Complex Comparison of Statistical and Econometrics Methods for Sales Forecasting

Oleksandr Kosovan(✉) and Myroslav Datsko

Ivan Franko University of Lviv, Lviv, Ukraine
{Oleksandr.Kosovan.AEKE,myroslav.datsko}@lnu.edu.ua

Abstract. Sales forecasting holds substantial significance in shaping decision-making processes in the retail industry. This study investigates the contemporary landscape of sales forecasting methods, aiming to provide empirical insights into the performance of various statistical and econometric models. By rigorously evaluating these models across diverse datasets, we identify stable methods that consistently demonstrate reliable predictive capabilities. Our research contributes to the field by offering baseline models that can furnish trustworthy forecasts, guiding practical applications and future research efforts. The paper details the study's methodology, results, and discussions, enabling a comprehensive understanding of the strengths, limitations, and implications of the evaluated forecasting methods.

Keywords: sales forecasting · retail · econometrics · time series

1 Introduction

Sales forecasting plays a pivotal role in both theoretical and practical domains, influencing critical decision-making processes in the retail industry. As outlined in Sect. 2, our investigation into the current landscape of the sales forecasting domain has revealed a need for empirical knowledge regarding the performance of various statistical and econometrics methods.

This study is driven by the central objective of addressing the challenge of selecting robust forecasting models for sales prediction across multifaceted business contexts. Through rigorous assessment of numerous statistical and econometric models across a spectrum of datasets, our research endeavors to pinpoint models that consistently exhibit dependable performance. The resulting insights will furnish recommendations for baseline models capable of furnishing reliable forecasts, thereby guiding both practical implementations and future research endeavors.

Detailed information regarding the utilized datasets can be found in Sect. 3, followed by comprehensive descriptions of the employed statistical and econometrics methods in Sect. 6. Our experimental design is outlined in Sect. 4, and

the evaluation metrics employed for assessing model performance are detailed in Sect. 5. The subsequent sections, namely Sects. 7 and 8, present a thorough exposition of the study's findings and subsequent discussions, leading to a holistic understanding of the efficacy and implications of the evaluated forecasting methods. Also, all experiment's code base is public in GitHub[1] and can be reproduced.

2 Background

There are various methods of research in the forecasting area, and one of them is through competitions like the M competitions. Rob J. et al. highlighted the transformative impact of the M competition series, a collection of major forecasting contests that have driven substantial advancements in theory and practice [15]. Other examples are competitions like the "M5 Forecasting Accuracy," hosted on Kaggle, which have yielded practical insights and methodological innovations, contributing to the field's progress [14,22]. Similar competitions, including the "Corporación Favorita Grocery Sales Forecasting" based on Ecuadorian retail data [8] and "HACK4Retail" focused on Ukrainian retail [19], further enrich the landscape.

Donna F. et al. conducted a study that explored the interplay of organizational factors in sales forecasting management, shedding light on the significance of integrated approaches [5]. Empirical investigations have expanded the horizon, examining diverse methodologies on specific datasets. For instance, Ensafi Y. et al. compared forecasting methods like Prophet, LSTM, and CNN [7]. Valls P. et al. extended this exploration to encompass RNNs and transformers [26], contributing to the methodological diversity.

Cross-dataset research also plays a role. Florian H. et al. engaged in a comparative analysis of machine learning solutions within sales forecasting, yielding valuable insights into model performance [12].

The primary objective of this study is to address the challenge of selecting effective forecasting models for sales prediction across diverse business contexts. By rigorously evaluating the performance of various statistical and econometric models across a range of data sets, this study aims to identify models that consistently demonstrate reliable performance. The outcomes will offer recommendations for baseline models that can provide trustworthy forecasts, guiding both practical applications and future research.

The central research question is: What is the relative performance of different statistical and econometric models in the context of sales forecasting across a diverse list of data sets? In line with this question, it is hypothesized that discernible patterns will emerge, highlighting models that consistently exhibit stable performance across diverse data sets. These models are expected to serve as valuable benchmarks for future investigations, contributing to the advancement of research in this critical domain.

[1] GitHub, https://github.com/OleksandrKosovan/complex-sales-forecasting.

