Time Series in Retail Analytics

Tripods/StemForAll 2023
August 3rd, 2023

Peter MacNeil, Moez Boussarsar, Colin Hascup, Jennifer Kim, Jiaming Lyu, Cooper Orio, Jiamu Tang, Yining Qian, Xinyi Liu

Table of Contents:

Introduction and Motivation

In the retail analytics world, being able to tell if a time series is forcastable (and forecast it) is an important skill. The industry has been using one measure for a while now. This measure is min(Q_1,Q_2) which we will explain later. We are motivated to find a new way to measure forecastability of time series.

Our proposed way is to use CKS complexity of a time series as a forecastability measure. And to do so, we consider the length of the compressed time series.

CO

Introduction - Industry Standards

Industry Standards for Complexity Measurement

Min of two quantities:

$$Q_2 = \frac{\sqrt{\frac{1}{N-1} * \sum_{i=1}^{N} (|f(i) - f(i-1)| - \mu_\Delta)^2}}{\mu_\Delta}$$

$$Q_1 = \frac{(\frac{1}{N} * \sum_{i=1}^{N} (f(i) - \mu)^2)^{1/2}}{\mu}$$

$$\mu_\Delta = \frac{1}{N-1} * \sum_{i=2}^{N} |f(i) - f(i-1)|$$

(Coefficient of Variation)

Should have a relationship with the error for a forecasting method.

Methodology

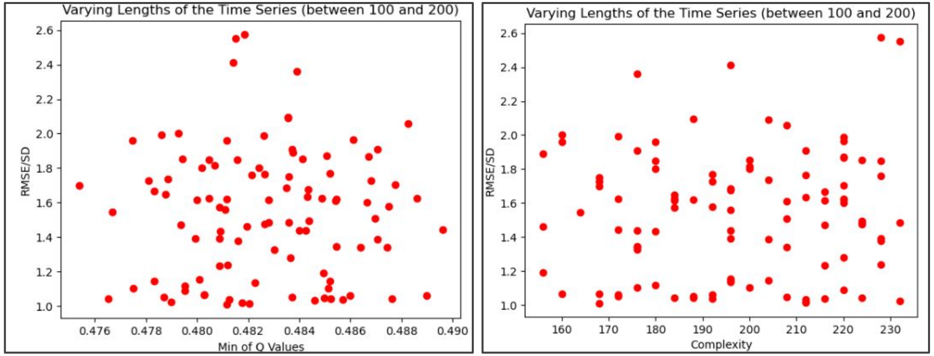
- Data Cleaning: First of all we had to go through all the sales data that we got and drop all the time series that weren't forecastable at all.
- Models trial: Then we tried to train our different models on 80% of our data and make it forecast the remaining 20% and compare it to the actual data.
- Time series discretization: It's a process that we're doing to make the time series more suitable for compression algorithms. It starts with a mapping from the time series values to [0,1], then we divide all the new values by m. Then we round the values to the closest j/m fraction. At the end we just multiply all these values by m to end up with integers
- Error .vs. Compressed length (CKS): We just graph both quantities to figure if there is any correlation.

