Model Deployment

Deploy & Manage Models to Maximize Business Impact

No matter how great your models are, realizing business value from data science work requires getting it into production, integrated with business processes and affecting live decisions.

Unfortunately, many companies struggle with this "last mile" of data science, for several reasons:

- **Models are often re-written** to overcome infrastructure or monitoring limitations, wasting time and causing delays.
- **DevOps challenges** make it hard (or impossible) to access GPUs to operationalize deep learning.
- Duplicated, siloed stacks make assets **hard to manage** on an ongoing basis.
- **Monitoring is ad hoc or nonexistent**, risking financial loss and poor customer experience as models' performance degrades over time.

Deploy and manage data science at scale, with four key capabilities:

Apps

Model APIs

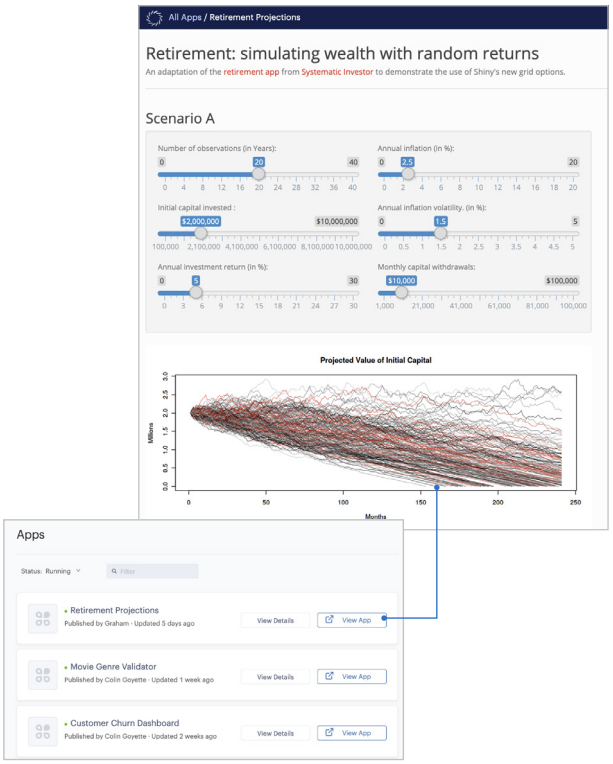
Scheduled Jobs

Management & Monitoring

Publish Apps

Data science often make interactive graphical applications for business analysts or other stakeholders. These "Apps" are as common and as valuable — if not more so — than real-time APIs, but data scientists often lack easy access to infrastructure to publish them. With Domino:

- Data scientists can **publish Apps** (e.g., Shiny, Dash, Flask) on self-serve production-grade infrastructure with one click. Or they can create **lightweight web forms** to let non-technical stakeholders run templatized analyses.
- Business stakeholders can **browse Domino App Gallery** to find relevant apps to use.
- Team leads and managers can see **usage statistics and trends** for different apps, to understand utilization and impact.



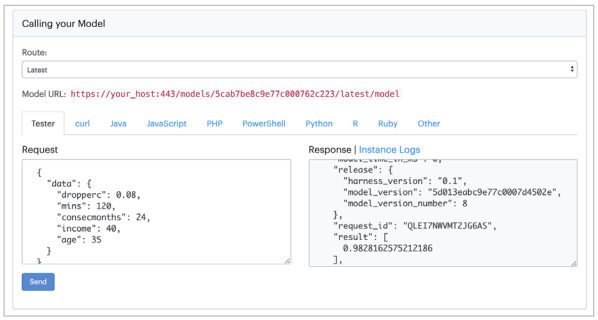
Publish & Host Model APIs

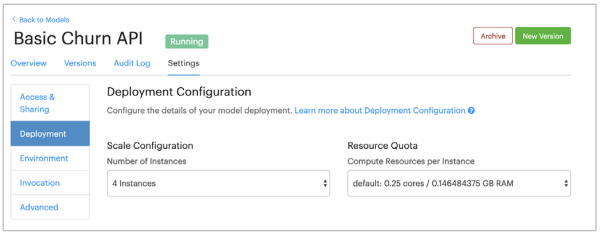
With a few clicks, Domino lets you turn your Python or R models into APIs that are suitable for most production use cases.