3 Data

In this study, we utilized three distinct and diverse data sets (see Table 1), each comprising a comprehensive retail history encompassing a wide list of products, geographical locations, and shops, all organized within a specific hierarchical structure. By incorporating these distinct data sets, we aimed to capture a comprehensive representation of the complexities and nuances present in retail environments. The summary statistics reveal that the data sets have different scales.

Table 1. The data set features

Feature Name	M5	Fozzy Group	Corporación Favorita
Country	USA	Ukraine	Ecuador
Products Count	3049	1961	4400
Start Date	Jan 29, 2011	Jan 1, 2020	Jan 1, 2013
End Date	Jun 19, 2016	Jul 19, 2021	Aug 15, 2017
Mean	1.29	0.21	8.58
Standard deviation	4.15	1.23	21.92
Minimum	0.0	0.0	0.0
Q_1 (25%)	0.0	0.0	2.0
Q_2 (50%)	0.0	0.0	4.0
Q_3 (75%)	1.0	0.0	9.0
Maximum	763	801	44142.0

One of the data sets utilized in our analysis is the M5 data set, made available by Walmart on the Kaggle competition. This data set encompasses the unit sales of 3049 distinct products sold across 10 stores situated in three states within the United States: California (CA), Texas (TX), and Wisconsin (WI). The product offerings are further categorized into three broad product categories, namely Hobbies, Foods, and Household, and subsequently disaggregated into seven more specific product departments. The temporal scope of the data spans from January 29, 2011, to June 19, 2016, comprising a duration of approximately 5.4 years or 1969 d [14,22].

The second data set was provided by the Fozzy Group (*Ukrainian retail company*) and covers the unit sales of products sold in Ukraine. This data set comprises three distinct components: the time series, geographical, and SKU data sets. The time series data set chronicles the sales history of 1961 unique stock-keeping units (SKUs) that span from January 1, 2020, to July 19, 2021. The SKU data set augments this information with metadata pertaining to each individual unit, including classifications into commodity groups, categories, types, and brand affiliations. The geographical distribution of sales is encoded across

515 distinct geo clusters, allowing for a comprehensive analysis based on both location and product attributes [19].

Another crucial data set underpinning our study is the Corporación Favorita Grocery Sales Forecasting data set, which encompasses detailed unit sales records. This data set, released by Corporación Favorita, a prominent Ecuadorian company with an extensive network of supermarkets across Latin America, was made available as part of a Kaggle competition in the year 2017. The data set chronicles daily sales records for a total of 4400 unique items, across 54 different Ecuadorian stores, encompassing a temporal span from January 1, 2013, to August 15, 2017. The accompanying data files provide additional pertinent information that holds potential utility in the development of predictive models. The comprehensive nature of this data set, coupled with its geographical specificity, presents an intriguing opportunity to explore and forecast the dynamics of grocery sales within the Ecuadorian retail landscape [8,26].

4 Experiments Design

In the research of exploring various statistical and econometric methodologies for sales forecasting, we defined an experimental design to ensure consistency across data sets, models, and time series. The primary objective of our experiment's design was to establish a uniform framework, facilitated by the utilization of the Open Source Time Series Ecosystem "nixtla," which integrates the robust "StatsForecast" package with implemented time series forecasting models [11].

To achieve this uniformity, we harmonized each data set by organizing it into the following structured columns:

- `unique-id` – a distinctive identifier for each time series.
- `ds` – the date of observation.
- `y` – the sales quantity corresponding to each date and time series.

It should be noted that we opted to exclude additional metadata present in the data sets, instead focusing on establishing a consistent data structure across all scenarios. We use a list of econometric and statistical models available within "StatsForecast" (see Sect. 6). To evaluate the performance of each model, we use a set of specialized evaluation metrics (see Sect. 5).

For our experimentation, we standardized the configuration across all data sets and individual time series. Our forecast horizon was set at **14 d** and we used a rigorous cross-validation methodology. The cross-validation process involves sliding windows across historical data and predicting subsequent periods. Leveraging the distributed capabilities of "StatsForecast," we executed cross-validation with **3 windows**, ensuring a rigorous assessment of models' performance across diverse temporal contexts. The step size for each window mirrored the forecasting horizon, which, in our case, was also set at **14 d**. Refer to Fig. 1 for an illustrative representation.

