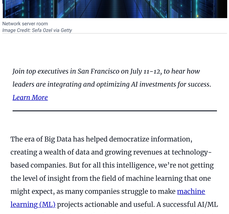
Community

How to make the most of your AI/ML investments: Start with your data infrastructure

Alexander Lovell, Fivetran
@alexander_sf

July 17, 2022 2:20 PM



Network server room
Image Credit: Sefa Ozel via Getty

Most Read

1 Cybersecurity meets AI: Augmenting and accelerating humans

2 Understanding the risks of generative AI for better business outcomes

3 4 reasons our future is decidedly virtual

Sign up for VB Daily

Want must read news straight to your inbox?

Subscribe

Join top executives in San Francisco on July 11-12, to hear how leaders are integrating and optimizing AI investments for success. Learn More

The era of Big Data has helped democratize information, creating a wealth of data and growing revenues at technology-based companies. But for all this intelligence, we're not getting the level of insight from the field of machine learning that one might expect, as many companies struggle to make machine learning (ML) projects actionable and useful. A successful AI/ML program doesn't start with a big team of data scientists. It starts with strong data infrastructure. Data needs to be accessible across systems and ready for analysis so data scientists can quickly draw comparisons and deliver business results, and the data needs to be reliable, which points to the challenge many companies face when starting a data science program.

The problem is that many companies jump feet first into data science, hire expensive data scientists, and then discover they don't have the tools or infrastructure data scientists need to succeed. Highly-paid researchers end up spending time categorizing, validating and preparing data — instead of searching for insights. This infrastructure work is important, but also misses the opportunity for data scientists to utilize their most useful skills in a way that adds the most value.

Challenges with data management

When leaders evaluate the reasons for success or failure of a data science project (and 87% of projects never make it to production) they often discover their company tried to jump ahead to the results without building a foundation of reliable data. If they don't have that solid foundation, data engineers can spend up to 44% of their time maintaining data pipelines with changes to APIs or data structures. Creating an automated process of integrating data can give engineers time back, and ensure companies have all the data they need for accurate machine learning. This also helps cut costs and maximize efficiency as companies build their data science capabilities.

Narrow data yields narrow insights

Machine learning is finicky — if there are gaps in the data, or it isn't formatted properly, machine learning either fails to function, or worse, gives inaccurate results.

When companies get into a position of uncertainty about their data, most organizations ask the data science team to manually label the data set as part of supervised machine learning, but this is a time-intensive process that brings additional risks to the project. Worse, when the training examples are trimmed too far because of data issues, there's the chance that the narrow scope will mean the ML model can only tell us what we already know.

The solution is to ensure the team can draw from a comprehensive, central store of data, encompassing a wide variety of sources and providing a shared understanding of the data. This improves the potential ROI from the ML models by providing more consistent data to work with. A data science program can only evolve if it's based on reliable, consistent data, and an understanding of the confidence bar for results.

Big models vs. valuable data

One of the biggest challenges to a successful data science program is balancing the volume and value of the data when making a prediction. A social media company that analyzes billions of interactions each day can use the large volume of relatively low-value actions (e.g. someone swiping up or sharing an article) to make reliable predictions. If an organization is trying to identify which customers are likely to renew a contract at the end of the year, then it's likely working with smaller data sets with large consequences. Since it could take a year to find out if the recommended actions resulted in success, this creates massive limitations for a data science program.

In these situations, companies need to break down internal data silos to combine all the data they have to drive the best recommendations. This may include zero-party information captured with gated content, first-party website data, and data from customer interactions with the product, along with successful outcomes, support tickets, customer satisfaction surveys, even unstructured data like user feedback. All of these sources of data contain clues if a customer will renew their contract. By combining data silos across business groups, metrics can be standardized, and there's enough depth and breadth to create confident predictions.

To avoid the trap of diminishing confidence and returns from an ML/AI program, companies can take the following steps.

1. **Recognize where you are** — Does your business have a clear understanding on how ML contributes to the business? Does your company have the infrastructure ready? Don't try to add fancy gilding on top of fuzzy data – be clear on where you're starting from, so you don't jump ahead too far.
2. **Get all your data in one place** — Make sure you have a central cloud service or data lake identified and integrated. Once everything is centralized, you can start acting on the data and find any discrepancies in reliability.
3. **Crawl-Walk-Run** — Start with the proper order of operations as you're building your data science program. First focus on data analytics and Business Intelligence, then build data engineering, and finally, a data science team.
4. **Don't forget the basics** — Once you have all data combined, cleaned and validated, then you're ready to do data science. But don't forget the "housekeeping" work necessary to maintain a foundation that will deliver significant results. These essential tasks include investing in cataloging and data hygiene, making sure to target the right metrics that will improve the customer experience, and manually maintaining data connections between systems or using an infrastructure service.

By building the right infrastructure for data science, companies can see what's important for the business, and where the blind spots are. Doing the groundwork first can deliver solid ROI, but more importantly, it will set up the data science team up for significant impact. Getting a budget for a flashy data science program is relatively easy, but remember, the majority of such projects fail. It's not as easy to get budget for the "boring" infrastructure tasks, but data management creates the foundation for data scientists to deliver the most meaningful impact on the business.

Alexander Lovell is head of product at Fivetran.

DataDecisionMakers

Welcome to the VentureBeat community!

DataDecisionMakers is where experts, including the technical people doing data work, can share data-related insights and innovation.

If you want to read about cutting-edge ideas and up-to-date information, best practices, and the future of data and data tech, join us at DataDecisionMakers.

You might even consider contributing an article of your own!

Read More From DataDecisionMakers