

Intermittent Demand Forecasting in Scale using Meta-Modelling (Deep Auto Regressive Linear Dynamic System)



Abhishek Sengupta
Staff Data Scientist, Walmart



Biswajit Pal
Director - Data Engineering, Analytics & Insights, Tata CLiQ



Shubhodeep Moitra
Senior Data Scientist, Walmart

Overview

❖ **Reliable Demand forecasts** of Items are integral to the health of Retail Operations



❖ Accurate forecasts lead to **improved decision-making** and outcomes in replenishment, capacity and resource planning

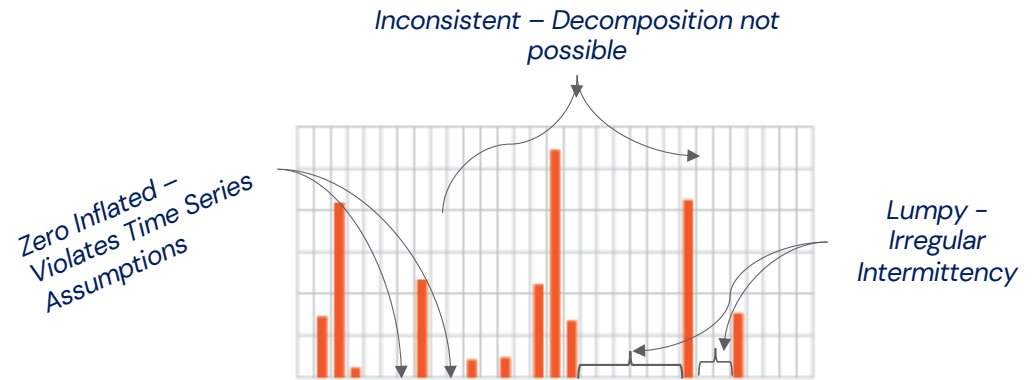


Overview

❖ Forecasting Models can be developed on **different levels of granularity**

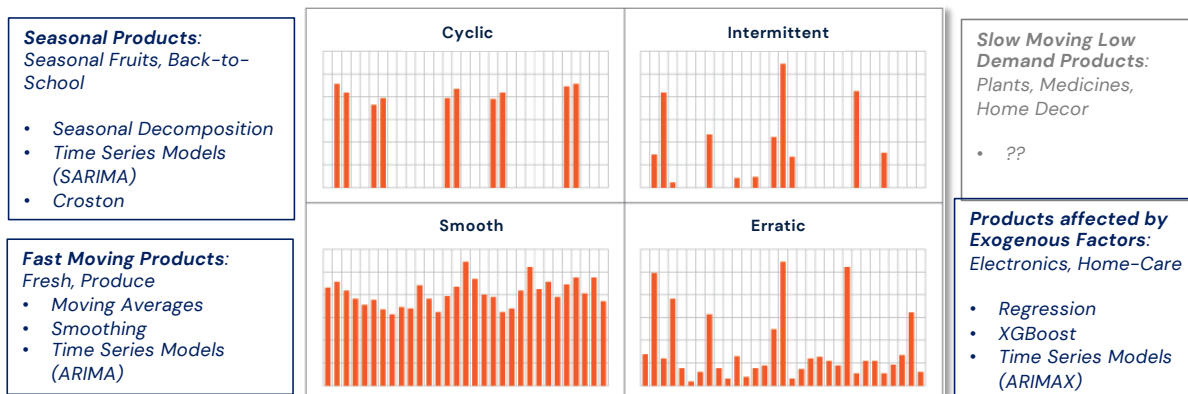


❖ Granular Forecasts (SKU-Week-Store Level) more valuable but lead to **temporal intermittenencies** for slow-moving items; leading to **prediction inaccuracies**



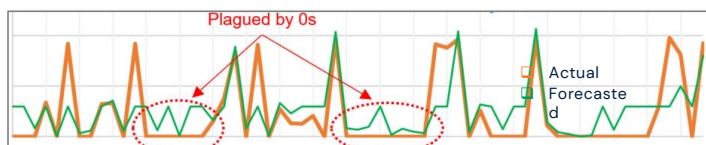
Problem Statement

- ❖ Demand Forecasting Patterns can be broadly classified into the below broad categories:



Traditional ML, Time Series approaches exist for 3/4 Categories

- ❖ For Intermittent Series, Time Series Assumptions break leading to over forecasts

