DS 5500 - Fall 2021

Capstone: Applications in Data Science

Walmart Sales Forecasting

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
- Initial Results
- Modeling Phase
- Key Learnings

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.

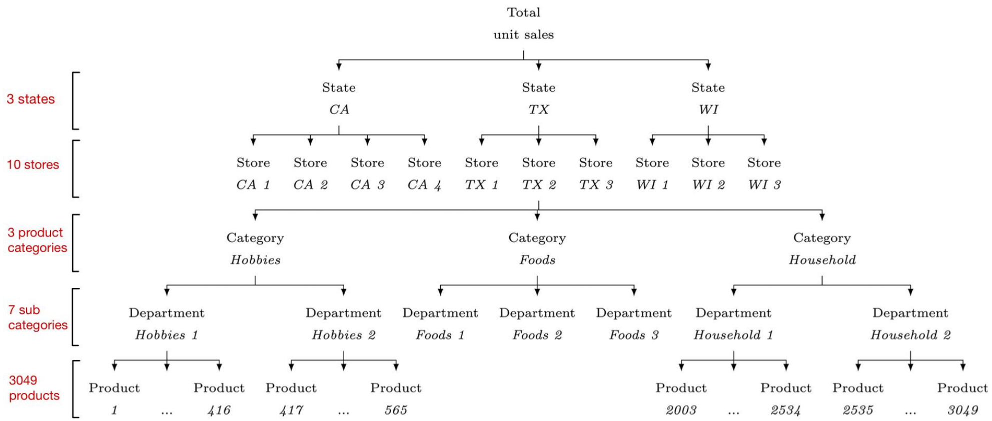
Data is hierarchically organized: stores are divided into 3 states, and products are grouped by categories and sub-categories

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



M5 Competition: https://mofc.unic.ac.cy/m5-competition/

Methodology

Standard statistical time-series forecasting strategies are important to establish a baseline for the model performance, which is our **primary goal for Phase 1**.

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.

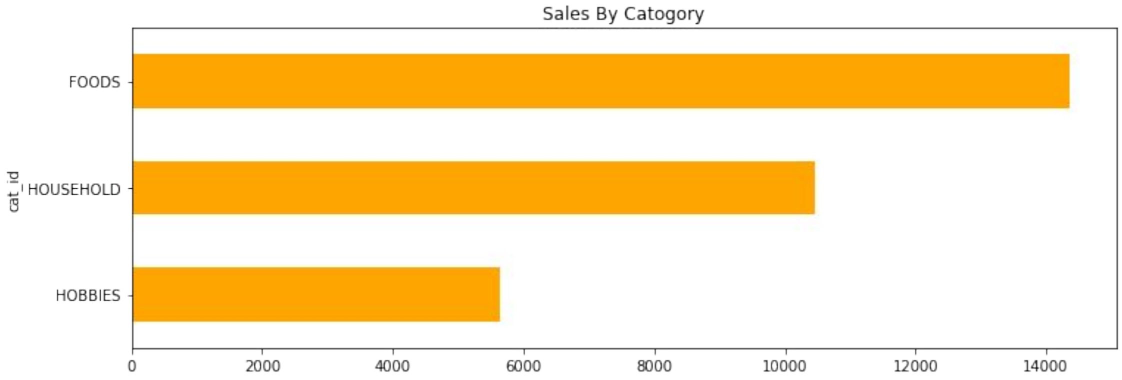
Causal Forecasting

- 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.

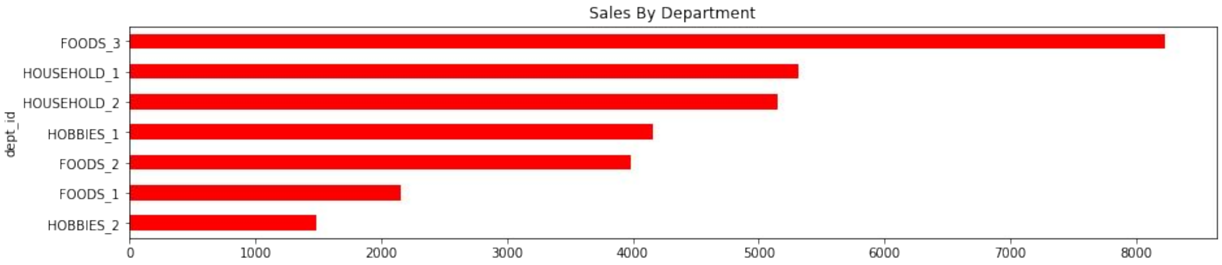
Deep Learning/Ensemble Models

- 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.

