Scaling Grocery Retail Pricing Modeling Using a Neural Net

1st Christopher Kuzemka
2025 Fordham Graduate School of Arts & Sciences - Data Science Focus
Fordham University
New York City, NY
ck56@fordham.edu

***Abstract*—This study is an attempt to scale a preexisting analysis, once completed, from an earlier semester at Fordham University, in which a Seasonal Auto-Regressive Integrated Moving Average with Exogenous variables considered based model (SARIMAX model in short) was implemented to forecast grocery sales for a store under the Favorita dataset offered on Kaggle [4]. This paper goes over a recap of the problem extensively by delivering context around the behavior of Favorita's business and it discusses new issues learned from the previous project that hurt the previous SARIMAX performance or otherwise misconstrued its results. With the intention to properly correct the mistakes, the author re-computes a chunk of the previous study and modularizes its processes with detail. With more precise and more careful data cleaning techniques implemented, improved SARIMAX models are shown sporting a different scoring metric known as Mean Absolute Scaled Error (MASE). The best performing SARIMAX model predicts sales for Store Number 45 (Grocery I category) and features a MASE index of 0.36 for forecasting the next 30 days. The results for the study were not submitted directly to Kaggle, due to the staleness of the competition and Kaggles strict choice of not releasing the testing labels to the public.**

Due to a variety of personal issues, it is with heavy heart to fully disclose that the project is severely incomplete. A neural net was working at some point, but pipeline issues for our data prevented any scaling. Time ran out. And as a result many corners had to be cut. While it is a gamble to write this within my submission, I as the author would like to extend an apology as this perhaps may be the most disappointing deliverable I have ever tried to output within the leftover time I have. It is a catastrophic failure to myself. I really wish I understood where it all went wrong on this project, but the most major setback on the project scope itself was isolating an extremely specific issue regarding introduced simulated economic crashes present from the leftover project before and buried under the extensive cleaning scripts.

This paper from here on will only feature mild comments on the work done and whatever possible deliverables that could be inserted before the final extended submission time. A large amount of notes are reflected within the source code's 'project_notebook.ipynb' which may serve slightly better for a better understanding of the project.

I. Introduction

With artificial intelligence (AI) developing at a very rapid pace, its inclusion in large-businesses for solving complex multi-variate problems is almost non-stop. Executive leaders of such businesses often stress the importance of leveraging AI in as many areas as possible, in an effort to keep up with industry and competitors. Relevant to such mention, now the term "AI" is almost considered to be a "buzz word" in this sense, as it umbrellas a vast amount of machine learning and modeling techniques that could be used to address a large variety of problems solvable with fast algorithmic computation. To some degree, this sometimes warrants the situation in which applying certain models to specific areas are not always the solution to understanding mass trends.

In this study, we focus on a well-known dataset housed on Kaggle known as the "Corporación Favorita Grocery Sales Forecasting" dataset [4]. This dataset was provided directly by Favorita, the largest native grocer in Ecuador [1], as part of a competition to create a generalized model to predict unknown sales data based on a variety of trainable features.

The dataset is considered to be very rich and inclusive of many features.

II. Background Information

A. Important Contextual Information Corresponding to Data Behavior

The previous report extensively covers much exogenous influence onto such sales data. For the sake of avoiding repetitiveness, we only cover some more important and directly relevant issues that may have spilled into our data.

B. Data Processing for Modeling

C. Data - Overview and Cleaning Summary

Re-highlighting the aim of our study, we were tasked with using historical sales data to make a model that could correctly forecast a 15-day period for Favorita. The study of Favorita's supply and demand began with five unique applicable datasets:

- **transactions**: columns = ['date', 'store_nbr', 'transactions']; shape = (83488, 3)
- **stores**: columns = ['store_nbr', 'city', 'state', 'type', 'cluster']; shape = (54,5)
- **oil**: columns = ['date', 'dcooilwtico']; shape = (1218,2)
- **holidays_events**: columns = ['date', 'type', 'locale', 'locale_name', 'description', 'transferred']; shape = (350,6)
- **training/testing**: columns = ['id','date', 'store_nbr', 'family', 'sales', 'onpromotion']; shape = (3000888, 6)

All of the datasets above were cleaned and merged together. Most notable cleaning includes several interpolations on oil pre merging, post merging, and during preprocessing for SARIMAX modeling.

