Prodapt Chase Extraordinary

Breaking the barrier between Machine Learning (ML) prototype and production

Leverage MLOps to scale and realize the ML use cases faster

Credits | Skanda Gurunathan | Dinesh Singh GC | Prashantkumar Maloo | Priyankaa A

“Launching ML pilots is deceptively easy but deploying them into production is notoriously challenging”- Gartner

Gartner

According to Gartner, “By the end of 2024, 75% of enterprises will shift from piloting to ***operationalizing AI***, driving a 5X increase in streaming data and analytics infrastructures”

Major challenges faced by Digital Service Providers (DSPs) in ML pipeline and operations

Inefficient data handling practices like manual data processing, validation and inference retrieval

Lack of standard change management process to address the change request in ML pipeline

Periodic manual re-training and deployment of the ML models to accommodate the data drift

Lack of in-depth visibility of the model's performance as they interact with real-world events

Inspite of spending more, DSPs face numerous ML operational challenges

While 63.2% reported they are spending between $500,000 and $10 million on their AI efforts, about **60.6%** continue to experience a variety of **operational challenges**

Despite the significant spend dedicated to AI, **64.4%** said that it is taking them between **7 and 18 months** to move AI/ML models from idea to production

28.4% stated that they **rebuild the models** every time they deploy them

Source: The State of Development and Operations of AI Applications

“

To overcome these challenges Digital Service Providers(DSPs) need to shift from the current method of model management to a faster and more agile format. **ML Operations (MLOps)** approach automates and monitors the entire machine learning lifecycle, enabling faster time to production of ML models

Most forward-thinking DSPs have started implementing MLOps to accelerate and scale their AI initiatives

A leading DSP in Latin America implemented MLOps for Reinforcement Learning (RL) based Personalized Offer Simulation

- RL-based offer simulation agent required **huge feature space** and **more than 6 models as input** where each model required manual training, validation, and hyperparameter tuning
- Identifying data drift and resolving them manually was time-consuming, which resulted in poor system quality

MLOps implementation enabled **auto-retraining and deployment** of models to accommodate the data drift. The **standardized change management** orchestrated by Cloud code management tools made the process transparent and effective

Reduced model training and deployment time from **8 hours to 1.5 hours**

A leading DSP in North America implemented MLOps for Network Event Prediction

The network event prediction degraded and resulted in less accuracy as the **number of devices increased**. Numerous models were developed which required manual training and best-fit model deployment to achieve high precision levels

Streamlining **MLOps** using MLFlow with an in-built model repository enabled **model versioning, auto-retraining**, and performance tracking of models. Further, it helped in parallel processing of multiple models and **automated best-fit model deployment**

Improved accuracy by **~98%** *using best-fit model deployment*

MLOps is the way to go forward, however implementing MLOps and achieving the best results is not easy. This insight details the key levers for the DSPs to have a **successful implementation of MLOps**

Key levers of MLOps approach to accelerate the AI initiatives of DSPs

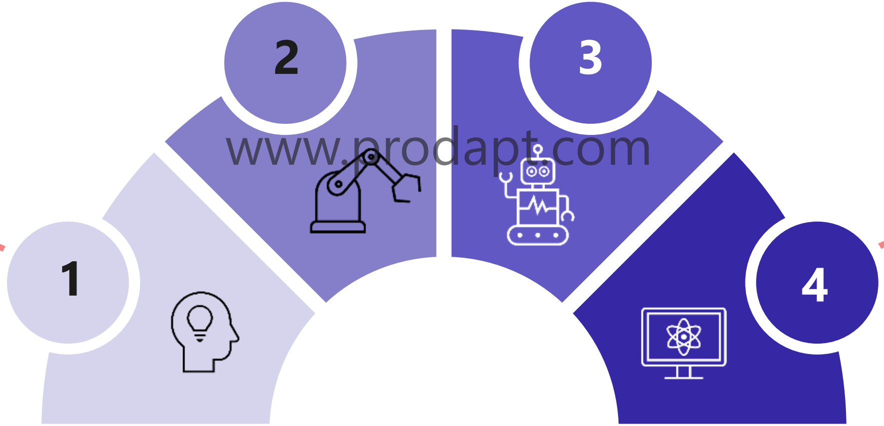
Reduce model training and deployment time by 70%

Standard and automated change management process for continuous integration and deployment of models

