

DS 5500 - Fall 2021

Capstone: Applications in Data Science

# Walmart Sales Forecasting

Final Presentation



Gourang Patel  
MS in Data Science



Hitashu Kanjani  
MS in Data Science



Sanjan Vijayakumar  
MS in Data Science



Sagar Singh  
MS in Data Science

# Agenda

- ❖ Summary
- ❖ Data Overview
- ❖ Methodology
- ❖ Project Timeline
- ❖ Best Model Benchmarks
- ❖ Deployment Demo
- ❖ Key Takeaways

# Summary

## Context :

- Ecommerce has been an ever-growing industry with projected revenue growth of **\$4.9 trillion** in 2021.
- Sales forecasting will help businesses understand changing customer demands, manage inventories and create a pricing strategy that reflects demand.

## Problem Goals :

- This project will present the right methodologies to analyze time-series sales data and predict **28 days** ahead point forecasts for Walmart to help take strategic decisions.
- We plan to leverage the traditional time series forecasting methods(**Phase 1**) as well as the modern forecasting methods(**Phase 2**), to analyze Walmart's sales data.

# Data Overview

The dataset contains **5 year historical sales** from 2011- 2016 for various products and stores.

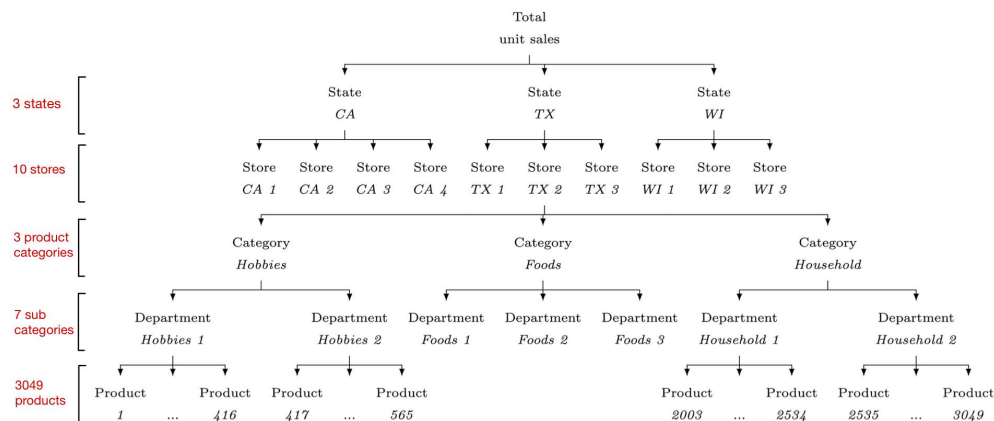
Data is hierarchically organized: stores are divided into 3 states, and products are grouped by categories and sub-categories

The dataset is organized in 3 CSV files :

**calendar.csv** - Contains dates on which the products are sold and events held on that day.

**Sales\_train\_evaluation.csv:** Contains historical daily unit sales of each product on each store

**Sell\_prices.csv:** price of products each week



# Methodology

We wanted to improve over the results from standard statistical time-series forecasting strategies from Phase 1, using advanced forecasting strategies which is our **primary goal for Phase 2**.

## Phase 1

### Traditional Time Series Models

- Involves historical analysis, finding dynamics of the data like cyclical patterns, trends, and growth rates.
- Three general ideas to tackle the forecasting problem would be Repeating/Static Patterns and Seasonal Trends.
- Exponential Smoothing(EA), ARIMA (Autoregressive integrated moving average) and SARIMA (Seasonal ARIMA) are some examples.

