ARIMA models with explanatory variables for time series forecasting on M5 Accuracy competition

Igor Faluhelyi [a], José Augusto Fiorucci [b]

[a] Department of Statistics, University of Brasília, Brasília, Brazil, Igor.Faluhelyi@aluno.unb.br

[b] Department of Statistics, University of Brasília, Brasília, Brazil, jafiorucci@unb.br

Abstract. This article delves into an analysis of time series forecasting, building upon existing research in this field by examining ARIMA models with explanatory variables and TBATS models in a particular modern competition context. At the conclusion, the ultimate goal is to present a comprehensive solution for the M5 Accuracy competition, the fifth instalment of the M competitions organized by the M Open Forecasting Centre (MOFC) at the University of Nicosia in 2020. Time series models were utilized to accurately forecast the daily sales volume of retail products sold by Walmart, in accordance with the framework of the M5 competition, since the historical sales data were generously provided by Walmart itself, for this globally open competition hosted on Kaggle (a renowned online community for data scientists and machine learning practitioners). Among the time series models evaluated, the TBATS models stood out, outperforming the ARIMA models and achieving commendable scores in the competition.

Keywords. Time series forecasting, dynamic regression models, TBATS models, M5 Accuracy competition

1. Introduction

Forecasting competitions have gained popularity as statistical, computational, and mathematical techniques are being more widely adopted and utilized worldwide. The 2020 M5 Forecasting - Accuracy competition is a prime example of this trend, as part of the successful series of M competitions organized by the MOFC since 1982, starting with M1. The M5 competition attracted around 6,000 participants from across the globe and awarded a total of $50,000 in prizes to the top five performers.

In this article, we will focus on utilizing time series models and the Python programming language to provide a comprehensive solution for the M5 competition. This particular competition involves using hierarchical sales data from Walmart, the largest company in the world by revenue, to forecast daily sales for the next 28 days. The dataset covers stores in three US states - California, Texas, and Wisconsin - and includes detailed information such as item level, department, product categories, and store details. Moreover, the dataset also includes explanatory variables such as price, promotions, day of the week, and special events, making it a rich resource for improving forecasting accuracy. We will delve into the intricacies of this competition and showcase how Python and time series modeling techniques can be leveraged to tackle the challenge of accurate sales forecasting in a complex retail environment.

2. Research Methods

2.1 Dynamic regression models

A brief introduction to Dynamic regression models will be provided, with further details available in the work of **Hyndman and Athanasopoulos (2018)**.

Dynamic regression models, also known as ARIMA models with explanatory variables, are essentially linear regression models where autocorrelation, which is transferred to the error component, is addressed by an ARIMA model.

Consider the regression model in the form:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_k x_{k,t} + \varepsilon_t \qquad (1)$$

where y_t is a linear function of k predictor variables $(x_{1,t}, \cdots, x_{k,t})$ and ε_t is assumed to be white noise in according to equation 1. However, it is possible that ε_t exhibits autocorrelation. To emphasize this, we replace ε_t with η_t, assuming that η_t follows an

ARIMA model. If η_t follows ARIMA(1,1,1), we have:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_k x_{k,t} + \eta_t,$$

$$(1 - \phi_1 B)(1 - B)\eta_t = (1 + \theta_1 B)\varepsilon_t \quad (2)$$

Where ε_t is a white noise according to equation (2). Note that the model has two error terms - the regression error denoted by η_t, and the ARIMA model error denoted by ε_t. Only the error of the ARIMA model is assumed to be a white noise.

2.2 Dynamic harmonic regression models

A brief introduction to Dynamic harmonic regression models will be provided, with further details available in the work of **Hyndman and Athanasopoulos (2018)**.