Initial Analysis: FOODS is the Top Category

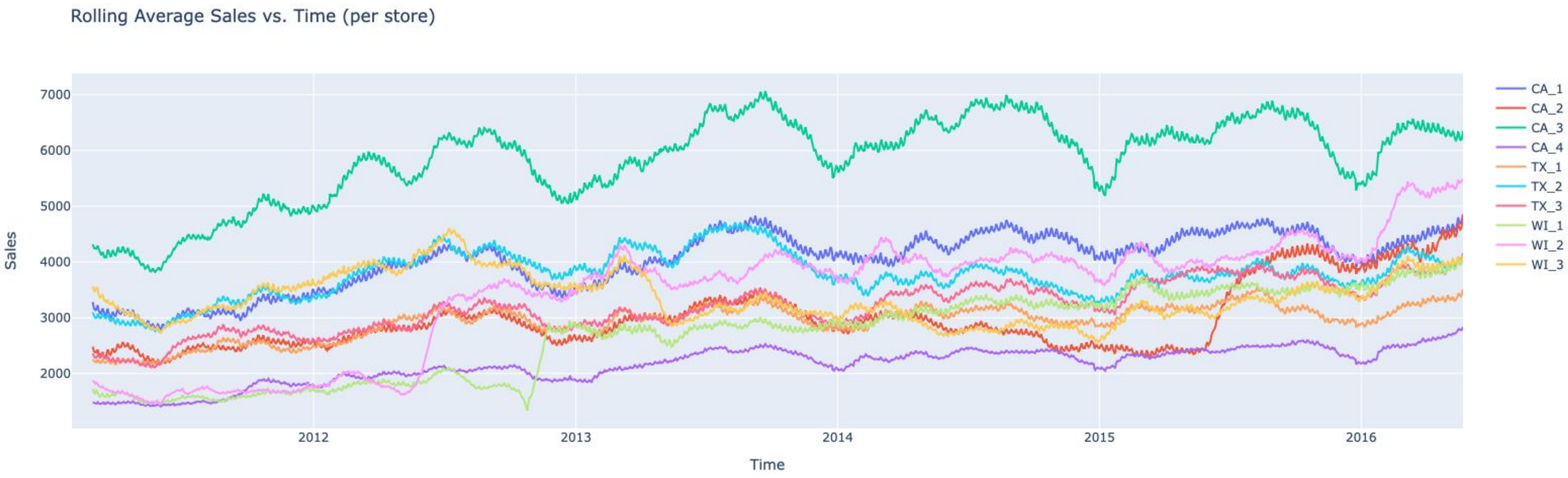


- FOODS category has the highest sales
- Precise forecasting on FOODS could be most beneficial



- Top 3 departments based on sales are FOODS_3, HOUSEHOLD_1 & HOUSEHOLD_2

Initial Analysis: CA has the highest overall sales



- Rolling average Sales with time for the store CA_3 situated in California is the highest amongst all
- We can also observe some seasonal high and lows in sales across all the stores

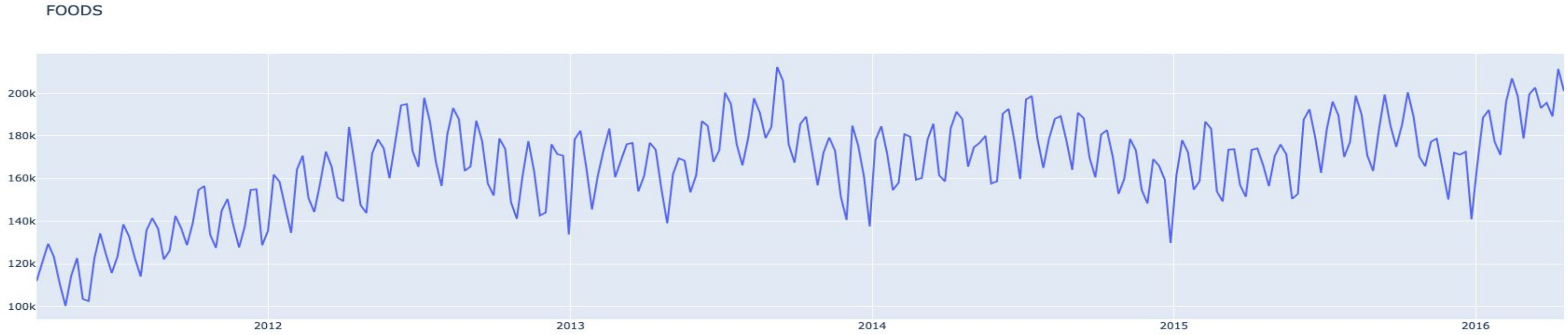
Modeling Initiation: Aggregated Weekly for Baseline

- Sales data is grouped by the different categories i.e FOODS, HOBBIES, HOUSEHOLD

- To reduce memory footprint, downcasting was performed along with weekly aggregation of sales

- Exogenous variables like holidays, weekends and paydays were created leveraging the calendar data

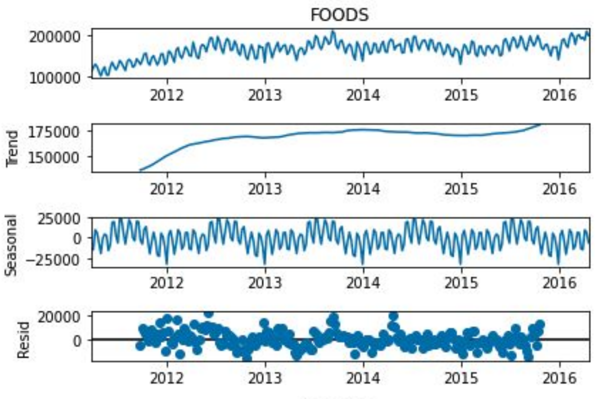
cat_id	FOODS	HOBBIES	HOUSEHOLD
d_1	23178	3764	5689
d_2	22758	3357	5634
d_3	17174	2682	3927
d_4	18878	2669	3865
d_5	14603	1814	2729

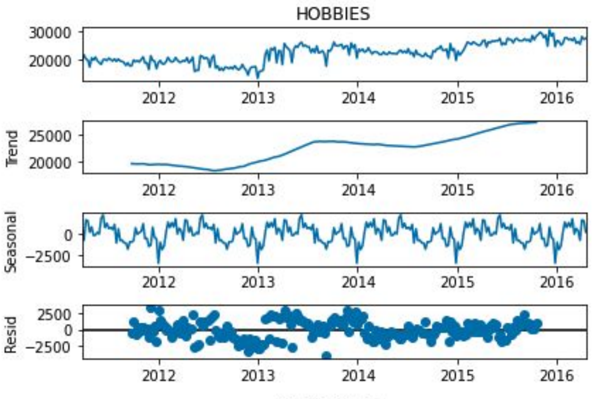


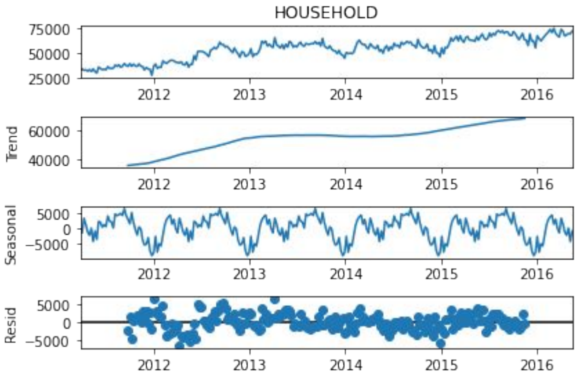
Total sales of the FOODS category across the entire timeline - Aggregated Weekly