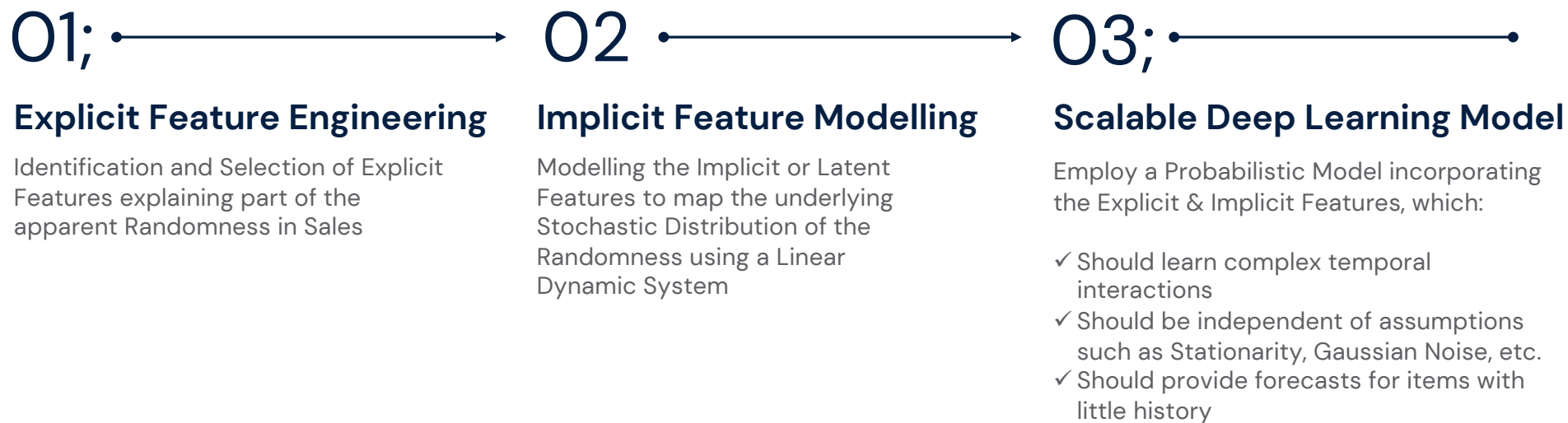


Problem at Hand:

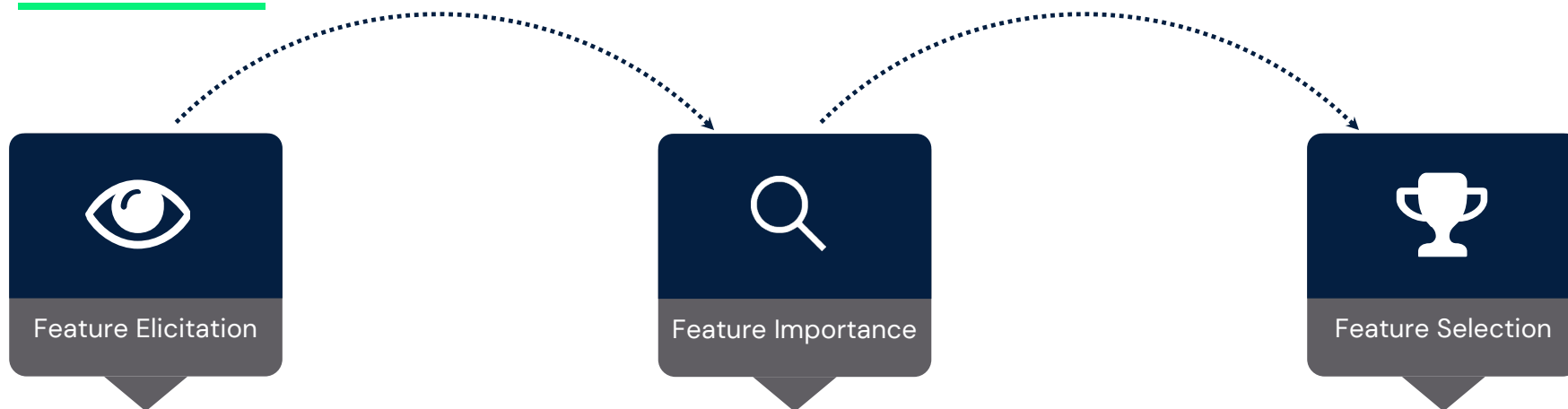
- ❖ Devise a Robust & Accurate Intermittent Time Series Forecast Solution at Scale:
 - ❖ Using Transaction Histories for 2 years (~10 Million Records)
 - ❖ Which runs within a 10-hour window by which the Forecasts should be generated
- ❖ Provide 14 Week Ahead Forecasts for all SKU-Store Combinations (Even for new items with little Transaction History)
- ❖ Achieved using a novel Meta-Model Architecture named as Deep ARLDS (Deep Auto Regressive Linear Dynamic System)

