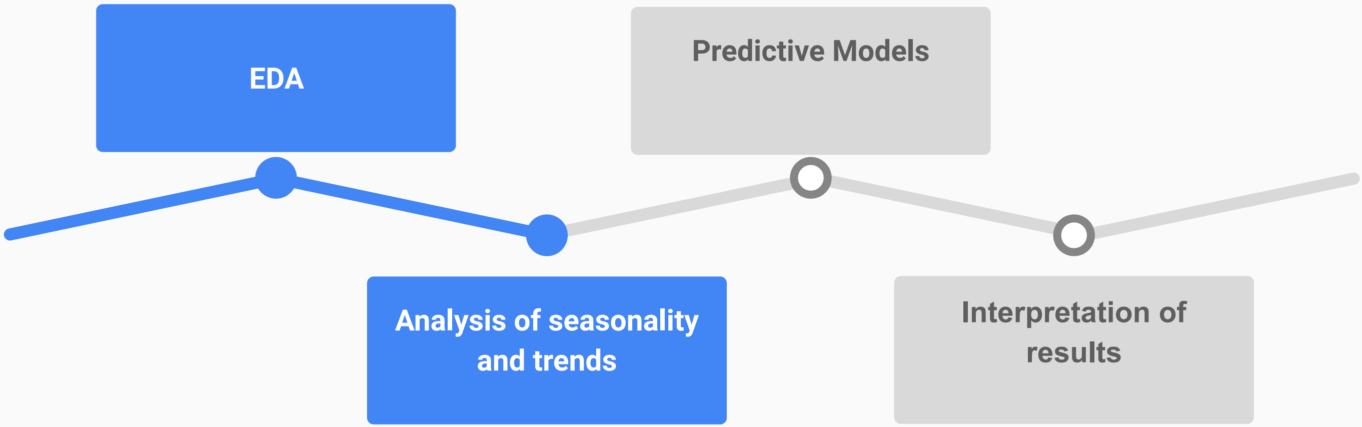
Forecast daily sales at the product (SKU)/store level

An accurate sales forecasting model will help reduce storage costs, minimize losses from unsold products and improve product availability in stores.

Problem

- The retail chain faces challenges in efficient inventory management due to the variability in product sales, which is influenced by multiple factors:
  - promotions,
  - seasonality
  - and special events.

Solutions Pipeline



EDA

Analysis of seasonality and trends

Predictive Models

Interpretation of results

Dataset: Exploratory Data Analysis

Data Explorations and Cleaning

- **Dependencies**
    - Pandas, Numpy, Seaborn, and Matplotlib libraries were utilized for exploratory analysis
    - Sklearn, statsmodels, calendar, LinearRegresor, SDG, XGBoost
    - rmse, mape, r2, mae

Data Ingestion and Cleaning

- csv files:
  - holidays
    - dtypes: bool(1), object(5), 350 entries
  - items
    - dtypes: bool(1), float64(1), int64(3), object(1), 67029200 entries
  - sample_submission
  - stores
    - dtypes: int64(2), object(3), 54 entries
  - transactions
    - dtypes: int64(2), object(1), 83488 entries
  - train
    - dtypes: bool(1), float64(1), int64(3), object(1), 125497040 entries
  - test
    - dtypes: bool(1), float64(1), int64(3), object(1), 67029280 entries

Exploratory Analysis - dataset

The Training dataset:

- The record starts from 2013-01-01 to 2017-08-15
- The onpromotion column contains: 21657651 null data
- There are 36810109 items that do not contain a record from the beginning of the period.

The Holidays dataset:

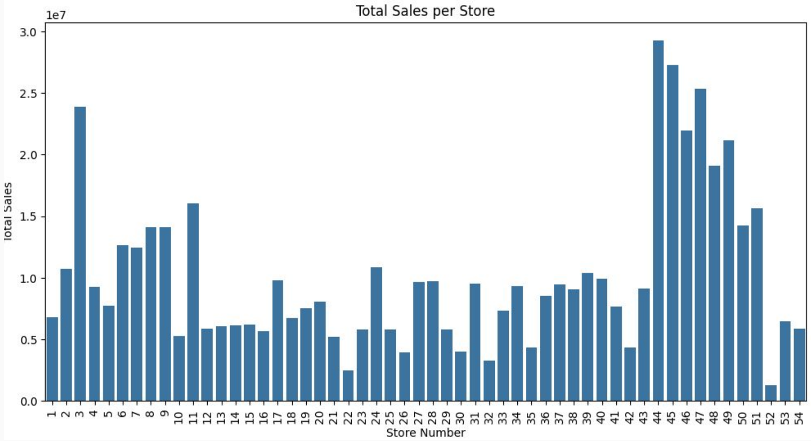
- The record starts from 2012-03-02 to 2017-12-26
- Holiday types: Holiday, Event, Additonal, Transfer, Bridge, Work Day
- Classification: National, Local or Regional