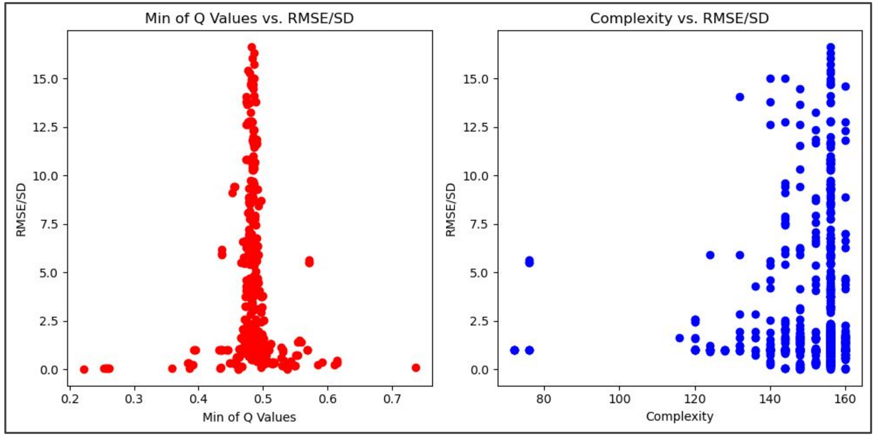
JK, CO

Synthetic Data

- Worked with synthetic data to try and better understand the effect that discretizing the time series might have on its proxy complexity and its forecastability in a more general sense.
- Attempted to compare the ability of complexity and the industry standard use of the minimum of the coefficient of variations of the series, and the consecutive differences between the series data points.
- For the triple exponential model, this was mostly inconclusive based on varying sizes, but varying frequencies did result in a weak trend showing error increasing as complexity increases.
- Enhanced function handles synthetic & real-life time series for LSTM & Prophet comparison
- Adapting code for TimeSynth enables varied synthetic data (AR, noise)
- Future potential: Generate surrogate data for deeper data analysis

(Varying sizes with sin(x) as the test function [below])





*(Varying frequencies with sin(a*x) as the test function [above])*

Models

-Triple Exponential

-LSTM

-Facebook Prophet

Data we used:

- Anonymous sales data from large retailer in csv format, from the 5th week of 2014 to the 3rd week of 2017
- Sales volume ranges from $0 to $127368
- Filters out rows whose last 10% of elements are 0
- 155 weeks between the time interval, forecast the last 20% of time series based on the previous 80%

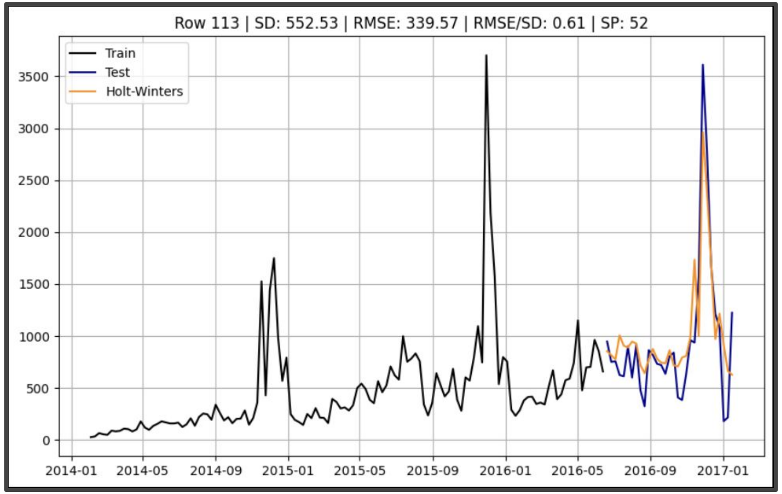
C

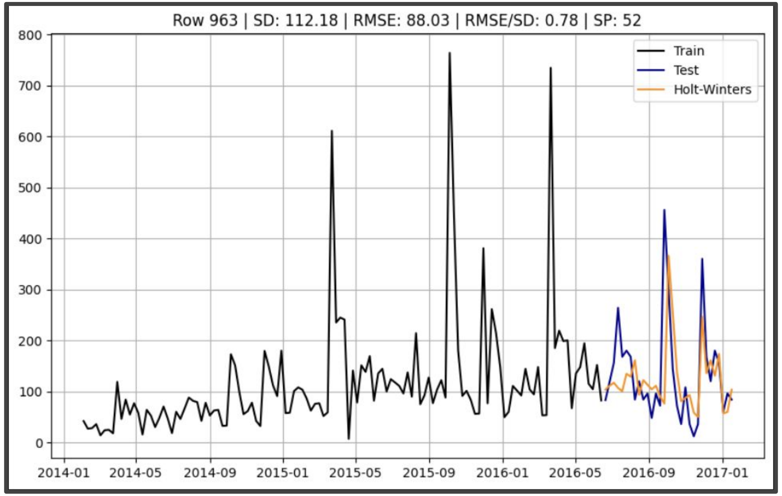
Triple Exponential

-Triple exponential model used on data sets without trailing 0's.