Meta Model Framework

Our Meta Model referred to as **Deep ARLDS** (Auto Regressive Linear Dynamic System) involves the below Key Steps:



Explicit Feature Engineering

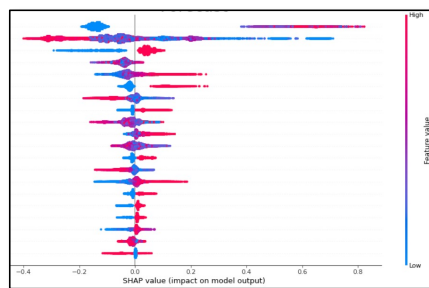


- ❖ Identify and List all possible Features (**300+**) which can explain Variability in the Series:

- ❖ Price
- ❖ Promotions/Markdowns
- ❖ Weather
- ❖ Store Demographics
- ❖ Product Lifecycle
- ❖ Item Attributes
- ❖ Calendar Events



- ❖ Train Vanilla XGBoost with all Features
- ❖ Evaluate Feature Importance from SHAP Plots



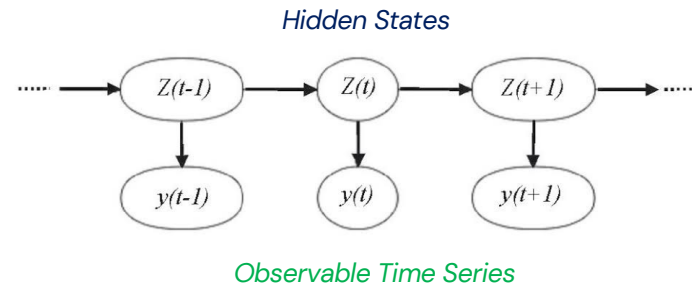
- ❖ Select Features (**65**) with overall Highest Importance:

- ❖ Promotion/Markdown Features
- ❖ Holidays/Special Events
- ❖ Temporal Variables (Fourier Transformed)
- ❖ Item Description Embeddings (GLoVE)
- ❖ Time since Item/Store Launch

Implicit Feature Modelling

- ❖ Train a Linear Dynamic System on Actual Sales y_t to model the underlying hidden signals indexed by a Temporal Difference Variable (Δ)

- ❖ LDS Equation:
$$\begin{aligned} z_t &= Az_{t-1} + B\Delta_{t,t-1} + \varepsilon_t \\ y_t &= Cz_t + \varepsilon_t \end{aligned}$$



- ❖ Δ denotes the time interval between two non-zero sales instances
- ❖ Including Δ in the state equation of the LDS ensures that the state depends on the time instant at which the previous observation was made
- ❖ Values of the hidden states (z_t) estimated using Expectation Maximization (EM) on Kalman Smoothing & Filtering Equations
- ❖ (z_t) s denote the Implicit Features governing the Stochastic Randomness of the Intermittency

Deep Learning Model

- ❖ The Meta Model Deep ARLDS combines the Explicit x_t and Implicit Features z_t (obtained from State Space Modelling) into a Deep Learning Framework
- ❖ An Auto Regressive Recurrent Neural Network (DeepAR) Architecture is used to determine Posterior Probability of the Demand of Item i at time t ($y_{i,t}$) using x_i and z_i as covariates:

- ✓ Implements an Encoder-Decoder setup sharing the same model architecture in the training and prediction range
- ✓ Models the conditional distribution $P(y_{i,t:t} | y_{i,t:t-1}, x_{i,1:T}, z_{i,1:T})$ of the future $[y_{i,t}, y_{i,t+1}, \dots, y_{i,T}]$ given its past $[y_{i,1}, y_{i,2}, \dots, y_{i,t-1}]$