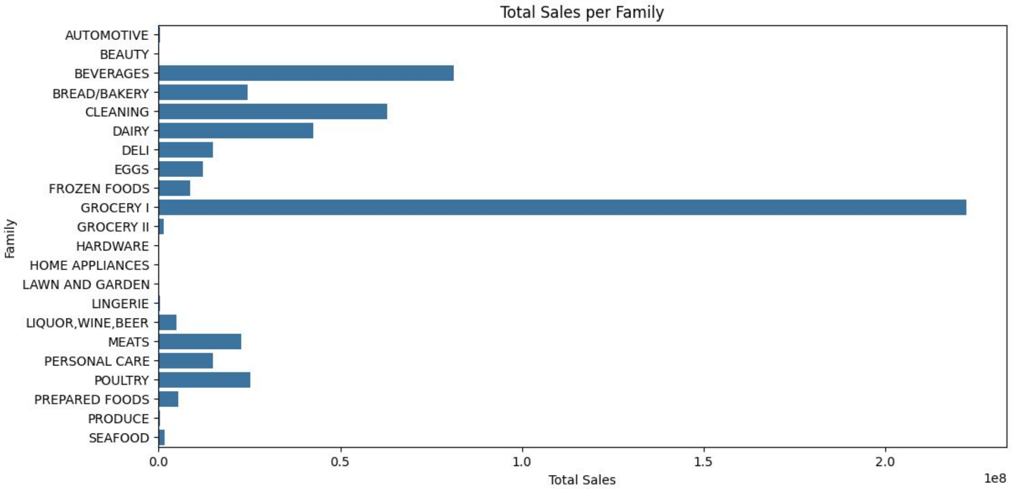
Exploratory Analysis - features

- The training dataset contains the features: id, date, store_nbr, item_nbr, unit_sales, onpromotion.

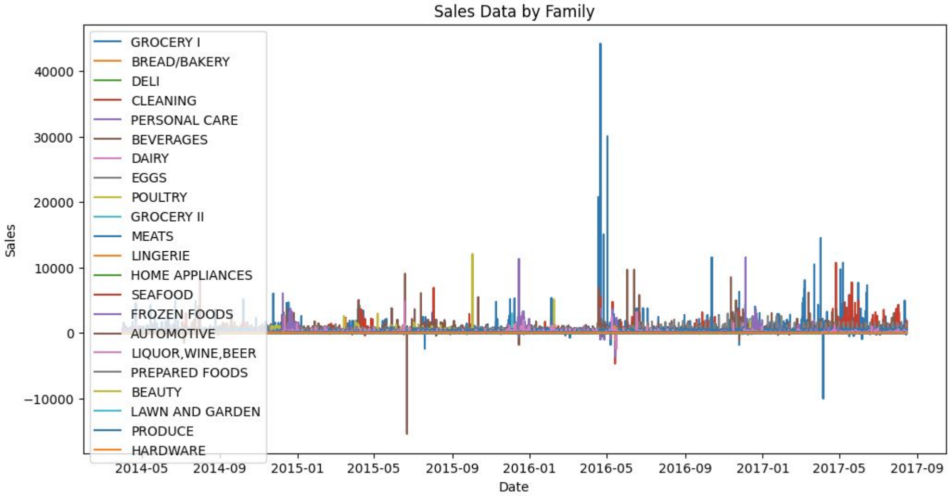
- New features added to the training dataset: 'family', 'perishable', 'type_store', 'cluster', 'type_holiday', 'transferred', weekday, year, month, day, payday, is_weekend, sales_lag_7, sales_lag_30, sales_roll_mean_7, sales_roll_mean_30, sales_ewm_alpha_095_lag_7, sales_ewm_alpha_095_lag_30, sales_ewm_alpha_09_lag_7, sales_ewm_alpha_09_lag_30, sales_ewm_alpha_08_lag_7, sales_ewm_alpha_08_lag_30