Model APIs in Domino come out of the box with:

- **Horizontal scalability & high availability,** with the ability to easily integrate GPUs for model inference of deep learning models
- **Audit logs** that track changes to the model and configuration settings
- **Flexible security** to control invocation
- **Scheduling capabilities** to support **recurring retraining and redeployment**

And, **Domino accelerates model validation and review** by tracking work during model development. Model validators can inspect and collaborate with data scientists before a model is deployed.

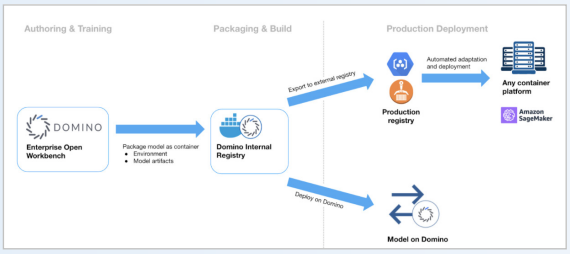




Export Model Images from Domino

You can export models as self-contained Docker images to integrate into CI/CD pipelines or deploy to your existing production environment. Export a SageMaker-compatible image for deployment in AWS, or deploy GPU-accelerated machine learning models to edge devices with NVIDIA FleetCommand™. You can even deploy and execute Python scoring code inside Snowflake's Data Cloud.

Domino approach to model packaging and deployment



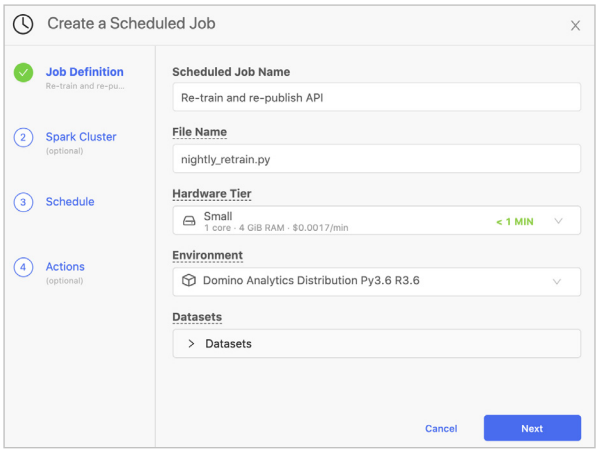
Schedule Batch Jobs & APIs

Domino makes it easy to setup scheduled jobs. These are perfect for:

- Batch scoring
- Automated report generation. Flexible notification features make it easy to use your data science artifacts (e.g., rendered Jupyter notebooks or RMarkdown) as elegant reports.
- Recurring data prep / ETL tasks

Batch jobs can automatically re-publish your model APIs, making them great for retraining.

