MIT-IBM Watson AI Lab

CSAIL

MIT DIGITAL
MIT INITIATIVE ON THE DIGITAL ECONOMY

Controlling AI cost through better visibility

Summary

With on-prem infrastructure, the cost of training and running machine learning models is usually hidden from the business i.e., there is a disconnect between the value generated from the model and what it takes in terms of engineering and compute resources to get there.

This case study, with MarketingPlatform Inc. (a pseudonym), shows that making the link between computational demand and cost transparent can help in creating incentives, and in reducing costs meaningfully.

Opportunity: Moving to the cloud to better leverage existing data

MarketingPlatform Inc. (MPI) helps retailers and nonprofits to improve the return on their marketing efforts, by predicting who will be most receptive to each campaign. A tiny improvement in accuracy, at scale, can be worth millions of dollars in additional sales or donations, so the stakes are high.

The datasets available for these models are enormous. MPI runs a data cooperative with thousands of members, with 25-40% of these members regularly contributing data on e.g., transactions, donations, or subscriptions. This is then combined with compiled 3rd party data on everything from e.g., demographics, lifestyle, census data, and household incomes. After feature engineering, the data comprises 12000 variables and covers almost the entire US population.

With such a rich dataset, MPI was at the limits of what their existing on-prem infrastructure could handle. The team was able train models only on their internal data, and even then, only a small portion of it (e.g., a 50-100k sample) at a time. MPI knew that if they could use more data, there was a huge opportunity to generate additional value.

Challenge: Initially this came with a significantly higher price tag

Moving to the cloud (IBM Cloudpaks for Data) increased MPI's ability to manage their data and leverage all their data assets, both offline and online. The additional scalability of compute resources also enabled training on 600k records (up from a max of 100k) and 800 features (up from 150-200). This, together with machine learning tools (e.g., XGBoost[1]), helped deliver a 20-30% lift in response rate, a dramatic increase in return for clients.

[1] XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

Initially, the scaling of compute resources (along with the initial switching cost and learning curve) meant that, to begin with, total costs increased. In a fully on-prem world, the cost of infrastructure had been independent of utilization, so data scientists were able to run whatever they wanted, constrained only by compute availability. With effectively unlimited scalability, experiments would have to have to be designed more thoughtfully.

Solution: Connecting to marginal incentives helped offset extra costs

Fortunately, moving to the cloud also enabled MPI to better understand, and ultimately to optimize their spend. The team was now able to generate a per-model cost of compute and incorporate these marginal incentives when structuring their data exploration and analysis.

The team was also able to increase the efficiency of their compute utilization by improving cluster allocation, data flow and the overall modelling pipeline. For the machine learning model itself, they ran tests to optimize the ~100 required model parameters and fixed several of them based on what worked well, to minimize the number of that would need tweaking each time.

Outcomes: Reduced training costs and a mandate for wider roll-out

The result is a dramatic reduction in training costs from $1500 per training run to $100s per training run, even with the 30% uplift in model performance.

MPI success in this area is now fueling a transformation of their data science capabilities, bringing the number of practitioners in the U.S. up from 40 to 4000 within the year.

Key takeaways from MPI’s experience suggest that:

- Visibility of training costs at a per-model run level of granularity can help make the cost of moving to new ML techniques less expensive than would be expected.
- Volume matters: even small gains in accuracy can have an enormous impact to an organization when it can be applied at scale.