WALMART SALES FORECAST PROJECT

CAPSTONE PROJECT 2

YOHEITA YOSHIMURA

MAY 27TH 2024

AGENDA

Introduction

Data acquisition & processing

Exploratory Data Analysis (EDA)

Methodology

Model Evaluation

Result

Conclusion

BACKGROUND

INTRODUCTION

Background information on the Walmart

- The largest company by the revenue in the world in 2023 (2nd place is Amazon)
- Fiscal year 2024 revenue of $648billion, a 6.0% increase from 2023
- Operates over 10,500 stores in 19 countries
- 210 distribution centers (9,000 tractors, 80,000 trailers and more than 11,000 drivers)
- Offering more than 75 million products (nearly 1.6 million products U.S. alone)

Referrance:
https://corporate.walmart.com/
https://stockanalysis.com/

INTRODUCTION

Project background

- Create a forecast model to predict the sales of each product in each stores.
- Use the data from M5 Forecasting Competition Project from Kaggle
- The data is provided by University of Nicosia
- We will find the most accurate model for the forecast
- We will use Python environment for this project

The data can be found at:
M5 Forecasting - Accuracy | Kaggle

APPROACH

APPROACH OF THIS PROJECT

Data acquisition and wrangling

- Input the data into python environment
- Understand data structures and analyze the provided data
- Data cleaning and handling missing values

Exploratory Data Analysis

- Data Visualization
  - Decompression of data for seasonality and trend
  - ACF and PACF graph to determine ARIMA(p,d,q) values
- Data transformation
- Integrate the data

Modeling

- Single product forecast with ARIMA and SARIMAX
- Apply AutoARIMA
- Apply Regression Model to forecast
  - Linear Regression, Decision Tree Regressor, XGBoost Regressor, and AutoML
  - Hyperparameter Tune the model
- Test with multiple products

Results

- Statical accuracy comparison
- Plot KDE graph
- Compare the accuracy of the prediction
- Feature Importance comparison

DATA ACQUISITION & PROCESSING

DATA ACQUISITION

Provided Tables

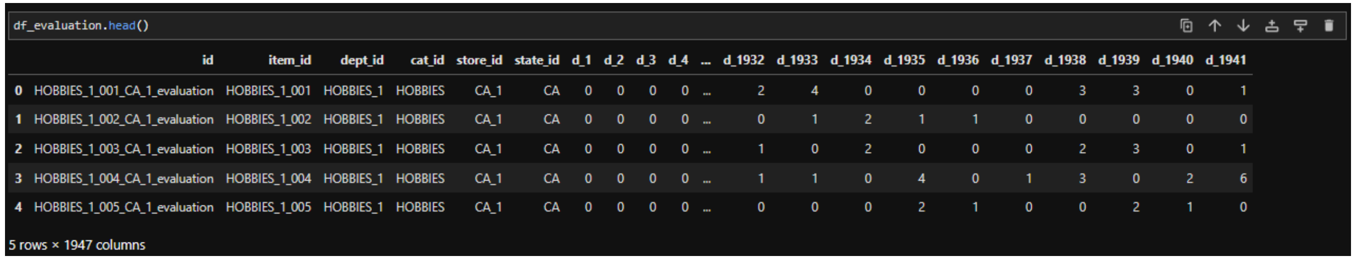
1. Calender.csv
   a. Contains date related information
      1) Dates (week day count, month, year)
      2) Events name and type of the event
      3) Supplemental Nutrition Assistance Program (SNAP) Availability date for each state
      4) Other date related data:
         a) Week ID (wm_yr_wk),
         b) d data (d_X where X is the number of date starting from 2011-01-29 -> 2016-06-19),
         c) weekly date count (Saturday = 1, Sunday = 7)

What is SNAP?
Program to benefits to eligible low-income individuals and families through an Electronic Benefits Transfer card, allowing recipients to purchase eligible food items at authorized retail food stores.
https://www.benefits.gov/benefit/361

DATA ACQUISITION

Provided Tables

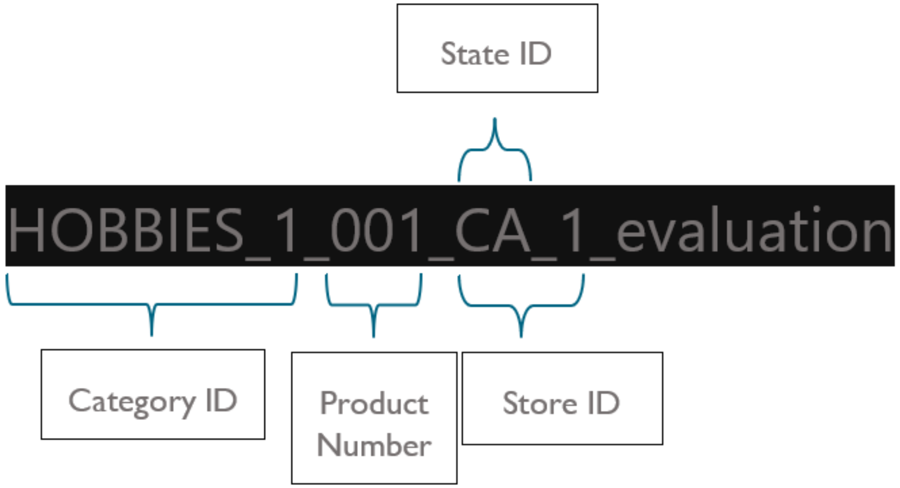
2. Sales_train_evaluation.csv
    a. Contains data of product id and daily sold quantity for each products
    b. Product id contains information on the Category ID, Store ID, State ID, and Product number
    c. Date of the daily sold quantity is represented with d value.



```
df_evaluation.head()
```

	id	item_id	dept_id	cat_id	store_id	state_id	d_1	d_2	d_3	d_4	...	d_1932	d_1933	d_1934	d_1935	d_1936	d_1937	d_1938	d_1939	d_1940	d_1941
0	HOBBIES_1_001_CA_1_evaluation	HOBBIES_1_001	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	...	2	4	0	0	0	0	3	3	0	1
1	HOBBIES_1_002_CA_1_evaluation	HOBBIES_1_002	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	...	0	1	2	1	1	0	0	0	0	0
2	HOBBIES_1_003_CA_1_evaluation	HOBBIES_1_003	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	...	1	0	2	0	0	0	2	3	0	1
3	HOBBIES_1_004_CA_1_evaluation	HOBBIES_1_004	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	...	1	1	0	4	0	1	3	0	2	6
4	HOBBIES_1_005_CA_1_evaluation	HOBBIES_1_005	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	...	0	0	0	2	1	0	0	2	1	0

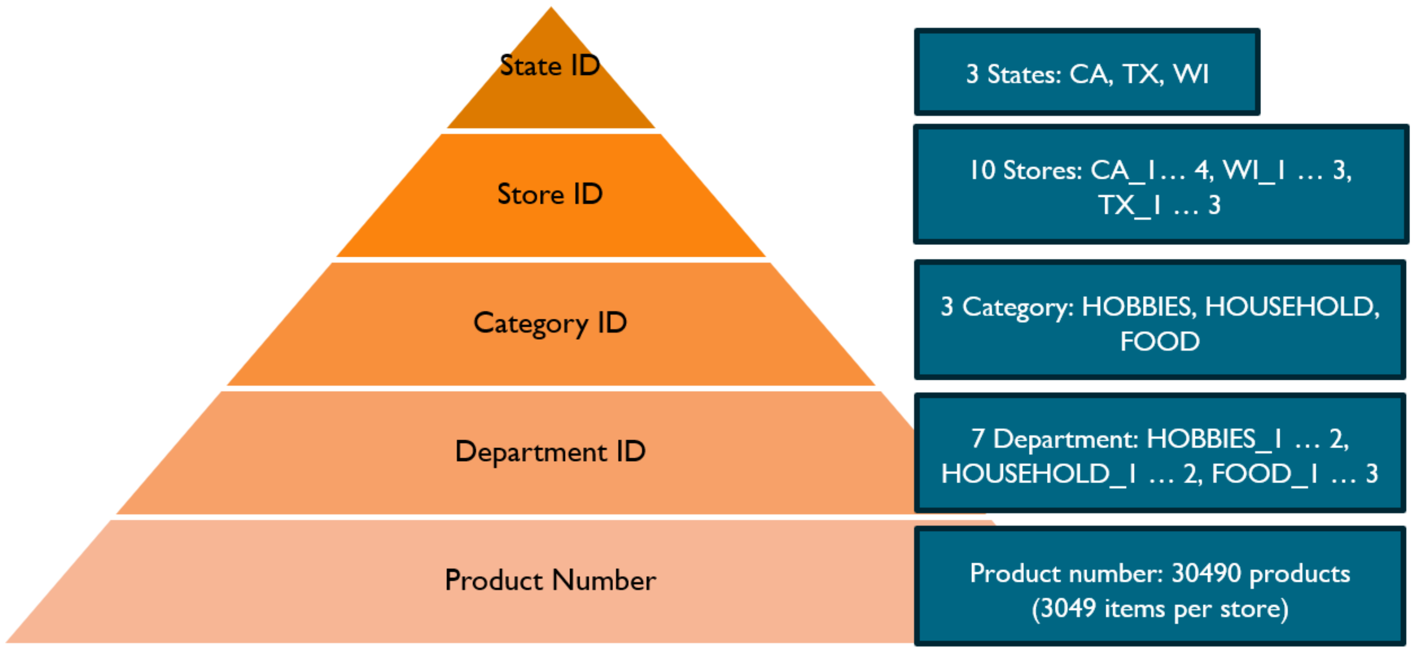
5 rows × 1947 columns



DATA ACQUISITION

Provided Tables

2. Sales_train_evaluation
   d. Data hierarchy
      1) 3 State data
         a) 4 Stores in CA
         b) 3 Stores in TX
         c) 3 Stores in WI
      2) 3 Category of products
         a) FOODS 1437
         b) HOBBIES 565
         c) HOUSEHOLD 1047
      3) 7 Departments (HOOBIES 1,2, HOUSEHOLD1,2, and FOODS1,2,3,4)
      4) Total of 30490 Products in this data



