Name – Dinesh Sonawane

Project – M5 Sales Forecasting

Date – May’2024

Project Details

Project Aim:

Aim is to forecast daily sales for next 28 days by using use hierarchical sales data for past 1941 days.

Inputs Provided:

- Expectations from provided data (problem statement)
- Datasets (.csv file) containing the hierarchical daily sales data
- Past references of data analysis and suggestions of forecasting methods

Project Expectations:

- Application of traditional time series methods
- Application of deep learning frameworks for time series data
- Conceptual clarity and approaches followed
- Report out detailing the problem, data and solutions

High Level Approach

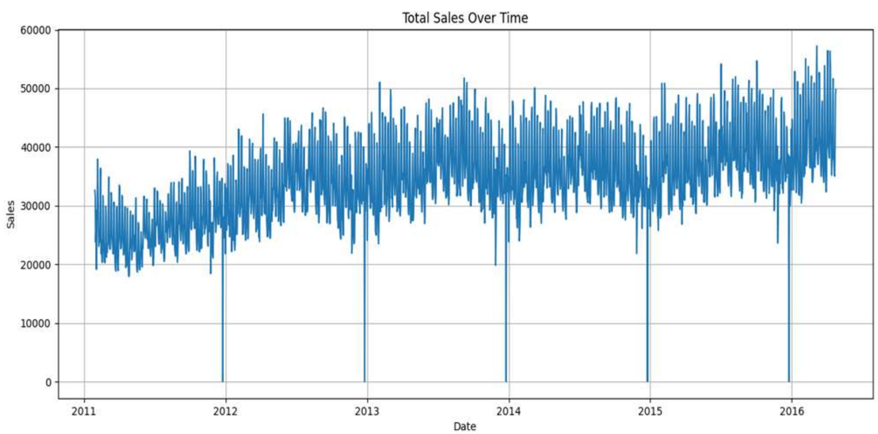
- Insights required from business owner's eyes:
  - Which states or stores to be focused for potential sales?
  - Lean Supply – Can I keep supply intact without increasing inventory cost?
  - Are these promotional offers really working?
  - Which products to be focused for promotional offers and which should not?
  - Suggestions on actions to improve overall sales

- Data information
  - Identify and treat the missing data
  - Identify and treat the duplicate data
  - Prepare data frame for deep dive
- Deep dive of data
  - Total sales trend over time
  - Total sales by category
  - Total sales by store
  - Impact of promotions on sales

