SPOTLIGHT

Sponsored by: Experian

This Spotlight provides an overview of challenges with implementing AI/ML capabilities for financial services use cases and overcoming these challenges through MLOps. The paper also profiles the Experian Ascend Ops platform that supports the full model deployment life cycle.

Accelerating AI-Enabled Financial Services Through Machine Learning Operations

September 2022

Written by: Raghunandhan Kuppuswamy, Research Manager, AI

Introduction

More enterprise organizations are leveraging artificial intelligence (AI) and machine learning (ML) capabilities to improve operational efficiencies, optimize capital investments, and gain competitive advantage through increased business agility and improved employer productivity. While more enterprise organizations are adopting AI capabilities, they are also facing challenges while implementing AI solutions.

A machine learning model goes through multiple stages (called a machine learning life cycle) such as data ingestion/preparation, model build, model evaluation, application integration, and model deployment while it moves from experimentation to production. Typically, data scientists, application developers, and IT operators work in different silos across these stages. Every stage poses unique challenges, such as challenges with data preparation, lack of data science and infrastructure management expertise, lack of operational excellence, designing an optimal feedback loop, and model monitoring, which are only exacerbated due to the lack of collaboration between personas, such as data scientists, ML engineers, and IT operations engineers.

IDC studies show that customers from financial services are at various levels of AI adoption, with use cases such as payment fraud detection, least-cost routing, credit risk management, and intelligent value and pricing showing higher adoption rates. Such use cases typically employ multiple data sources, real-time and batch deployments, and stringent compliance obligations. As a result, enabling machine learning capabilities for these use cases is more challenging.

AT A GLANCE

KEY STATS

- In an IDC survey, more than 50% of customers cited cost, lack of skilled personnel, and lack of adequate volumes and quality of data as top challenges with implementing AI solutions.
- IDC studies show that it takes about 290 days on average — from start to finish — to fully deploy a model into production.

Challenges with AI Implementation

While global adoption of AI capabilities among enterprise organizations is increasing, it is not without challenges. In IDC's *AI Strategies BuyerView Global Survey*, CY21, respondents cited cost, lack of skilled personnel, lack of adequate volumes and quality of data, machine learning operations, and trust and ethics as the top challenges during AI implementations (see Figure 1).

Cost

While enterprises are leveraging AI capabilities to gain a competitive advantage by bringing innovations to market faster, they often are unable to accelerate as much as they would like to. Challenges such as excessive cost of model building and training, difficulties in moving models from experimentation to production, and complexity of managing multiple tools/platforms across stages of the ML life cycle impede AI acceleration. In the aforementioned IDC survey, 56% of respondents cited cost and lack of automation (end-to-end and data operations) as their top challenges during AI implementations, reducing their ability to introduce new capabilities in a timely manner.

Lack of Skilled Personnel

Implementing AI capabilities requires expertise across various domains and technologies such as data engineering, data science, agile model development, and IT operations. Data engineering expertise is required to prepare and manage data for model building and training. Data science expertise is required to select the right learning algorithms to use, build models, and fine-tune models to achieve the required levels of accuracy. Since model training and model inferencing usually require accelerated hardware, allocating the right computational resources for these tasks both during experimentation and in production requires infrastructure operational expertise. Setting up a data science platform to work on large-scale modeling problems and working with the newest methods for machine learning require specialized skill sets. Enterprise organizations may not have sufficient in-house expertise in these areas — so much that about 55% of respondents cited lack of expertise as one of their top challenges with AI implementation. It is also costly to have data scientists cover more operational tasks throughout the life cycle instead of being able to focus on building and testing models.

Lack of Adequate Volumes and Quality of Data

About 53% of respondents cited lack of quality data as one of the top challenges with AI implementations. Customers also indicated challenges with procuring adequate volumes of data and labeling them.

Machine Learning Operations

Respondents cited difficulty with managing large-scale data efficiently, including versioning and model data life cycle management. Other challenges include model management, model governance, model performance monitoring, and moving applications to production.

Respondents also identified performance, long development times, and selecting the right algorithms as top challenges during model development. Scale, maximizing GPU utilization, and experiment management/tracking were identified as top challenges during model training.

FIGURE 1: ***Challenges with AI Implementation***

n = 2,000

Source: IDC's AI Strategies BuyerView Global Survey, CY21

How Model Velocity Affects Time to Market

Model velocity (MV) refers to the time taken to move machine learning models from experimentation to production. Model velocity directly impacts an organization's ability to roll out new product features — the slower the model velocity, the longer it takes to deliver the capability to market. Hence, is it important to accelerate model velocity to the fullest degree possible.