Automated retraining and deployment of models to accommodate data drift and reduce model retraining time

De-coupled architecture for parallel execution of ETL and training of ML models

Continuous monitoring to track the performance of the model in terms of its accuracy and drift



This insight deep dives into the 4 key levers of MLOps approach and provides best practices for its effective implementation with **Personalized Offer Simulation (Next Best Offer Recommendation)** as a sample use case

Decoupled architecture for parallel ETL and training of ML models

Decoupled system for Personalized Offer Simulation enables cost savings of up to 50%

Coupled architecture

- Difficult to scale up the ML pipeline for each newly identified ML use case due to resource dependencies
- **Sequential ETL execution, corpus creation and model training** leads to increased time, cost and complex code management

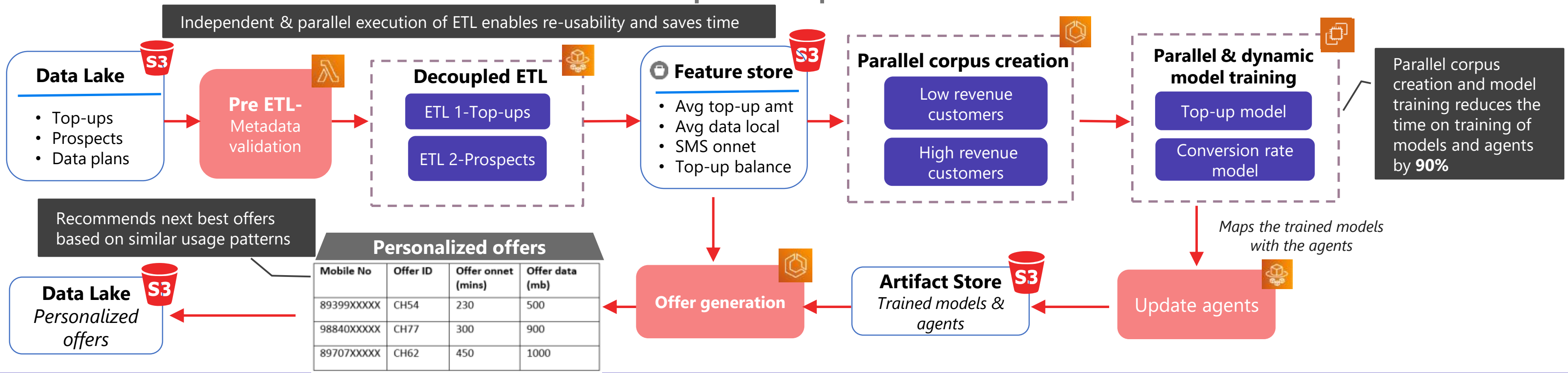
Decoupled architecture

- Components from the **existing pipeline can be reused** for a new ML use case. For example, the *output of ETL1 from offer simulation can be reused for churn analysis*
- **Parallel ETL execution, corpus creation and model training** saves the time, effort and cost required to orchestrate the ML pipeline

Key recommendations

- Implement services like **AWS Lambda** or **Google Cloud Functions** to **validate the metadata** and ensure whether necessary configurations are met before proceeding to ETL. It avoids validation issues during the model run, thereby reducing time and cost
- Develop an **AWS Glue** or **Google Dataproc homologation script** to handle the changes when data is transferred from the Data Lake to the ML engine
- Leverage **AWS Fargate** or **Google Cloud Run** for small scripts like updating the agents where memory usage is less, enabling **5X cost savings**

Sample use case-Decoupled architecture for Personalized Offer Simulation & Recommendation



Mobile No	Offer ID	Offer onnet (mins)	Offer data (mb)
89399XXXXX	CH54	230	500
98840XXXXX	CH77	300	900
89707XXXXX	CH62	450	1000

Standard and automated change management process for continuous integration and deployment of ML assets

" Standardizing and automating the change management process enables continuous deployment of models with ***reproducibility, security, and code version control***

In a typical ML model operationalization, DSPs are challenged to set up a pipeline where the changes are continuously built and made ready for production, seamlessly and accurately. Further, it has various limitations such as:

- Lack of structured ways for defining and maintaining configurations
- Inefficient tracking of resources that are responsible for approving the deployment
- Lack of code version and model artifacts control

Sample use case – Standardized Change Management for rollout of new data plan model to simulate offers for new geography customers