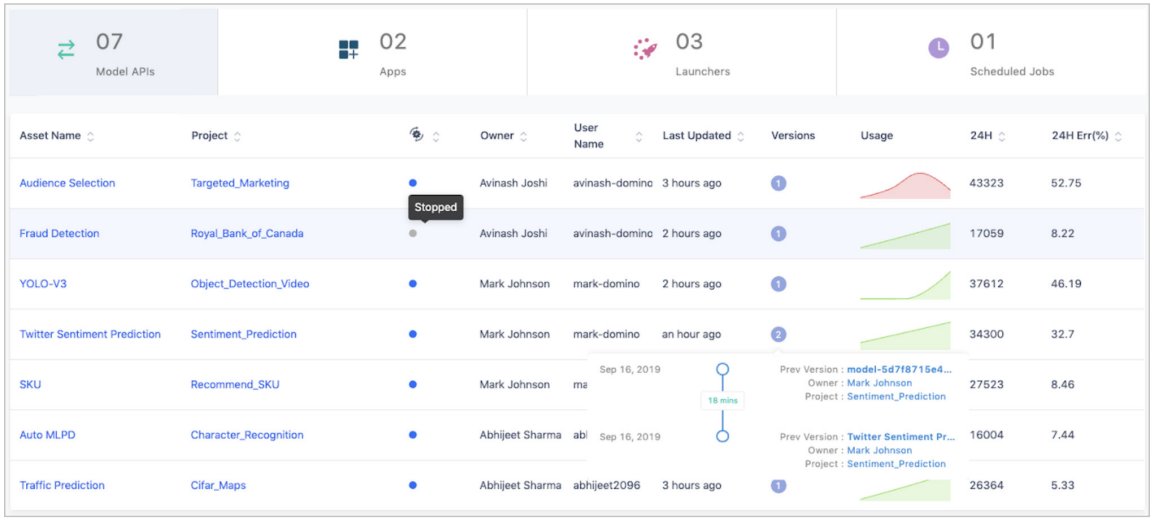
And Domino lets you programmatically trigger batch jobs from your own systems (e.g., Airflow) so you can integrate Domino jobs into existing processes.



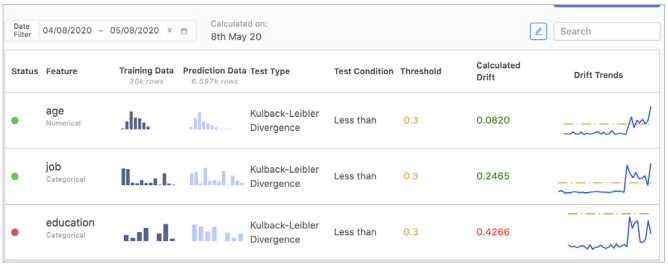
Manage Production Assets & Monitor Models

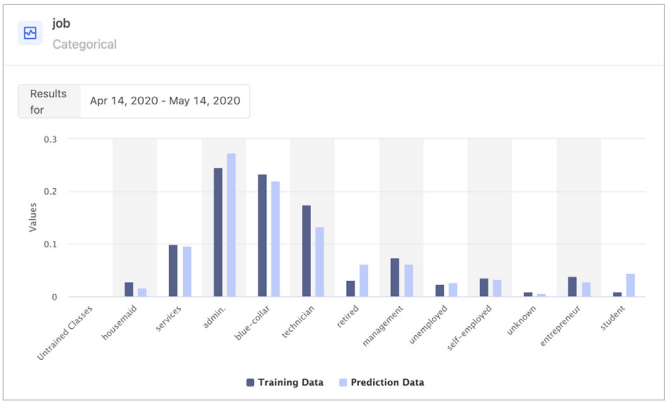
Getting a single model or App deployed can be an accomplishment in its own right — but managing dozens, hundreds, or thousands of assets is another matter entirely.

With Domino, you can see all your production assets — APIs, Apps, Scheduled Jobs — at a glance, with visibility into how often they are used and when they were last updated.



Integrated Model Monitoring takes this a step further, to enable proactive monitoring and alerting when your models drift, so you can retrain or fix issues quickly before they cause serious financial impact. It's the one place to monitor all your models — even ones developed or hosted outside of Domino (e.g., in Snowflake's Data Cloud).





- Detect drift in input features and output predictions using advanced statistical checks.
- Register groundtruth to detect performance drift.
- Schedule automated checks and alerts to save data scientists' time and more rapidly respond to drift.
- Explore data using Automated Insights to diagnose drift quickly.
- Accelerate model remediation with easy access to the original development environment.
- Scale model monitoring capacity infinitely with Domino's Elastic Monitoring Engine to support the most demanding monitoring requirements.