## Phase 2

### Causal Forecasting

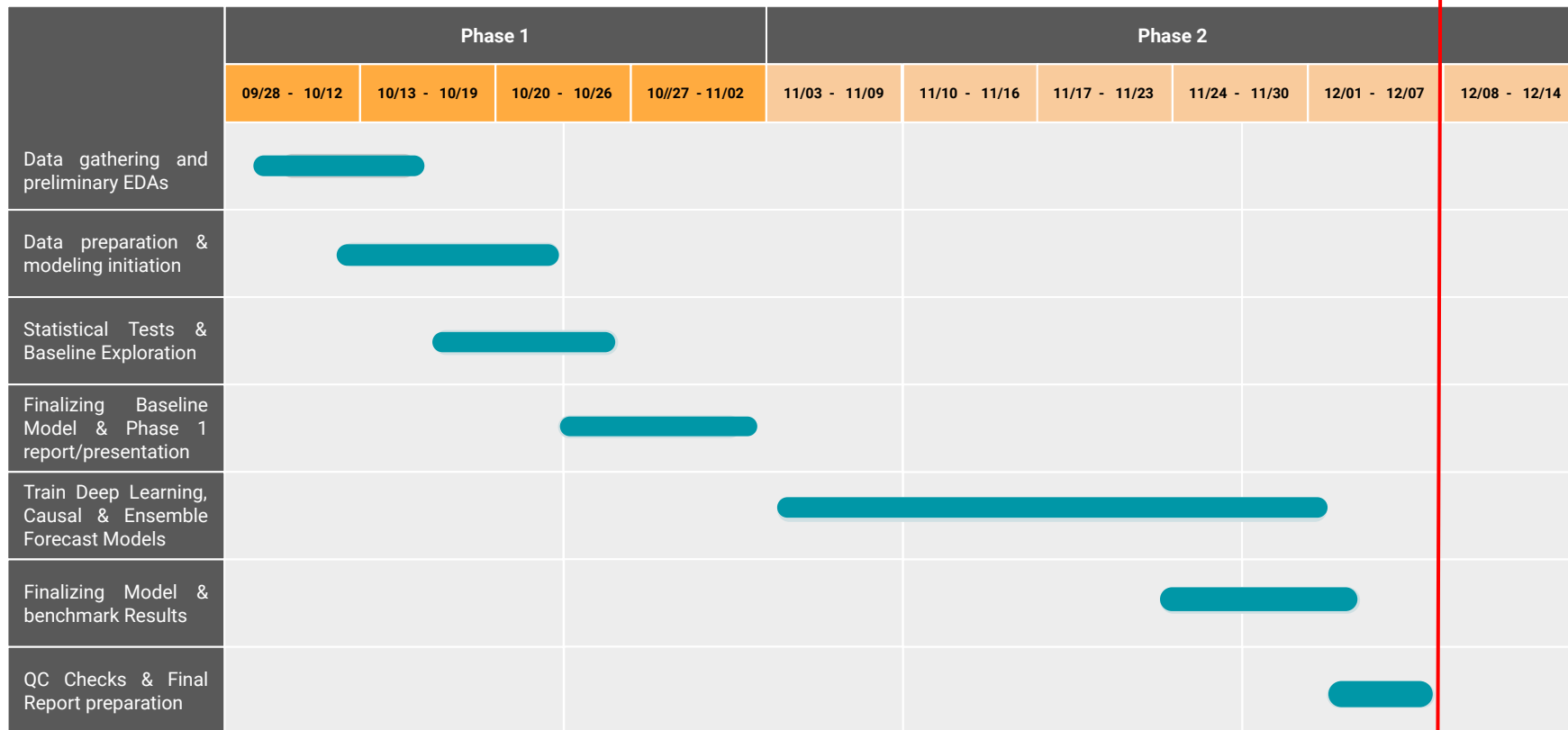
- Assumes the variable to be forecasted has a cause-effect relationship with one or more independent variables.
- For sales, it can be used to forecast at a much granular level i.e., by product, product category, subclass etc.
- Regression model and Econometric model are some examples we will explore.

### Deep Learning/Ensemble Models

- Neural Networks can learn inherent patterns in different time series without bothering about breaking up the trend and seasonality patterns.
- Forecasting can also be significantly improved using techniques like Gradient Boosted Trees.
- DeepVanilla LSTM (Long Short Temporal Memory), XGBoost and LightGBM are some examples.