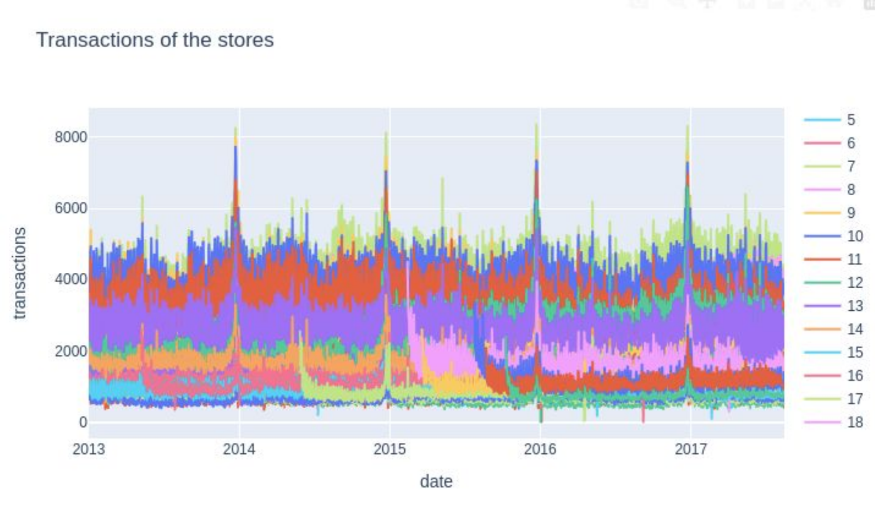
Exploratory Analysis



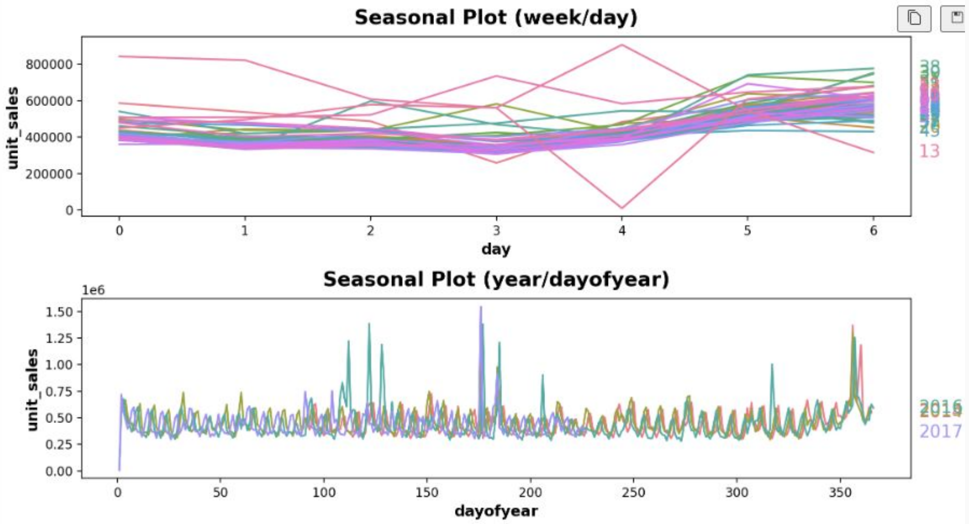


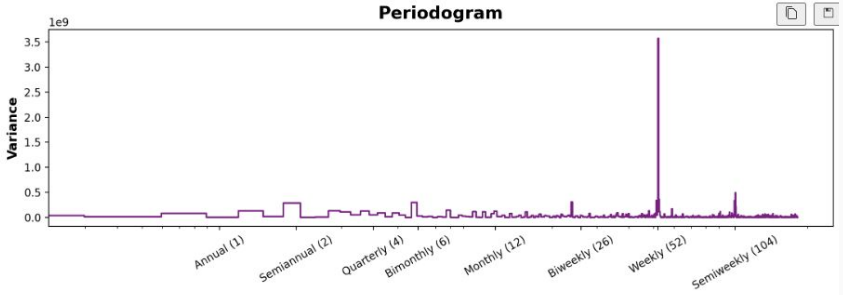
Exploratory Analysis



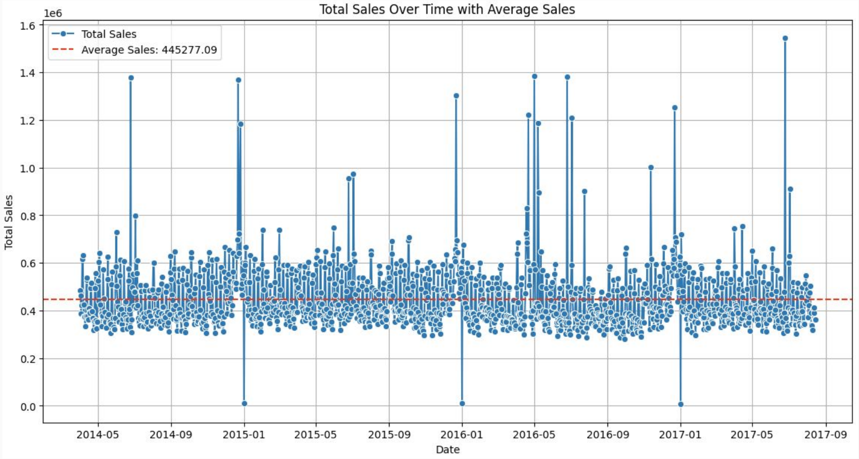


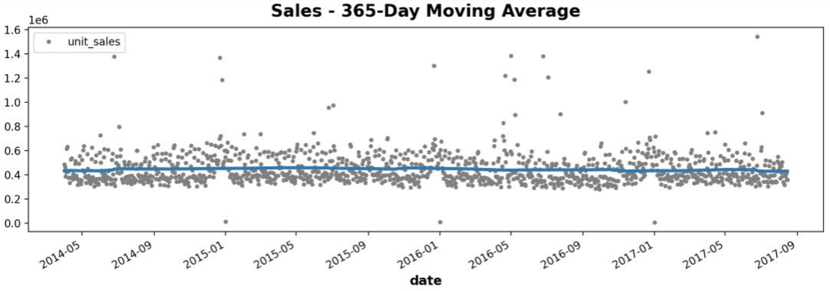
Seasonality Analysis

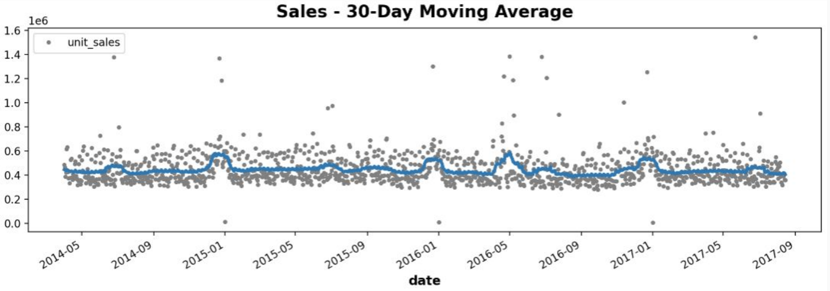




Trend Analysis