DATA ACQUISITION

Provided Tables

3. sell_prices.csv
   a. Contains data of store id and item id to distinguish products and week's id (wm_yr_wk) and the price of the product on that week

	store_id	item_id	wm_yr_wk	sell_price
0	CA_1	HOBBIES_1_001	11325	9.58
1	CA_1	HOBBIES_1_001	11326	9.58
2	CA_1	HOBBIES_1_001	11327	8.26
3	CA_1	HOBBIES_1_001	11328	8.26
4	CA_1	HOBBIES_1_001	11329	8.26

DATA PROCESSING

Data Preprocessing

Data Cleaning and Feature Engineering

a. Contains no skipping data of date

b. NaN in calendar for the event_type and event_name was replaced with non_NaN values

c. Adding Extra Features:

   1) Additional temporal Information.

      a) "is_weekend": A indicator denoting whether the date falls on a weekend.

      b) "Month_day": The day of the month.

      c) "week_number": The week number of the year.

      d) "day": The sequential day count throughout the dataset.

   e) To capture historical patterns: Lags, Rolling Mean, Rolling STD, Gap Indicators.

DATA PROCESSING

Data Preprocessing

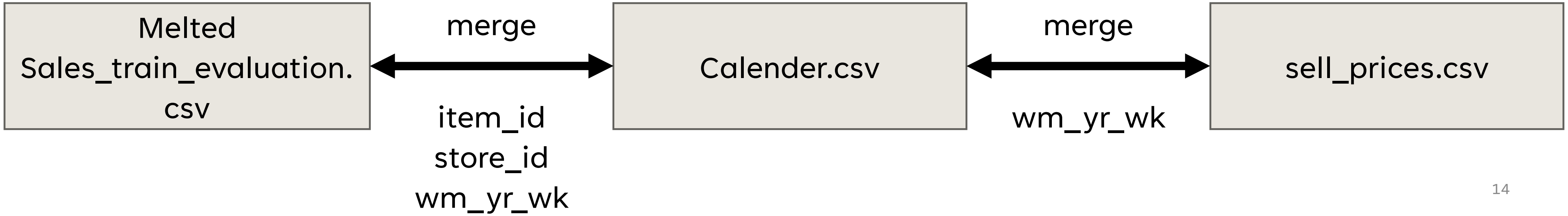
Creating the final table

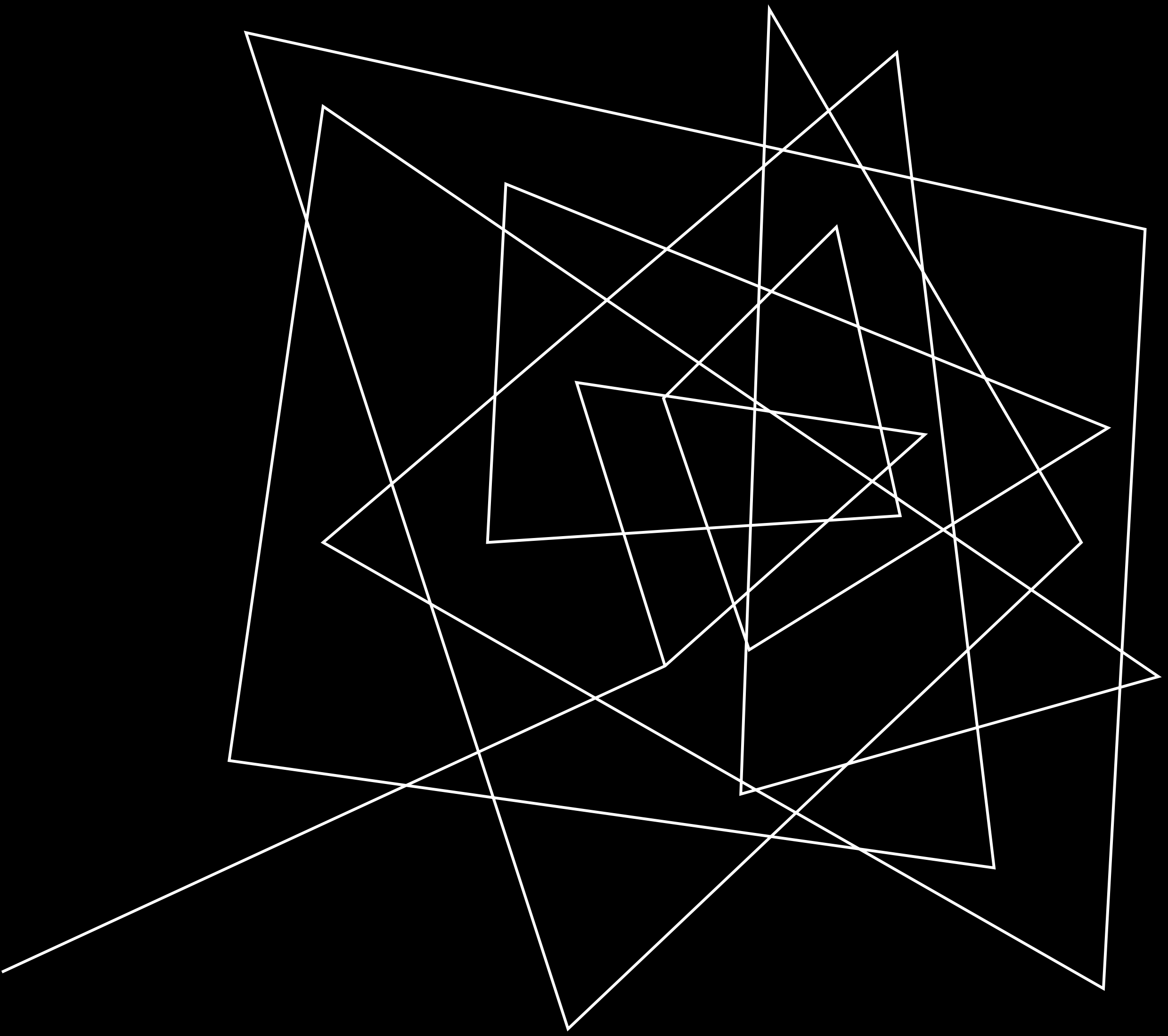
d. Merged all three tables into one for simplified data management and consolidate relevant information

e. Calculating the 'Daily_Sell' by multiplying the quantity sold by the price of the product on that day provides a useful metric for understanding daily sales revenue.

f. Adding Lag and Rolling Features:

- Including lag features (lag1 ~ 49) captures the historical behavior of the target variable
- Adding rolling median and rolling standard deviation features (rolling_median7 ~ 56, rolling_std7 ~ 56) helps to smooth out noise and capture trends in the data over a specified window of time.

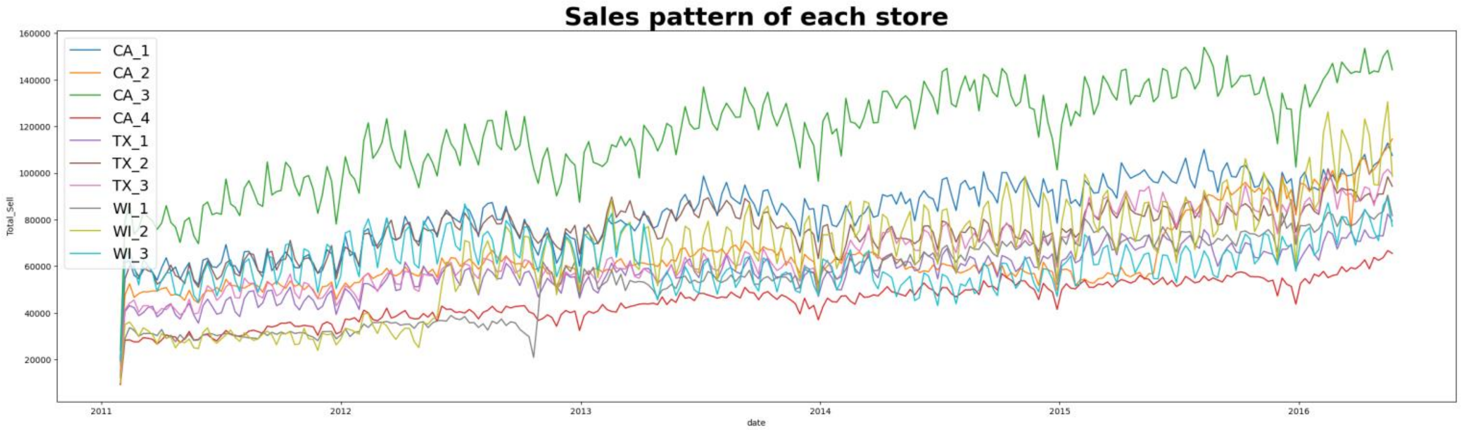




EXPLORATORY DATA ANALYSIS (EDA)