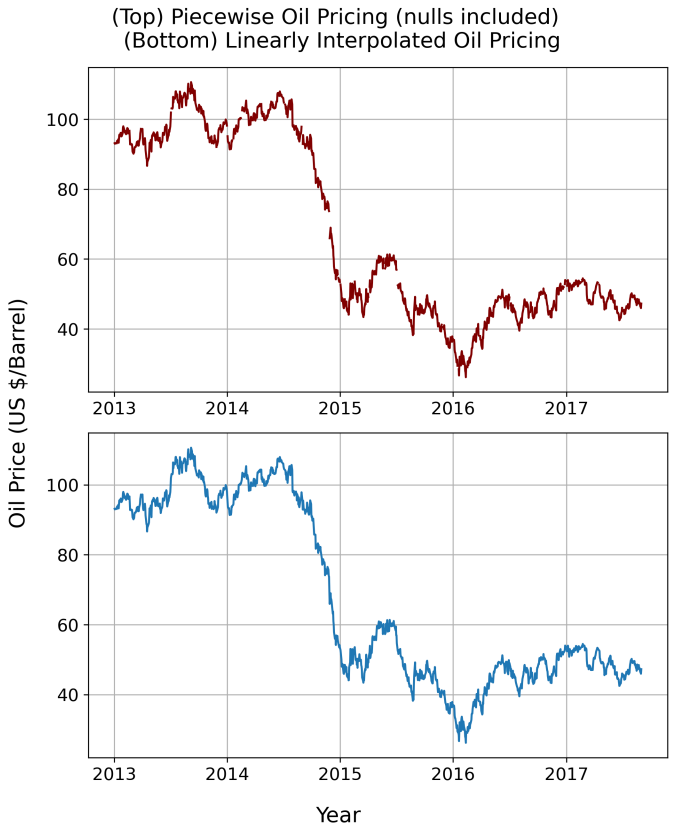

Fig. 1. [Top] The piecewise representation of our base oil pricing data. [Bottom] The transformation of oil pricing data using linear interpolation.

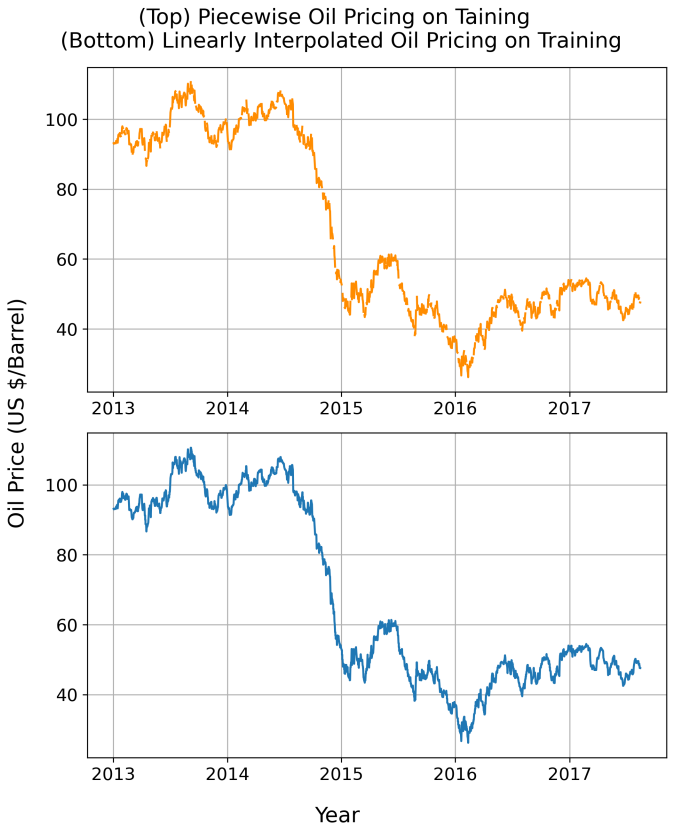

Fig. 2. [Top] The piecewise representation of our transformed oil pricing data against the training set. [Bottom] The repeated transformation of oil pricing data off our training set using linear interpolation.

Above is a plot showcasing a pre-merged oil dataframe leveraaging linear interpolation between local boundary values with forward filling and back filling to satisfy edge cases.

Above is the second interpolation done on the oil when merged on the date index of the training. The training data had higher granularity which is why more nulls appeared.

