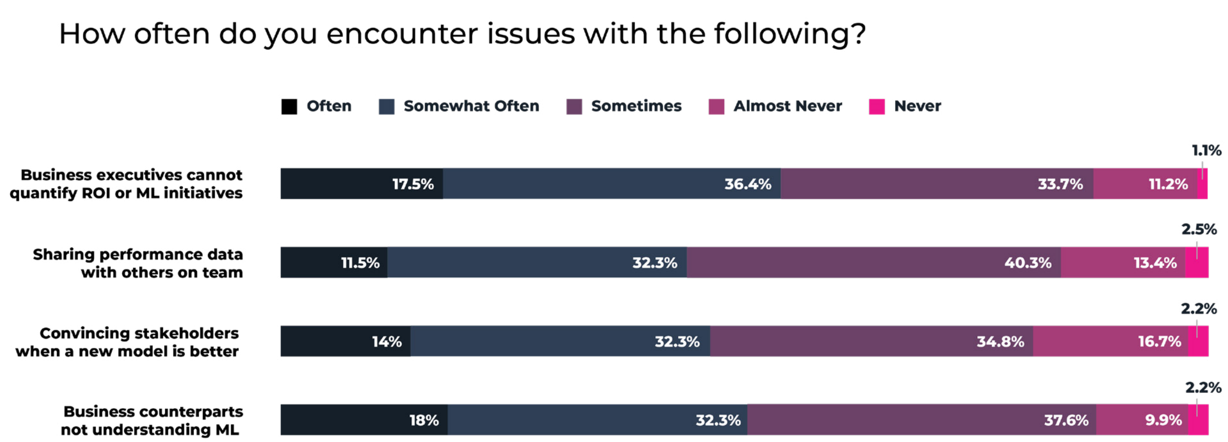
The desire for deeper capabilities ML monitoring and observability is also consistent across roles. Interestingly, technical executives tend to place higher importance on explainability than ML engineers. Among ML engineers – who are actually the ones who actually deploy models into production in many cases – the ability to monitor and troubleshoot model and data drift is a higher priority than explainability, which they rank on par with monitoring performance and data quality issues in terms of importance.



84.4% of data scientists and ML engineers cite the time it takes to detect and diagnose issues with a model as a challenge at least sometimes. Concerningly, 42.1% say they lack access to protected data to root out bias or ethics issues often or somewhat often.

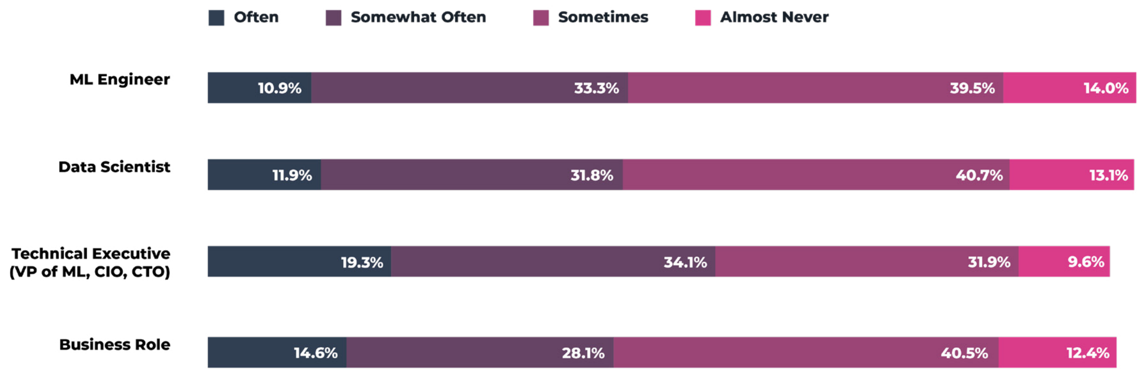


Team dynamics are another consistent challenge, particularly navigating the gulf between business executives and ML teams. 54.0% of data scientists and ML engineers report that they encounter issues often or somewhat often with business executives not being able to quantify the ROI of ML initiatives, with nearly as many reporting that business executives simply do not understand machine learning. Likely contributing to this disconnect is the fact that "sharing data with others on the team" and "convincing stakeholders when a new model is better" remain issues at least sometimes for over 80% of ML practitioners.



One reason why business executives lack visibility into AI ROI might simply be because they are not seeing the data. According to the survey, 53.4% of technical executives encounter issues with sharing performance data with others on the team often or somewhat often – more than any other group surveyed.

How often do you encounter issues with: sharing performance data with others on team

