fabrique.ai

the engine for AI / ML / Data driven business

Technical Whitepaper

Enabling Reliable Machine Learning Solutions
for Mission Critical Business Applications

Table of Contents

Executive Summary

In the last decades, as massive amounts of relevant, timely and high quality data finally have become available for most modern enterprises, and as computer processing power has soared, it finally became possible to use sophisticated machine learning models to automate various business processes. Even more so, modern enterprises are contemplating automating some mission-critical tasks that have historically been very labour intensive. And they are doing it with machine learning, a technology that has matured over the years and now there are many successful cases when complex models have dramatically increased business efficiency. However, modern machine learning models are quite complex, may consist of sophisticated pipelines of algorithms, consume a lot of resources and may appear to the business as complete black boxes. The former problem is particularly challenging, since it raises concerns about relying on these models for decision making without really understanding how decisions are made.

Overall, mission critical applications require much more than good models to operate. They need an environment that is both reliable and will shield the application from external attacks, fraud, faults in the upstream and external systems and dramatic sudden changes in operating conditions. Furthermore, in mission critical applications there should be multiple models available on standby, which include weaker, but more robust models. All the feature extractions pipelines need to be carefully managed and versioned, such that models can be easily compared and model improvement is a straightforward process. The data that the models operate on must be complete and of highest possible quality. Model features and outputs must be intelligently monitored and automatic alerts need to be set up in order to quickly detected changes in market conditions, fraud waves or technical malfunctions in the upstream systems. Intelligent monitoring and an ability to fall back to simpler and more robust models can mitigate the issue with using complex black-box models such as neural networks.

Prolabs has developed its machine learning platform Fabrique to address these issues, including shielding the application from various security risks.

Introduction

Organizations all over the world are currently capitalizing on the promises of Artificial Intelligence, Big Data and Machine Learning, trying to leverage their data as a strategic competitive advantage. Indeed, after many decades of research in data management and machine learning technologies, a lot of labour intensive business processes can now be automated and performed by intelligent software agents. However, there are a number of serious technical challenges that need to be addressed in order to take full advantage of this new promise. It all starts with a problem that business are facing overwhelmingly - finding highly qualified data scientists that are able to extract and clean up all the relevant data in order to build high quality machine learning models. These data scientists are expected to combine data wrangling skills, business acumen and top notch machine learning and statistics skills to get their job done. However, building a machine learning model turns out to be the tip of the iceberg in automating business processes. Its is a relatively simple job to obtain historical data for modelling purposes, but building reliable data pipelines that enable high volume real time data streams is quite a different challenge. And this is precisely what is need in modern enterprises that are beginning to take advantage of Big Data and IoT technologies. And when such models are deployed in critical business processes, they need to be carefully monitored and retrained on a constant basis. The process of model deployment needs to be simple, transparent and reliable. Moreover, advanced machine learning techniques that achieve top results often result in complex models that appear as complete black boxes to the business stakeholders. But these models do perform well, making use of factors with high variability that were traditionally ignored by all decision making processes. In order to make use of such complex models, there must be a mechanism to intelligently monitor the performance of such black box models and to be able to roll back to simpler and robust models quickly in case when the variables that the model depends on suddenly change dramatically. Finally, models need to be improved with new data sources and features, and this process of constant model evolution needs to be made reliable as well, where teams of data scientists can collaborate on jointly building the models and making incremental changes, instead of forcing a single data scientist to own the entire process. Due to all these extra requirements, recruiters are starting to look for data scientists with solid engineering skills and experience in building production services, piling even more responsibilities on this already overburdened role. Prolabs takes a cardinally different approach - instead of overburdening data scientists and engineers, we have built Fabrique, a platform for deploying machine learning models that mitigates these challenges. Fabrique offers a clear and reliable pathway to develop data pipelines and models, to reliably build and deploy complex models and monitor them intelligently. Anomalies in all model features and various combinations of features are automatically detected in Fabrique and the business can receive early warning signs and switch to simple reliable models, while investigating issues with complex but higher performing models. New data sources and features can be safely added with version control without breaking existing models and bringing about chaos to the data science teams. Finally, the platform enables both traditional business processes with minimal latency requirements as well as modern real-time business processes, including automating online business processes.