EXPLORATORY DATA ANALYSIS

Sales comparison of each state

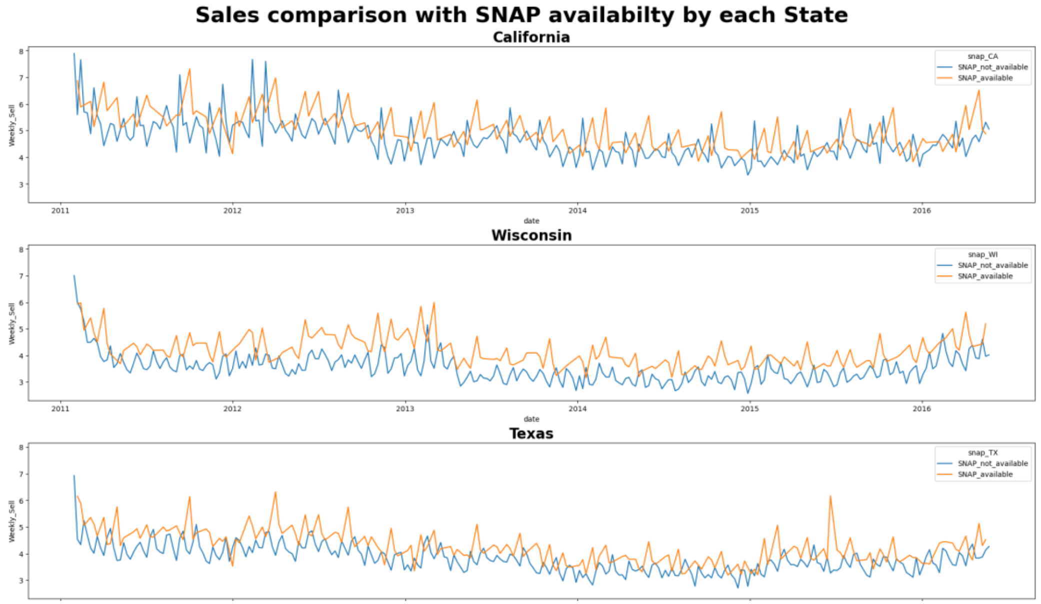


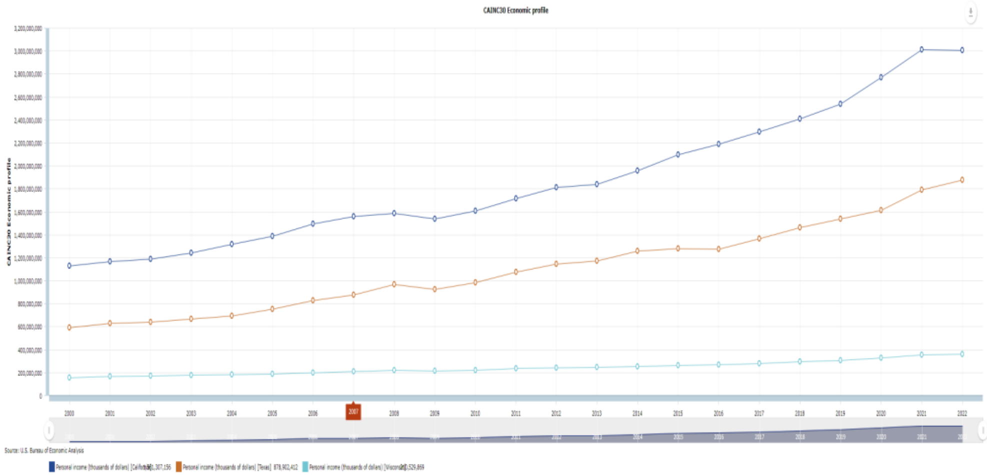
1. Sales plot for the each stores
    a. Highest revenue store seems to be Store CA_1 and the lowest is CA_4
        a. Possibly explaining the economy gap within the CA
    b. Overall WI stores seems to have lower sell trend than other stores (Excluding than CA_4)

EXPLORATORY DATA ANALYSIS

Sales comparison with SNAP availability

2. Average Sales comparison of SNAP Availability
    a. Able to observe greater gap between Wisconsin than Texas or California.
        a. This shows higher needs at WI than other two states (TX and CA) showing the level of economy difference between three states.
    b. The below graph is from https://apps.bea.gov/ displaying the Personal income difference between CA, TX, and WI.
        a. Where income is (CA > TX > WI)