# Timely completion of project resulted in best results

We are here



# Phase 2 Analysis



## Shortcomings of Phase 1

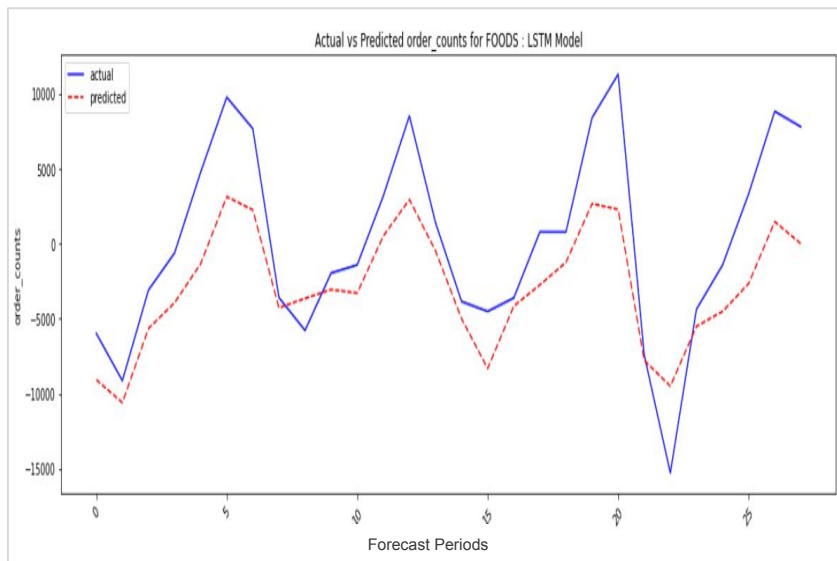
- Statistical models showed **poor fit when trained on the daily data**
  - Unable to capture the trend/seasonality
  - Data was **aggregated weekly** to smoothen some of underlying volatility
- SARIMAX had **high confidence range**, despite reduction in RMSE wrt baseline



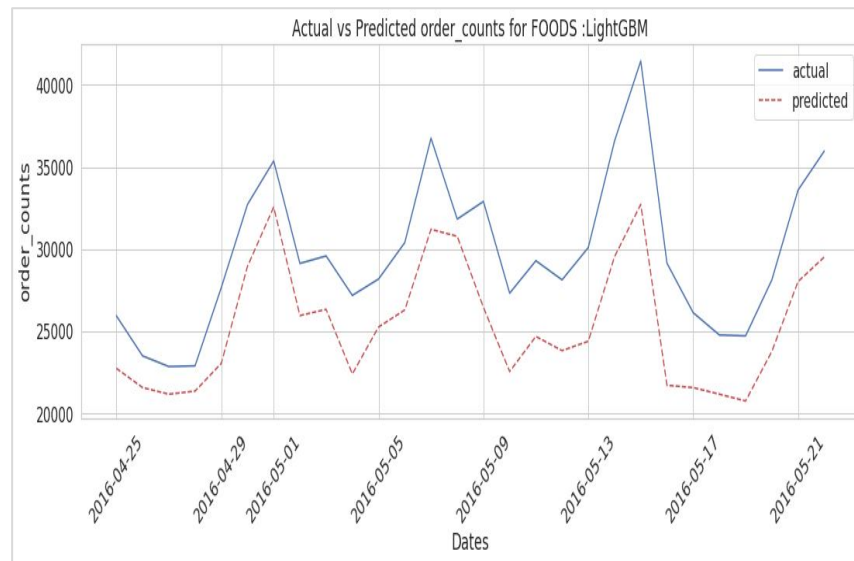
## Target for Phase 2

- Train models like RandomForest, **LightGBM** and **LSTM** on daily data
- Causal forecasting to model cause and effect relationship within time series
- Build a **frontend UI** for the complete project and deploy it