Time series Decompose: Highlights Trend & Seasonality

Based on Trend and Seasonality plots, we hypothesize that data is **non-stationary**





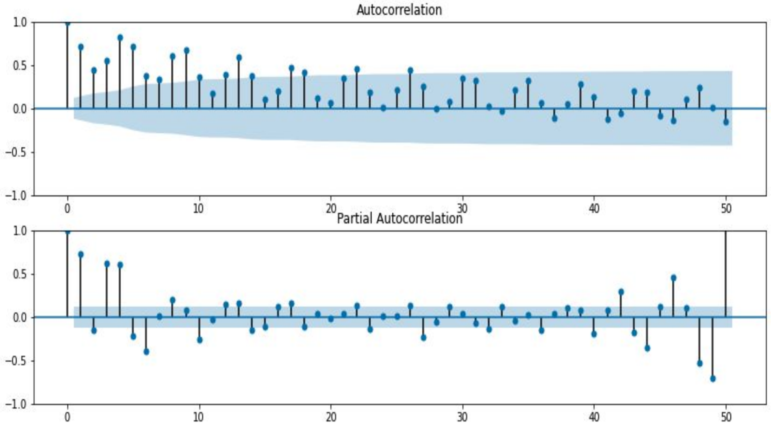


Statistical Tests: Augmented Dickey-Fuller Stationarity Test

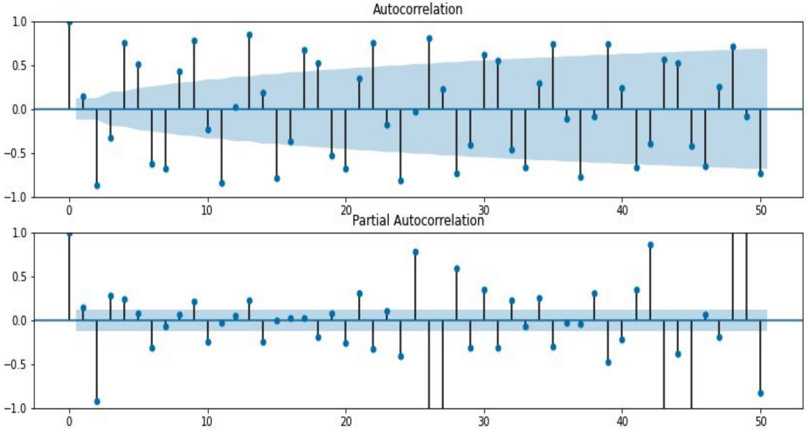
P-value of **<0.05** indicates data is stationarized and **suited for forecasting**

```
Results of Dickey-Fuller Test for :  FOODS
Test Statistic                  -2.521061
p-value                          0.110434
#Lags Used                      16.000000
Number of Observations Used    248.000000
Critical Value (1%)             -3.460000
Critical Value (5%)             -2.870000
Critical Value (10%)            -2.570000
dtype: float64
++++++++++++++++++++++++++++++++++++++
```

```
Results of Dickey-Fuller Test for :  FOODS
Test Statistic                  -4.734638
p-value                          0.000072
#Lags Used                      13.000000
Number of Observations Used    249.000000
Critical Value (1%)             -3.460000
Critical Value (5%)             -2.870000
Critical Value (10%)            -2.570000
dtype: float64
++++++++++++++++++++++++++++++++++++++++
```