New geography with a different customer segment → Rollout of new data plan model to simulate offers for new customers → Developer updates the code artifact →

Code Commit
- Code repo
- Dynamic model config.yml

Automated Change Management module

Triggers mail requesting for approval of changes for new data plan model → Approver reviews & approves the code change → Update code changes of new model → Time based deployment of new data plan model to ML Prod pipeline

Standard and auto-change management process helped accelerate the rollout of models for new plans by 60-70%

Key recommendations

Encode services in programmer friendly languages

Implement services like **Kubeflow pipelines or AWS Cloud Development Kit (CDK)** for defining the resources in familiar languages. It auto-creates an equivalent YAML file, resulting in easier maintenance of the huge MLOps codes

Leverage a unified code repository

Leverage a **single source code repository like AWS Code Commit or Google Cloud Build** to develop and release multiple ML artifact versions, resulting in reproducible code updates

Enable plug and play of ML models

Implement **plug and play of ML modelling** with different algorithms **using AWS SageMaker and Elastic Container Registry(ECR),** which makes code management easier

Create a dynamic model config file

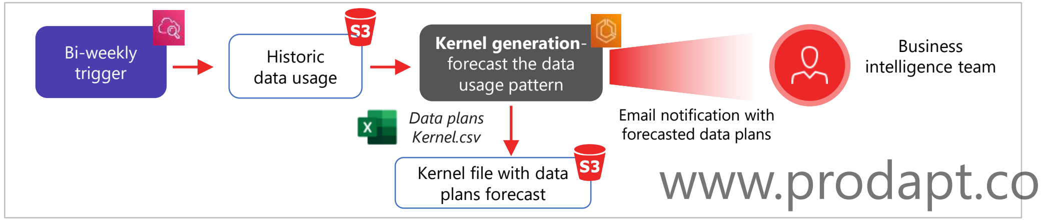
Develop a **dynamic model configuration file** with features and hyperparameters which helps in scaling out of models by making just few tweaks instead of altering 500+ configurations

Data quality validation for streamlined detection of data drift

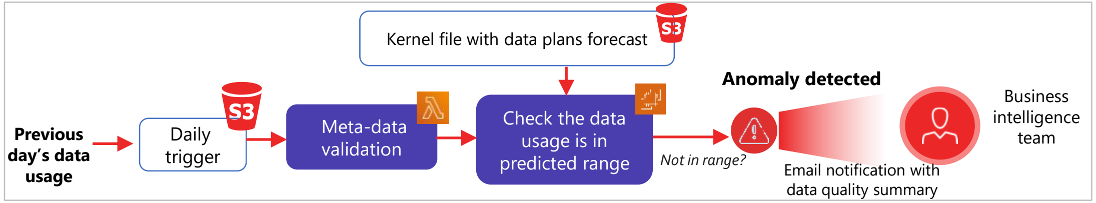
DSPs face frequent data drift as customer behavior is constantly changing and **80% of the data drift occurs due to unexpected events or occasions.** Analyzing the data quality regularly assists the DSPs to detect the data drift at an early stage and decide on the next best actions

Recommendations

- Generate a data quality kernel to pull the previous month's data usage and generate projections for different users
- Implement an anomaly detection model to **forecast the data plans** based on the historical data



- Track the data usage on daily basis and validate it with the range in the kernel file. For instance, when the **data usage is not in the defined range of 300-500 MB**, it is an **anomaly** that should be removed
- Send a mail report to the BI team with the list of anomalies. When the number of **anomalies exceeds the defined threshold** (e.g., 20% of the total data), **retrain the ML model** to accommodate the **data drift**



Sample monitoring reports of Personalized Offer Simulation

Kernel file showing Data Plans Forecast

Date	Day	Forecast_Low (mb)	Forecast_High (mb)
2021-06-17	Thursday	100	200
2021-06-18	Friday	115	180
2021-06-19	Saturday	340	500
2021-06-20	Sunday	380	525

Data quality report - Anomaly detected on 20/6/21

Anomaly! =>(380</=600</=525)

Data quality summary report from 16/8/21 to 22/8/21

Exec_Date	File_Date	Prospects	Plans	DNA	Topups
2021-08-16	2021-08-15	None	False	False	False
2021-08-17	2021-08-16	None	False	False	False
2021-08-18	2021-08-17	None	False	False	False
2021-08-19	2021-08-18	None	False	False	False
2021-08-20	2021-08-19	None	False	False	False
2021-08-21	2021-08-20	None	False	False	False
2021-08-22	2021-08-21	None	False	None	False