# LSTM showed ~8.9% improvement wrt SARIMAX [baseline] on FOODS



(i) LSTM

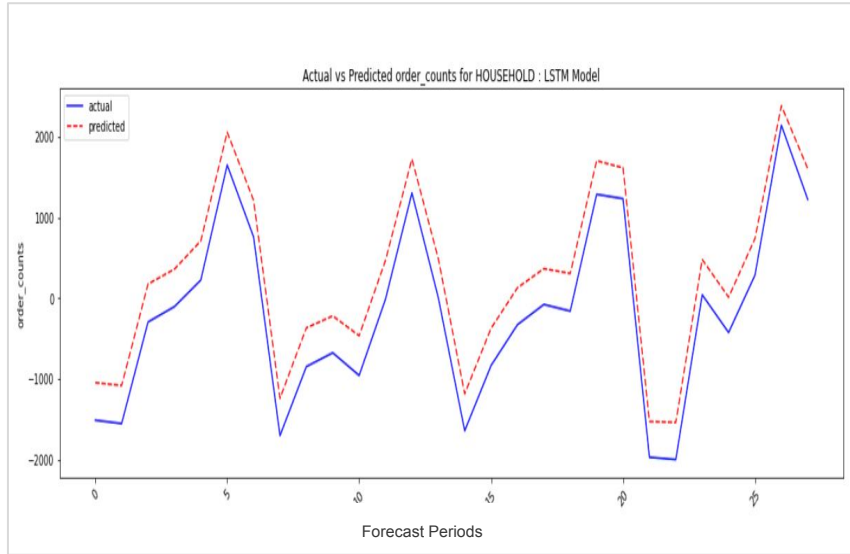


(ii) LightGBM

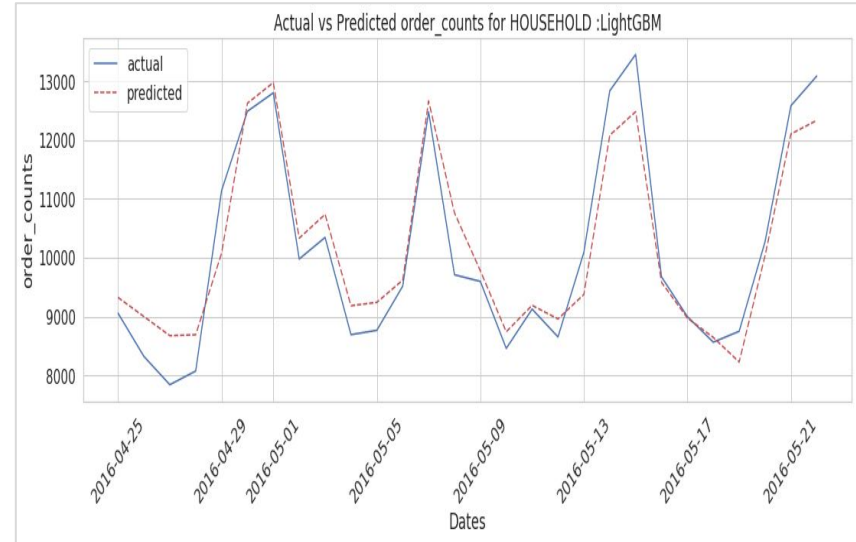
- Unlike Statistical models, advanced models were able to **fit on daily data** and forecast is for **28 days**
- **Exogenous variables** like snap, event v/s no events, paydays and holidays were also considered
- LSTM performed better than LightGBM with **~7.1% improvement** on RMSE Score