A machine learning model goes through various stages across the machine learning life cycle, including data ingestion, data preparation, model exploration, model build/train, model acceptance testing, application integration/model deployment, and monitoring. As these stages employ different teams, tools, and processes that most often operate in different silos, they slow down the model velocity, with end users often finding it difficult to move models to production. Machine learning use cases also often involve multiple models, and enterprises tend to enable multiple use cases at the same time.

IDC studies show that it takes about 290 days on average — from start to finish — to fully deploy a model into production. Such prolonged delays slow down time to market, thereby impacting the organization's potential to improve the bottom line. Organizations that can modernize and accelerate their model velocity can garner a competitive advantage and reduce risk. IDC also observes that as more models are deployed into production, end users are facing challenges with model performance, model drift, and bias. Enterprises also face risk by not implementing a proactive process to track and monitor models in production, which inhibits awareness of how they perform over time.

ML for Financial Services: Use Cases

Enterprise organizations are leveraging AI/ML capabilities for a variety of use cases across various industries, including manufacturing, retail, financial services, healthcare, and IT operations. Machine learning capabilities enable financial institutions to provide enhanced user experience through improved accuracy, personalized recommendations, and

customer care while being able to explain the model results. These capabilities also enable financial institutions to provide innovative capabilities previously not possible. For example, financial institutions can provide customized loans/insurance quotes or investment recommendations tailored to the needs of the customer and customer profile. In various IDC studies, financial institutions have indicated increased investments in AI for use cases such as for credit decisioning, fraud detection, AI-assisted customer care, personalized recommendations, and real-time financial advice.

Leveraging machine learning capabilities to enhance a financial institution's customer journey is not a straightforward process. It usually involves building and managing machine learning models, serving models as programmatic endpoints, and integrating endpoints with end user-facing financial applications. Other aspects such as quality data preparation, user authentication, and agile model development also impact model accuracy and the pace of innovation. The following are attributes of machine learning systems that are relevant for financial services:

- **Integration — identity, APIs, and use cases.** Financial institutions' customer journeys leverage machine learning capabilities across various touch points, including in person, self-service kiosks, web, and mobile applications, to meet customers where they are. Financial applications need to ensure the right identity of the end customer to provide secure and safe access to the capabilities. These applications need to be integrated with machine learning capabilities that serve prediction endpoints through right authorization mechanisms. Versioning and tracking of models are also required for governance or when multiple models are leveraged.
- **Data management — from ingestion to production.** Model accuracy is directly impacted by the availability of quality data. With more uses cases that serve personalized features, financial applications can leverage a multitude of data sources in different formats. For example, personalized recommendations may use real-time data such as location and nonconventional data sources such as social media feeds. Model life-cycle management itself uses a variety of data types to manage models, track experiments, and monitor model performance. Such diversity of data sources and formats needs better data management capabilities across all stages of the ML pipeline, from data ingestion to production.
- **Heterogeneous deployments.** With financial applications being served at multiple locations, machine learning predictions also need to be served from multiple locations to enable better performance, low latency, and improved customer experience. IDC research shows that about 70% of currently deployed machine learning software is in on-premises and private cloud locations, less than 20% is in the public cloud, and less than 10% is in edge locations (IDC's *Future Enterprise Resiliency and Spending Survey, Wave 7,* August 2021). With the increasing adoption of ML techniques, expanding use cases, and increased investments in edge deployments, edge-based deployment of ML software is expected to increase.
- **Model velocity.** Model velocity directly impacts the ability to bring product capabilities to market. With the rise of digital-first financial services providers and nonconventional options for managing finances/investments, financial institutions are under constant pressure to retain and attract new customers through a better user experience. There is significant pressure on providers to increase their model velocity as a way to improve their chances of innovating faster than the competition. Agile development practices and CI/CD pipelines enable continuous delivery of application capabilities; similarly, establishing agile model development practices can accelerate model velocity through continuous integration/continuous model training.

Accelerating AI Adoption Through Machine Learning Operations

Machine learning operations (MLOps) refers to tools and technologies that enable model deployment, model management, and model monitoring at scale. An MLOps platform should ideally be able to support deploying models across various locations after rigorous development and user acceptance testing (UAT), serving models as endpoints, tracking model lineage, setting model performance metrics, monitoring model performance, troubleshooting model drift, enabling model governance, and providing feedback loops into various stages of the ML life cycle. It should provide flexibility and choice with model building/training/inferencing platforms, programming environments, libraries, data ingestion mechanisms, data types, and data management platforms.

An MLOps platform should further enable collaboration among data scientists through model repository/registry and between all the personas across the ML pipeline — data scientists, ML engineers, and IT operations engineers. It should also provide for model artifacts and model versioning, which can help create an auditable trait that enables institutions to implement model governance. Through these capabilities, MLOps platforms enable setting up scalable ML pipelines through which multiple models can be managed, tracked, and monitored simultaneously. In case of model drift or performance degradation, an MLOps platform can help identify the issue and trigger workflows to identify problems, helping companies take corrective action.