Reliable Business Process of Deploying Machine Learning Pipelines

The business process of deploying machine learning pipelines for business-critical decisions must be reliable, repeatable and manageable. Fabrique integrates with BPM tools to deliver a robust business process for machine learning deployments. Before a model is deployed to production, it must pass all the required steps of the business process that guarantees that the model had been developed correctly, is of appropriate quality and has been sufficiently tested.

Managing High-Volume Streaming Data

Even though Data Management is an established field with enterprise grade solutions for many problems, managing high-volume streaming data, aggregating it and providing this data to machine learning models for scoring with real-time guarantees is still a very challenging task. Enterprises have traditionally avoided switching to real-time data streams for a good reason: these systems tend to be expensive and hard to engineer, yet the benefits were not immediately clear. Currently, with the emergence of online internet companies that have massively capitalized on real-time data about their customers and business processes, enterprises are starting to pay attention. Of course Fabrique does not require the enterprise to switch completely to a real-time streaming architecture, the system can be deployed in a traditional setting as well. But when an online process or a process with low latency requirements needs to be automated with machine learning, Fabrique can be used to set-up real-time data pipelines for this process. And an additional benefit of the approach we employ in Fabrique is that machine learning models don’t have to manage and assemble complex data objects from the data stream. Instead, view over event data can be set up and maintained incrementally, and models can access aggregated views via Fabrique.

Building and Deploying Machine Learning Models

Even though developing high quality machine learning models is a challenge in itself, there are other serious obstacles to a reliable service based on machine learning. Modern machine learning techniques can find subtle correlations in huge volumes of data, and while such models tend to perform well, they are typically based on inputs with high variability. In case when market conditions suddenly change or the features that the model was based upon abruptly become irrelevant, the model can quickly degrade and incur significant losses to the business. Furthermore, if a high variability model is starting to underperform, there should be an option to scale back to a more robust model right away. However, for this approach to be viable, multiple models have to be run in parallel, monitored and continuously checked for anomalies. Ideally, the enterprise should have a line of models starting with the most basic linear models on the low end and advanced models such as deep learning models on the high end. When the high-end models start to deviate significantly from their historical behaviours, simpler models need to take over the business process while data scientists analyze and update the complex models.

Fabrique solves this challenge by monitoring any number of models concurrently, checking for anomalies and issuing alerts when significant anomalies appear to be under way.

Using Global Features

Most machine learning problems are traditionally viewed as individual classification tasks. Typically a number of historical individual examples are provided to train the learning algorithm, and then the model makes decision one object at a time. This is a simple and very practical model of machine learning, however, in many cases it has a serious drawback of not considering the general features of the entire flow of objects. Analyzing and being able to draw features from the entire data flow can be extremely useful in a number of applications. One of such areas is fraud detection, where a vulnerability in the system is massively exploited in a short burst of time. For example in financial services, a fraud wave might include stolen identities from a local area, which would result in a massive flow of applications from the same geographic area. Such an event would trigger Fabrique's anomaly detection system. Additionally, machine learning models can receive aggregate features over the flow of applications.

Anomaly Detection for Machine Learning Models

When multiple models are deployed in Fabrique, we monitor each input of the model and each output over a number of dimensions to be sure to detect anomalies even when they are hidden. Fabrique mostly uses unsupervised learning for anomaly detection, which means that each feature and model output behaviour is checked against historical behaviour and anomalies are flagged, when a significant deviation is uncovered. Fabrique users can manually adjust sensitivity levels of the alerts and check what alerts would be generated on historical data. Also, model features and outputs can be monitored for anomalies across multiple dimensions, detecting anomalies that would pass undetected due to low volumes. Consider issuing online loans to customers, an anomaly might manifest itself when an unusual number of applications from a certain age group in a specific region enter the system. However, this anomaly might be undetectable in the overall flow of applications, and could only be found out when viewed across the corresponding dimensions. This situation is common in Big Data scenarios, where the features and model outputs are applied to large data streams and various anomalies might pass undetected due to low relative volumes.

Scalability and Fault Tolerance

The final challenge for a machine learning platform is to scale well with any volume of incoming data and recover from system failures seamlessly. Fabrique has been designed from the ground up as a distributed fault-tolerant platform that can be easily scaled up to handle growing data volumes and that has no single points of failure. Scalability is achieved simply by adding more servers to the platform. All internal systems in Fabrique, including database systems, are also distributed and replicated to achieve load balancing and fault tolerance.