Upon completion of the cross-validation step, our data frame was enriched with two additional columns: the model's predictions and the "cutoff" timestamp, denoting the final datestamp or temporal index for each window. This

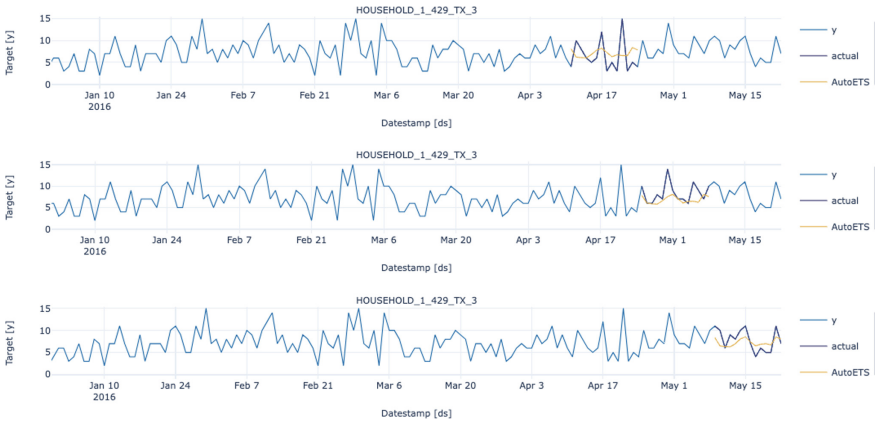

Fig. 1. Example of cross-validation

augmentation facilitated an in-depth evaluation of models and individual time series for each window. Such analysis included examining the distribution of model quality metrics in various contexts.

5 Metrics

The metrics for forecasting performance can be classified as scale-dependent and scale-independent. Scale-dependent measures, such as MSE, RMSE, and MAE, are sensitive to the scale of the data. These measures are particularly relevant when comparing forecasts made on data sets of similar scales. However, caution is warranted when comparing forecasts across series with differing scales, as scale-dependent measures can yield misleading results [3,9,18].

Scale independence has been identified as a key characteristic of effective metrics [18,21], especially when dealing with heterogeneous data sets. Consequently, our primary focus lies on the use of scale-independent metrics, namely **Mean Absolute Percentage Error** (MAPE) (Formula 1) and **Symmetric Mean Absolute Percentage Error** (sMAPE) (Formula 2). These measures are designed to provide meaningful comparisons across series with varying scales, making them suitable candidates for our evaluation.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \times 100\% \tag{1}$$

where: n – total number of observations, Y_i – actual value of the i-th observation, $\hat{Y}_i$ – predicted value of the i-th observation.

$$sMAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i - \hat{Y}_i|}{(|Y_i| + |\hat{Y}_i|)/2} \times 100\% \tag{2}$$

where variables have the same meanings as in Eq. 1.

6 Models

We use different statistical and econometrics methods that are available in the StatForecast package. In this section, we describe models from different families: baseline models, exponential smoothing models, theta models, etc.

6.1 Baseline Models

The **Historic Average method** (referred to as `HistoricAverage`), also known as the mean method, operates by setting the forecasts of all future values equal to the average of the historical data [16]. If we denote the historical data by $y_1, \ldots, y_T$, the forecasts can be expressed as Formula 3.

$$\hat{y}_{T+h|T} = \bar{y} = \frac{y_1 + \cdots + y_T}{T}. \tag{3}$$

where $\hat{y}_{T+h|T}$ represents the estimate of y_{T+h} based on the data $y_1, \ldots, y_T$.

The **Window Average method** (referred to as `WindowAverage`) calculates forecasts by taking the average of the last N observations, where N represents the length of the window [16]. The forecasts can be expressed using the Formula 4.

$$\hat{y}_{T+h|T} = \frac{y_{T-N+1} + \cdots + y_T}{N}, \tag{4}$$

where N is the window length.