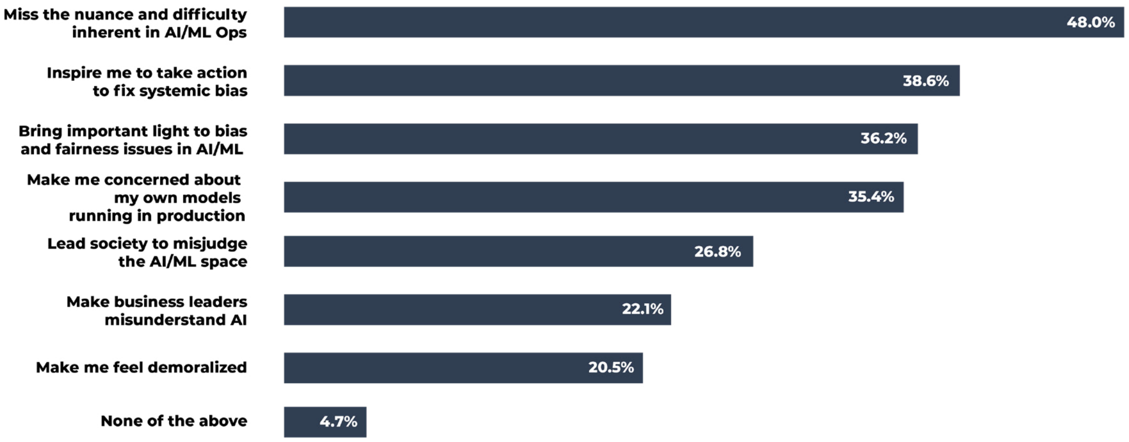


Looking Ahead

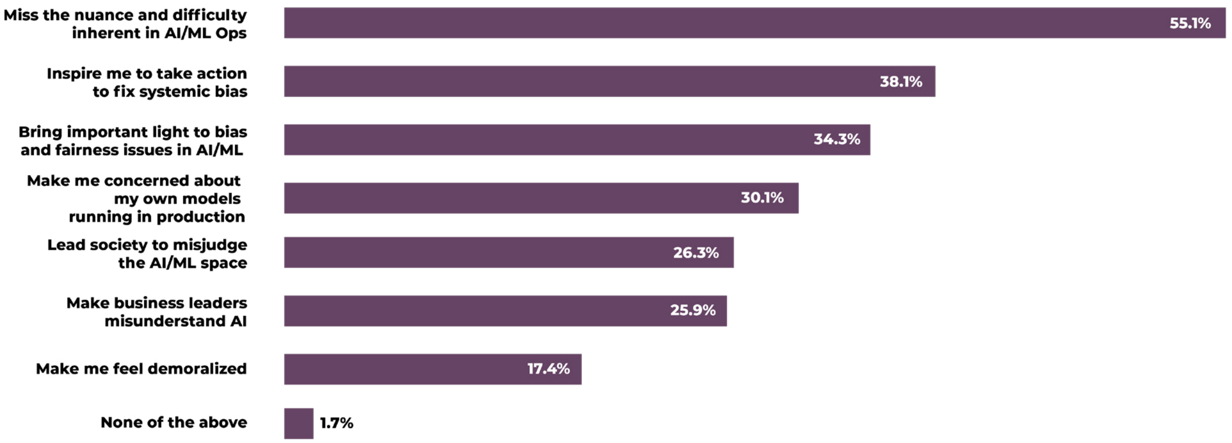
While burnout and stress are common in many industries, data scientists and ML engineers also face another challenge: rampant misunderstanding and skepticism about AI.

Most ML practitioners express frustration that negative headlines about AI miss the nuance and difficulty of MLOps. However, many are inspired to do something about it. In all, 38.5% of ML engineers and 30.1% of data scientists say negative headlines inspire them “to take action to fix systemic bias.” Over one-third of both groups also say such headlines make them “concerned about their own models,” likely speaking to conscientiousness in terms of addressing any issues in their domain.

ML Engineers: Negative headlines/stories about AI (i.e. around bias, ethics and fairness) — Select all that apply



Data Scientists: Negative headlines/stories about AI (i.e. around bias, ethics and fairness) — Select all that apply

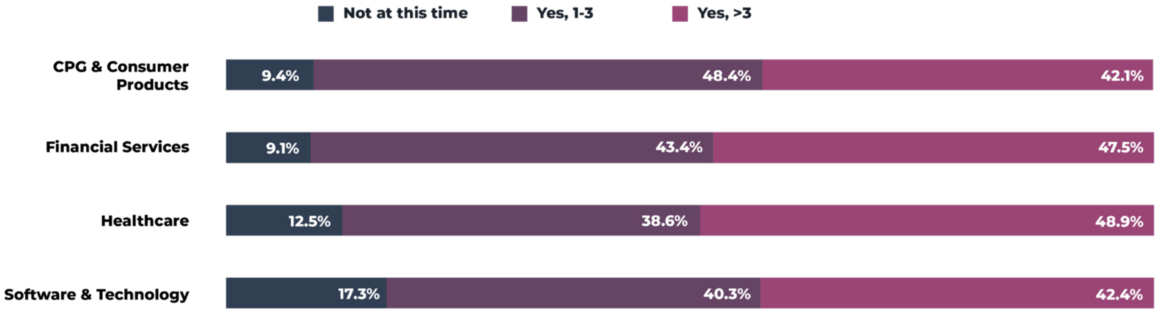


Most data scientists and ML engineers (59.0%) agree that the industry needs to do a better job with model transparency and explainability. Encouragingly, most also report that their organization makes ethics a core value. One-third believe companies need to adopt ethical governance around AI, showing that there is still a ways to go in terms of practitioner support for operationalizing AI ethics.

When it comes to building more ethical frameworks for ML, agree with the following — Select all that apply

While opinions on the future of the industry vary, one area where there is little divergence is around hiring. According to the survey, 81.9% of respondents are currently hiring data scientists and ML engineers – with financial services brands hiring the most.

Is your team hiring ML engineers and/or data scientists? If so, how many:



Conclusion

Compared to devops or data engineering, MLOps is still relatively young as an industry. That said, the importance of AI teams — who influence everything from the future of cancer care to whether retailers survive supply chain disruptions — cannot be overstated. Hopefully this survey begins to clarify what it will take for this rapidly-expanding group of professionals to thrive in an increasingly AI-driven world.