D. Model Scoring Techniques

Mean Anbolute Scaled Error (MASE) is a scoring metric good for our timeseries data and forecasting approaches as it ratios the Mean Absolute Error (MAE) of a model against its naively assumed MAE (called this MAE_α). Thus the resulting equation for MASE is as follows:

$$\text{Eq. (1) MASE} = \frac{\text{MAE}}{\text{MAE}_\alpha}$$

We focus first on a naive MAE which is computed based on naive forecasting. Naive forecasting is is a simplistic approach towards forecasting which makes the assumption that a future value remain mostly unchanged from the last previous value - ergo it does not take into account exogenous conditions and featuress that may influence the trend of a value over time. Naive forecasting is done often for establishing a benchmark or baseline.

For the purpose of this project, we are going to construct our own function that we can use for measuring the MASE of our predictions.

We start with showcasing the formula for the naive MAE, MAE_α. This specific formula looks over an entire duration of timesteps...

$$\text{Eq. (2) MAE}_\alpha = \frac{1}{N-1}\sum_{i=2}^{N} \text{abs}(y_i - y_{i-1})$$

Now if we incorporate a seasonal variation from the forecast, we would then equate the forecast to an actual value from the period, but from the previous season corresponding to the current period (i.e. this year vs last year)...

Thus we have:

$$\text{Eq. (3) MAE}_\alpha = \frac{1}{N-m}\sum_{i=m+1}^{N} \text{abs}(y_i - y_{i-m})$$

Above we cover the MAE for the naive forecast. To continue building our MASE metric, we would also leverage an MAE on our new algorithm meant to construct predictions (call then o_n)

Therefore:

$$\text{Eq. (4) MAE} = \frac{1}{N}\sum_{j=1}^{N} \text{abs}(o_i - y_i)$$

Finally, the MASE is then constructured from these error metrics as:

$$\text{Eq. (1) MASE} = \frac{\text{MAE}}{\text{MAE}_\alpha}$$

As a ratio, MASE is measured around the value of 1. When the MAE from our model performs better than the MAE of the naive assumption, the MASE ratio goes below the value of 1. Above the value of one indicates our model's error is much stronger than the naive error, indicating the model performs very poorly.

E. Model Pre-proccesing - Cyclic Feature Encoding and One-hot Encoding

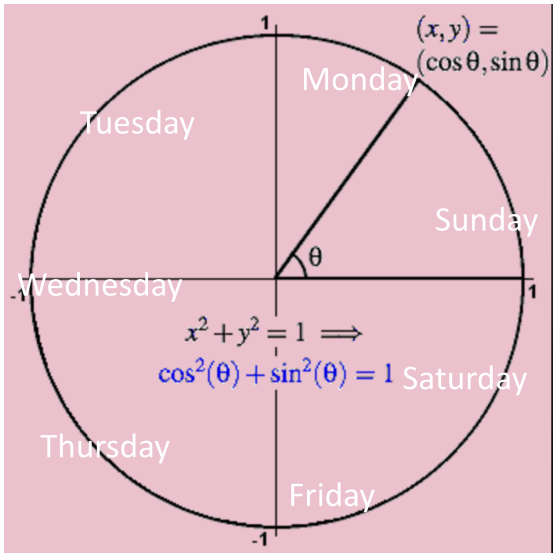

Fig. 3. A modified image by the author showcasing graphically how cyclic features could be redimensionalized. Use of a circle is meant to describe the sine and cosine relationship and its mapping to the 'day_of_week' column.

The equations below are ones used for cyclic encoding on the dataset. Since days of the week felt like an important feature and could not be set on a hierarchy for any model, a cyclic encoding was necessary to describe the categorical relationship between days that influence shopping behaviors.

One hot encoding days of the week (a feature engineered column) would explode features for our model. It would sustain uniqueness of days, but remove the cyclic and sequential relationship of day progression. Choosing to strictly apply a number representing each day would imply that days have a hierarchy whereas application of an integer value to represent a day has no arbitrary meaning. A model may believe that if you say "Saturday = 7; Sunday = 1", then "Saturday is greater than Sunday" even though Sunday comes after Saturday.