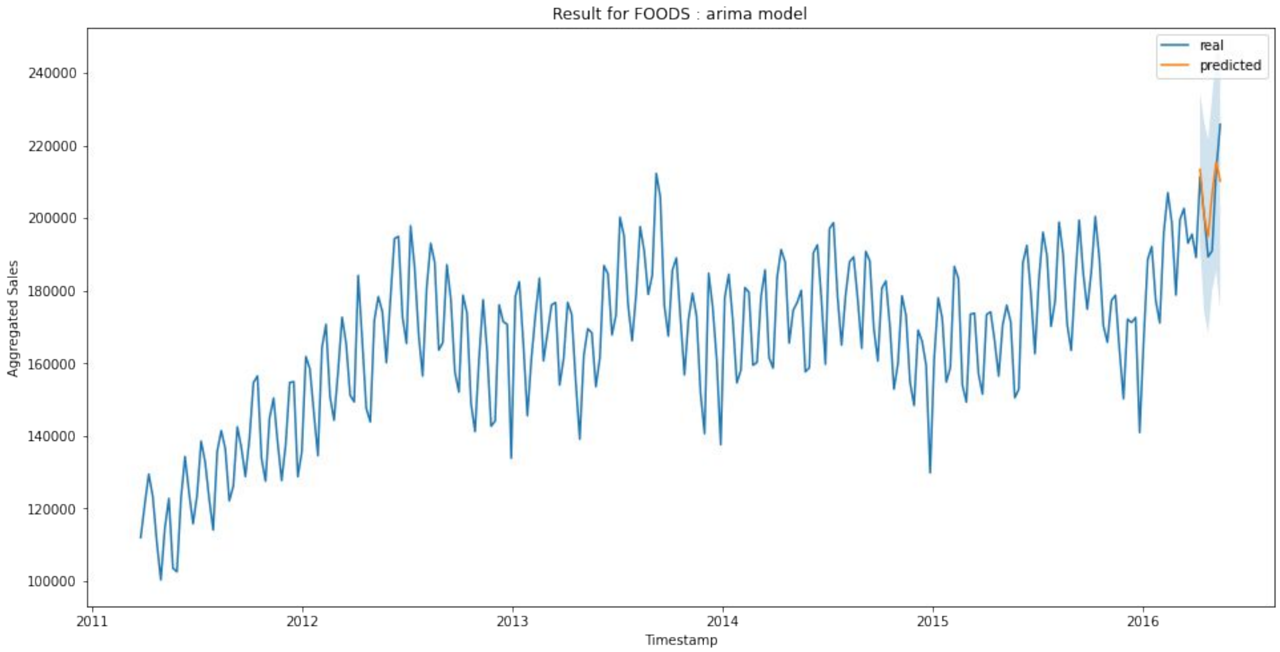


ACF/PACF plots before differencing



ACF/PACF plots after differencing

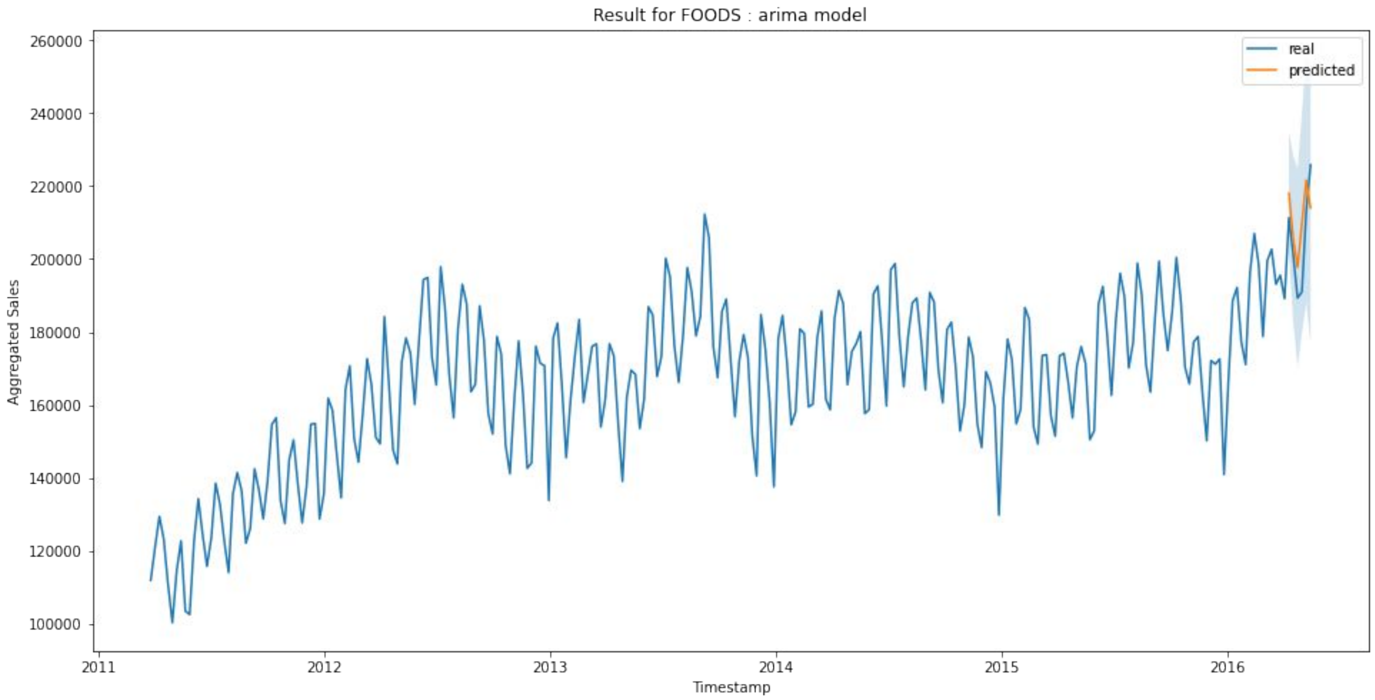
FOODS: Baseline ARIMA performance was not optimal



Results:

- Model tuning was performed using **Auto ARIMA.**
- The test size was kept at 6 weeks to be closer to 28 days forecast.

