Model Development

Haldun Köktaş

2022-11-26

Objective

The goal is to create a model that can forecast sales 30 days into the future.

Pre-process

- Dropped duplicates.
- Created a "Revenue" column which will be the primary time series.
- Resampled by daily frequency to create daily sums.
- The last 30 observations are kept out for testing.
- The train set was further split into train and validation set by the last 30 observations.

Metrics

Mean absolute error (MAE) is used as the main metric. Mean absolute percentage error can't be used because the series has 0 values. That's why

mean absolute error / validation series mean

is used as the proxy for measuring error in percentage.

The Base Model

Projecting the last 30 observations in the train set is used as the base model.

MAE: 159138.54

Percentage error: 64%

ARIMA

The series is stationary and an ARMA process. But I used ARIMA for ease of implementation by setting "d" to 0. A 10x10 grid does a parameter search for **p** and **q**.

The best model is (7,0,7) (p,q,d), and the proximity of residuals to white noise is statistically significant.

MAE: 71789.72

Percentage error: 29%

Prophet

Results of Prophet forecasting with default parameters.

MAE: 73089.46

Percentage error: 29%

Feature-Based Forecasting

54 features were created for each timestep. These include rolling statistics (mean, std, var, min, max, 1st quantile, median and 3rd quantile), expanding statistics, daily features (month, day of month, day of week) and 30-day lag.

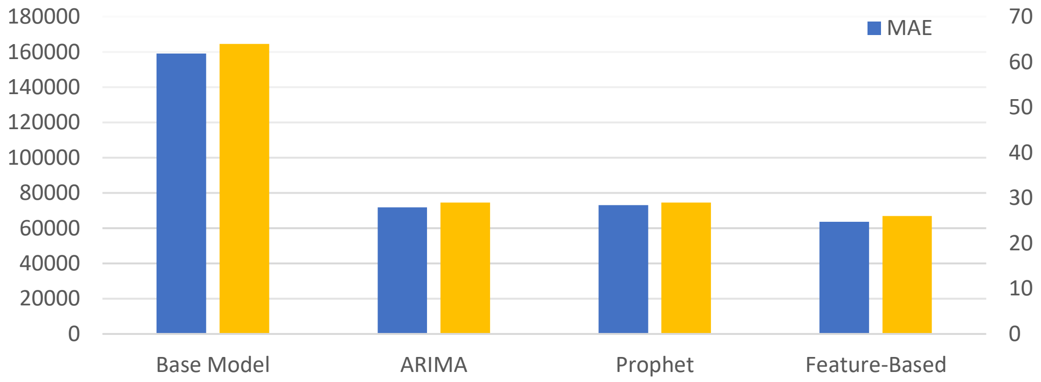
A chain of GradientBoostingRegressor is used for recursive forecasting. 1 row of features make the forecast for the next 30 days.

MAE: 63643.67

Percentage error: 26%

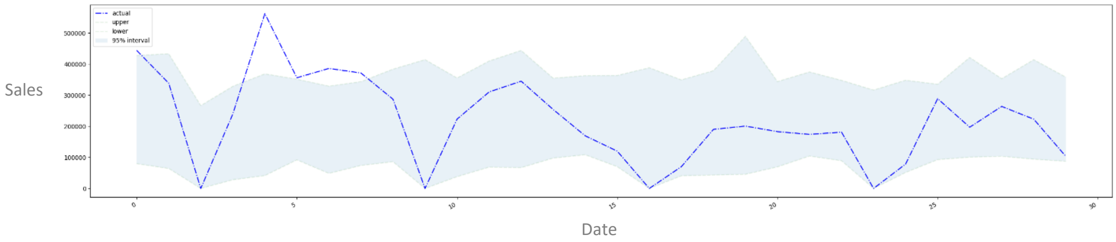
Results

The comparison of the five forecasting models:



Feature-Based forecasting is the best model with the least MAE and the percentage error. But it still has 26% error rate. To reduce this, I incorporated confidence intervals instead of point forecasting and trained two more regressors for upper and lower bounds by setting alpha to 0,95 and 0,05. Meaning 95% of the actual values will fall within the area as shown in the figure:

Prediction with 95% Confidence Interval


actual
upper
lower
95% interval
Sales
Date