# LSTM showed ~66.1% improvement wrt SARIMAX [baseline] on HOUSEHOLDS



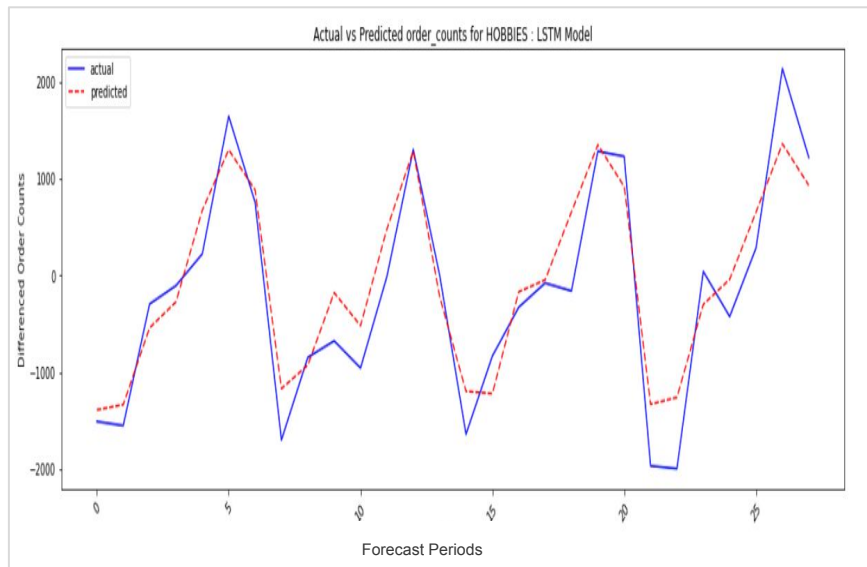
(i) LSTM



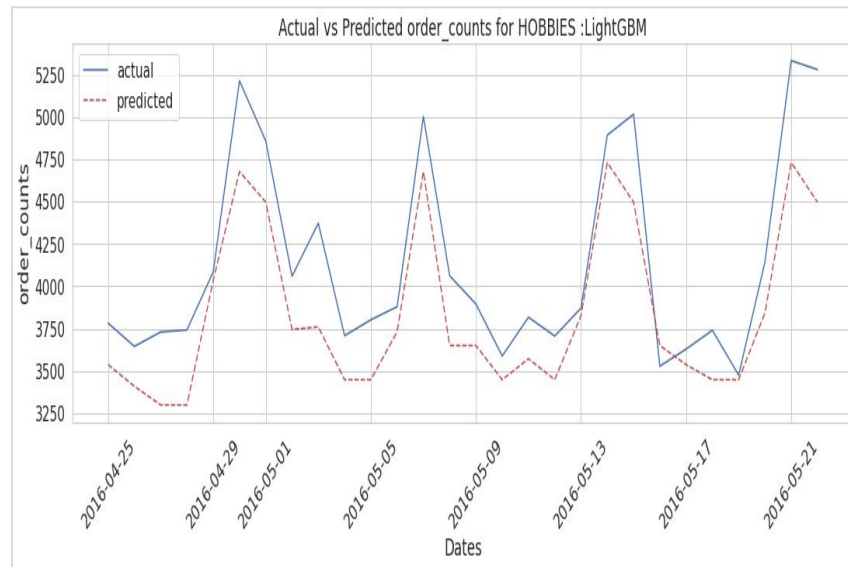
(ii) LightGBM

- Unlike Statistical models, advanced models were able to **fit on daily data** and forecast is for **28 days**
- **Exogenous variables** like snap, event v/s no events, paydays and holidays were also considered
- LSTM performed better than LightGBM with **~16.9% improvement** on RMSE Score

# No improvement on RMSE score for HOBBIES wrt SARIMAX [baseline]



(i) LSTM



(ii) LightGBM

- Unlike Statistical models, advanced models were able to **fit on daily data** and forecast is for **28 days**
- **Exogenous variables** like snap, event v/s no events, paydays and holidays were also considered
- No improvement on RMSE score for HOBBIES wrt LightGBM

# Model Benchmarks - RMSE Score Evaluation

MODELS / CATEGORIES		FOODS	HOUSEHOLD	HOBBIES
Phase 2	ARIMA	9,802.53	2,532.63	712.39
	ARIMA with Gridsearch	6,217.83	2,130.08	466.83
	SARIMAX with Gridsearch	4,802.55	1,314.61	215.28
	Random Forest	6,514.09	2,704.61	988.62
	LightGBM	4,705.59	536.88	356.64
	Deep Vanilla LSTM	4,373.36	446.29	646.36

Phase 1

# Learning and Outcomes

## Deprioritized Cloud Architecture

- Was incurring extra cost to us
- Streamlit, an open Source app was used for deployment (will be demoed)