True : Anomaly detected
False : No anomaly detected
None : There isn't data quality executed

Seamless tracking of data quality assists the DSPs to stand on top of anomalies and resolve the issues faster

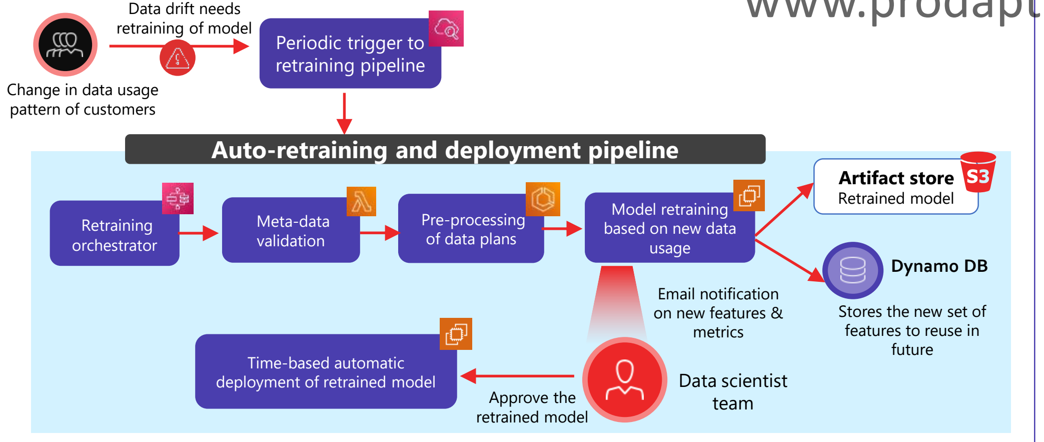
Automated retraining and deployment of ML models to accommodate data drift and reduce model retraining time by 70%

Once the data drift is identified, it is vital to retrain the model based on the new data. Resolving the data drift by manually retraining the model is cumbersome and time-consuming for the DSPs

Challenges in handling data drift manually

- Collecting the latest data usage patterns manually in local storage and building an efficient model for the new set of data
- Resolving the data drift manually and achieving good offer predictions with frequent changes in the data usage patterns
- Managing the maintenance window manually and deploying the newly trained model into production

Sample use case: Auto-retraining and deployment of models accommodating drift in customer's data usage



Sample report- Features before and after drift identified by auto-retraining pipeline

	Current model	New retrained model
Important features	• avg _phone calls • avg_top-up amt • top3_app usage ratio • data local • last activity days	• avg_phone calls • avg_top-up amt • data local • phone calls_month1 • sms onnet
Accuracy	84.2%	91.36%

New features due to drift ,resulting in improved accuracy

Current model features that are no longer important

avg_phone calls, avg_top-up amt, activity days, avg_freq, balance_main_acc, data_local_mb days_since_plan, n_boosters, n_plans_month1

Key recommendations

- Store the retrained models in Cloud storage like **Amazon S3** or **Google Big Query** enabling better **traceability** and **reusability** of the models for future predictions
- Set up an **AWS SageMaker** instance to ease the start and stop of Jupyter notebooks, enabling quick model deployments and controlled pricing
- Send an automated mail summarizing the **change of features due to data drift**, for the data scientist team to decide on the next best actions

Continuous monitoring to track the performance of the ML pipeline

Track the performance to gain real-time visibility of personalized offers

Since the DSPs' machine learning models often interact with various real-world events, the model predictions and accuracy can degrade over time. As they process new data, the models in production require continuous monitoring to make sure they are performing as per expectations

Recommendations

- Implement a monitoring pipeline to track the performance of the model whenever the model is retrained. For e.g., If the model gets retrained every week to generate offers, the monitoring pipeline should track and capture the performance of the predicted offers for the previous week
- Aggregate the performance of the ML pipeline from different systems to generate **reports on end-to-end utilization of the use case**. This helps in analyzing how the pre-processing, business logic and predictions of the offer simulation model performed
- Track the model metrics seamlessly to retrain and tune the model as and when needed