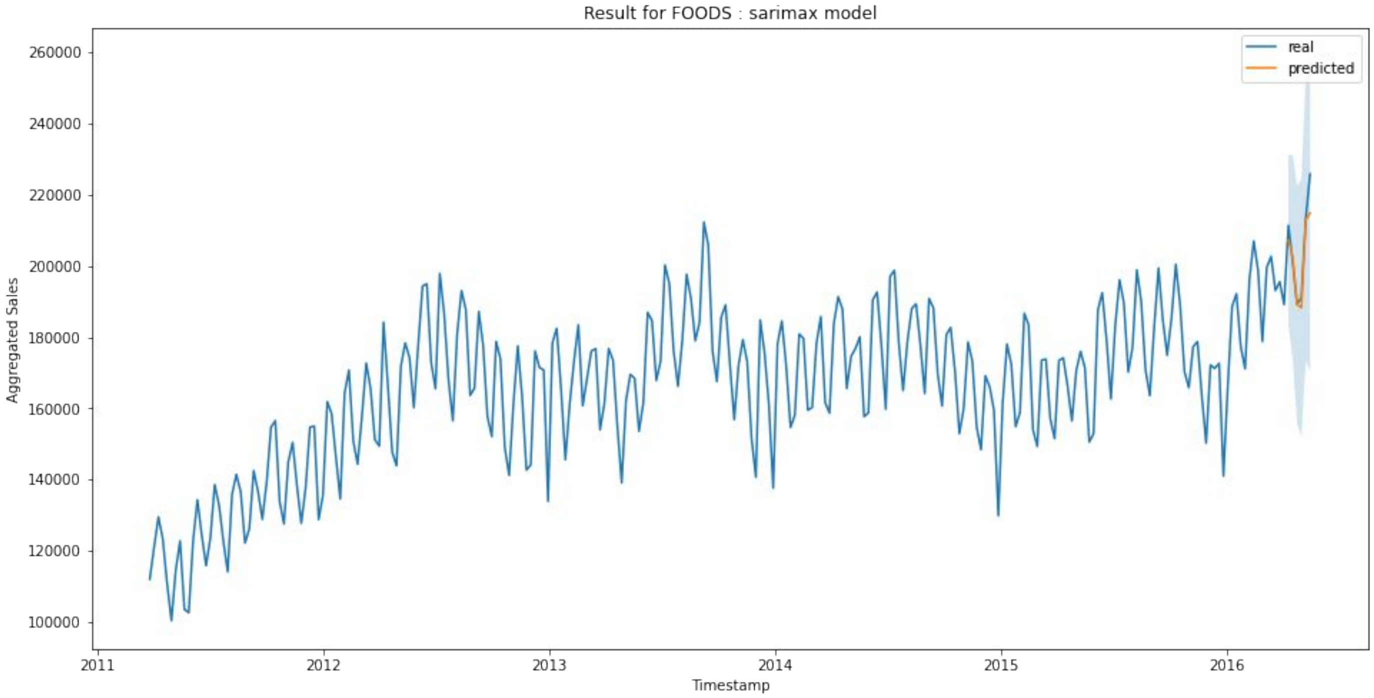
FOODS: Gridsearch optimized Baseline ARIMA by 36%



Results:

- The test size was kept at 6 weeks to be closer to 28 days forecast.
- As an alternative to k-fold CV, **Walk Forward validation** was used to perform backtesting.
- GridSearch gave RMSE of 6217.83, a **~36.5%** improvement over baseline ARIMA

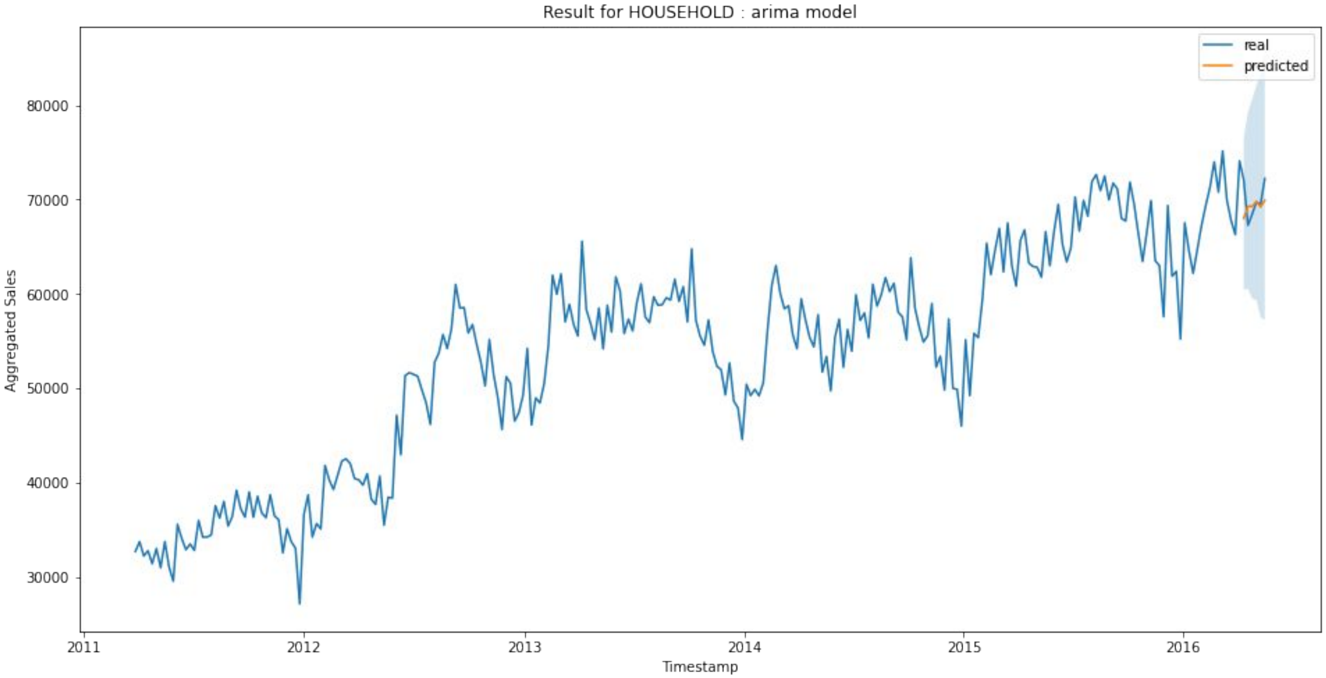
FOODS: SARIMAX improved ARIMA model by further 20%



Results:

- **Exogenous variables** like holidays, paydays and weekend were modeled to improve trend capture.
- Improved RMSE score of 4969.28, further improved the forecasting RMSE by **~20.1%.**