-Used 80% as training data and 20% as testing data. RMSE/SD was calculated for each.

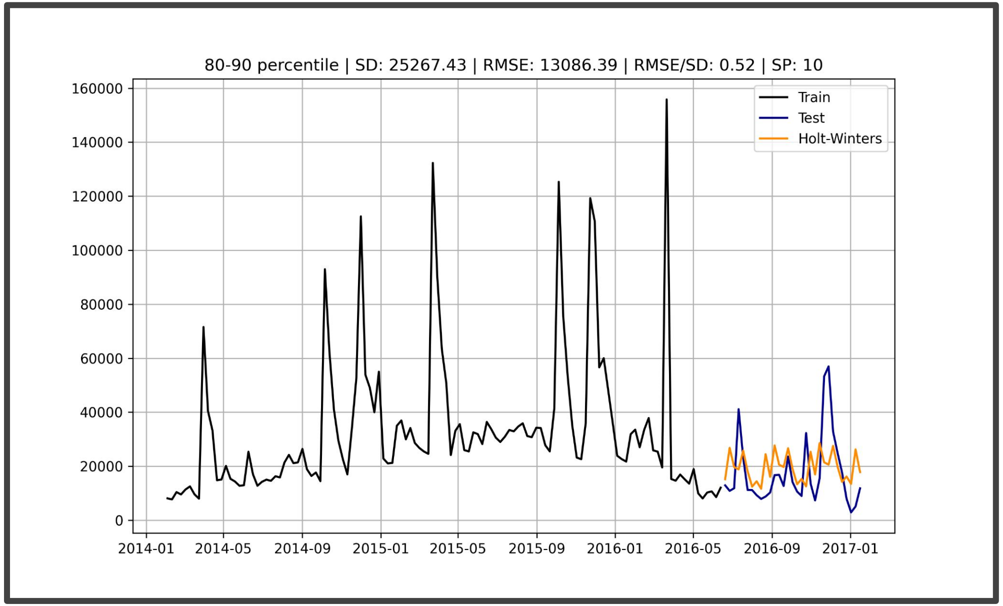
-Gave decent predictions on very seasonal data. (We tried different SP values to find the most accurate)





C

Triple Exponential With Aggregation by Total Sales



-Used Excel to sort and combine data sets into 10 percentiles.

-Ran triple exponential as before.

-Lowest percentiles gave worse predictions (less data).