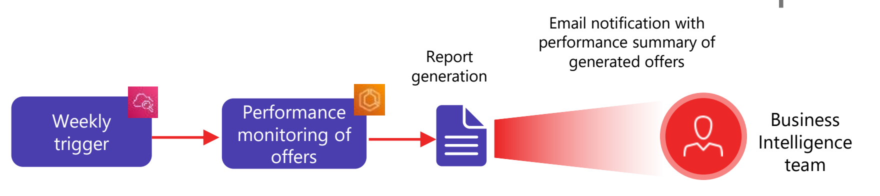


Fig: Performance monitoring pipeline to track the offer simulation

Sample report – Pipeline performance summary for past week generated offers

Date	RL offers	Sent offers	No offer	Different offer	Mismatch ratio %
2021-08-16	116715	108681	8034	0	6.88
2021-08-17	84866	75977	8889	0	10.47
2021-08-18	69332	60309	9023	0	13.01
2021-08-19	47453	39011	8442	0	17.79
2021-08-20	82978	73957	9021	0	10.87
2021-08-21	102521	91534	10987	0	10.72
2021-08-22	96403	86352	10051	0	10.43

From 2021-08-16 to 2021-08-22 RL generated **600268 offers and 89.3% of the recommended offers** were sent.

Continuous model monitoring provides the DSPs with **real-time and in-depth visibility of the models** and helps to identify potential issues before they impact the business

Benefits achieved by a leading DSP in Latin America by leveraging the MLOps approach as described in this insight

Implementing the key levers of MLOps as discussed in this insight, resulted in the following benefits

- **50%** reduction in data pre-processing and model prediction time
- **70%** reduction in re-training time and time-to-market of AI/ML models
- **50%** OpEx savings due to decoupled systems and dynamic spawning of resources
- **~75 to 85%** consistent improvement in the baseline accuracy of the ML use cases

THANKS!

Get in touch

USA

Prodapt North America, Inc.
Oregon: 10260 SW Greenburg Road, Portland
Phone: +1 503 636 3737

Dallas: 1333, Corporate Dr., Suite 101, Irving
Phone: +1 972 201 9009

New York: 1 Bridge Street, Irvington
Phone: +1 646 403 8161

CANADA

Prodapt Canada, Inc.
Vancouver: 777, Hornby Street,
Suite 600, BC V6Z 1S4
Phone: +1 503 210 0107

PANAMA

Prodapt Panama, Inc.
Panama Pacifico: Suite No 206, Building 3815
Phone: +1 503 636 3737

CHILE

Prodapt Chile SPA
Las Condes: Avenida Amperico Vespucio Sur
100, 11th Floor, Santiago de Chile

UK

Prodapt (UK) Limited
Reading: Suite 277, 200 Brook Drive,
Green Park, RG2 6UB
Phone: +44 (0) 11 8900 1068

IRELAND

Prodapt Ireland Limited
Dublin: Suite 3, One earlsfort centre,
lower hatch street
Phone: +44 (0) 11 8900 1068

EUROPE

Prodapt Solutions Europe &
Prodapt Consulting B.V.
Rijswijk: De Bruyn Kopsstraat 14
Phone: +31 (0) 70 4140722

Prodapt Germany GmbH
Münich: Brienner Straße 12, 80333
Phone: +31 (0) 70 4140722

Prodapt Digital Solution LLC
Zagreb: Grand Centar,
Hektorovićeva ulica 2, 10 000

Prodapt Switzerland GmbH
Zurich: Muhlebachstrasse 54,
8008 Zürich

Prodapt Austria GmbH
Vienna: Karlsplatz 3/19 1010
Phone: +31 (0) 70 4140722

Prodapt Slovakia j.s.a
Bratislava: Plynárenská 7/A, 821 09

SOUTH AFRICA

Prodapt SA (Pty) Ltd.
Johannesburg: No. 3, 3rd Avenue, Rivonia
Phone: +27 (0) 11 259 4000

INDIA

Prodapt Solutions Pvt. Ltd.
Chennai: Prince Infocity II, OMR
Phone: +91 44 4903 3000

"Chennai One" SEZ, Thoraipakkam
Phone: +91 44 4230 2300

IIT Madras Research Park II,
3rd floor, Kanagam Road, Taramani
Phone: +91 44 4903 3020

Bangalore: "CareerNet Campus"
2nd floor, No. 53, Devarabisana Halli,
Phone: +91 80 4655 7008

Hyderabad: Workafella Cyber Crown 4th Floor,
Sec II Village, HUDA Techno, Madhapur

THANK YOU!