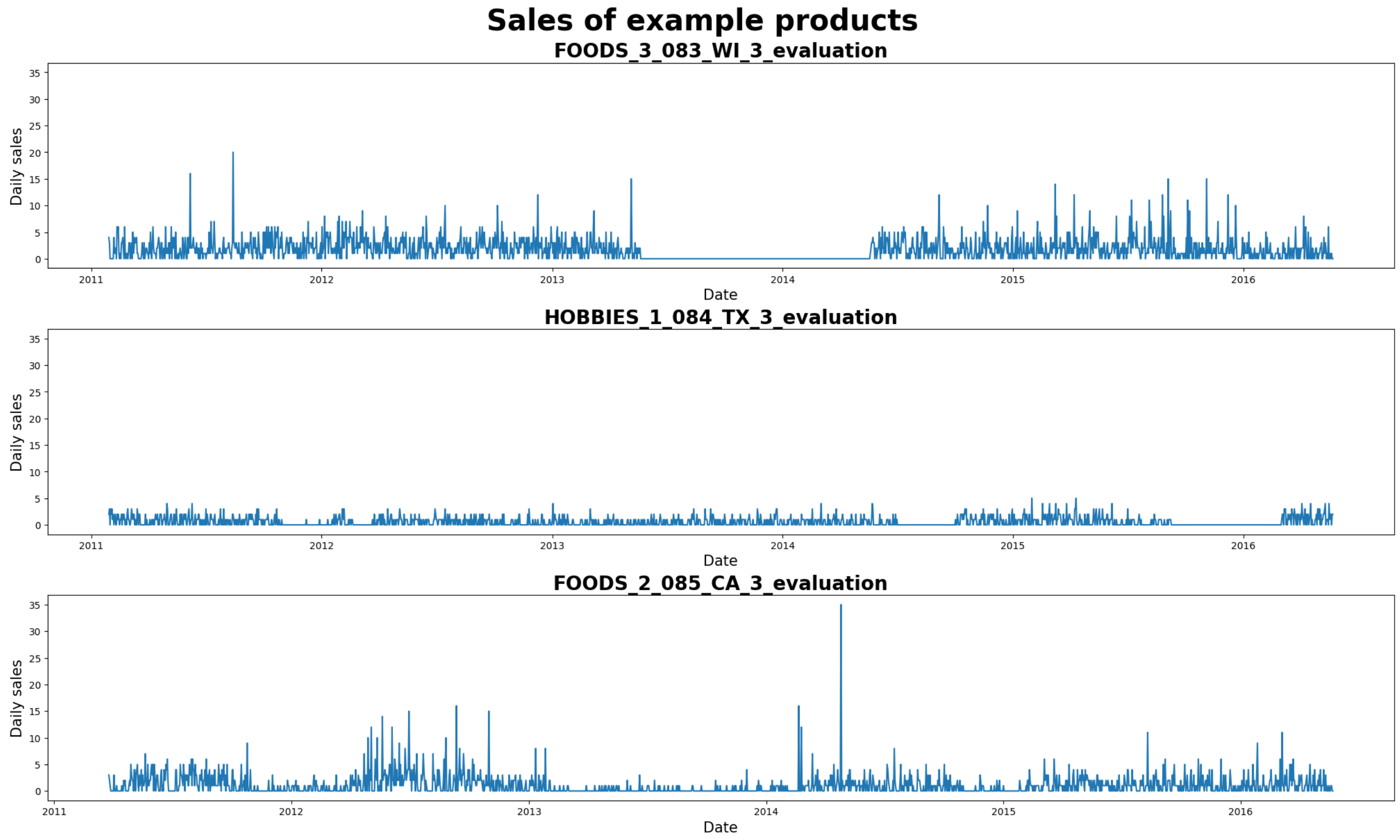


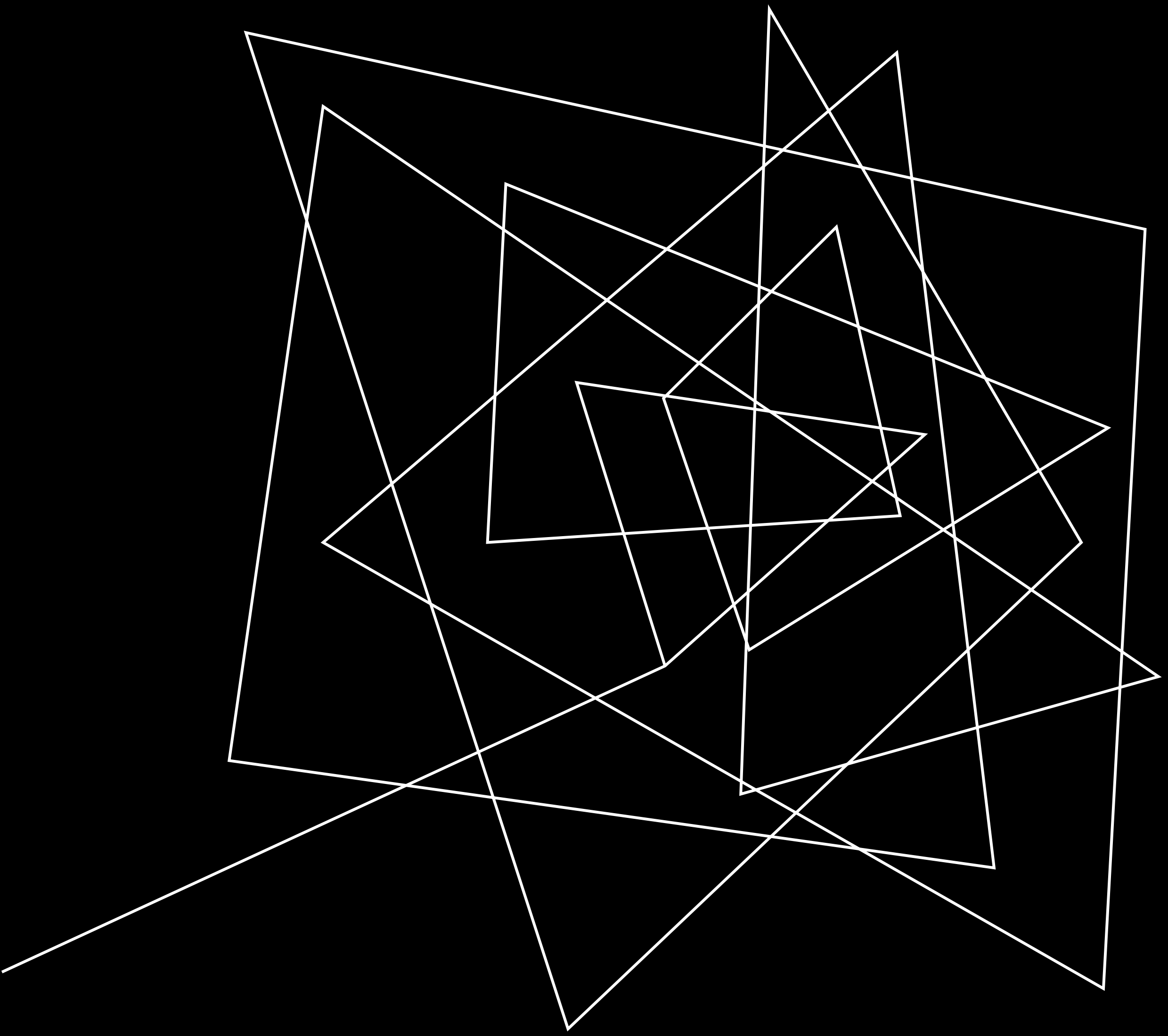


EXPLORATORY DATA ANALYSIS

Overview of product sales pattern

3. Random product data
    3. Plotted 3 random products
        3. Able to see a gap in between some of the data.
        4. In order to avoid the large gap to be trained, we will add a column that will flag unusual large gap.





BASELINE MODELING

BASELINE MODELING

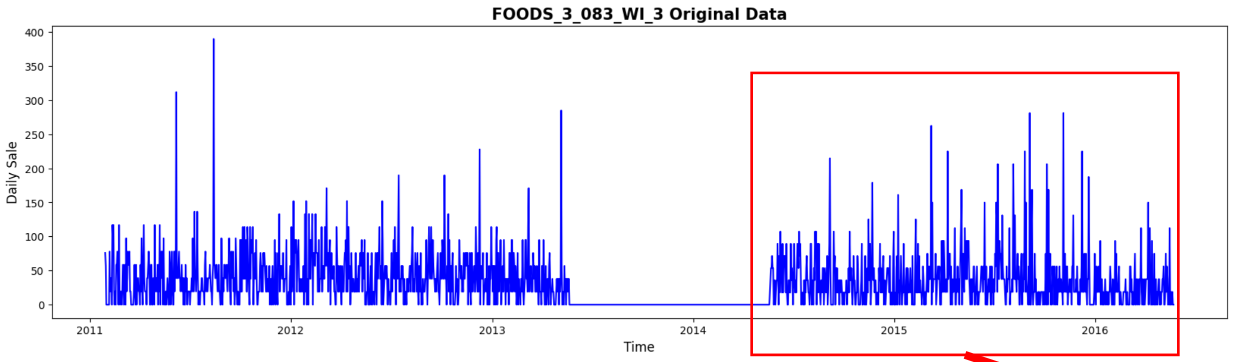
ARIMA Forecast

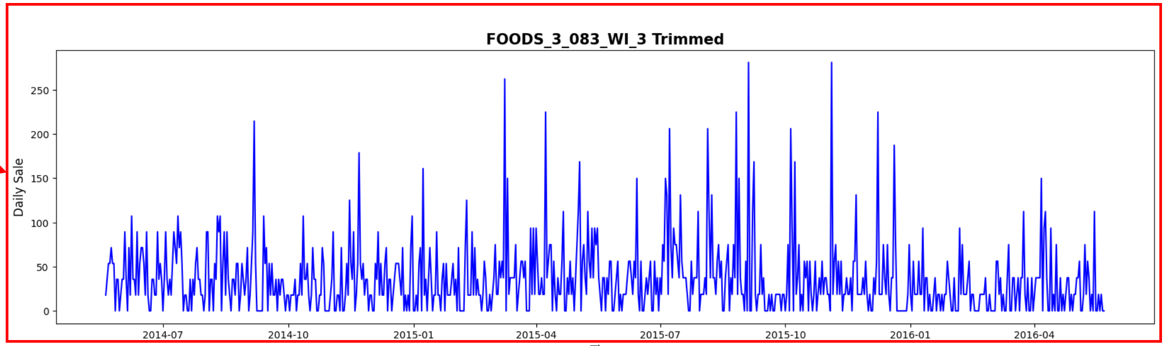
1. ARIMA(Autoregressive Integrated Moving Average)
   a. One of the popular time series forecasting model
   b. Procedure:
      1) Use Dickey-Fuller Test to test for stationary of the data
      2) Decompose the data for seasonality and trend
      3) Plot ACF and PACF to find p,d,q parameters
      4) Train ARIMA model
2. SARIMAX
   a. We will use another ARIMA type to time series forecasting model to include Exogeneous variable

METHODOLOGY

ARIMA Forecast

For testing purpose, we will train the data with the gap and data after the large gap



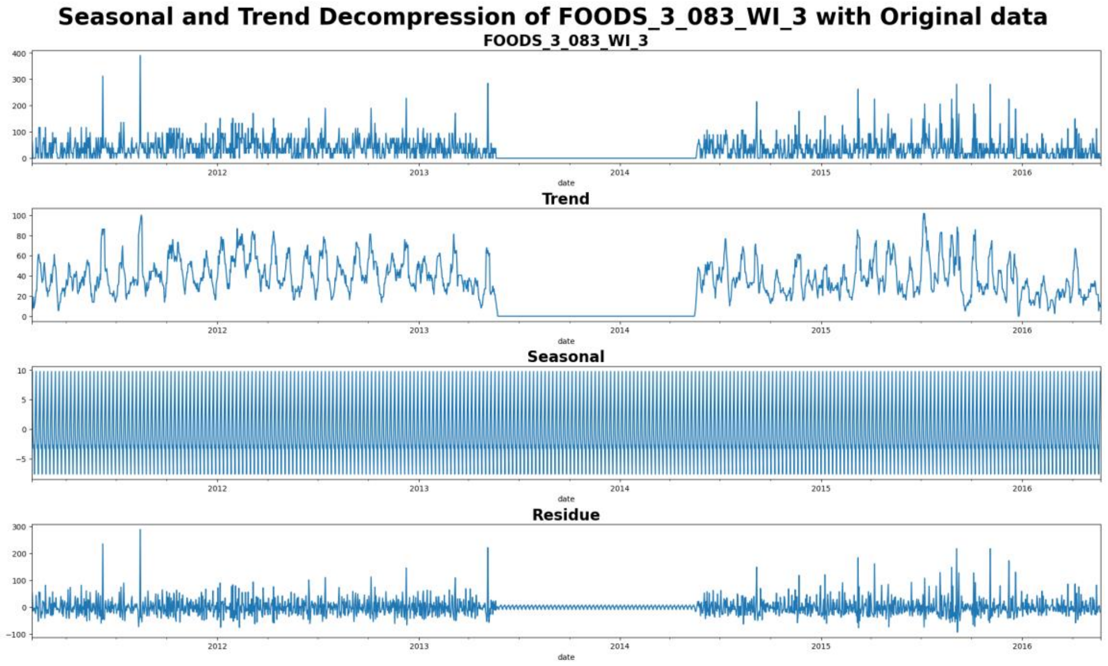


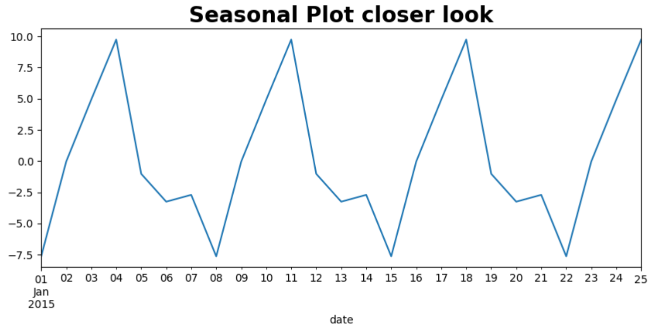
METHODOLOGY

ARIMA Forecast

Decompressing data for Seasonality and Trend

- Trend seems to be in negative trend, but very slightly and difficult to tell
- Seasonality is fluctuating with 7 days frequency
- Residue is difficult to tell the pattern. It seems there are many white noises