Total Sales Over Time with Average Sales


Sales - 365-Day Moving Average

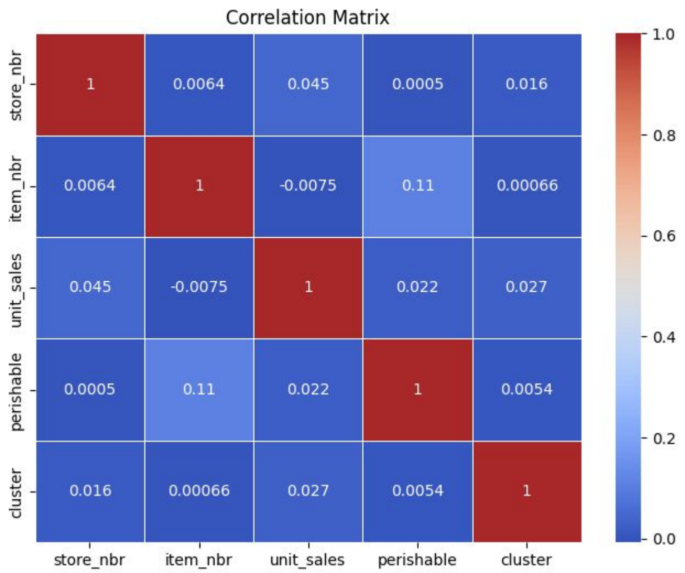

Sales - 30-Day Moving Average

Data Preparation and Feature Selection

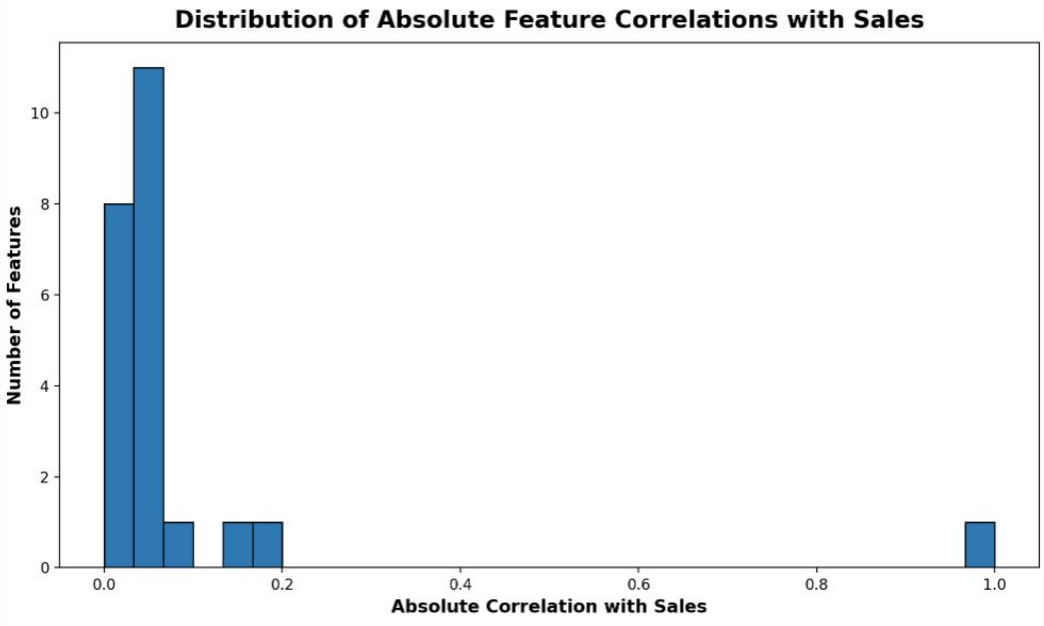
Data Preparation

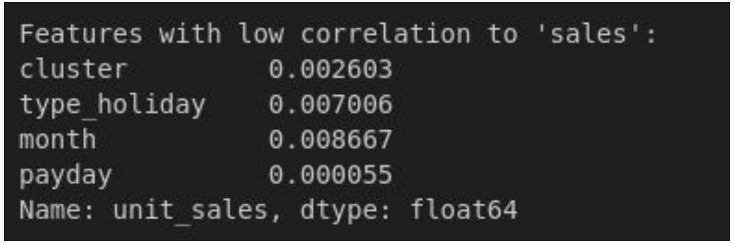
The data frame has been divided into two 90-10 sets:

- train: 24 columns and 61029280
- test: 23 columns and 6029562 data



Feature Selection





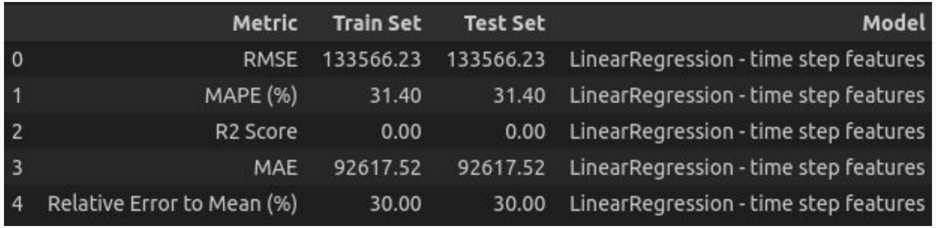
```
Features with low correlation to 'sales':
cluster         0.002603
type_holiday    0.007006
month           0.008667
payday          0.000055
Name: unit_sales, dtype: float64
```