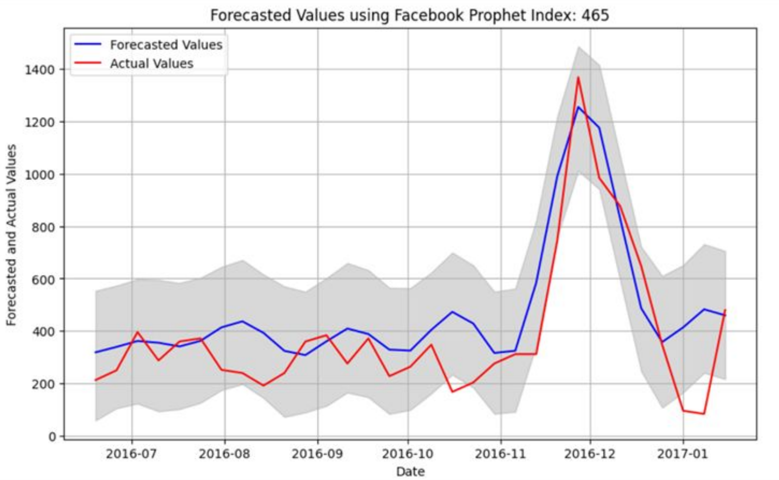
JL

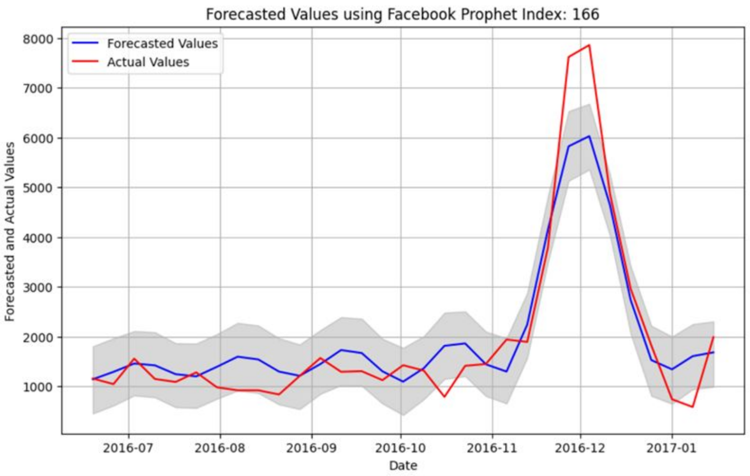
Facebook Prophet Forecast

- Predicts well with only one peak (less seasonal data) with default hyperparameters
- Adding regressors (e.g Consumer Price Index) and hyperparameters (e.g seasonality, holidays) could potentially improve the model

Product id	Location id	Error (RMSE/SD)
1753	550	0.554976489

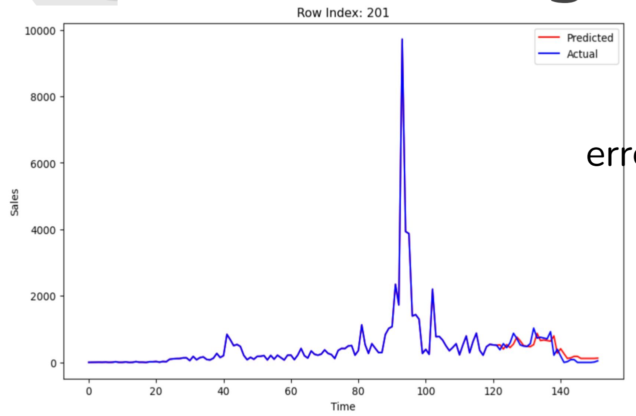
Product id	Location id	Error (RMSE/SD)
645	516	0.458250809





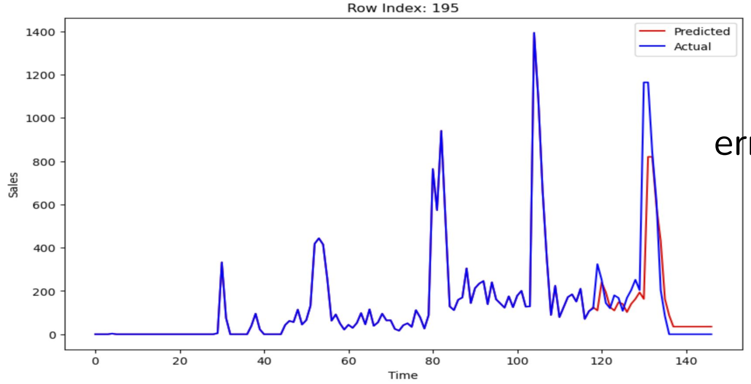
Chan

Forecasting results using LSTM



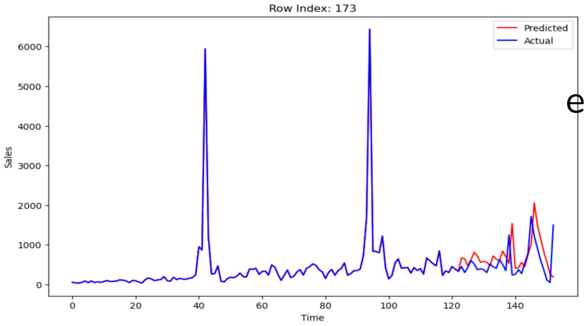
error:0.21

Testing RMSE for Row Index 201: 197.33159124093882
0.2105882522475097
1/1 [==============================] - 0s 141ms/step