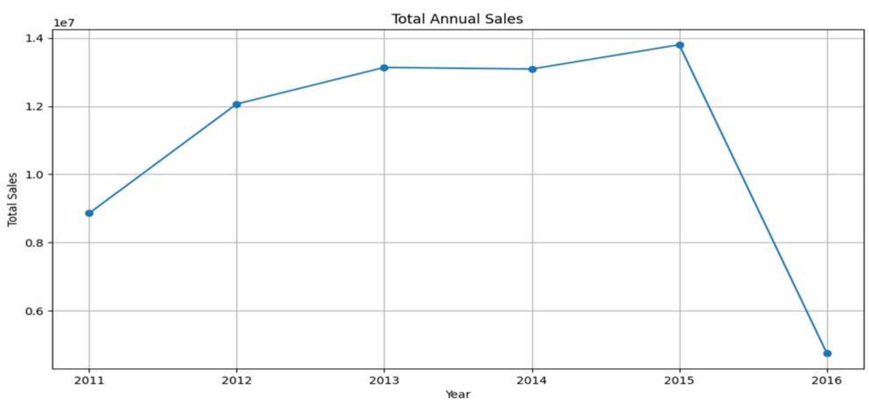
Exploratory Data Analysis

Aim – Sales growth and inventory management

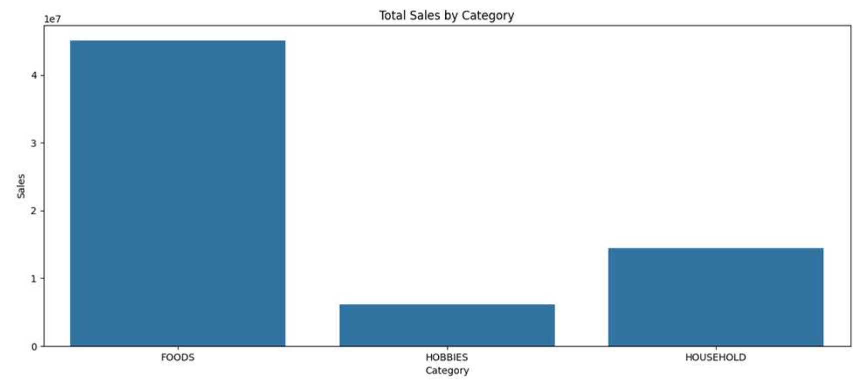
➢ Total sales trend over time

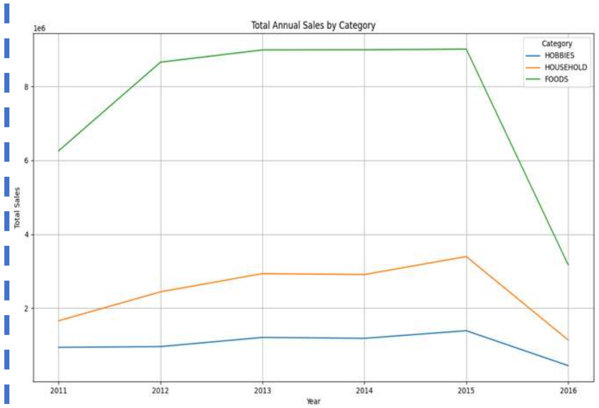


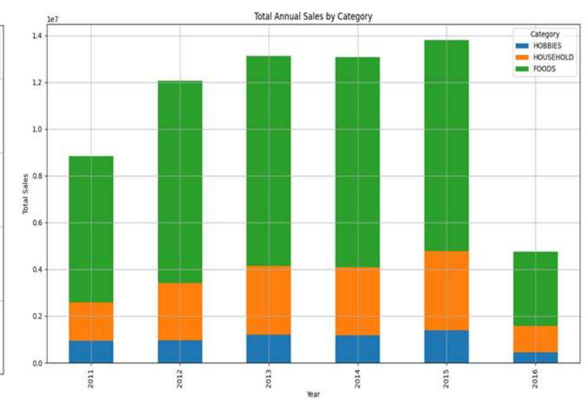
➢ Total **annual** sales trend over time