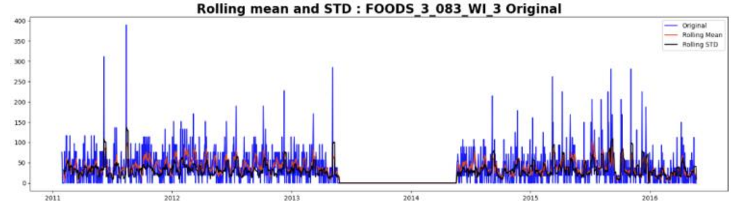




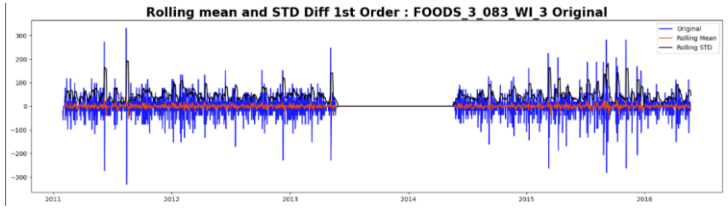
METHODOLOGY

ARIMA Forecast (Whole data)

Stationarity of the data

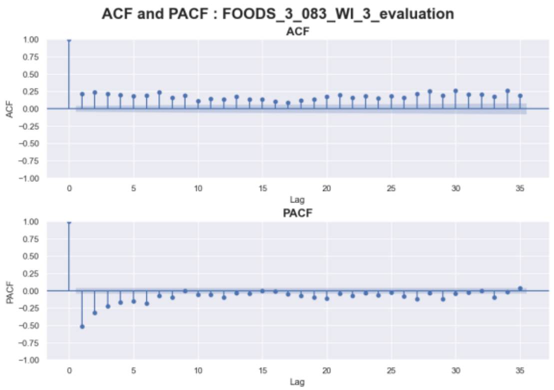


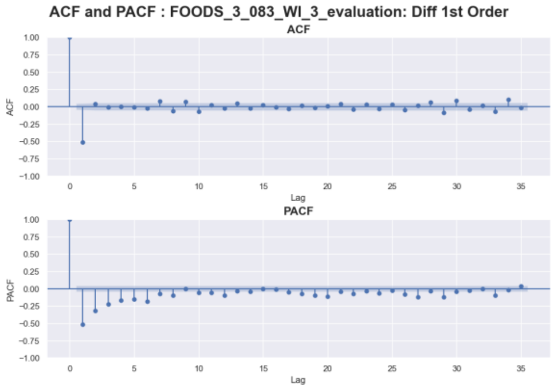
Dickey-Fuller Test:	
ADF Test Static	-3.621516
p-value	0.005363
# of Legs used	26.000000
Number of Observation Used	1914.000000
Critical Value (1%)	-3.433771
Critical Value (5%)	-2.863051
Critical Value (10%)	-2.567575



Dickey-Fuller Test:	
ADF Test Static	-1.701619e+01
p-value	8.521376e-30
# of Legs used	2.600000e+01
Number of Observation Used	1.913000e+03
Critical Value (1%)	-3.433773e+00
Critical Value (5%)	-2.863052e+00
Critical Value (10%)	-2.567575e+00

ACF and PACF Graph





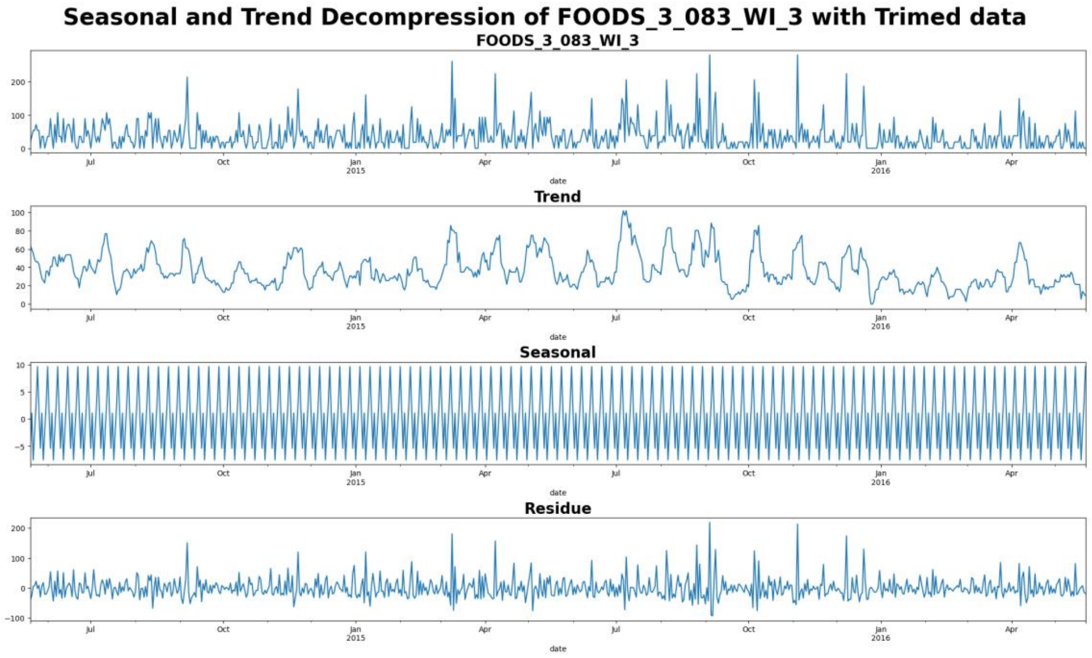
1st order of differential was applied since the ACF was not decaying as the lag increased.
High spike can be seen on ACF lag 1 in negative direction.
We will use ARIMA(p,d,q) = ARIMA(1,1,0)

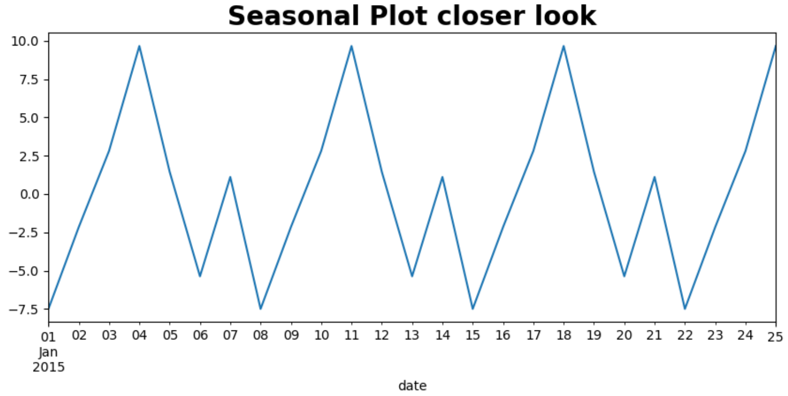
METHODOLOGY

ARIMA Forecast (Using data after gap)

Decompressing data for Seasonality and Trend

- Similar with data with whole data, trend seems to be in negative trend, but very slightly and difficult to tell
- Trend plot has some kind of pattern where the sales the sales increases around the beginning to first quarter of the month
- Seasonality is fluctuating with 7 days frequency
- The residue plot has peak at the beginning of April, the beginning of September and the middle of December, this tells that there might be seasonal pattern in yearly.

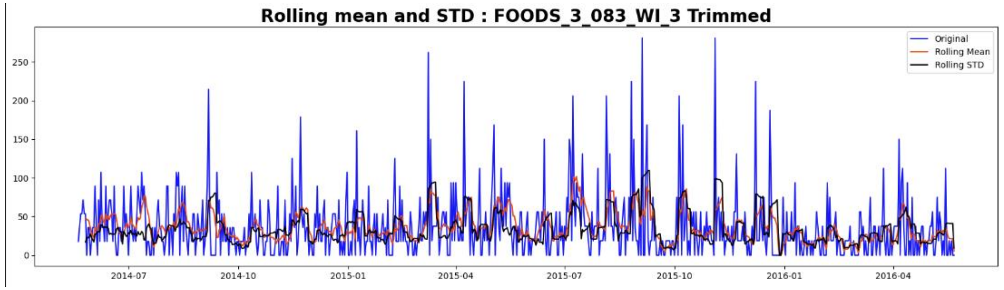




METHODOLOGY

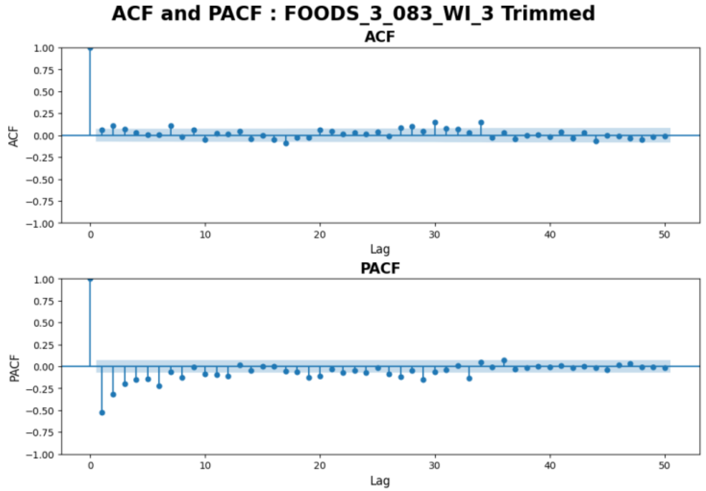
ARIMA Forecast (Using data after gap)

Stationarity of the data




```
Dickey-Fuller Test:
ADF Test Static                 -8.373310e+00
p-value                          2.626541e-13
# of Legs used                   6.000000e+00
Number of Observation Used       7.270000e+02
Critical Value (1%)             -3.439377e+00
Critical Value (5%)             -2.865524e+00
Critical Value (10%)            -2.568891e+00
```

ACF and PACF Graph



No need to apply 1st order of differential since the data was stationary and able to see ACF decay

High spike can be seen on PACF lag 6 in negative direction.