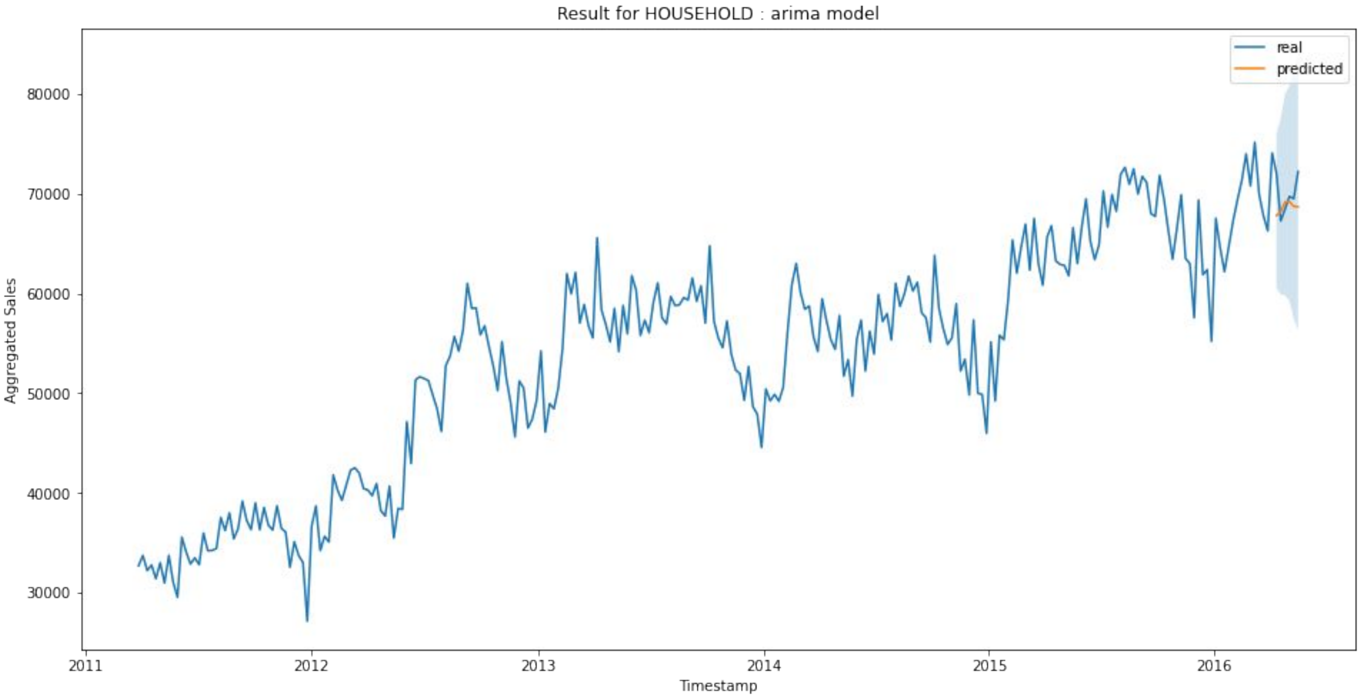
HOUSEHOLD: Baseline ARIMA model showed a poor fit



Results:

- Model tuning was performed using Auto ARIMA.
- **Log Transformation** was applied to account for variance
- The test size was kept at 6 weeks to be closer to 28 days forecast.

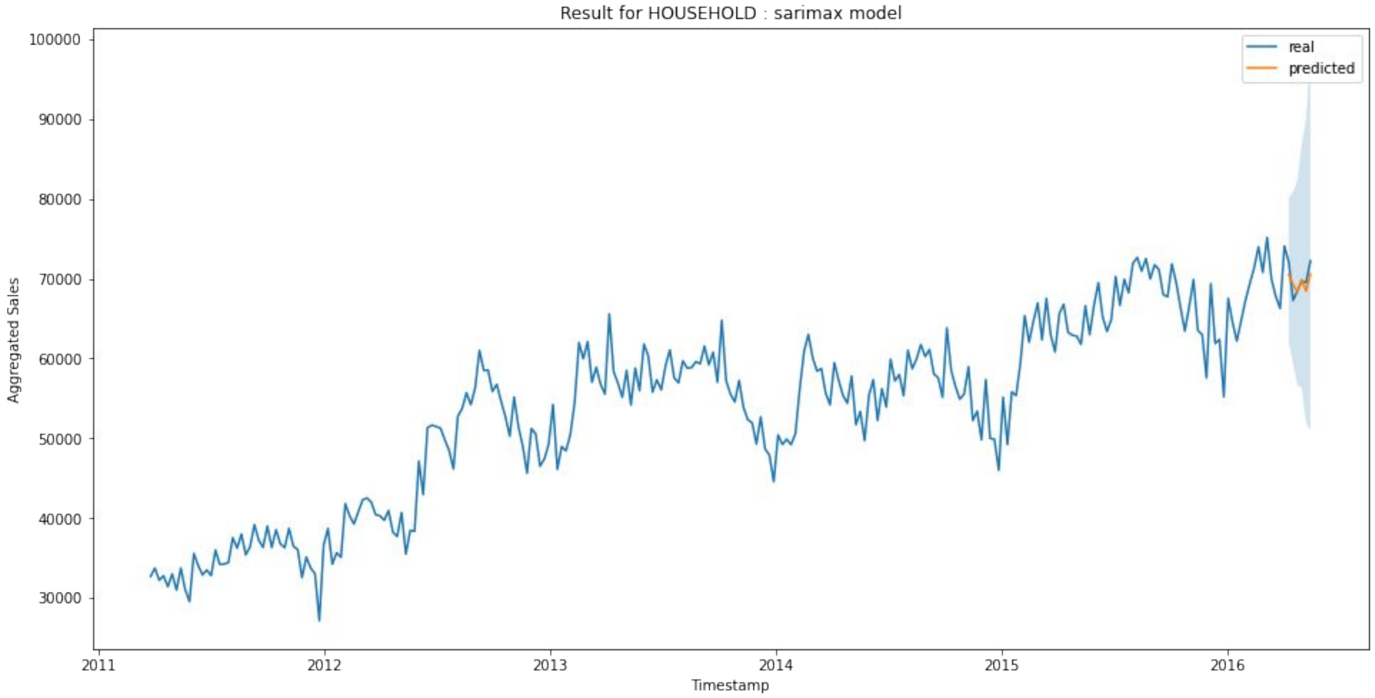
HOUSEHOLD: ARIMA model shows poor fit despite gridsearch



Results:

- The test size was kept at 6 weeks and backtesting was performed using Walk Forward validation.
- **Log Transformation** was applied to account for variance
- RMSE of 2130.08 after gridsearch, only showed a minor **~15.9%** improvement over baseline ARIMA.

