JUNE 2023

GROWTH ACCELERATION PARTNERS

DON'T FALL IN LOVE WITH YOUR DATA: WHY AN AGNOSTIC PARTNER WORKS FOR YOUR DATA SCIENCE PROJECT

If You Can't Find a Data Scientist Who Handles NLP and Computer Vision with Ease — Here's Why You Hire GAP

DATA SCIENCE

As your organization cries out for faster and greater business insight to make competitor-beating decisions, chances are you will fall for the idea of hiring a data scientist. But don't fall too hard or you could fall flat on your face.

Many organizations are in a similar boat: according to the U.S. Bureau of Labor Statistics (BLS), data scientist is one of the 20 occupations with the highest projected percent change of employment between 2021 and 2031. Yet the difference between needs and expectations can sometimes cause roadblocks, both for the data scientists and the organizations hiring them.

Put simply, data science projects are marathons. For any such undertaking, it requires comprehensive training to prepare for the race itself. Not doing any training is inadvisable to the point of foolhardiness; having a training partner alongside you who has run many marathons is often recommended.

Because the mistakes organizations make on data science projects can mean you stumble before you've even reached the starting gate. To illustrate just how basic, let's briefly outline here the difference between data engineers, data analysts, and data scientists:

- **DATA ENGINEERS**

Are in charge of building the infrastructure for data to flow correctly, as well as enable aspects such as data consolidation, the combining and storing of varied data in a single place. The analysts and scientists then consume what the engineer prepared.

- **DATA ANALYSTS**

Use statistical and logical techniques to evaluate data, create visualizations and provide insights.

- **DATA SCIENTISTS**

Process and develop data models, and then deploy them into applications using, among other things, machine learning algorithms. Data scientists have advanced knowledge of languages such as Python, R and SAS, as well as have a good understanding of statistics.

Yet there have been various examples of organizations hiring data scientists without understanding what they do; without having a robust set of processes for their data; without a suitable infrastructure in place to get value out of artificial intelligence; without having a data-driven culture and governance principles from which to work.

It may be a long way from companies just wanting "a chart to present in their board meeting each day" or perhaps having an AI solution for better decision-making. They can visualize the potential of building models, but they don't have the technical foundations to give the data scientists the tools they need.

There are various stages in getting a data science project off the ground. First, the organization needs to determine whether machine learning and modeling is required, or whether the objective can be satisfied with an analytics product. Then it becomes a question of data. Does the organization have the data to answer this question? If not, can the data be built to answer the question, and can the infrastructure be set up to consume this data?

Once the development environment and access to the data is secured, the exploratory analysis begins.

- How do each of the variables behave?
- What are their distributions?
- Is every column complete?
- Does each column accurately capture the phenomena you expect from the business behavior you are trying to study?

Once this all makes sense, then the data scientist will start prototyping, and deciding which type of model the business problem fits into; supervised learning, which utilizes labeled input and output data; and unsupervised learning, which looks to analyze and cluster unlabeled data sets. Semi-supervised learning, as the name suggests, relates to where some of the input data are labeled, or partially annotated. A simple example to illustrate is if you want to segment your customer base by those who are able or unable to pay a loan. You would classify a new observation, and this would be a supervised learning problem. Once you get solid results, you go to the stakeholders. And once you get approval, it can go to production.

Assuming the brief is understood, there are roadblocks when it comes to implementation. One challenge organizations face can be with regard to data drift and model drift, where machine learning models in production change — and degrade — over time. This is the result of a gradual change in the statistical properties of the data used to train a model.

Take the example of a ML model that predicted the likelihood of customer purchasing based on age and income. If the distribution of ages and incomes of the customers changed and did not correlate over time, the model would become ineffective at accurately predicting the likelihood of a purchase.

Regulatory expertise is another minefield for the unsuspecting. If you handle data, particularly in the U.S., it is subject to industry, federal and/or state level regulations. Two good examples here are financial and medical data. Data scientists need to know what they can and cannot do with such data, as well as how they are handled internally. Data enrichment is also something that needs to be handled carefully; third-party data can have its own regulatory concerns.

Another issue is with regard to the variety of technical disciplines within data science — and what that means for projects. If your project involves computer vision, natural language processing (NLP), or audio signal processing, then these are specializations in their own right. Would it make sense to hire a single full-time data scientist expecting them to do everything if your upcoming projects require these different disciplines?

The good news is that working with a partner like GAP on data science projects means you can get the best of the best: you can get support across all specializations, from NLP to computer vision, without breaking the bank. Even if you have an in-house data science team, GAP can provide greater experience with regard to more stringent industries such as finance and healthcare, as well as a wide variety of models to incorporate within projects.

Like other cloud-native projects, the varied machine learning offerings from Amazon Web Services (AWS), Microsoft Azure and Google Cloud Platform can be daunting. If you opt for AWS, you can purchase an ML offering specifically around computer vision and image recognition (Amazon Rekognition), to give freer rein to develop custom solutions with Amazon SageMaker. Outside of the hyperscalers, you can use the likes of Databricks and Snowflake, which can then integrate with a Jupyter Notebook so data scientists can create and share documents with live code, equations and other resources.

Above all else, there can be wider, less tangible benefits to having a disinterested party in charge of building out a data science project. It is a well-known piece of business advice to not fall in love with one's own idea or product, as you may be too resistant to change course if it does not work out. Remember the advice not to fall too hard, lest you fall flat on your face?There is a similar theory at work here. If you have not properly understood the brief or the data, or you have not transformed the data to make them applicable to the correct process, then it does not matter how good your chosen model is.

Model building may be the most romantic part of data science, where all the dreams of real-time, competitor-beating insights live. Yet, to revisit the running analogy, it is more like the last minute of a marathon, where you reap the rewards of putting all the hard work in before. With GAP, you can train, prepare for and then run the marathon in a record time — and, crucially, not break down halfway along the course.

To find out more, please visit WeAreGAP.com