Managing ML Models @ Scale
Intuit’s ML Platform
Srivathsan Canchi
Tobias Wenzel
Who we serve

Consumers

Small businesses

Self-employed
ML in Action
ML driven experiences

✅Turbotax

Investment interest expenses

Margin interest investment expense

Interest expense investment

Investment interest expense carryover

Where do I enter a 1099-G for a state or local tax refund?
Important: If your 1099-G shows an amount in Box 1, click or tap here for alternate instructions, as the steps below won’t...

Where do I enter Form 1099-S?
You may get a 1099-S if you sold your home, a rental property, stock in a co-op or any other real estate, including land,...

Self-help
in TurboTax

QuickBooks

Forecasts cash flow
in QuickBooks

Mint

Automatic categorization
in Mint
Managing Models in Production is hard
We spend more time bringing the model to production than developing and training it

– Data Scientists, 2018
“Hidden Debt” of ML

Data Science

ML Operationalization

Infrastructure Management

Model Development
Feature Engineering
Model Metadata
Deployment
Orchestration
Security
Cloud Infrastructure
Feature Store
Data Processing

How can ML Practitioners focus on their craft?

ML Practitioners

ML Platform

Data Science

Model Development
Feature Engineering

Model Metadata
Deployment
Orchestration
Security
Cloud Infrastructure
Feature Store
Data Processing

ML Operationalization

Infrastructure Management
Principles of Intuit ML Platform

Collaboration

Self-Service

Security
Powering insights and experiences with ML

Data Scientist
“I can quickly train and deploy models that improve the customer experience”

ML Engineer
“I can quickly iterate and make models performant”

Data Analyst
“I can quickly train and deploy models that inform stakeholders how the business is doing”

Product Manager
“I can easily run experiments that rely on models to get a customer benefit”

Marketer
“I can easily target particular users that have a certain characteristics”
Intuit’s ML Platform Today

Ingest and Curate Data

Explore

Feature Processing

Feature Store

Online & Offline

Train

Evaluate

Infer

(Online and Batch)

Model Efficacy Feedback

Orchestration Tools

Model Lifecycle Management

Data and Model Metadata

Data Processing Infrastructure
Technologies in use

- AWS SageMaker
- Kubernetes
- Argo
- beam
- Spark
- Spring
Feature Store

- Store Features in offline (durable) and online (transactional) stores
- Metadata about Features
- Find and re-use Features
- Feature access during model inference and model training
**Feature Processing**

**Streaming**

- Kafka
  - Beam Feature Processor
  - Kafka
  - Online (DynamoDB)

**Batch**

- Data Lake
  - Spark Feature Processor
  - Offline (S3)
  - Kafka
Model Training
Model Training

```python
pipeline_type: sagemaker-training
pipeline_version: master
artifact_url: test.docker.repo/creditcard-approval-model/
project_name: creditcard-approval-model
kms_key_arn: arn:aws:kms:us-west-2:1234567890:key/key
email: firstNameLastName@intuit.com
training_data_config:
data_uri: s3://creditcard-approval-input/data(training/)
content_type: csv
validation_data_config:
data_uri: s3://creditcard-approval-input/data/validation/)
content_type: csv
additional_data_configs:
- data_config_name: test
data_uri: s3://creditcard-approval-input/data/validation/
content_type: csv
output_bucket_url: s3://creditcard-approval-output/
tags:
- Key: intuit:billing:fp
  Value: 3po4uih3-9084h49iu3fn3-894n
resource_config:
training:
  instance_count: 1
  instance_type: ml.m5.xlarge
  max_run_time_sec: 86400
  ebs_volume_gb: 1
```
Self Service Interfaces

Statistics about productivity increases

non-prod environments for customers
Automate the simple things

- Single-click Deploy
- Versioning
- Live Metrics, Alerts
- Model's Runtime Status
- Running Performance Tests
With great speed, comes cost

Visibility and Self-service

Platform success = Lots of experiments

Cost Transparency
• Upfront pricing
• Cost dashboards

Cost Assignment
• Tagging
• Business Unit assignments
  • Platform ensures correct BU chargebacks
Cost transparency
Cost transparency
What can go wrong?

`jellyfish 0.7.1`

`pip install jellyfish`

A library for doing approximate and phonetic matching of strings.

**Project description**

Jellyfish is a python library for doing approximate and phonetic matching of strings.

Written by James Turk (<dev@jamesturk.net> and Michael Stephens.

See [github.com/jamesturk/jellyfish/graphs/contributors](https://github.com/jamesturk/jellyfish/graphs/contributors) for contributors.

See [http://jellyfish.readthedocs.io](http://jellyfish.readthedocs.io) for documentation.

Source is available at [http://github.com/jamesturk/jellyfish](http://github.com/jamesturk/jellyfish).

Jellyfish > 0.7 only supports Python 3, if you need Python 2 please use 0.6.x.

Security

ML Models are as vulnerable as any other piece of software

ML Model Vulnerabilities
• ML models vulnerable for attacks
• Need security gates at every stage of development

Quality Gates
• Source image validations
• Standardized framework for container images
• Image hardening, certification and signing

Security Gates in our AWS setup
• Secure VPC endpoints
• VPC flow logs - monitor traffic
• Security groups - restrict outbound access
• KMS encryption - secure data at rest
• Least privilege IAM roles - secure management of the service
Compliance

Complex regulations

- The well known
  - CCPA, GDPR

- More common
  - PCI

- The arcane (very specific to tax compliance)
  - INDOR, NIST

Platform enforces compliance

- Any customer data that is used in the ML model lifecycle is managed centrally

- Achieving compliance for all models reduces complexity in the process for applications using them

- Automating the necessary controlling actions for the platform ensures that future models remain compliant
Learnings

Models as Software
- Accelerated deployment through self service
- Using a GitOps approach allows for declarative development of models
- Production monitoring and alerting
- Security, compliance built-in

Automated Workflows
- Workflows for standard ML operations help with complexity
- Hosting a workflow engine allows for easy extensibility
- Small threshold for customers starting out with the platform through templates with short configuration

Curated Feature Store
- Model quality Feature quality
- Curated feature store offers library of meaningful features
- Search, explore, share features across models enables acceleration of model development
In Future...

- **Custom ML Operator** to orchestrate and manage ML Resources (Spark, SageMaker Training/Deploy, Feature Store)

- **Declarative Management** of MDLC such as re-training of models, monitoring of features etc.

- **Managed Notebook Service**
Contact Us

Tobias Wenzel
tobias_wenzel@intuit.com

Srivathsan Canchi
srivathsan_canchi@intuit.com