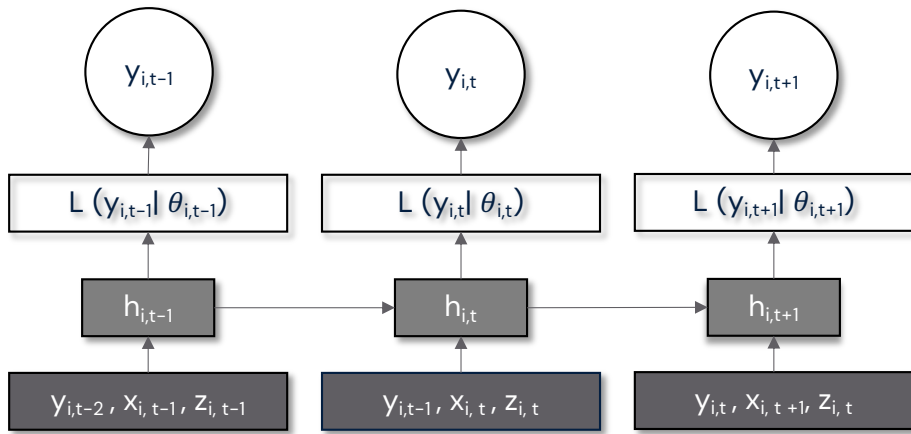
$$P_{\theta}(y_{i,t:t} | y_{i,t:t-1}, x_{i,1:T}, z_{i,1:T}) = \prod_{t=t_0}^T P_{\theta}(y_{i,t} | y_{i,t:t-1}, x_{i,1:T}, z_{i,1:T}) = \prod_{t=t_0}^T L(y_{i,t} | \theta(h_{i,t}, \theta))$$

$$h_{i,t} = h(h_{i,t-1}, y_{i,t-1}, x_{i,t}, z_{i,t})$$

h : function implemented by a multi-layer recurrent neural network with LSTM cell

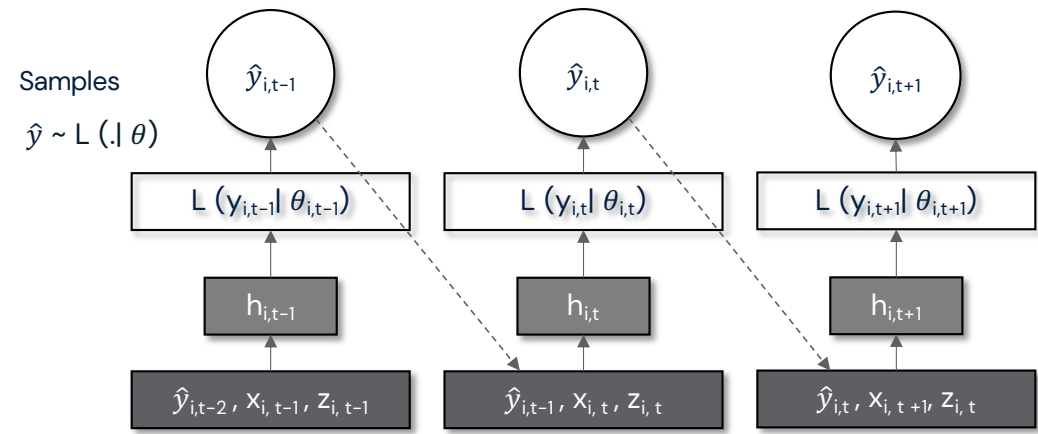
- ✓ $L(y_{i,t} | \theta(h_{i,t}, \theta))$ is assumed to be Negative Binomial to account for Zero Inflated Time Series
- ✓ Model Parameters θ shared across all SKUs within a Category, enabling Forecasts for Items with Low Training History
- ✓ Sample Quantiles from $\hat{y}_{i,t:t} \sim P_{\theta}(y_{i,t:t} | y_{i,t:t-1}, x_{i,1:T}, z_{i,1:T})$ (0.75 Percentile in this Scenario) is used to Obtain Final Forecasts