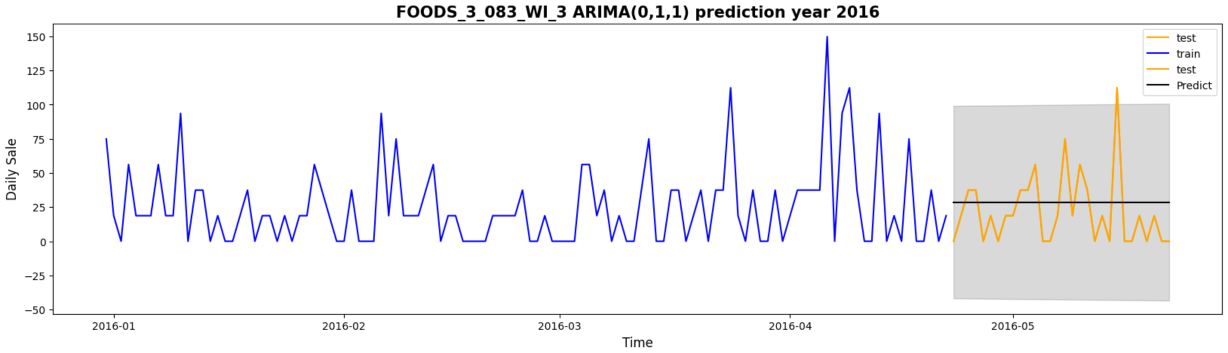
We will use ARIMA(p,d,q) = ARIMA(0,0,6)

METHODOLOGY

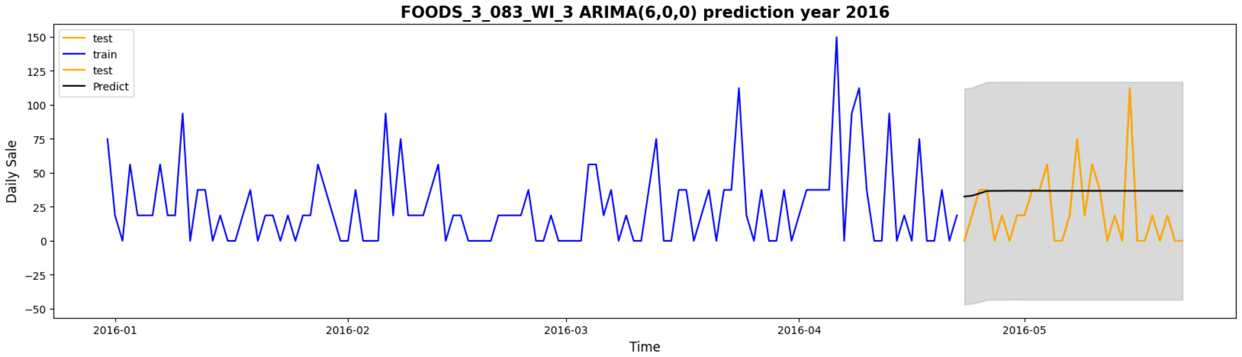
ARIMA Forecast comparison

There is slight difference between ARIMA using the whole data and using the data after the gap.
The data used after the gap was able to show small pattern but not enough to show the pattern

ARIMA model forecast with whole data

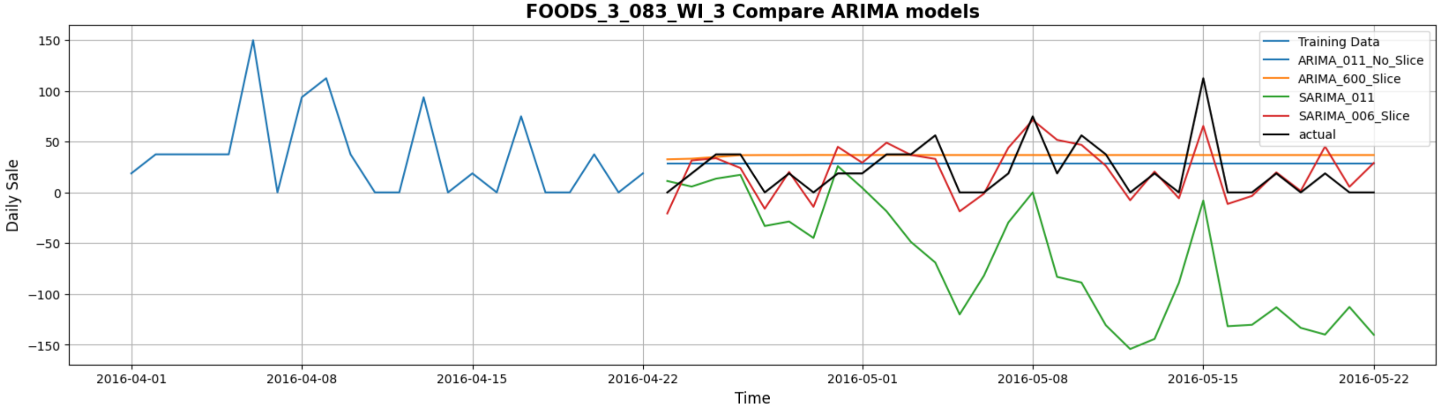


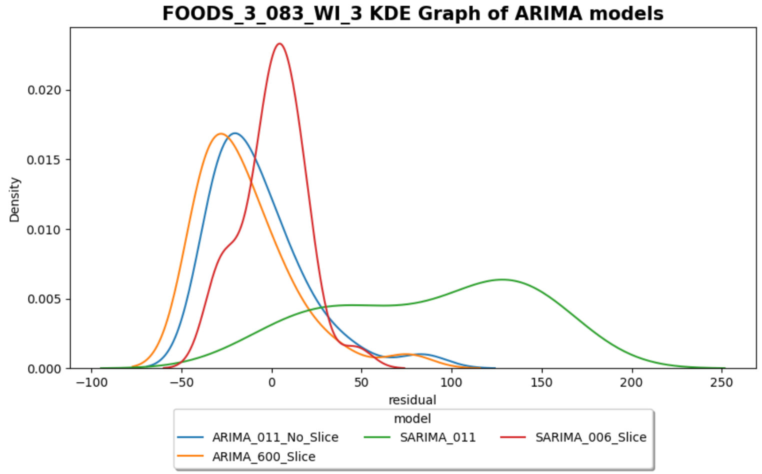
ARIMA model forecast with data after the gap