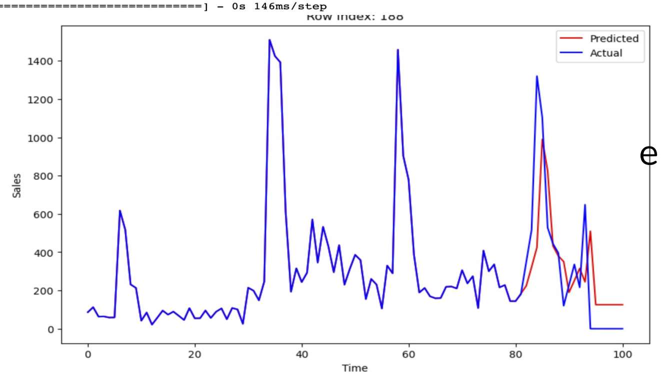
error:0.87

Testing RMSE for Row Index 195: 209.30971607998544
0.8658369703534585
1/1 [==============================] - 0s 146ms/step



error:0.67

Testing RMSE for Row Index 173: 490.3280592163528
0.6750334434887835
1/1 [==============================] - 0s 132ms/step



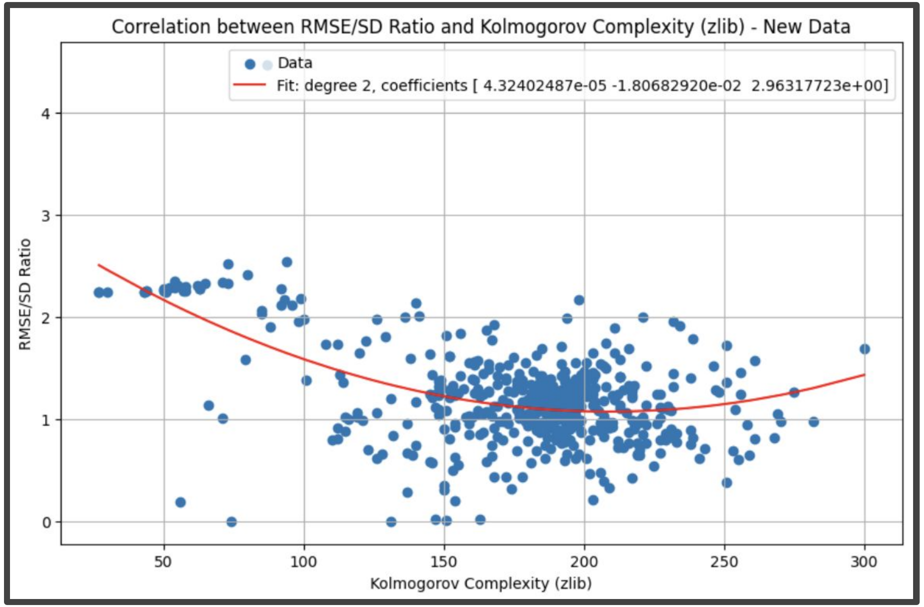
error:0.83

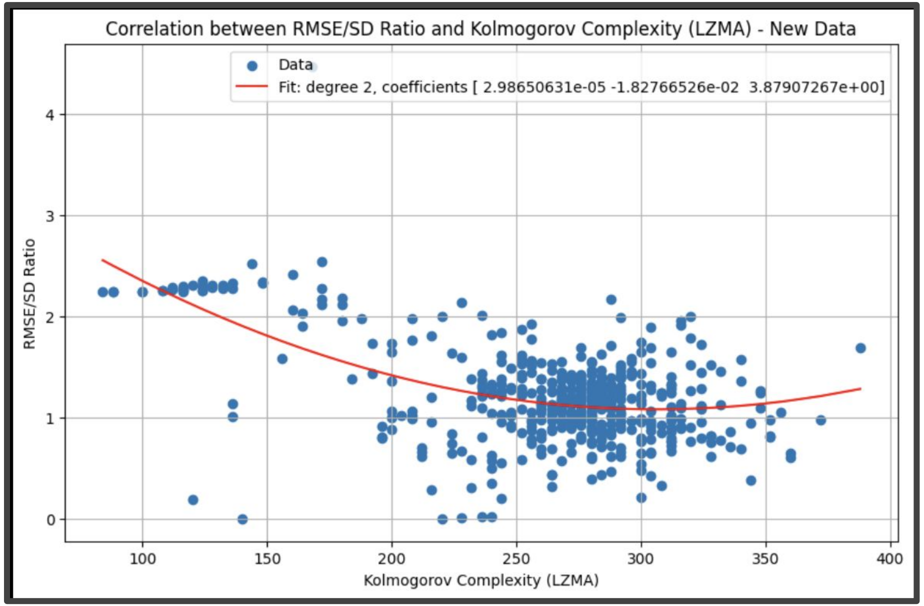
Testing RMSE for Row Index 188: 277.9500707184537
0.862974888070411
1/1 [==============================] - 0s 136ms/step

Findings

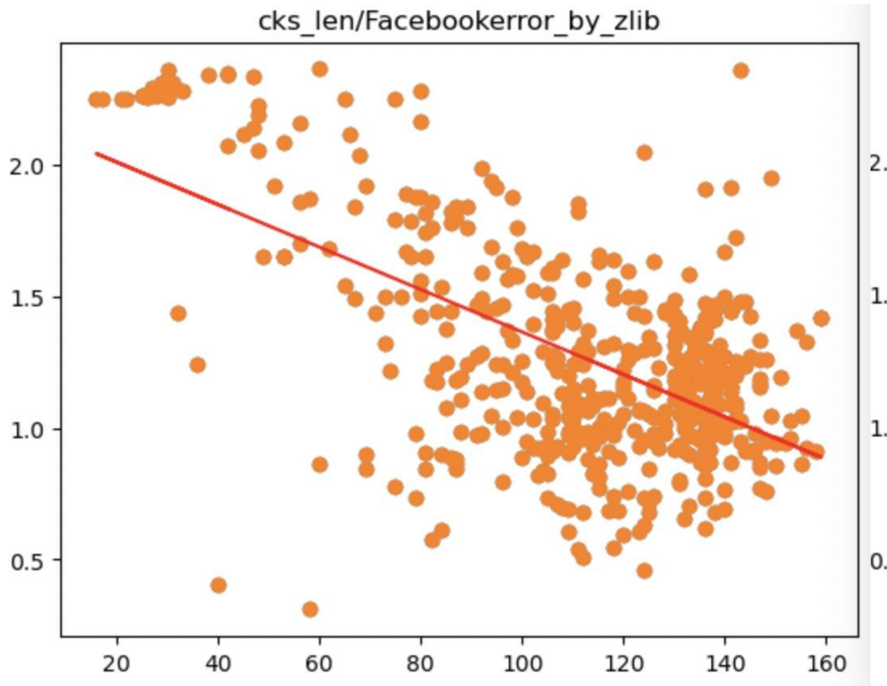
Does our proxy for CKS upper bound correlate to forecastability of time series?