Overall, an MLOps platform can help accelerate AI adoption by providing end-to-end automation of the ML pipeline, increasing model velocity through agile model development processes, and breaking silos between personas through a consistent, connected layer.

Considering Experian's Ascend Ops Platform

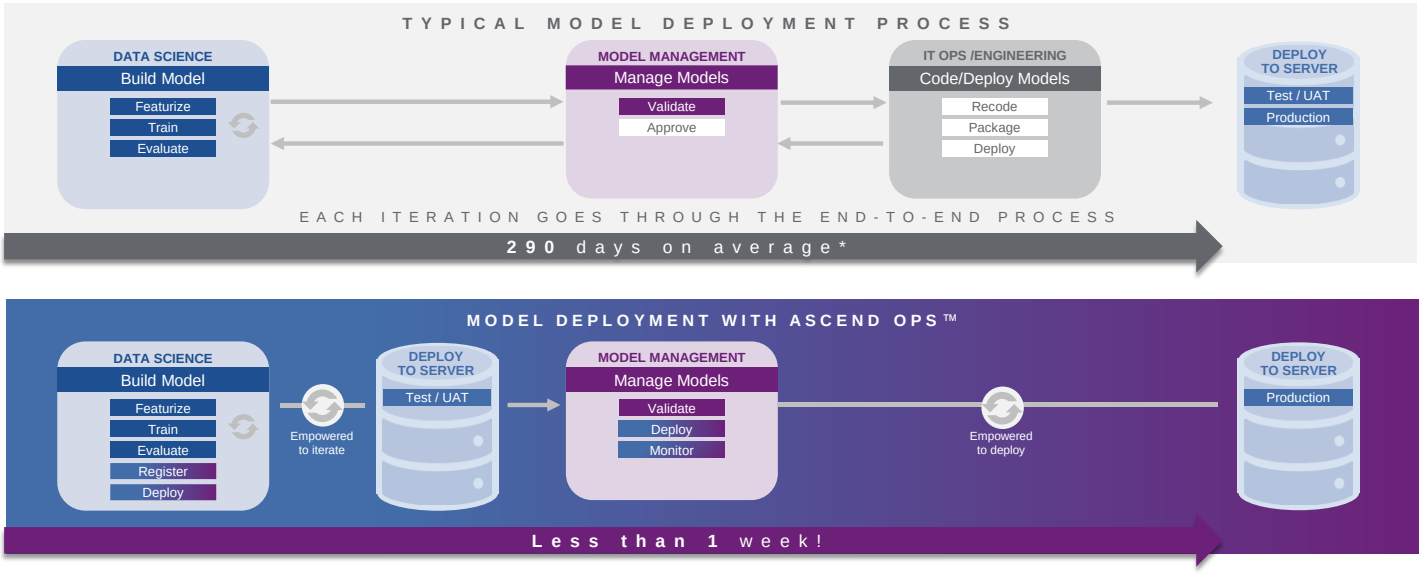
Experian's Ascend Ops platform, part of the company's Ascend Technology Platform, is an MLOps platform that accelerates model building.

Salient features of the Ascend Ops platform include:

- **Support for heterogeneous model deployments.** The Ascend Ops platform supports deploying machine learning models across cloud environments (including Amazon Web Services, Microsoft Azure, and Google Cloud Platform) and across development, staging, and production. While the Ascend Technology Platform supports model development with a range of sandbox options (Ascend Analytical Sandbox) to build, test, and train models with full access to Experian data and prebuilt analytics, Ascend Ops allows a user to move models into staging and production environments, regardless of how those models were developed or the language (R, Python) or platform used. Staging environments provide production-formatted data. Once in production, a single model can be used for multiple use cases and can be reused for both batch and real-time processing, enabling a "code once, use anywhere" approach that saves time and reduces complexity. The Ascend Ops platform is a self-service platform, with a user-friendly interface (Ascend Ops Manager) that provides the ability to access a full range of features supporting model testing, model development, model management, and model monitoring. Experian also offers a managed services option for the full model development and deployment life cycle.