METHODOLOGY

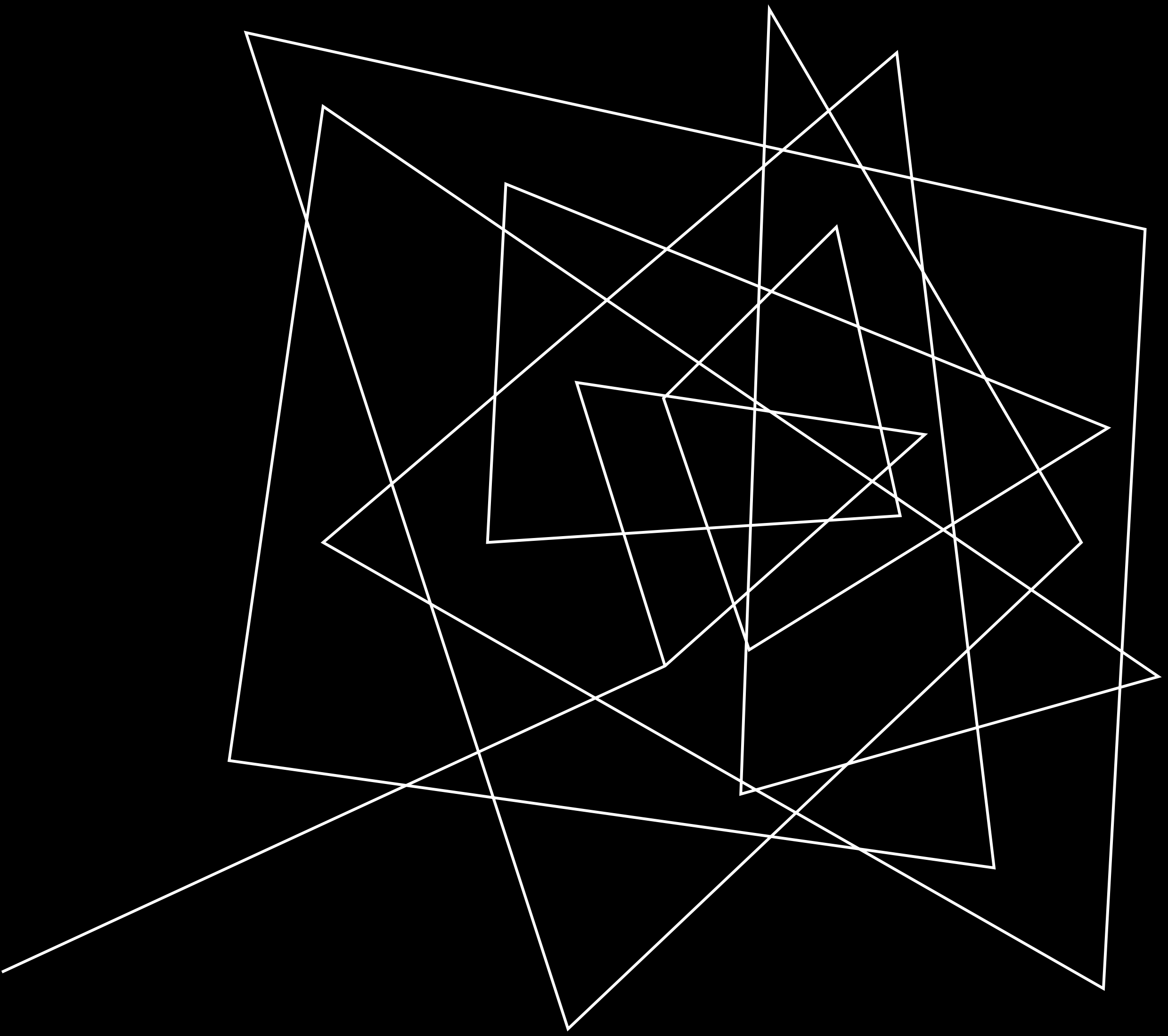
Forecast result with ARIMA and SARIMAX





Best data is SARIMAX(0,0,6) with exogenous variables.

- Able to capture the trend with some peaks
- KDE Graph shows forecasted data was able to capture most data overall due to higher density residue around zero.
- Shows importance of applying exogeneous variables and not using the gap that seemed uncommon.



EXTENDED MODELING

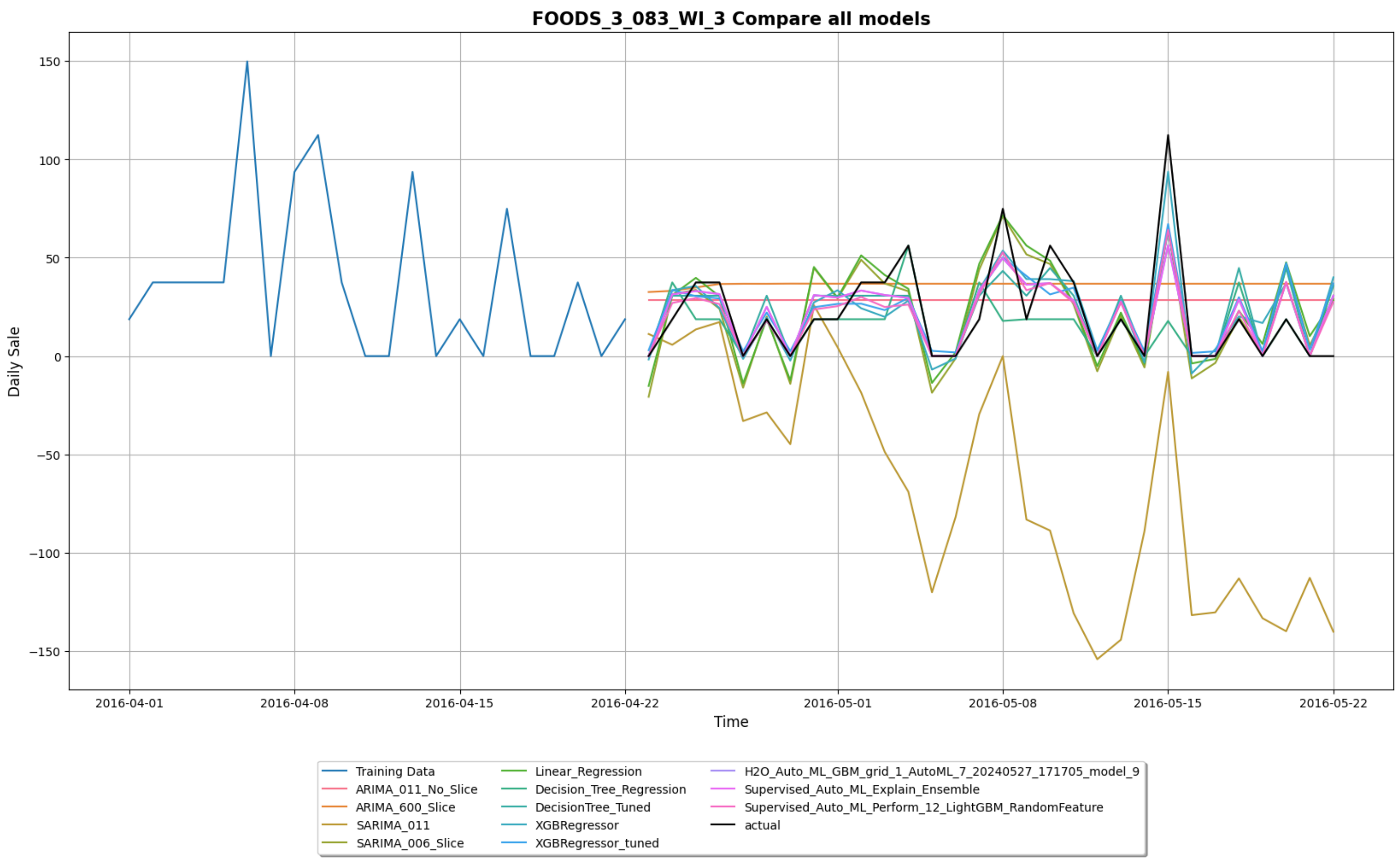
EXTENDED MODELING

Using other ML models

We have made this model as regression model by applying exogeneous variables.

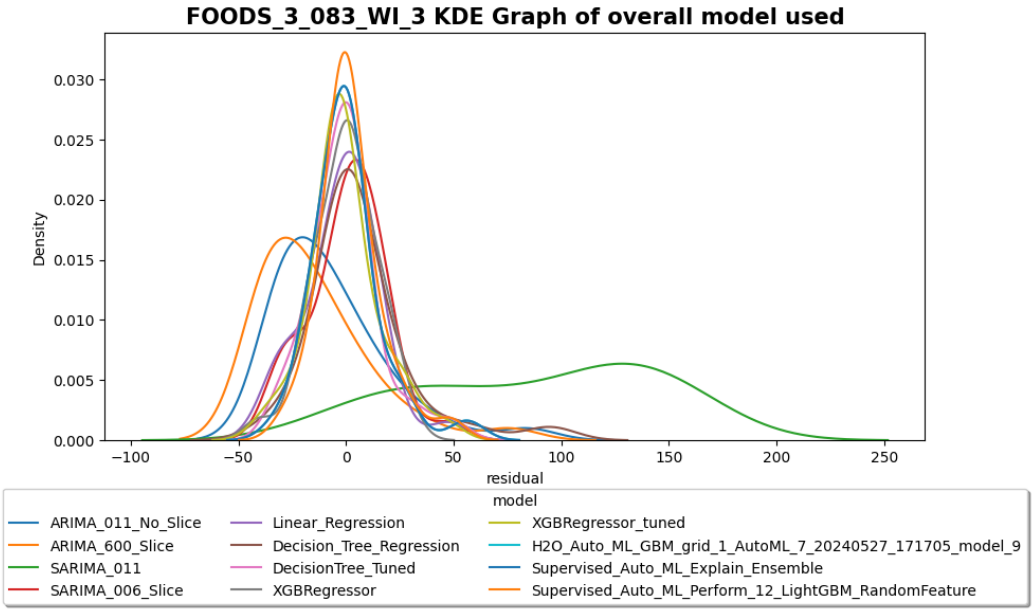
Applied Regression ML:

- Linear Regression
- Random Forest
- XGBoost Regressor
- Mljar-supervised AutoML
- H2O AutoML (Without DeepLearning)



EXTENDED MODELING

Using other ML models



- By comparing by the MAPE, the best result was with Decision Tree Regression with MAPE 24.27.
- Best RMSE result was with be XGBoost Regressor with RMSE 14.33
- the best KDE plot seems Supervised AutoML that used LightGBM with Random Features.