Modeling

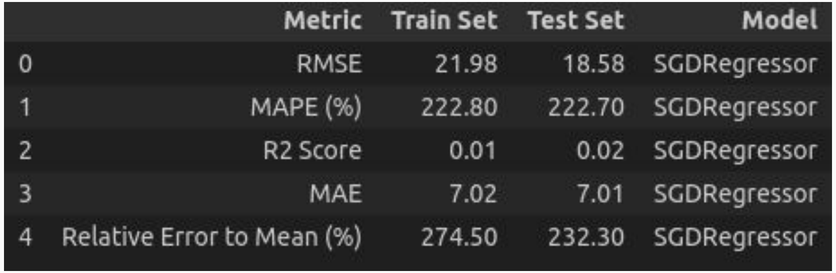
Model architecture

- Linear Regressor Model: Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data.
- SVG Model: Predicting a continuous output variable, also known as the dependent variable, from one or more input data, also known as independent variables
- XGBoost Regressor: Is a powerful approach for building supervised regression models

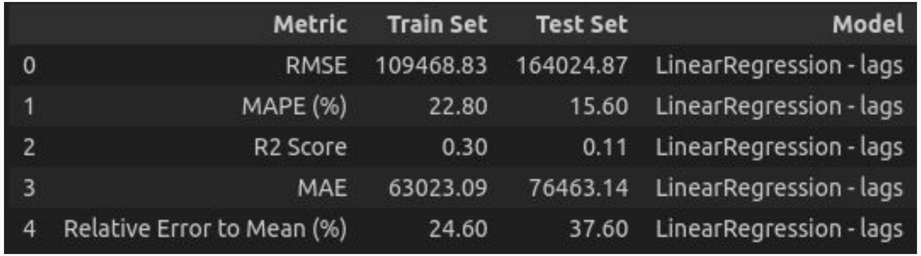
Model results



	Metric	Train Set	Test Set	Model
0	RMSE	133566.23	133566.23	LinearRegression - time step features
1	MAPE (%)	31.40	31.40	LinearRegression - time step features
2	R2 Score	0.00	0.00	LinearRegression - time step features
3	MAE	92617.52	92617.52	LinearRegression - time step features
4	Relative Error to Mean (%)	30.00	30.00	LinearRegression - time step features