Dynamic harmonic regression is based on the principle that a combination of sine and cosine functions can approximate any periodic function.

$$y_t = \beta_0 + \sum_{k=1}^{K} [\alpha_k s_k(t) + \gamma_k c_k(t)] + \varepsilon_t \qquad (3)$$

Where:

$$s_k(t) = \sin \frac{2\pi kt}{m} \qquad (4)$$

$$c_k(t) = \cos \frac{2\pi kt}{m} \qquad (5)$$

Where m is the length of the seasonal cycle, K is the number of harmonics needed for approximation, α_k and γ_k are regression coefficients, ε_t is modeled as a non-seasonal ARIMA process, accorging to equations 3 to 5.

2.3 TBATS models

A brief introduction to TBATS models will be provided, with further details available in the work of **Livera, Hyndman e Snyder (2011).**

BATS is an acronym for Box-Cox transform, ARMA errors, Trend, and Seasonal components. The model is also denoted as BATS$(\omega, \phi, p, q, m_1, m_2, \cdots, m_T)$ to indicate the Box-Cox parameter, Damping or dumping, ARMA(p, q) components, and the seasonal periods $(m_1, \cdots, m_T)$. The model structure follows with the Box-Cox transformation, ARMA errors, and T seasonal patterns.

The notation ${y_t}^{(\omega)}$t is used to represent the Box-Cox transformation applied to the observed series with parameter ω.

$${y_t}^{(\omega)} = \frac{y_t - 1}{\omega}, \omega \neq 0$$

$${y_t}^{(\omega)} = \log(y_t)$$

$${y_t}^{(\omega)} = l_{t-1} + \phi b_t + \sum_{i=1}^{T} \left[{s_{t-m_i}}^{(i)} \right] + d_t, \quad (6)$$

$$l_t = l_{t-1} + \phi b_{t-1} + \alpha d_t$$

$$b_t = (1 - \phi)b + \phi b_{t-1} + \beta d_t$$

$${s_t}^{(i)} = {s_{t-m_i}}^{(i)} + \gamma_i d_t$$

$$d_t = \sum_{i=1}^{p} [\phi d_{t-i}] + \sum_{i=1}^{q} \theta_i \varepsilon_t$$

Where $(m_1, \cdots, m_T)$ denote the seasonal periods, l_t is the local level at time t, b is the long-term trend, b_t is the short-term trend at time t, ${s_t}^{(i)}$ represents the i-th seasonal component at time t, d_t denotes an ARMA(p, q) process, and ε_t is white noise. The smoothing parameters, denoted by α, β, and γ_i for $i = 1, \cdots, T$ are used for the smoothing process, according to equation 6.

The TBATS model is a variation of the BATS model that incorporates the seasonal component using a trigonometric formulation.

3. Exploratory data analysis

The M5 competition dataset includes the sales history of nearly 3075 different items, which are sold in 10 different stores located across three US states - California, Texas, or Wisconsin. Additionally, the dataset includes weekly prices for these items, which may indicate an increase or decrease (due to promotions) in price over the weeks. The historical sales data captures the daily sales of these items in their respective stores. Along with the sales data, a calendar file is provided, containing valuable information on holidays, special occasions, day of the week, and accepted forms of payment for each day over a timeframe of more than 5 years, from 2011-01-29 to 2016-06-19. These comprehensive datasets collectively serve as a rich source of information for developing time series models for accurate daily sales forecasting in this context. In the following sections, we will explore the steps and techniques employed in leveraging these datasets, along with Python's powerful data analysis and modeling capabilities, to develop a solution for the M5 competition.

As the data provided for the M5 competition is analyzed, graphical representations can offer valuable insights. Visualizing the data through plots and charts can help us identify patterns, trends, and characteristics of the time series data. In this analysis, we will present Figures 1 to 4, which are original visualizations created using Python's powerful data visualization libraries, such as Matplotlib and Seaborn. These visualizations will provide us with a visual overview of the data, aiding in our understanding of the M5 competition dataset and informing our approach to developing accurate sales forecasting models.