The **Seasonal Window Average method** (referred to as `SeasWA`) functions similarly to the **Window Average method**, with the addition of an extra parameter: the season length. This method computes forecasts by averaging the last N observations within each seasonal period, where N corresponds to the length of the window. The inclusion of the season length parameter allows the Seasonal Window Average method to consider the specific periodicity of the time series data, making it suitable for forecasting seasonal patterns [16].

The **Naive model** (referred to as `Naive`) employs a straightforward forecasting approach, where all future forecasts are set equal to the value of the most recent observation. This uncomplicated method assumes that the future behavior of the time series will mirror its most recent data point (Formula 5). The Naive model can serve as a baseline for comparison against more sophisticated forecasting techniques [16].

$$\hat{y}_{T+h|T} = y_T. \tag{5}$$

The **Seasonal Naive model** (referred to as `SeasonalNaive`) is similar to the **Naive method** and proves especially valuable when dealing with highly seasonal data. In this approach, each forecast is set equal to the last observed value from the corresponding season of the previous year, effectively leveraging the seasonal patterns within the time series [16]. The mathematical representation of the Seasonal Naive model for forecasting at time $T + h$ is defined as:

$$\hat{y}T + h \mid T = yT + h - m(k + 1), \tag{6}$$

where m signifies the length of the seasonal period, and k corresponds to the integer part of $(h-1)/m$ - representing the number of complete years in the forecast period leading up to time $T+h$.

The **Random Walk With Drift** (referred to as `RandomWalkWithDrift`) presents a nuanced variant of the naive method, allowing forecasts to exhibit gradual increases or decreases over time. This approach incorporates a concept known as drift, which signifies the average change observed in historical data. Consequently, the forecast for time $T+h$ is expressed as:

$$\hat{y}T + h \mid T = y_T + \frac{h}{T-1}\sum t = 2^T(y_t - y_{t-1}) = y_T + h\left(\frac{y_T - y_1}{T-1}\right). \quad (7)$$

In essence, the Random Walk With Drift model is akin to drawing a line connecting the initial and final observations and extending it into the future. By capturing the average change in the historical data, this model accommodates gradual trends in the forecasts, providing a method for forecasting that maintains a level of realism [16].

6.2 Exponential Smoothing Models

The **Simple Exponential Smoothing** method (referred to as `SES`) is a forecasting approach that employs a weighted average of past observations, with weights diminishing exponentially as they extend into the past. This method is particularly useful for data that lacks discernible trends or seasonality. Given t observations, the one-step forecast is calculated as follows:

$$\hat{y}t + 1 = \alpha y_t + (1-\alpha)\hat{y}t - 1. \quad (8)$$

Here, the parameter $0 \leq \alpha \leq 1$ represents the smoothing parameter, which determines the rate at which weights decline. When $\alpha = 1$, `SES` simplifies to the naive method [13].

The **Holt method** (referred to as `Holt`) is an extension of exponential smoothing designed to accommodate time series data with trends. This method takes into account both the level and the trend of the series. By integrating an exponential smoothing factor for both the observed level and trend, the Holt method offers an effective way to forecast series characterized by varying trends [13,16].

6.3 Sparse or Intermittent Models

The **Aggregate-Dissagregate Intermittent Demand Approach** (referred to as `ADIDA`) leverages temporal aggregation to mitigate the impact of zero observations. It applies optimized Simple Exponential Smoothing (SES) at the aggregated level and then disaggregates forecasts using equal weights. ADIDA is tailored for sparse or intermittent time series with minimal non-zero data points, offering a specialized solution for improved forecasting in such challenging scenarios [24].

The **Intermittent Multiple Aggregation Prediction Algorithm** (referred to as `IMAPA`) extends the concept of `ADIDA` by incorporating multiple aggregation levels to account for diverse data dynamics. It employs optimized SES for generating forecasts across these levels and combines them through a straightforward average aggregation [2].

The **Croston method** (referred to as `CrostonClassic`) is a technique for forecasting time series characterized by intermittent demand. It involves decomposing the original time series into two components: non-zero demand sizes denoted as z_t, and inter-demand intervals denoted as p_t [4]. The forecast is then given by:

$$\hat{y}_t = \frac{\hat{z}_t}{\hat{p}_t}. \tag{9}$$

where $\hat{z}_t$ and $\hat{p}_t$ are forecasted using SES. Both components are smoothed with a common smoothing parameter of 0.1.