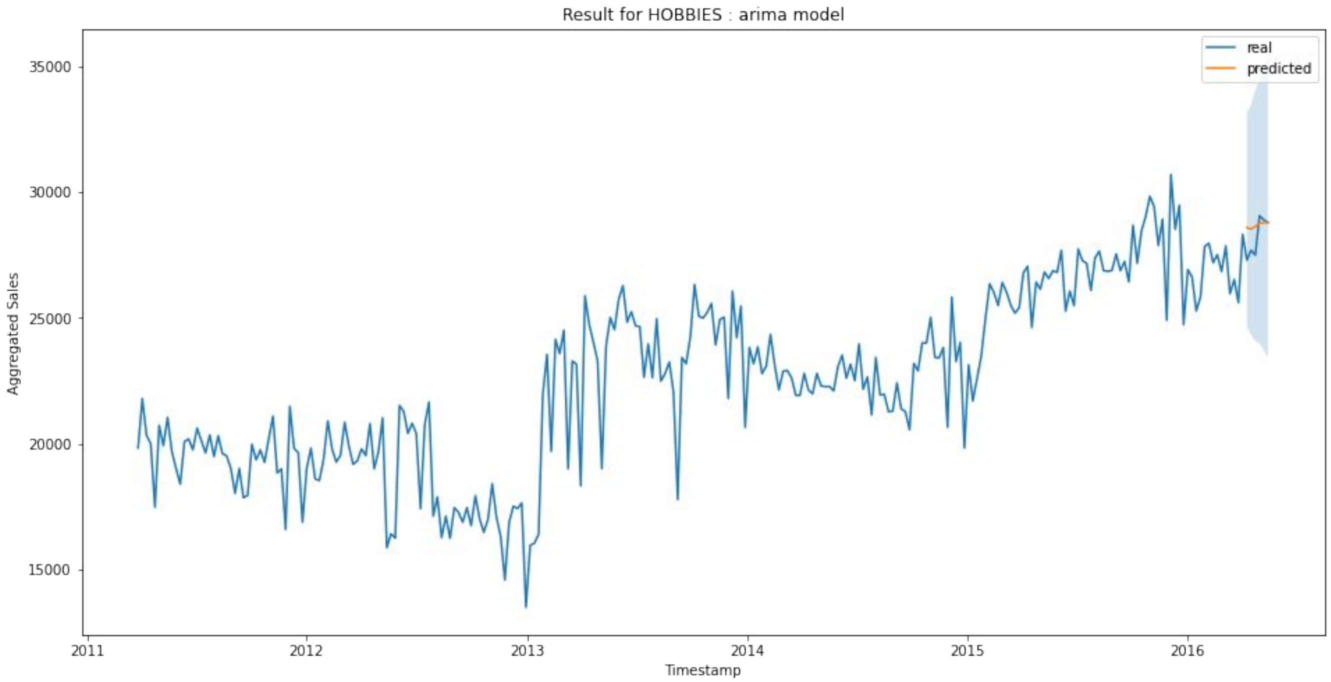
HOUSEHOLD: SARIMAX improved ARIMA model fit by 38%



Results:

- **Exogenous variables** like holidays, paydays and weekend were modeled to improve trend capture.
- RMSE score of 1314.61 after gridsearch optimization, showed a significant improvement of **~38%** over best ARIMA model.

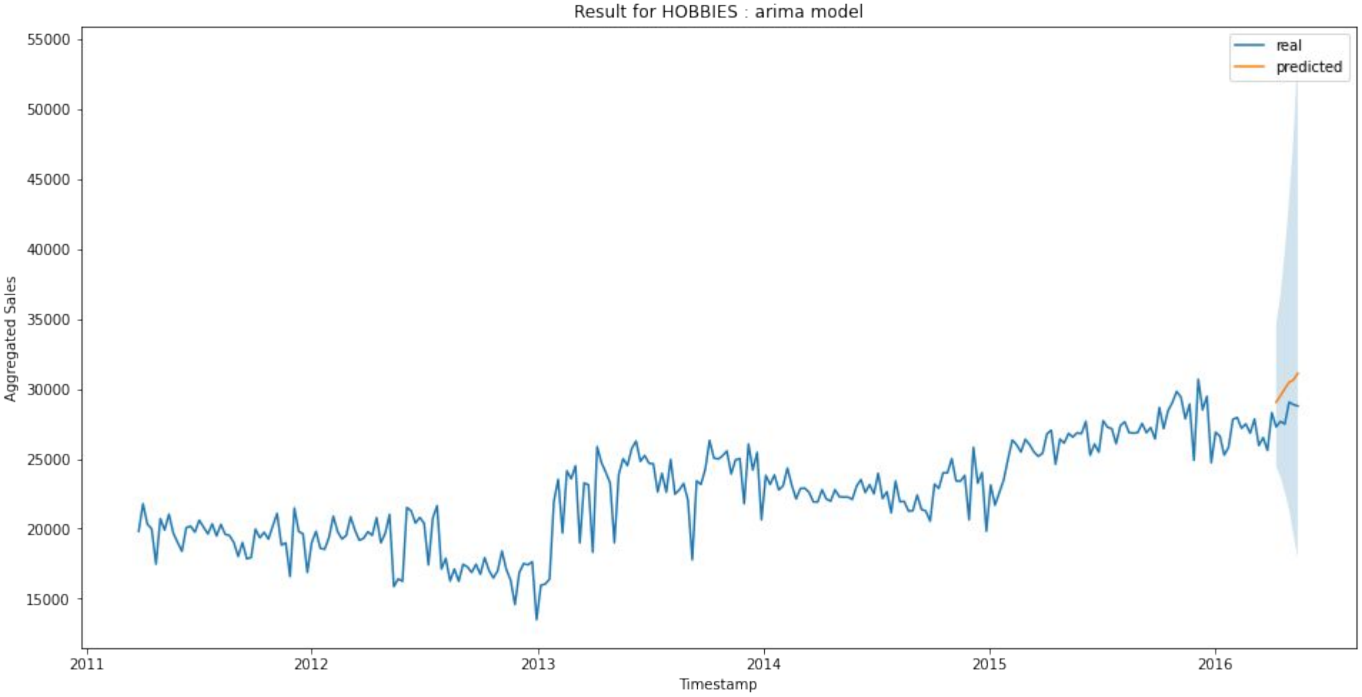
HOBBIES: Baseline ARIMA model showed almost no trend capture