➢ Total Sales by Category



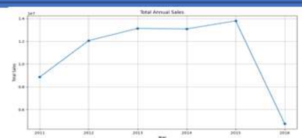


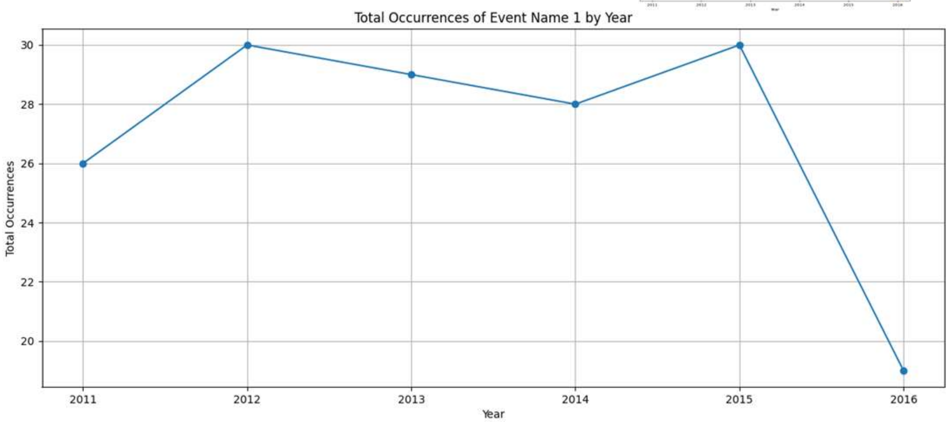


Exploratory Data Analysis

Aim - Sales growth and inventory management

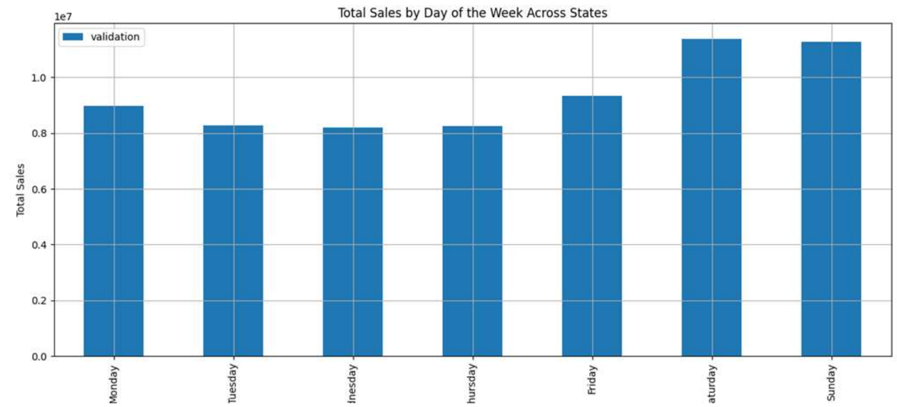
➤ Total number of event_1 per year



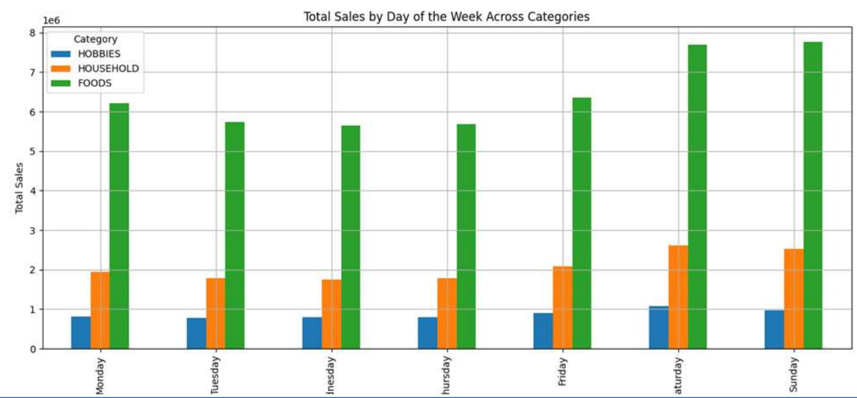


- We clearly observe that number of event across stores were dropped in 2012 and in 2013
- Total sales was slightly increased in 2012 and it was flat in 2013.
- So there is possibility that event_1 may not be boosting sales, this needs further investigations

- Sales as per day of the week



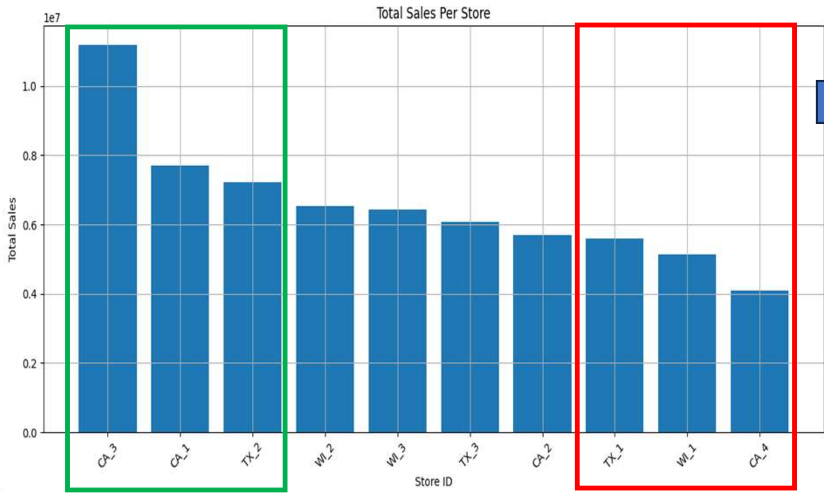
- Sales as per day for the week for each category