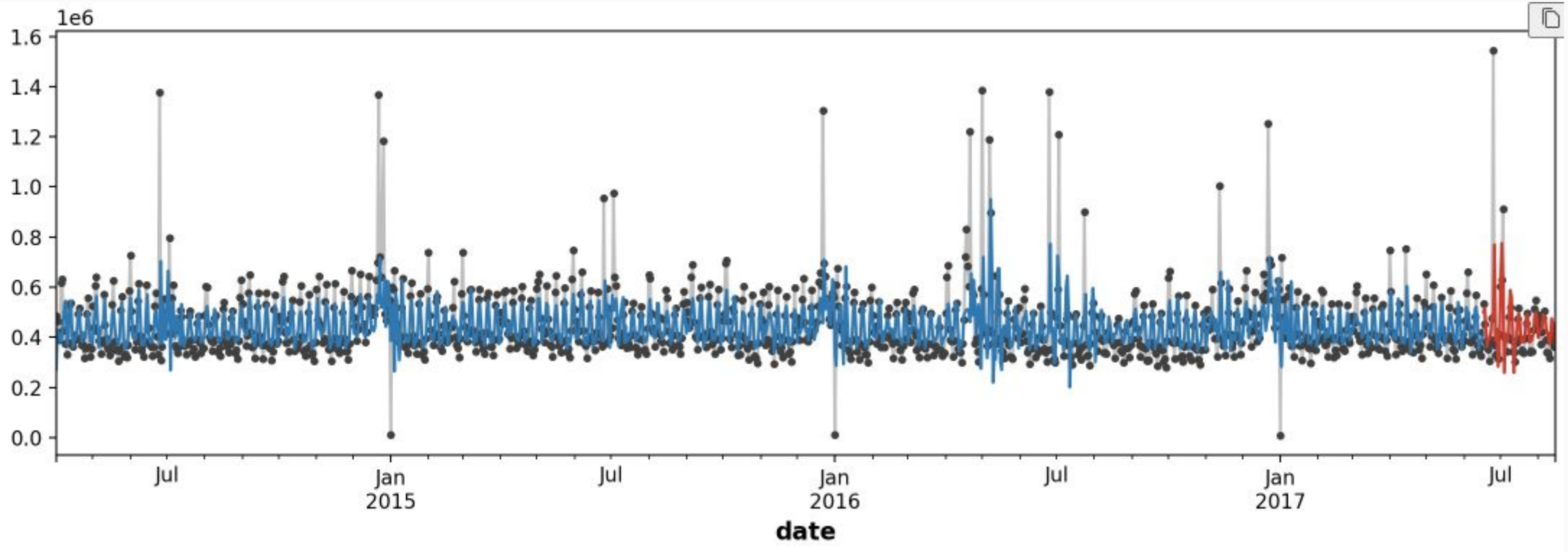


	Metric	Train Set	Test Set	Model
0	RMSE	21.98	18.58	SGDRegressor
1	MAPE (%)	222.80	222.70	SGDRegressor
2	R2 Score	0.01	0.02	SGDRegressor
3	MAE	7.02	7.01	SGDRegressor
4	Relative Error to Mean (%)	274.50	232.30	SGDRegressor

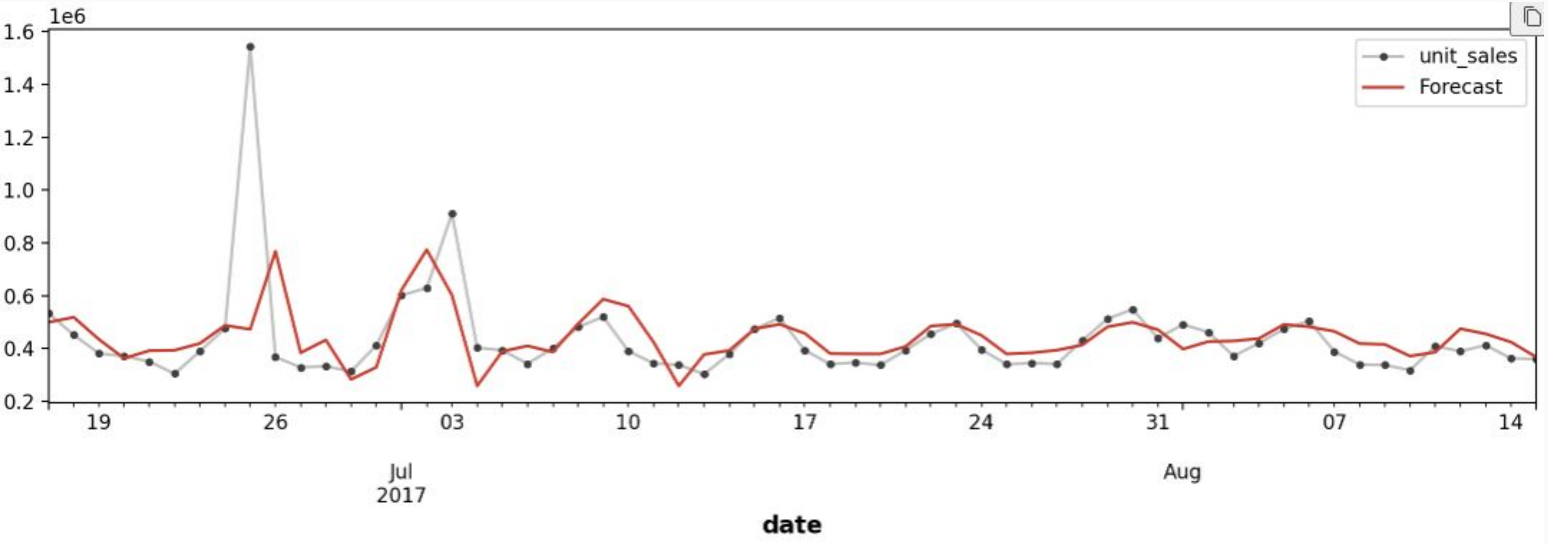


	Metric	Train Set	Test Set	Model
0	RMSE	109468.83	164024.87	LinearRegression - lags
1	MAPE (%)	22.80	15.60	LinearRegression - lags
2	R2 Score	0.30	0.11	LinearRegression - lags
3	MAE	63023.09	76463.14	LinearRegression - lags
4	Relative Error to Mean (%)	24.60	37.60	LinearRegression - lags

Model results



Model results


1e6
unit_sales
Forecast
Jul
Aug
date