Success Story

Automating AI/ML model training and deployment for a Singapore-based retail software provider

Customer	Country	Industry
ETP Group	Singapore	Retail/ Technology

About The Client

ETP stands out as a leading retail software provider in Singapore, extending its services to market leaders in 24 countries across the Asia Pacific, India, and the Middle East. With a successful track record of over 500 enterprise software projects for 300+ brands and 35,000+ stores, ETP's unique value lies in its ability to consistently deliver enterprise-grade omnichannel solutions to its customers.

Technology Stack

Apache Airflow	Jenkins	python
MySQL	mongoDB	mlflow

Services Used

AI consulting   AI development service

Business Situation

The client faced a significant challenge in handling multi-tenancy, as they were reliant on manual deployment methods for configuring AI/ML models tailored for each tenant. This involved the complex process of individually setting up and configuring the deployment environment for each tenant. The manual effort invested in this procedure was notably time-consuming, susceptible to errors, and inefficient.

To overcome these challenges and accommodate their expanding customer base, ETP sought our assistance, recognizing our expertise and standing in the field of Artificial Intelligence. Our key mandate was to optimize the client's infrastructure, automating the training and deployment processes for AI/ML models. This initiative aimed to ensure a streamlined onboarding experience for tenants, aligning with the client's objectives.

The key requirements that the client entrusted Daffodil to fulfill include:

✅ Employ advanced automation to minimize the risk of errors and inefficiencies, ensuring the reliability of AI/ML models.

✅ Implement standardized deployment procedures to promote uniformity and reduce the risk of operational disparities.

✅ Engineer scalable system to remove deployment bottlenecks, facilitating seamless integration as the client's customer base expands.

✅ Accelerate onboarding new tenants by automating setup and configuration processes, minimizing delays, and adapting to diverse organizational needs.

✅ Enhance security and compliance protocols by automating deployment configurations, ensuring robust data protection in various organizational contexts.

The Solution

The client, having already developed in-house AI/ML models, sought our assistance in automating processes for two key use cases:

1. **Anomaly detection model**

2. **Recommendation system**

✅ **Anomaly detection**

For the anomaly model, team Daffodil adopted the extended isolation forest model, due to its effectiveness in detecting anomalies across diverse datasets. We received a notebook containing the model details, which we translated into code, and utilized MLflow for experiment tracking and data engineering during the training phase.

1. **Training automation:**

Our solution aimed at creating a streamlined, automated training workflow. We systematically identified models for training based on company details stored in the database. Apache Airflow was implemented to schedule and automate the training process on a weekly basis for each tenant.

2. Deployment automation:

Following the training phase, the Daffodil team set up the deployment environment, which involved configuring the hardware and software infrastructure necessary to host and run the AI model. Deployment automation was achieved through Airflow DAGs, which facilitated the deployment phase seamlessly.

Notably, in this use case, predictions followed immediately after training. The real-time anomaly prediction occurred every 20 minutes, utilizing a batch processing approach. These predictions were efficiently stored in the Google Cloud Platform (GCP) buckets.

The system was designed for dynamic tenant onboarding, with models trained and predictions activated when a new tenant subscribed. MLflow, acting as a central model registry, not only facilitated effective visualization of predictions but also added an extra layer of transparency to the entire process. This transparency enables stakeholders to gain deeper insights into the model's impact.

The automation extended further, covering configuration setups for new tenants and eliminating manual interventions. Furthermore, To ensure system hygiene, cleanup scripts were put in place to remove redundant data within an hour of prediction completion, enhancing efficiency and resource utilization.

✅ **Recommendation system**

Moving to the recommendation system, we proceeded to automate both training and deployment. Three recommendation types were implemented: item recommendation from the user, user recommendation from an item, and item-to-item recommendation.

1. Item recommendation from user:

This recommendation type delves into a user's historical interactions and preferences, using this data to suggest items that closely align with their interests. The goal is to create a more personalized content experience, ensuring users receive recommendations tailored to their specific tastes.

2. User recommendation from Item:

In this type, the system examines item features and popularity to find users who might like a particular item. By recommending items based on what users prefer, this type boosts user engagement with content that aligns with their interests.

3. Item-to-Item recommendation:

Aimed at encouraging user exploration, this type suggests related items based on their interactions. By pointing out items often chosen together or with similar characteristics, the system prompts users to explore more content that complements what they've already chosen.

In the recommendation system, our approach to automating the training and deployment process mirrored that of the anomaly detection case. However, there were some key differences and enhancements tailored to the specific requirements of this use case.

Training automation:

Similar to anomaly detection, training automation in the recommendation system relied on identifying tenants in the database and scheduling training based on their preferences. While in the anomaly detection case, we set weekly training schedules for all companies, this time, we introduced a more flexible approach. Each company now has the autonomy to choose the frequency of model retraining according to their specific needs. This customization is facilitated by a new column in the database that identifies which models are eligible for training each day.

To execute this, a DAG is triggered daily to identify eligible companies for training based on the updated database entries. Once identified, the models are trained, and upon completion, another DAG is dynamically called to initiate deployment.

Deployment automation:

Previously, predictions were obtained from the Airflow framework. However, we evolved our approach by creating a dedicated backend engine for each tenant. This means that whenever a new tenant subscribes to ETP's service, a new backend engine is provisioned specifically for them. This segregation at the server level ensures data isolation and prevents any form of contamination between tenants.

To achieve this, we deploy each tenant in a separate namespace, guaranteeing data integrity and security. This entire process is fully automated through the seamless integration of Airflow and Jenkins, a CI/CD tool.

Additionally, the predictions for this use case were executed in real-time. Unlike anomaly detection, the recommendation system involved creating APIs to expose the above-mentioned three recommendation types through a back-end service. This enhancement allowed for seamless integration and usage of the recommendations through API calls.

The Impact

The successful automation of Anomaly Detection and the Recommendation System at ETP has brought about a substantial boost in operational efficiency. By replacing manual processes with advanced automation, we minimized errors and streamlined the onboarding of new tenants. Impressed by the outcomes, ETP has even extended their trust in Daffodil by assigning the automation of another critical module, the "Forecast model". This further collaboration is an exciting opportunity for ongoing innovation within ETP's retail software ecosystem. We are enthusiastic about the continued positive impact and advancements that our partnership can bring to their business operations.

Have a software product vision in mind?

Setup a personalized consultation with our technology expert.

Let's Talk