Results:

- Model tuning was performed using **Auto ARIMA.**
- **Log Transformation** was applied to account for variance
- The test size was kept at 6 weeks to be closer to 28 days forecast.

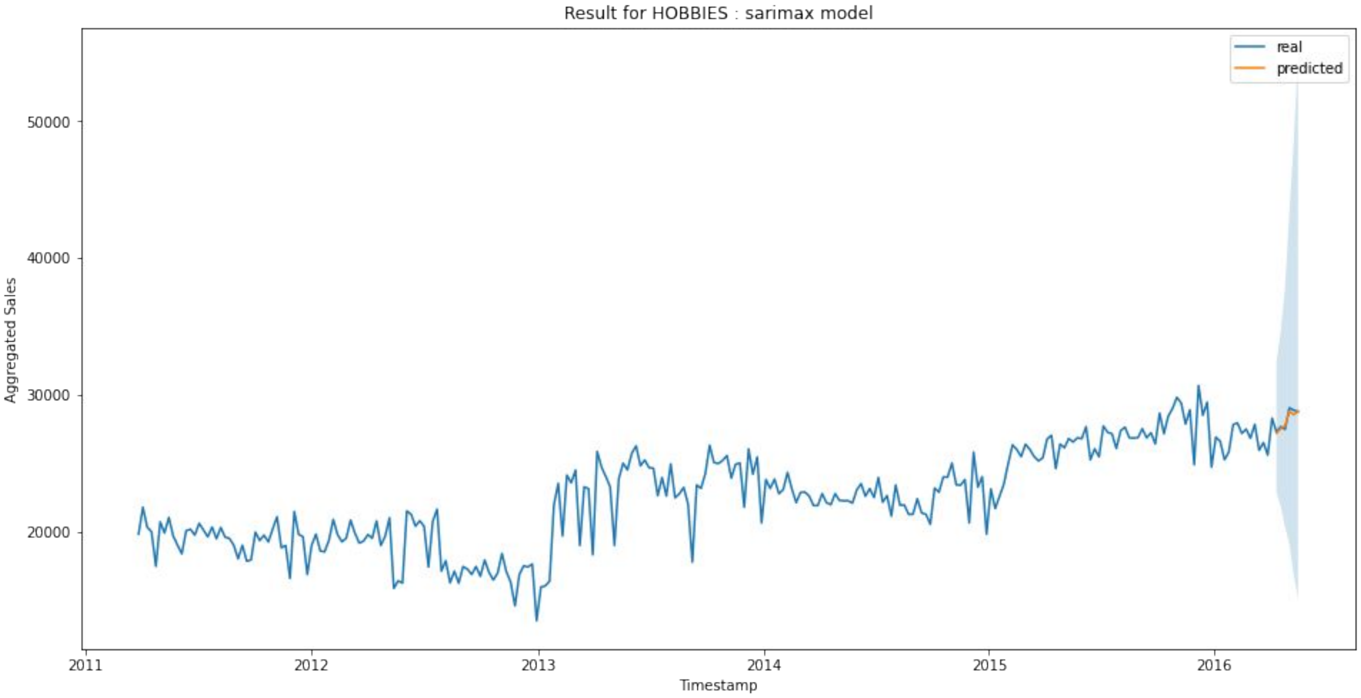
HOBBIES: ARIMA showed a very poor fit despite Gridsearch



Results:

- The test size was kept at 6 weeks and backtesting was performed using Walk Forward validation.
- Log Transformation was applied to account for variance.
- ARIMA model was unable to show a good fit even after applying gridsearch based optimizations.

HOBBIES: SARIMAX improved ARIMA model fit by 53%



Results:

- Exogenous variables like holidays, paydays and weekend were modeled to improve trend capture.
- RMSE score of 215.28 after gridsearch optimization, showed an impressive **~53%** improvement over best ARIMA model.

Phase 1 Goals were successfully met

Key Learnings

- ➔ Downcasting is a good alternative to Spark and other Cloud based architecture when training on larger data.
  - ◆ Reduced memory usage by almost ~80%.
- ➔ Walk Forward Validation is a good alternative to k-fold CV for backtesting.
- ➔ SARIMAX performed better than ARIMA across all the categories, due to better Trend and Seasonality capture.
- ➔ Log Transformation is a good step when the data has lot of variance (eg: in case of HOBBIES and HOUSEHOLD).
- ➔ While Gridsearch performs much slower than Auto ARIMA, it showcased much better results on almost all the 3 categories.

Despite lower RMSE, larger confidence interval in SARIMAX results is a compelling reason to try modern forecasting techniques, which will be the primary goal for Phase 2.