Leveraging a unit circle and a sine cosine relationship on a numerical representation for each day solves the issue of numerically representing days of the week while preserving the cyclic ordinality.

$$\text{Eq. (5) } x = \sin(\frac{a \times 2\pi}{\max(a)})$$

$$\text{Eq. (6) } x = \cos(\frac{a \times 2\pi}{\max(a)})$$

III. Results and Future Expansions

A. SARIMAX

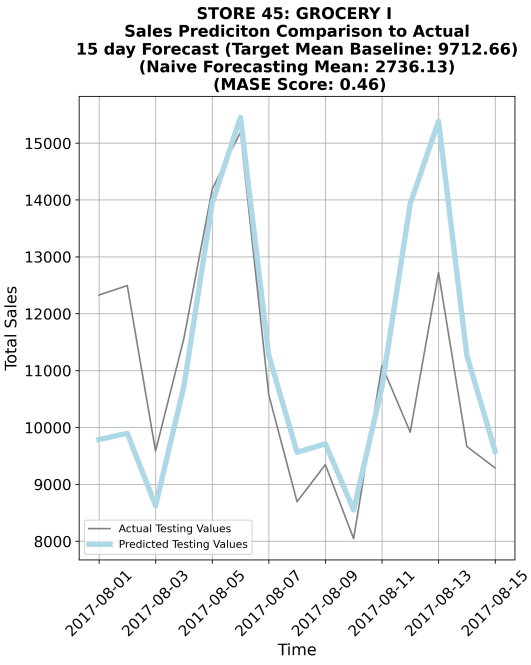

Fig. 4.

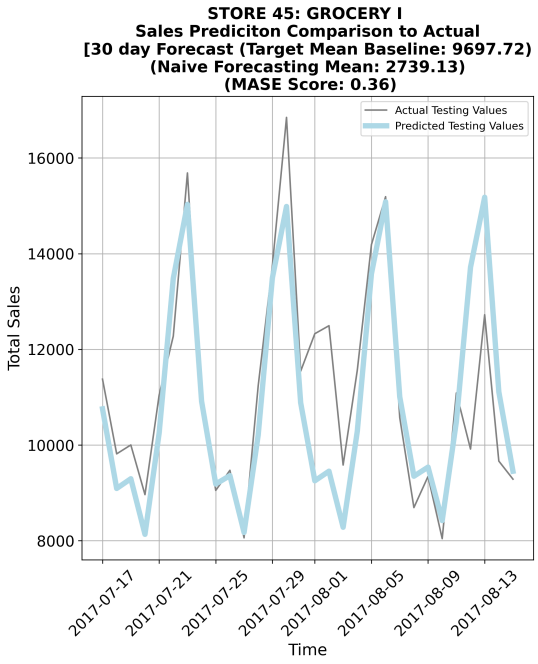

Fig. 5.

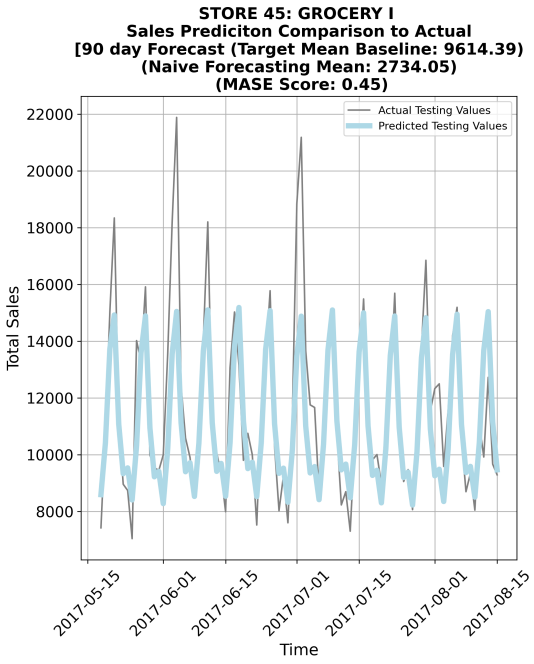

Fig. 7.

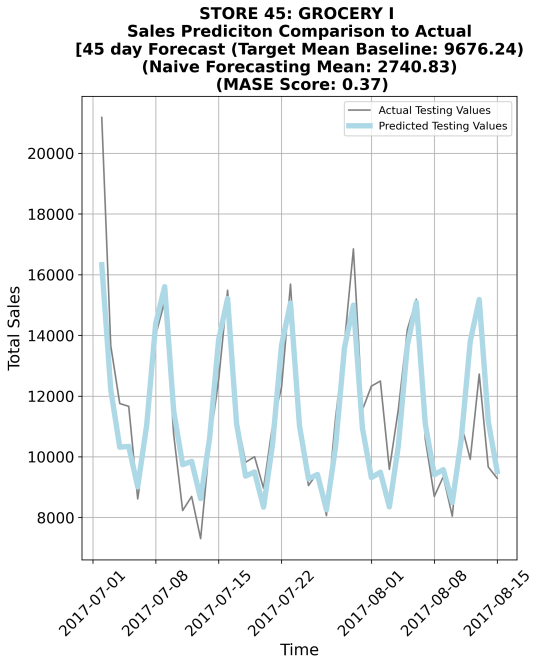

Fig. 6.