Deep Dive of Data

Aim - Sales growth and inventory management

➢ Sales per store



Low sales Store

High sales Store

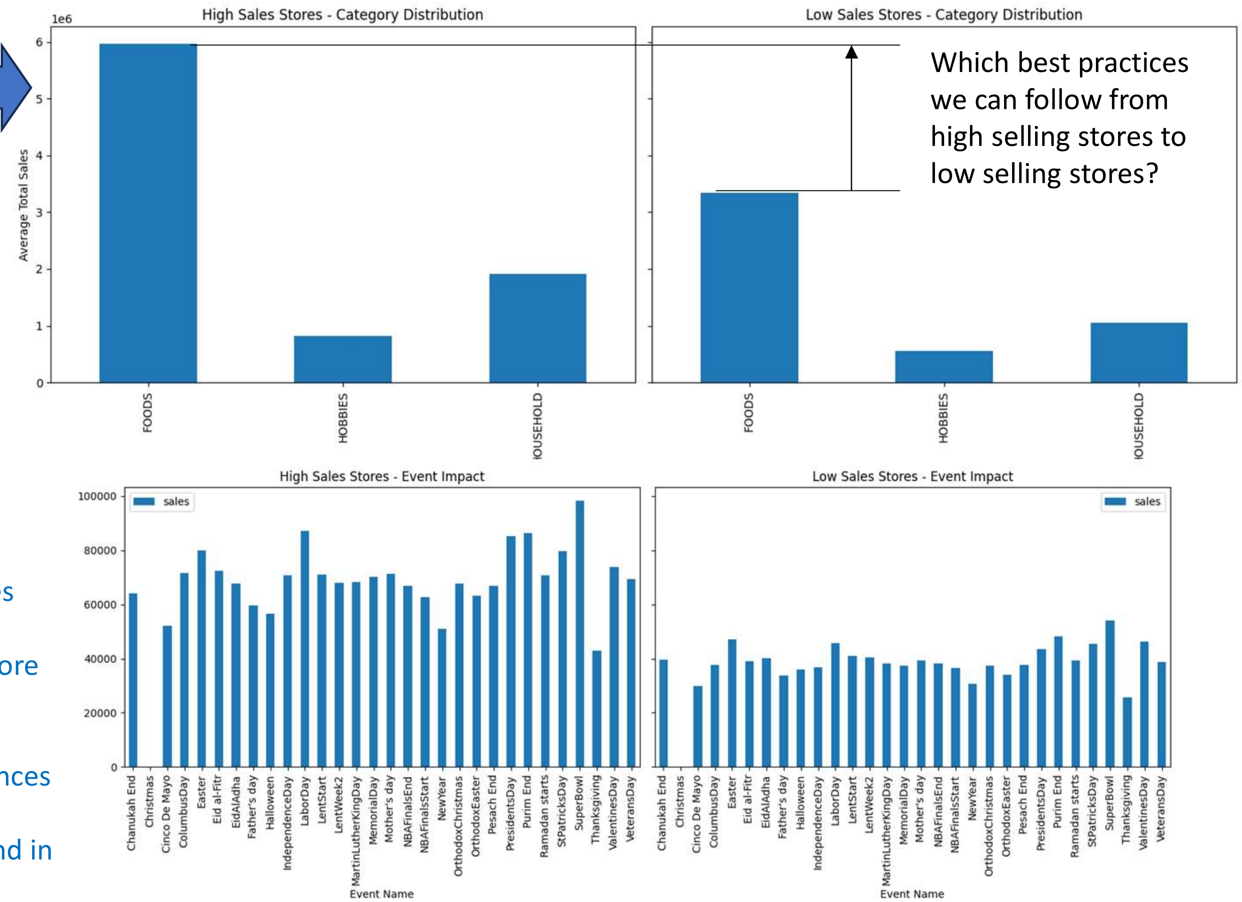
Action 1 - Store Layout and Customer Experience:

- Evaluate the store layout and customer experience of high sales stores.
- Implement similar layouts, customer service training, and in-store promotions in low sales stores.

Action 2 - Local Market Analysis:

- Conduct local market analysis to understand customer preferences and demographics.
- Tailor inventory and marketing strategies to match local demand in low sales areas.