Based on figure 1, it can be concluded that a higher number of sales occur during weekends

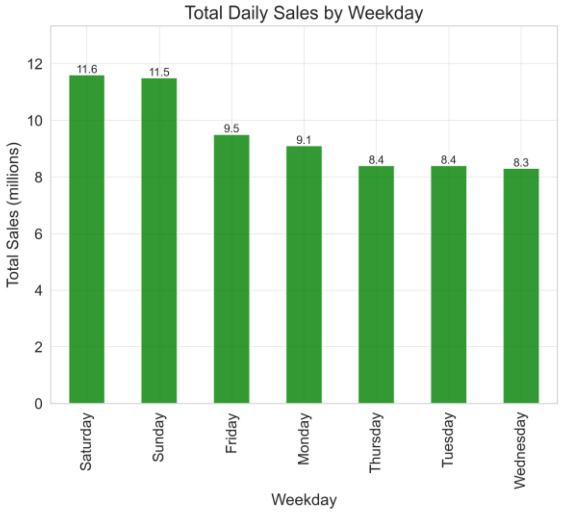

Fig. 1 – Total daily sales by weekday.

Based on figures 2 and 3, it can be concluded that, on average, days with special events tend to have higher sales compared to days without special events. Additionally, it appears that religious and sports events result in higher sales compared to national and cultural events.

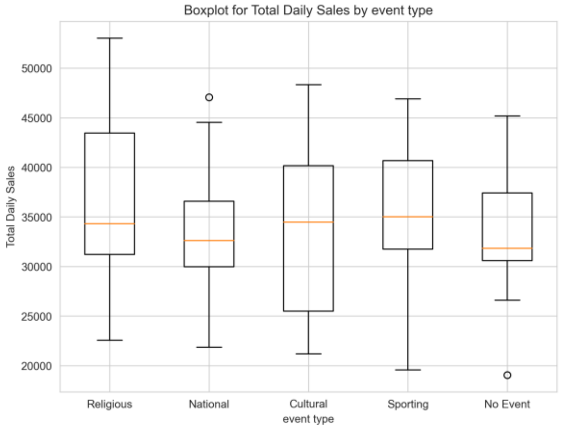

Fig. 2 – Boxplot for total daily sales by event type.

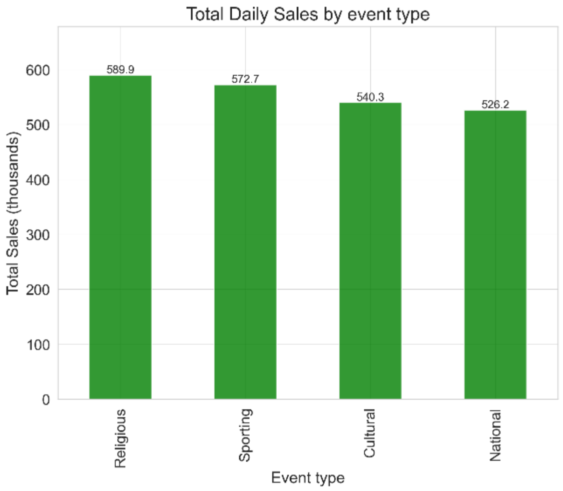

Fig. 3 – Total daily sales by event type.

Based on figure 4, it can be concluded that weeks with lower percentage changes in sell prices of items tend to have higher sales volumes during the same week.

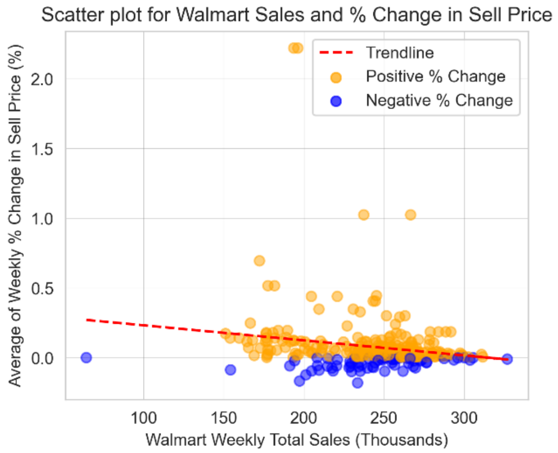

Fig. 4 – Scatter plot for weekly Walmart sales and weekly sell price percentage change.