The **Croston SBA method** (referred to as `CrostonSBA`) is short for Croston Seasonal Exponential Smoothing with Backshift Adjustment. This approach merges the Croston method with exponential smoothing, enhancing forecast precision for time series data showcasing both trend and seasonality [4].

The **Teunter-Syntetos-Babai method** (referred to as `TSB`) is akin to the Croston method but uses demand probabilities d_t instead of demand intervals to estimate demand sizes. It is particularly suitable for time series data characterized by extended periods of zero demand. Demand probabilities are defined as:

$$d_t = \begin{cases} 1 & \text{if demand occurs at time } t \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Consequently, the forecast is expressed as:

$$\hat{y}_t = \hat{d}_t \hat{z}_t$$

Both $\hat{d}_t$ and $\hat{z}_t$ are estimated using SES. The smoothing parameters for each component may differ, similar to the optimized Croston's method [25].

6.4 Multiple Seasonalities

Multiple Seasonal-Trend decomposition using LOESS (referred to as `MSTL`) designed for time series with multiple seasonal cycles, MSTL is an automated algorithm that extends the STL decomposition. It iteratively estimates multiple seasonal components, controlling their smoothness and separating variations. For non-seasonal series, MSTL determines trend and remainder components [1].

6.5 Theta Family

The **Theta method** (referred to as `Theta`) is utilized for non-seasonal or deseasonalized time series, often achieved through multiplicative classical decomposition. This approach transforms the original time series into two new lines using

theta coefficients, maintaining the same mean and slope but adjusting local curvatures based on the coefficient's value [10].

6.6 ARCH Family

The **Autoregressive Conditional Heteroskedasticity model** (referred to as `ARCH`) is a statistical model used in time series analysis to characterize the variance of the current innovation based on past error terms' magnitudes, often considering their squares [6,23]. It assumes that at time t, y_t is expressed as:

$$y_t = \epsilon_t \sigma_t \tag{11}$$

where ϵ_t is a sequence of random variables with zero mean and unit variance, and σ_t^2 is defined by:

$$\sigma_t^2 = w_0 + \sum_{i=1}^{p} a_i y_{t-i}^2 \tag{12}$$

Here, w and a_i, $i = 1, ..., p$, are coefficients that must satisfy nonnegativity conditions, and $\sum_{k=1}^{p} a_k < 1$.

6.7 ARIMA Family

The **Autoregressive model** (referred to as `AutoRegressive`) is a fundamental time series model that predicts future values based on their linear dependence on past observations [17].

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t \tag{13}$$

where ϕ_i are the coefficients, c is a constant term, and ϵ_t represents the white noise error term.

The **Autoregressive Integrated Moving Average model** (ARIMA) is a widely used time series forecasting method that combines autoregressive (AR), differencing (I), and moving average (MA) components. It captures the dependencies between past observations, and differences to make the series stationary, and incorporates the influence of past forecast errors. The model is denoted as ARIMA(p, d, q), where p, d, and q represent the orders of the AR, differencing, and MA components respectively [17].

The **AutoARIMA model** employs an automated approach to select the optimal ARIMA model based on the Akaike Information Criterion (AICc), a well-known information criterion [17].

7 Results

The error analysis of forecast accuracy for three different data sets and the list of methods using MAPE and sMAPE metrics revealed interesting empirical insights. For the M5 data set, the mean MAPE and sMAPE were 31.48 and 136.83, respectively, showcasing a moderate overall accuracy (see Table 2). Contrasting this, the Fozzy Group data set revealed a notable degree of variability. It displayed a mean MAPE of 8.65, while the sMAPE mean was considerably higher at 157.75. Remarkably, the Fozzy Group data set exhibited substantial standard deviations in both metrics, hinting at the potential presence of outliers or significant variations (see Table 3). Turning to the Corporación Favorita data set, a different pattern emerged. The mean MAPE registered at 191.39, while the sMAPE mean was relatively lower at 59.62 (see Table 4). These findings underscore the diversified performance of forecasting methodologies across distinct data sets, offering valuable insights for optimizing parameters and model selection.