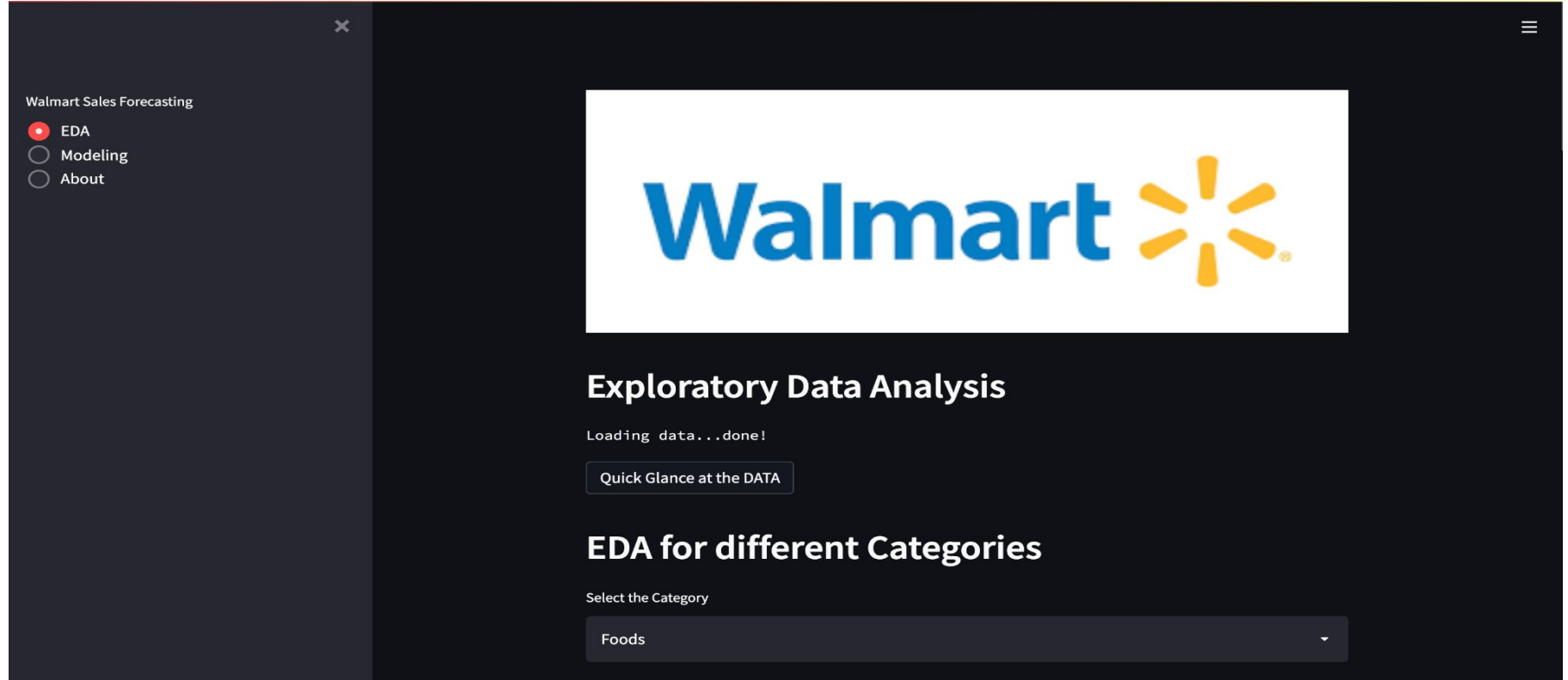
## LSTM Outperformed on all categories

- Models were able to fit on daily sampled data
- For HOBBIES the RMSE was higher but showed better trend capture

## Causal Forecasting wasn't possible

- Data was masked and Econometric fields like CPI was missing for Granger Causality tests
- State-wise CPI data was not easily feasible as it's purposely created with a focus on buying habits of urban customers

# Application UI - A Quick Glance



**Thank you!**