➢ Sales per store – category distribution



Time Series Forecasting – Sales Pred

Aim - Sales growth and inventory management

Overall Approach Followed –

1. Start with prediction at store level aggregate sales
   - Use the traditional approach for sales prediction of one store, CA_1 – SARIMAX model
   - Use automated approach using SARIMA model for CA_1 store – pmdarima model
   - Use same automated approach extended to all other stores – pmdarima model on multiple series

2. Sales prediction using deep learning framework –
   - Use DeepAR Estimator for daily sales prediction
   - Use DeepAR Estimator for store level sales prediction
   - Use LSTM time series forecasting for store level sales prediction

Traditional approach for sales prediction

For single Series

Aim - Sales growth and inventory management

Prepare dataset

date	CA_1
2011-01-29	4337
2011-01-30	4155
2011-01-31	2816
2011-02-01	3051
2011-02-02	2630
...	...
2016-03-23	3770
2016-03-24	3970
2016-03-25	4904
2016-03-26	6139
2016-03-27	4669

Check if the series has auto correlation or not - Durbin Watson Test

- The Durbin-Watson test is used to detect the presence of autocorrelation
- The test statistic ranges from 0 to 4
- Interpretation:
  - **Value of 2**: no autocorrelation.
  - **Value near 0**: strong +ve autocorrelation.
  - **Value near 4**: strong –ve autocorrelation.
- Presence of auto-correlation is necessary to use time series models like ARIMA
- Durbin Watson test show Autocorrelation in CA_1 series data

Check the time series components, decompose the series

- Why Decompose a Time Series?
  - Understanding Patterns
  - Anomaly Detection
  - Simplifying Analysis
- Presence of strong seasonality in CA_1
- Upcoming trend in CA_1 (non-linear)
- Presence of noise

Check whether time series data is stationary or not, Augmented Dickey Fuller Test

- A time series is considered stationary if its statistical properties, such as mean, variance, and autocorrelation, remain constant over time
- time series forecasting models, such as ARIMA, assume that the series is stationary
- The ADF test is a statistical test used to determine if a time series is stationary
- CA_1 time series found to be non-stationary
- Series made stationary with differencing period of 30 days

Identify the p,d,q parameters for Trend and Seasonality in time series

- **p - Autoregressive (AR) Order:** Represents the number of lag observations included in the model
- **d - Difference Order**: Represents the number of times the data needs to be differenced to achieve stationarity
- **q - Moving Average (MA) Order:** It specifies the number of past forecast errors that are used to predict the current value
- **P,D,Q and m**: Seasonality parameters over period of m
- For CA_1 series:
  - Trend - p=1, q=1, d=1
  - Seasonality - P=1, Q=1, D=1

Iterate over diffrent values of p,d,q parameters and choose those giving minimum AIC value

```
import itertools
p = d = q = range(0,2)
pdq = list(itertools.product(p,d,q))
seasonal_pdq = [(x[0], x[1], x[2], 30) for x in pdq]
```

- Parameters were assigned any value from 0,1,2 and allowed to create combinations
- These combinations passed over SARIMAX model and observed resulted AIC
- Final parameters with min AIC received are: (1, 0, 1), (1, 1, 1, 30)

For multiple time series modeling, utilized pmdarima package which auto-calculates the parameters. However, single time series is used for training independent of other time series'. Demonstrated the application of pmdarima and received good results.