Table 2. Statistical Summary of Forecasting Performance on M5 Data Set

Statistic	Value (MAPE)	Value (sMAPE)
Number of time-series	1737930	1737930
Mean	31.48	136.83
Standard deviation	47.48	52.85
Minimum	0.00	0.00
Q_1 (25%)	12.76	94.32
Q_2 (50%)	23.34	147.55
Q_3 (75%)	39.06	185.88
Maximum	44204.92	200.00

Table 3. Statistical Summary of Forecasting Performance on Fozzy Group Data Set

Statistic	Value (MAPE)	Value (sMAPE)
Number of time-series	1656504	1656504
Mean	8.65	157.75
Standard deviation	3020.11	75.57
Minimum	0.00	0.00
Q_1 (25%)	0.00	170.34
Q_2 (50%)	0.00	200.00
Q_3 (75%)	6.54	200.00
Maximum	3263154	200.00

The comparative summary statistics for the M5 data set (see Table 5) show that models such as `AutoRegressive`, `CrostonSBA`, and `CrostonClassic` exhibit

Table 4. Statistical Summary of Forecasting Performance on Corporación Favorita Data Set

Statistic	Value (MAPE)	Value (sMAPE)
Number of time-series	2910458	2910458
Mean	191.39	59.62
Standard deviation	147333.13	30.23
Minimum	0.00	0.00
Q_1 (25%)	49.00	42.29
Q_2 (50%)	71.79	52.85
Q_3 (75%)	113.93	66.79
Maximum	251348620	200.00

similar mean MAPE values (*ranging from 25.98 to 26.48*), indicating consistent predictive accuracy with relatively moderate variability. However, the `ARCH` model deviates with a substantially higher mean MAPE of 67.99, accompanied by a notable standard deviation of 158.8, suggesting less stable performance. The quartiles (Q_1, Q_2, Q_3) further illustrate the spread of performance across the models, showcasing variations in accuracy levels.

Table 5. Comparative Summary of Forecasting Model Performance on M5 Data Set

Model	Mean	std	Min	Q_1 (25%)	Q_2 (50%)	Q_3 (75%)	Max
AutoRegressive	25.98	23.27	0.0	12.24	21.25	33.1	622.8
CrostonSBA	26.18	22.46	0.0	12.58	21.8	33.33	580.25
CrostonClassic	26.48	23.87	0.0	12.32	21.32	33.31	612.29
ADIDA	26.77	25.24	0.0	12.44	21.39	33.18	826.81
WindowAverage	26.77	24.35	0.0	12.24	21.43	34.25	771.94
IMAPA	26.85	25.46	0.0	12.37	21.36	33.37	826.81
Holt	26.87	25.65	0.0	12.44	21.37	33.29	829.04
HistoricAverage	27.01	25.68	0.0	12.23	22.11	34.7	654.45
Theta	27.01	25.97	0.0	12.23	21.37	33.8	842.8
ARIMA	27.01	25.68	0.0	12.23	22.11	34.7	654.47
SeasWA	27.35	22.97	0.0	13.27	23.31	35.8	697.37
AutoARIMA	27.36	23.87	0.0	12.95	22.35	35.11	669.82
TSB	27.46	26.48	0.0	11.94	21.43	34.93	816.72
SES	29.25	29.53	0.0	11.97	22.32	37.19	1004.96
MSTL	32.45	28.26	0.0	14.31	27.04	43.11	782.8
SeasonalNaive	38.06	33.1	0.0	14.29	30.95	52.38	766.19
Naive	40.53	50.26	0.0	11.9	28.57	50.0	1953.15
RWD	40.77	50.44	0.0	11.9	28.73	50.97	1961.34
ARCH	67.99	158.8	0.0	21.88	54.77	103.62	44204.92

Analyzing the results per model for the Fozzy Group data set (see Table 6) offers valuable insights. Notably, a lot of models like `SeasWA`, `CrostonClassic`, and `CrostonSBA` exhibit minimum MAPE values of 0.0%. This suggests instances of near-perfect forecasts, which can be attributed to the prevalence of zero sales in the time series of the Fozzy Group data set. For example, the model `AutoRegressive` shows a Q_1 MAPE of 0.0%, indicating that at least 25% of the time series have forecasts with no error. However, the same model has a Q_3 MAPE of 6.73%, signifying that 75% of the time series have forecasts with a MAPE of 6.73% or lower. This comprehensive evaluation proves the challenges caused by zero sales in forecasting accuracy and highlights the varying degrees of predictive capability across different models. These findings can aid in selecting appropriate forecasting techniques tailored to scenarios with a significant presence of zero values in time series data.