C

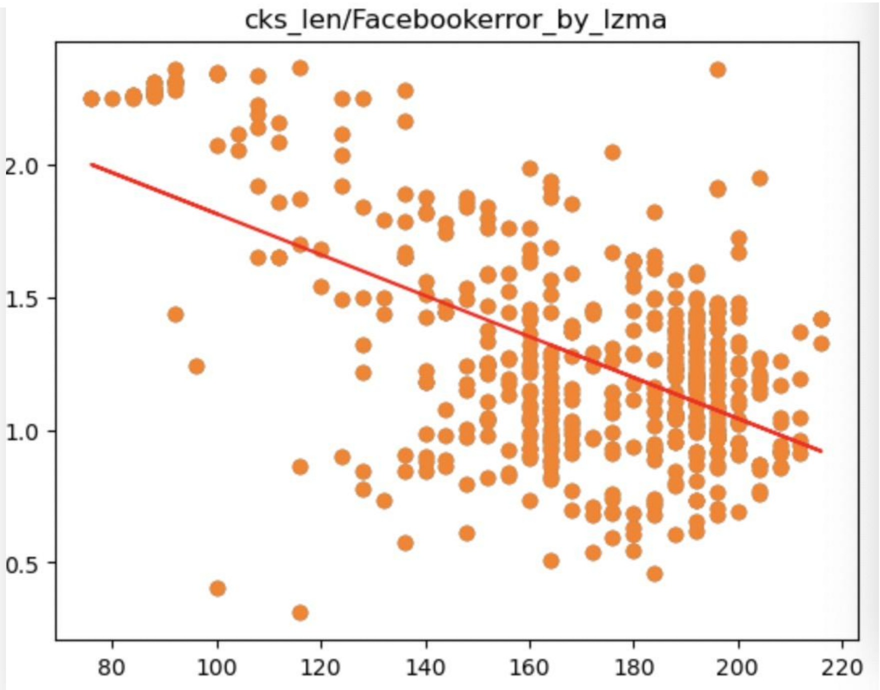




Correlations for the Triple Exponential

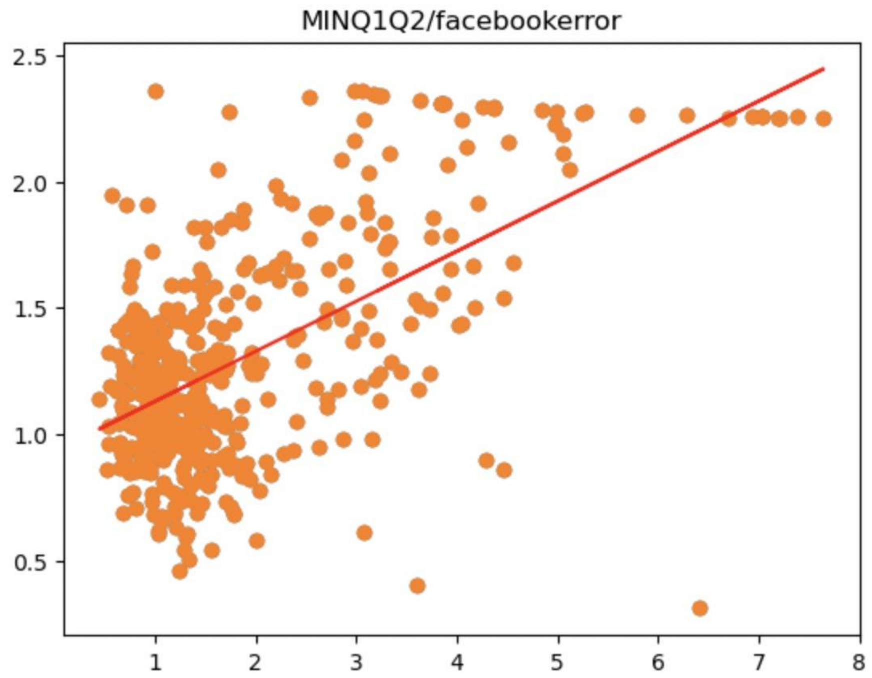


Correlation coefficient: -0.60



Correlation coefficient: -0.59

Correlations for the Facebook Prophet

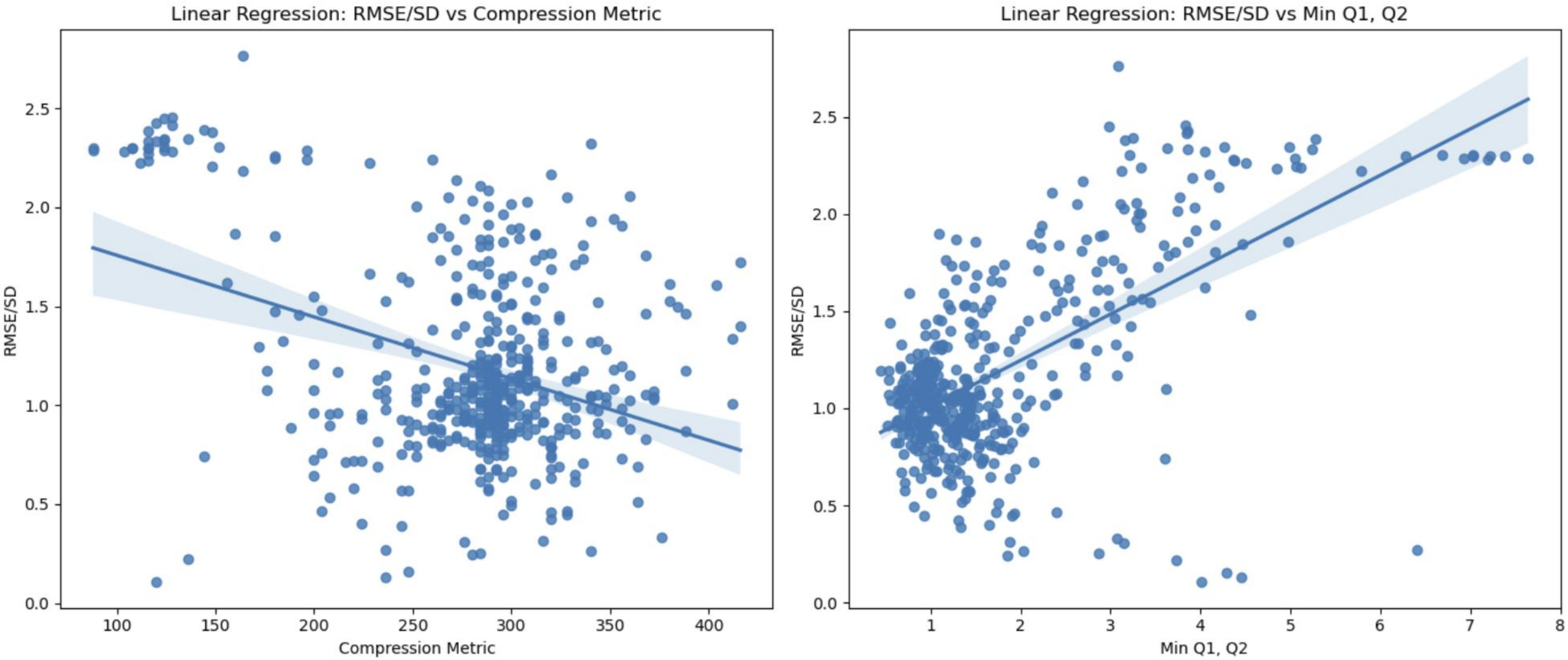


Correlation coefficient: 0.59

Industry Standard Prediction of Forecastability FB Prophet

P

Complexity and Industrial standard LSTM - No leading 0 removed



P

Complexity and Industrial standard LSTM - leading 0 removed

JK

Future Directions

- Explore more compression and discretization mapping techniques
- Explore regressors that are highly correlated with sales, e.g. CPI adjustment
- Fine-tuning hyperparameters for Facebook Prophet

End