4. Results

The models introduced in Section 2 were fitted to the data, and their respective accuracies were evaluated using the sMAPE (symmetric Mean Absolute Percentage Error) and MASE (Mean Absolute Scaled Error) metrics. For more details on the metrics used, refer to **Hyndman and Koehler (2006)**. Furthermore, the paper presents the private score obtained in the M5 competition, which ranks and determines the competitors as winners based on the lowest score. In other words, the closer the score is to zero, the better the performance in the competition. In table 1, the best-performing time series model in the M5 competition is highlighted in bold.

Tab. 1 – Results of the solution provided in the M5 competition.

Model	sMAPE	MASE	Score
Dynamic Regression	0.148	1.402	1.05334
Dynamic Harmonic Regression	0.076	0.791	0.68533
TBATS	**0.071**	**0.739**	**0.64666**

The actual Walmart daily sales data is depicted in Figure 5, which also includes forecasts based on different model fits for the next 28 days.

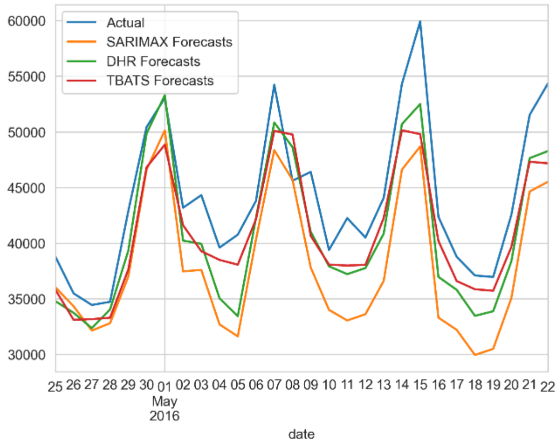

Fig. 5 – Walmart daily sales: forecasts and actuals.

5. Discussion

The winning solution of the competition was developed by competitor Yeonjun In (from South Korea) using a machine learning technique known as LightGBM. The solution achieved a M5 private score of 0.52043, demonstrating its effectiveness. Among the top seven solutions, five of them utilized the LightGBM technique, displaying its superiority in terms of M5 private score compared to the time series modeling approach employed in this project.

6. Conclusions

In addition to the monetary prize, competitors are also awarded medals based on their rankings on the official competition leaderboard on Kaggle. Gold medals are awarded to the top 21 competitors, Silver medals to competitors ranked up to 277, and Bronze medals to competitors ranked up to 555. The TBATS forecasts would have ranked at position 294 with private score of 0.64666, while the Harmonic Dynamic Regression forecasts would have ranked at position 500 with private score of 0.68533, out of a total of 5558 competitors worldwide.

Based on the results obtained, both the Harmonic Dynamic Regression and TBATS models would have resulted in a Bronze medal, which is a commendable achievement within the competition ranking.

7. Acknowledgement

This article is a re-edition of the text titled "Modelos ARIMA com variáveis explanatórias", originally authored by Igor Faluhelyi and Prof. José Augusto Fiorucci as part of the Institutional Program of Scientific Initiation Scholarships - PIBIC (CNPq) 2021/2022, in which we were recipients of the scholarships. We express our gratitude to University of Brasilia (UnB) for their support.

8. References

[1] Hyndman, R. J.; Athanasopoulos, G. *Forecasting: Principles and Practice.* 2. ed. Australia: Otexts, 2018.

[1] Livera, A. M. D.; Hyndman, R. J.; Snyder, R. D. Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, Taylor Francis, v. 106, n. 496, p. 1513–1527, 2011

[3] HYNDMAN, R. J.; KOEHLER, A. B. Another look at measures of forecast accuracy. *International Journal of Forecasting*, v. 22, n. 4, p. 679–688, 2006