Highest MAPE

	id	model	adjust_r2	mape	r2	rmse
5	FOODS_3_083_WI_3_evaluation	Decision_Tree_Regression	0.083680	26.047636	0.368055	24.271428
0	FOODS_3_083_WI_3_evaluation	ARIMA_011_No_Slice	-0.351559	27.615579	0.067891	26.951169
13	FOODS_3_083_WI_3_evaluation	XGBRegressor_tuned	0.194988	30.105138	0.444819	16.211480
4	FOODS_3_083_WI_3_evaluation	DecisionTree_Tuned	0.270323	32.239183	0.496775	16.005696
1	FOODS_3_083_WI_3_evaluation	ARIMA_600_Slice	0.007167	35.001550	0.315288	29.831409
11	FOODS_3_083_WI_3_evaluation	Supervised_Auto_ML_Perform_12_LightGBM_RandomF...	0.192126	37.265888	0.442845	14.494863
6	FOODS_3_083_WI_3_evaluation	H2O_Auto_ML_GBM_grid_1_AutoML_7_20240527_17170...	0.187311	41.506044	0.439525	15.826093
10	FOODS_3_083_WI_3_evaluation	Supervised_Auto_ML_Explain_Ensemble	0.187311	41.506044	0.439525	15.826093
12	FOODS_3_083_WI_3_evaluation	XGBRegressor	0.556164	73.397855	0.693906	14.334725
7	FOODS_3_083_WI_3_evaluation	Linear_Regression	0.864611	87.044692	0.906629	17.327096
2	FOODS_3_083_WI_3_evaluation	Auto_ARIMA_no_zero_mod	0.946537	91.708215	0.963129	17.349737
8	FOODS_3_083_WI_3_evaluation	SARIMA_006_Slice	0.946537	91.708215	0.963129	17.349737
3	FOODS_3_083_WI_3_evaluation	Auto_ARIMA_sliced	1.732753	95.426566	1.505347	37.947214
9	FOODS_3_083_WI_3_evaluation	SARIMA_011	23.537154	249.770264	16.542865	103.543732

Highest RMSE

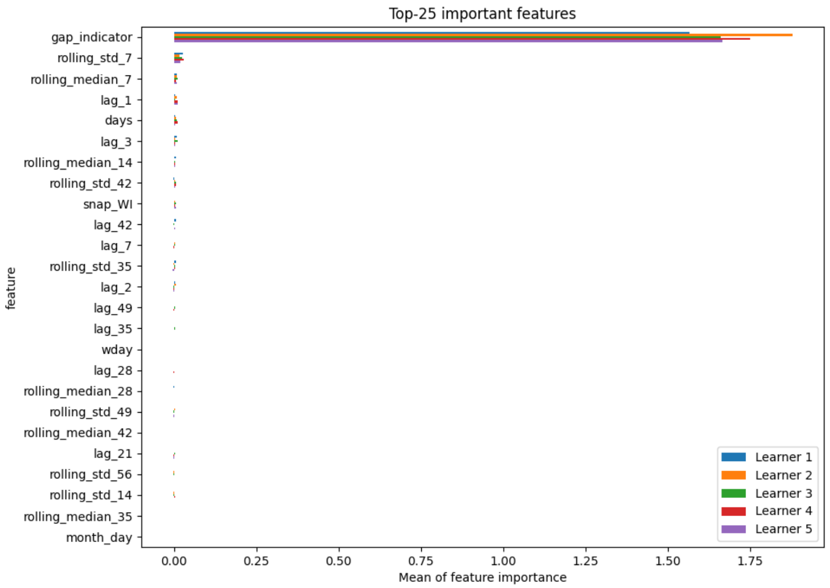
	id	model	adjust_r2	mape	r2	rmse
12	FOODS_3_083_WI_3_evaluation	XGBRegressor	0.556164	73.397855	0.693906	14.334725
11	FOODS_3_083_WI_3_evaluation	Supervised_Auto_ML_Perform_12_LightGBM_RandomF...	0.192126	37.265888	0.442845	14.494863
6	FOODS_3_083_WI_3_evaluation	H2O_Auto_ML_GBM_grid_1_AutoML_7_20240527_17170...	0.187311	41.506044	0.439525	15.826093
10	FOODS_3_083_WI_3_evaluation	Supervised_Auto_ML_Explain_Ensemble	0.187311	41.506044	0.439525	15.826093
4	FOODS_3_083_WI_3_evaluation	DecisionTree_Tuned	0.270323	32.239183	0.496775	16.005696
13	FOODS_3_083_WI_3_evaluation	XGBRegressor_tuned	0.194988	30.105138	0.444819	16.211480
7	FOODS_3_083_WI_3_evaluation	Linear_Regression	0.864611	87.044692	0.906629	17.327096
2	FOODS_3_083_WI_3_evaluation	Auto_ARIMA_no_zero_mod	0.946537	91.708215	0.963129	17.349737
8	FOODS_3_083_WI_3_evaluation	SARIMA_006_Slice	0.946537	91.708215	0.963129	17.349737
5	FOODS_3_083_WI_3_evaluation	Decision_Tree_Regression	0.083680	26.047636	0.368055	24.271428
0	FOODS_3_083_WI_3_evaluation	ARIMA_011_No_Slice	-0.351559	27.615579	0.067891	26.951169
1	FOODS_3_083_WI_3_evaluation	ARIMA_600_Slice	0.007167	35.001550	0.315288	29.831409
3	FOODS_3_083_WI_3_evaluation	Auto_ARIMA_sliced	1.732753	95.426566	1.505347	37.947214
9	FOODS_3_083_WI_3_evaluation	SARIMA_011	23.537154	249.770264	16.542865	103.543732

EXTENDED MODELING

Evaluating AutoML mljar-supervised

Highest MAPE models found during AutoML

	name	model_type	metric_type	metric_value	train_time	single_prediction_time
28	12_LightGBM_RandomFeature	LightGBM	rmse	30.247796	26.29	0.0561
68	Ensemble	Ensemble	rmse	30.537247	0.94	0.4879
15	12_LightGBM	LightGBM	rmse	30.652571	34.87	0.0591
34	26_LightGBM	LightGBM	rmse	30.652571	39.72	0.0527
50	42_LightGBM	LightGBM	rmse	30.675910	45.18	0.0536
...	...	...	...	...	...	...
9	22_NeuralNetwork	Neural Network	rmse	39.455583	12.45	0.1581
14	23_NeuralNetwork	Neural Network	rmse	40.387608	16.57	0.2262
49	41_NeuralNetwork	Neural Network	rmse	40.579153	28.33	0.1814
66	58_NeuralNetwork	Neural Network	rmse	42.736069	35.80	0.1726
19	24_NeuralNetwork	Neural Network	rmse	45.558246	16.89	0.1733

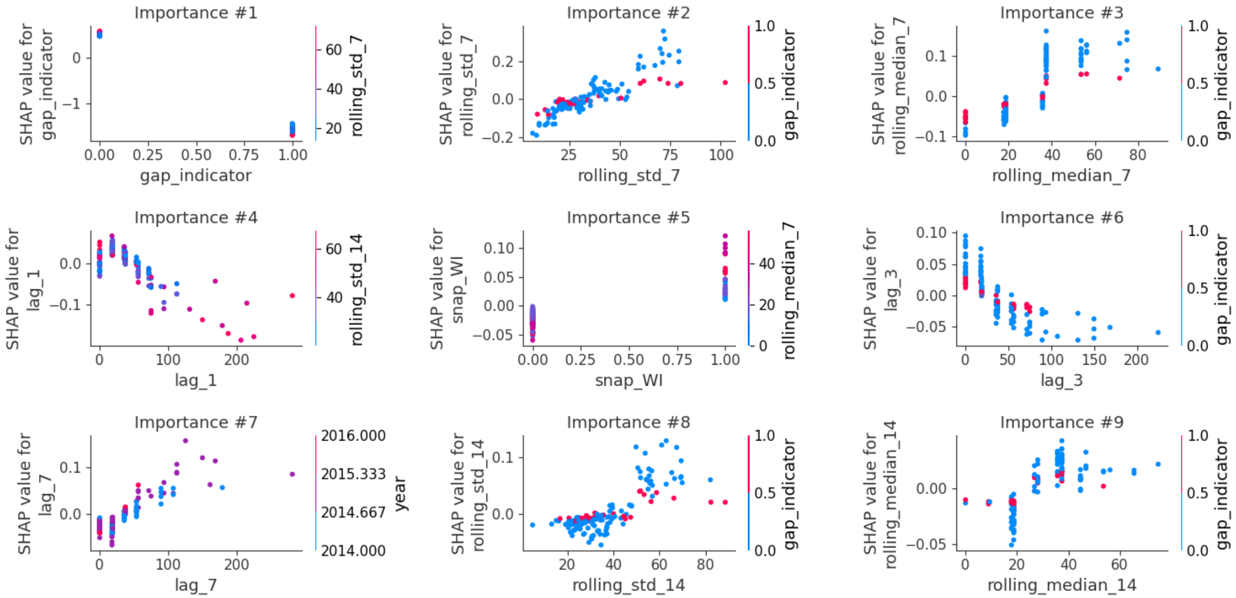


- LightGBM with Random Features:
  - The Feature importance shows the highest with ‘dap_indicator’ where I specified if the gap was within 7 days, or between 7-70 day or over 70 days.

EXTENDED MODELING

Evaluating AutoML mljar-supervised

- This is the SHAP dependance plot for 12_Light_GBM
  - rolling_std_7 and rolling_median_7 mostly show positive SHAP values indicating their importance in capturing patterns or trends in the data.
  - The highest importance is gap_indicator, but rolling_std_7, rolling_median_7 are showing mostly showing as positive SHAP values.
  - Predicting small value seems to work well with rolling std (such as with 7 and 14) ad higher value was predicting well with rolling median.



CONCLUSION

- Regression models outperformed ARIMA and SARIMAX due to the presence of zero values in the data.
- Large zero gap can cause .
- Explore different data resolutions and incorporate hyperparameter tuning in AutoML models.
- Investigate the application of deep learning techniques for improved forecasting accuracy.

FUTURE WORKS

- Explore different data resolutions.
  - Changing the resolution to weekly mean data (reducing zero gap in the data)
  - Transition the project into a pipeline format for seamless integration with cloud computing services to explore with different AutoML models that are not available as open source.
  - Azure AutoM, AWS, Google Cloud
- Investigate the application of deep learning techniques for improved forecasting accuracy.

THANK YOU

Brita Tamm

502-555-0152

brita@firstupconsultants.com

www.firstupconsultants.com