Deep ARLDS Network Architecture



Training

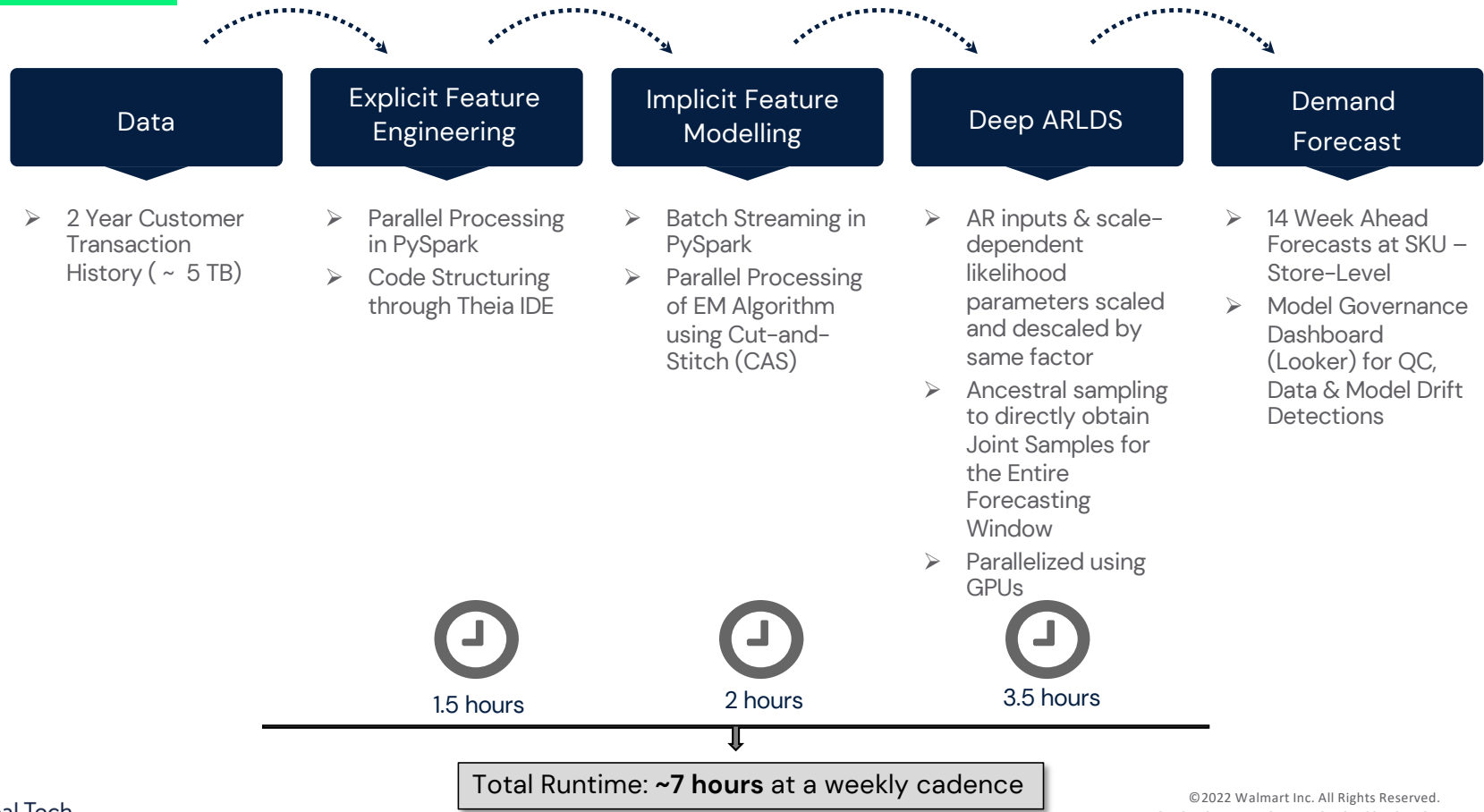
- ❖ At each step t , the Network Inputs are:
 - ✓ The Explicit Covariates $x_{i,t}$
 - ✓ The Implicit Covariates $z_{i,t}$
 - ✓ The Target value at the previous time step $y_{i,t-1}$
 - ✓ The previous network output $h_{i,t-1}$
- ❖ Network Output $h_{i,t} = h(h_{i,t-1}, y_{i,t-1}, x_{i,t}, z_{i,t}, \theta)$ used to estimate $\theta_{i,t}$ of the likelihood $(y|\theta)$, which is used for training the model parameters