- **End-to-end ML pipeline automation.** The Ascend Ops platform supports end-to-end automation of the ML pipeline from the data ingestion phase to deploying and running models in production. A user registers a model and provides selected artifacts, which kick off an automated process to containerize the model. Once a model is in production, a daily process automatically monitors models and alerts users to model drift or recalibration needs. When a model result is needed in a business process, the Ascend Ops platform automates the identification of the API endpoint, gathers the data that the model requires, formats the data (and determines the payload), calculates the model output, and delivers the model output to the end application to satisfy the business request. The Ascend Ops platform supports model development, model training, model management (through model registration, model onboarding, model lineage), model scoring, model deployment (both real-time and batch deployments), and model monitoring. It also supports third-party tools and allows models, regardless of their language or development platform, to be managed in a single environment. The platform is well integrated with data feature stores (Ascend Big Data Lake Feature Store/Ascend Big Data Lake Trade and Inquiry Data) and Experian downstream applications such as decisioning and marketing solutions.
- **ModelOps and model life-cycle management.** The Ascend Ops platform supports end-to-end model life-cycle management across various stages of the model life cycle through an intuitive user interface. The platform makes it easy to deploy models to endpoint servers, where they are ready to provide model outputs 24 x 7. By enabling programmatic access of these capabilities and support for third-party source code management tools, the Ascend Ops platform (see Figure 2) can enable CI/CD pipelines of machine learning models. Ascend Ops Manager enables containerizing models, which facilitate interoperability of models across build, staging, and production environments. It also acts as the central repository of model attributes (model registry) and the central repository of all models registered and deployed through Ascend Ops.

FIGURE 2: ***The Ascend Ops Platform***



Source: Experian, 2022

Key advantages of the Ascend Ops platform include:

- **Increased model velocity.** The Ascend Ops platform, through its support for model operations, model reuse, and end-to-end ML pipeline automation, can significantly improve custom feature and model velocity, thereby improving the ability to bring innovative financial applications to market sooner.
- **Simplified integration.** Ascend Ops simplifies model integration with Experian data, multibureau data, proprietary data, and downstream applications through APIs in both real time and batch. It also enables seamless integration with credit, fraud, and alternative data sources. The richness and quality of this data help eliminate some of the common data quality issues faced in the testing and deployment process.
- **Efficient model and pipeline management.** Users can see all models in their inventory and identify those currently in production and how they are performing. The containerization process enables use of an evolving set of development tools and sources and supports deployment without recoding. Once models are registered, a full range of model documentation, data, and artifacts can be stored in one place to ease the review and governance processes as well as support ongoing audit needs.
- **Ability to leverage Experian's credit and expanded Fair Credit Reporting Act (FCRA) data capabilities.** The Ascend Ops platform enables financial institutions to make use of thousands of model-ready prebuilt attributes and credit scores, including Premier Attributes, Trending 3D attributes, and a wide range of expanded FCRA data including ML models and scores, without the need for much customization.
- **Flexibility and choice.** Experian offers support for model development through the Ascend Analytical Sandbox, which provides flexibility and choice in model development for popular ML languages, tools, and libraries (including, but not limited, to Python, TensorFlow, JupyterHub, H2O.ai, and RStudio). The sandbox can also be delivered as a data service for firms that have invested in building their own analytics environment. Enterprises can use the Ascend Ops platform to containerize models developed across a range of model-building platforms, and users can choose to move models to production using a self-service tool (Ascend Ops Manager) or by working with Experian as a managed service. Such flexibility and choice enable model developers to leverage the Ascend Ops platform regardless of their size, or the number of models they manage, while continuing to build models on their platform of choice.

Challenges

The AI/ML life-cycle software ecosystem is a crowded space with a variety of vendors vying to get a share of the market. While the Experian Ascend Ops platform focuses exclusively on serving the financial services industry, it faces a credible challenge from vendors that are known to cater to multiple industry verticals.

Conclusion

MLOps platforms improve collaboration between multiple personas across the ML development and deployment life cycle and dramatically increase model velocity, thereby enabling faster time-to-market product innovations. This enables organizations to improve operational efficiencies, optimize capital investments, and gain a competitive advantage.

About the Analyst

Raghunandhan Kuppuswamy, *Research Manager, AI*

Raghunandhan is Research Manager, AI within the Artificial Intelligence and Automation Group. His core research covers innovative AI applications and solutions across industries and business processes as well as collaboration with the IDC Tracker team to develop total accessible market (TAM) reports for these rapidly expanding market segments.

MESSAGE FROM THE SPONSOR

More About Experian

Experian offers a robust MLOps platform that significantly improves the speed, efficiency, and reliability of the model deployment process. Ascend Ops™ delivers end-to-end pipeline automation that streamlines model operations and supports model reuse, all while leveraging thousands of pre-built features based on Experian data. Ascend Ops removes obstacles in the model lifecycle, accelerates model velocity and the ability to bring innovative financial applications to market at the speed of today's business.

IDC Custom Solutions

The content in this paper was adapted from existing IDC research published on www.idc.com.

IDC Research, Inc.
140 Kendrick Street
Building B
Needham, MA 02494, USA
T 508.872.8200
F 508.935.4015
Twitter @IDC
idc-insights-community.com
www.idc.com