Table 6. Comparative Summary of Forecasting Model Performance on Fozzy Group Data Set

Model	Mean	std	Min	Q_1 (25%)	Q_2 (50%)	Q_3 (75%)	Max
SeasWA	3.91	11.85	0.0	0.0	0.0	0.00	725.0
CrostonClassic	4.45	10.93	0.0	0.0	0.0	5.93	372.4
CrostonSBA	4.46	10.70	0.0	0.0	0.0	6.00	351.6
IMAPA	4.61	11.05	0.0	0.0	0.0	6.53	377.8
ADIDA	4.61	10.99	0.0	0.0	0.0	6.51	365.9
Theta	4.65	11.32	0.0	0.0	0.0	6.57	385.6
Holt	4.65	11.75	0.0	0.0	0.0	6.50	638.3
TSB	4.69	11.55	0.0	0.0	0.0	6.66	441.1
WindowAverage	4.70	11.51	0.0	0.0	0.0	6.63	416.7
HistoricAverage	4.74	10.75	0.0	0.0	0.0	6.47	463.1
ARIMA	4.74	10.75	0.0	0.0	0.0	6.47	463.1
SES	4.84	12.24	0.0	0.0	0.0	6.88	507.6
MSTL	5.22	13.26	0.0	0.0	0.0	6.82	620.2
Naive	5.92	17.72	0.0	0.0	0.0	7.14	1020.9
SeasonalNaive	5.99	15.54	0.0	0.0	0.0	7.14	540.4
RWD	6.12	18.09	0.0	0.0	0.0	7.24	1045.1
ARCH(1)	8.27	63.24	0.0	0.0	0.0	7.23	17909.4
AutoRegressive	69.23	12813	0.0	0.0	0.0	6.73	3263154

In another context, the comparative summary statistics for the Corporación Favorita data set (see Table 7) reveal that models such as `CrostonSBA`, `AutoARIMA`, and `SeasWA` exhibit mean MAPE values in the range of 87.08 to 89.02. This range demonstrates consistent accuracy with relatively moderate

variability. Conversely, models like `ARCH` and `AutoRegressive` present considerably higher mean MAPE values of 216.93 and 1751.39, respectively. These models are accompanied by notable standard deviations, indicating more diverse and potentially less reliable performance. It's noteworthy that `CrostonSBA`, `AutoARIMA`, and `SeasWA` emerge as the most stable models in the context of the maximum MAPE, showcasing their ability to consistently produce accurate forecasts.

Table 7. Comparative Summary of Forecasting Model Performance on Corporación Favorita Data Set

Model	Mean	std	Min	Q_1 (25%)	Q_2 (50%)	Q_3 (75%)	Max
CrostonSBA	87.08	190.11	0.07	45.13	62.69	92.28	37194.1
AutoARIMA	88.94	212.74	0.00	46.93	64.72	93.81	34626.7
SeasWA	89.02	115.02	0.00	45.32	70.12	106.70	8156.2
WindowAverage	92.78	240.42	0.00	46.43	65.02	96.52	53182.3
CrostonClassic	92.82	200.46	0.02	47.63	66.85	98.76	39154.7
IMAPA	93.45	237.63	0.00	47.34	66.49	98.44	39154.7
ADIDA	93.45	237.63	0.00	47.34	66.49	98.44	39154.7
TSB	93.65	222.70	0.00	46.57	65.74	99.54	35972.9
Theta	93.79	264.91	0.07	47.27	66.31	98.15	44498.9
SES	95.27	232.87	0.00	45.95	65.62	101.96	34841.9
Holt	98.88	352.02	0.04	48.51	68.23	100.97	76787.1
SeasonalNaive	105.70	243.48	0.00	53.27	77.13	113.69	28471.3
ARIMA	107.78	139.49	3.52	54.54	79.88	122.80	19461.0
HistoricAverage	107.78	139.49	3.52	54.54	79.88	122.80	19461.0
Naive	108.47	317.03	0.00	40.83	63.81	116.83	64910.0
MSTL	109.15	301.54	3.55	54.41	76.73	112.23	54314.6
RWD	110.08	337.87	0.00	42.01	64.27	116.96	71091.0
ARCH(1)	216.93	471.62	89.84	128.83	155.20	203.02	40463.2
AutoRegressive	1751.39	642209.24	0.64	52.83	74.32	109.39	251348620