Prediction

- ❖ The Historical time series $y_{i,t}$ is fed in for $t < t_0$ (Training Range)
- ❖ For $t \geq t_0$ (Prediction Range) a sample $\hat{y}_{i,t} \sim L(\cdot|\theta_{i,t})$ is drawn and fed back for the next point until the end of the prediction range T ; generating one sample
- ❖ Repeating the process yields many samples from the joint predicted distribution, whose sample quantiles are used for Final Prediction

Scaling

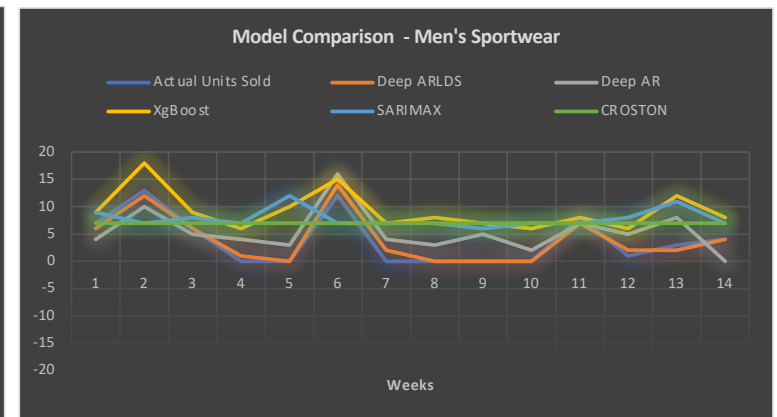
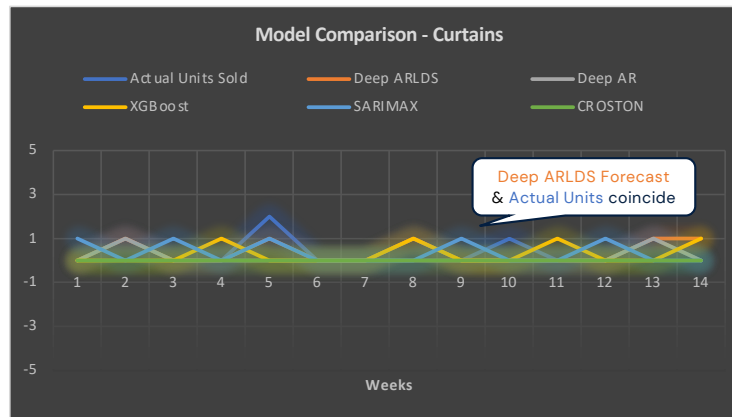


Results

- ❖ Solution scaled for 32 Departments (~30000 SKUs) across 210 Walmart Stores for 14 weeks ahead Demand Forecasts
- ❖ **Deep ARLDS** scores the lowest error Metrics across all KPIs, except Underforecasting % (although the Total Overforecast + Underforecast % is vastly reduced)

Error KPIs	Deep ARLDS	Deep AR	XGBoost	SARIMAX	CROSTON
MAD (Overall)	0.89	1.42	1.48	4.27	9.89
MAD (Zero sales week)	0.17	1	1.28	2.56	4.24
MAD (Non-Zero Sales Week)	1.6	1.84	1.68	5.99	13.28
SMAPE	60.91	76.75	88.12	100.02	85.37
Overforecast	42%	48%	66%	76%	22%
Underforecast	31%	29%	27%	19%	73%

Sample Forecast Outputs at SKU-Store-Week Level



Conclusion and Future Scope

Summary

- ❖ The Meta Model Architecture of **Deep ARLDS** allows for more accurate demand predictions at granular levels, by successfully modelling the intermittency and temporal interactions through LDS and Deep Learning respectively
- ❖ The Model is easily scalable due to parallel implementations using PySpark and GPUs
- ❖ Can be generalized for Forecast Solutions where the assumptions of Time Series Modelling fail. (Yields accurate results for Standard Time Series as well, however Traditional Forecast Models will yield faster results in those scenarios)

Future Scope

- ❖ Further Speed up Deep ARLDS to achieve online live forecasts
- ❖ Add a Segmentation Module which allows for training multiple Deep ARLDS models within the same category for varying within-category item behaviour

DATA+AI
SUMMIT 2022

Thank you



Abhishek Sengupta

(<https://www.linkedin.com/in/mr-abhishek-sengupta/>)