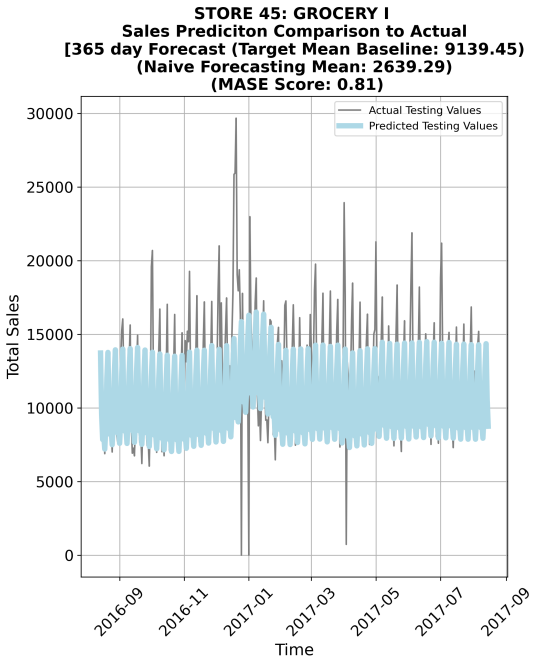

Fig. 8.

Each SARIMAX model offers an attractive MASE score in that it is better performing than the Naive forecast. However note that tuning is needed should the forecasting of days increases. These models are also solely limited to one store and one category. A neural net's consideration for this project was to have the ability to leverage a single model to understand multivariate patterns.

B. LSTM

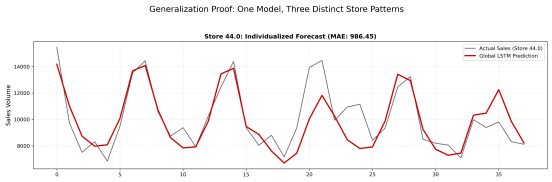

Fig. 9.

Above we showcase the implementation of our LSTM and the only successful image generated showcasing performance on store 44. Originally we intended to scale this to 3 stores to start, but stores 45 and 46 were being lost in the data process pipeline and the issue persist. As of now the neural net code does not work.

REFERENCES

A note to the reader. The paper ends here. I was unable to clean the paper up for a respectable presentation. Some sources labeled below carried over into this project's discussion however not all sources are recorded here due to the time constraints and not all of the sources here are called upon in some of the paragraphs above. There is no intent to plagiarize whatsoever. AI was used to help guide the project's directions and address syntax and code errors, but no AI was leveraged in writing this paper.

Again....I sincerely apologize for any difficulty created in grading this paper.

REFERENCES

[1] Taylor, James. "James Taylor Helps Ecuador's Largest Retailer Accelerate Innovation to Achieve Exponential Growth." *Corporación Favorita* 2019

[2] Mead, Dave. "The 20214 Plunge in import Petroleum Prices: What Happened?" *US Bureau of Labor Statistics*

[3] Lazzeri, Francesca. "Introduction to Feature Engineering for Time Series Forecasting." *Data Science at Microsoft* 2021, October 5th

[4] "Corporación Favorita Grocery Sales Forecasting." *Kaggle* 2017, October 19th

[5] "Crude Oil Prices: West Texas Intermediate (WTI)." *Federal Reserve Economic Data* 2024, December 16th

[6] "Oil-Prices." *Github* 2024, December 11th

[7] Baakwa. "Time Series Forecasting Predictive Model Building for Store Sales at Favorita Grocery Retailer." *Medium* 2024, Februrary 4th

[8] Blyth, Russell. "The Amazing Unit Circle: The Fundamental Trigonometric Identity." *MathMistakes* 2019

[9] Matsumoto, Shiro. "Understanding the Capabilities of Cyclic Encoding." *Medium* 2023, October 17th

[10] Alharbi, Fahad R. "A Seasonal Autoregressive Integrated Moving Average with Exogenous Factors (SARIMAX) Forecasting Model-Based Time Series Approach." *Inventions* 2022, October 16th

[11] Hyndman, Rob J. "Forecasting: Principles and Practice" 2nd ed. (Ch. 8.1 ARIMA models - Stationary and differencing). *OTexts*, 2018, May 8th

[12] Kuzemka, Christopher. "store_sales_time_series_forecasting". *Github* 2024, December 16th