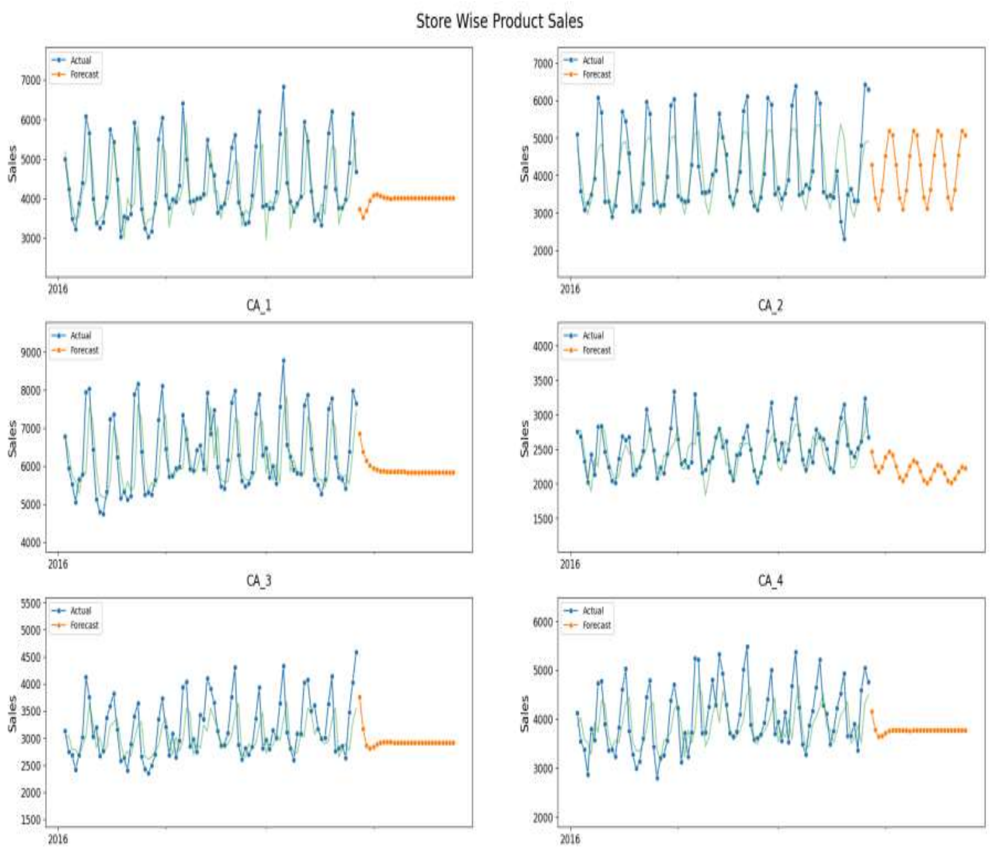
REF - Exploring Auto ARIMA in Python for Multiple Time Series Forecasting

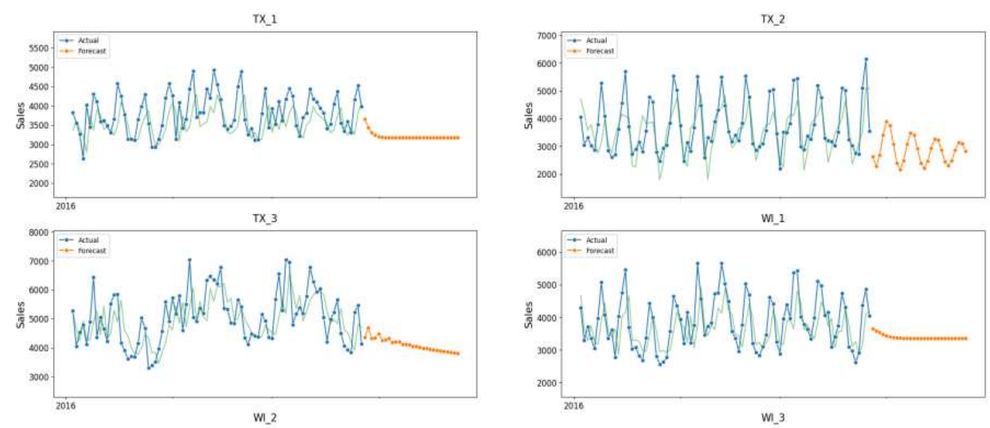
Traditional approach for sales prediction

For single Series

Aim - Sales growth and inventory management

SARIMA for using pmdarima (auto-arima) for multiple time series at store level



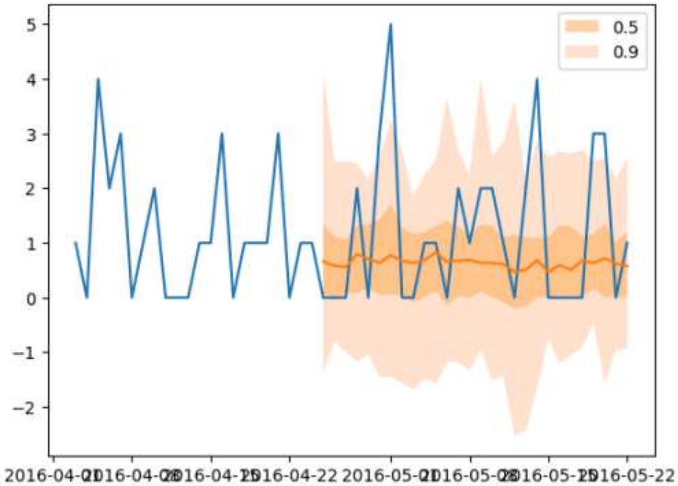


Deep Learning Framework for sales prediction

For Multiple Series

Aim - Sales growth and inventory management

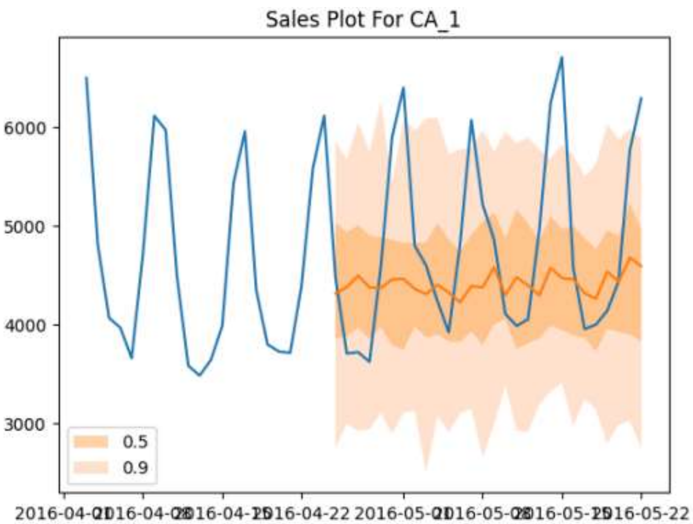
DeepAR Estimator for multiple time series at product level



Actual Vs Forecasts with probabilistic predictions for one of the series

Overall RMSE – 1.89

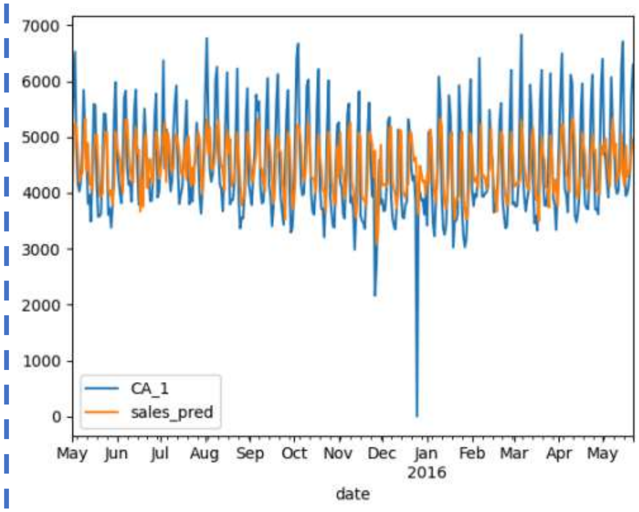
DeepAR Estimator for multiple time series at store level



Actual Vs Forecasts with probabilistic predictions for CA_1 sales series

Overall RMSE – 847

LSTM for multiple time series at store level



Actual Vs Forecasts predictions for CA_1 sales series

Training RMSE – 0.1044
Testing RMSE – 0.1081

Concluding Summary

Aim - Sales growth and inventory management

- Data Analysis Observations:
  - Overall trend for total sales is trending up. Same behaviour is shown by category sales too
  - There is possibility that event_1 may not be boosting sales, this needs further investigations
  - Saturday and Sunday shows increased sales relative to other weekdays. Same behaviour is shown by categories also.

- Recommendations:
  1. Store Layout and Customer Experience:
     - Evaluate the store layout and customer experience of high sales stores
     - Implement similar layouts, customer service training, and in-store promotions in low sales stores
  2. Local Market Analysis:
     - Conduct local market analysis to understand customer preferences and demographics.
     - Tailor inventory and marketing strategies to match local demand in low sales areas.
  3. LSTM model can be utilized for predicting sales of a store. This will help to estimate growth of business from particular store and actions can be planned upfront for further expansion
  4. DeepAR model can be utilized for product level prediction, which will help to manage inventory specifically for food items