In conclusion, our analysis identifies CrostonSBA, CrostonClassic, ADIDA, and IMAPA as the most stable forecasting methods across diverse data sets. These models consistently exhibit accurate predictive capabilities, suggesting their suitability as reliable baseline models for guiding future research endeavors in the field of sales forecasting.

8 Discussion and Future Work

In this section, we delve into the insights garnered from the analysis of different forecasting models across multiple datasets, while also addressing the limitations and potential avenues for future research in the realm of sales forecasting.

Our utilization of statistical and econometrics methods has led to a heightened level of interpretability compared to machine learning or neural network approaches [20]. However, it is noteworthy that certain models, such as

AutoARIMA, exhibit complexities in interpretation, marking a limitation of our study.

Our findings have illuminated the varying performance of forecasting methods across different datasets. Notably, specific models demonstrate superior performance on distinct datasets, and the stability of certain models across varied datasets, like CrostonSBA and CrostonClassic, underscores their potential as baseline models for future research endeavors. However, we acknowledge that the task of identifying universally stable methods remains intricate.

The evaluation of time series forecasting remains a challenging endeavor, and we recognize room for enhancement in the evaluation process. Utilizing real-time data in our research can augment the practical applicability of our findings, offering a more comprehensive understanding of forecasting performance.

The limitations of our study encompass the nature of the dataset used. A future direction could involve incorporating live sales data and accounting for a broader spectrum of factors, both internal and external, that could significantly influence sales patterns.

Further investigations into outliers in model performance are warranted to deepen our understanding of their effects. Additionally, the necessity of periodic reevaluation and updating of forecasting models in light of new data cannot be understated, ensuring their ongoing efficacy.

In practical implementation, our study presents the notion of employing the list of stable models as a resource for selecting baseline models. This proposition holds value for both researchers embarking on forecasting research and businesses initiating their sales forecasting processes. These baseline models offer a stepping stone for evaluation and improvement, acting as empirical anchors in the dynamic field of sales forecasting.

9 Conclusion

In conclusion, this study has provided a comprehensive exploration of the current state of the sales forecasting domain through empirical analysis of different statistical and econometrics methods across diverse datasets. The findings underscore the importance of stable and interpretable methods, with models such as CrostonSBA, CrostonClassic, ADIDA, and IMAPA identified as particularly robust choices for baseline sales forecasting.

The inherent stability and interpretability of these methods hold great significance for the decision-making processes within the retail industry, offering reliable insights for informed strategies. While this study contributes valuable insights, it is imperative to acknowledge its limitations and avenues for future research, as discussed in Sect. 8.

By conducting this comparative analysis, the study sheds light on the strengths, weaknesses, and stability of different forecasting models, enriching our understanding of their applicability and potential. As the field of sales forecasting continues to evolve, we encourage readers to delve into the evolving landscape of forecasting methodologies, building upon the insights provided by this research.

References

1. Bandara, K., Hyndman, R.J., Bergmeir, C.: MSTL: a seasonal-trend decomposition algorithm for time series with multiple seasonal patterns. arXiv. 2021. https://doi.org/10.48550/ARXIV.2107.13462. https://arxiv.org/abs/2107.13462
2. Boylan, J.E., Syntetos, A.A.: Intermittent Demand Forecasting: Context, Methods and Applications. Wiley (2021). https://www.ifors.org/intermittent-demand-forecasting-context-methodsand-applications/
3. Chatfield, C.: Apples, oranges and mean square error. Int. J. Forecasting **4**(4), 515–518 (1988). https://doi.org/10.1016/0169-2070(88)90127-6
4. Croston, J.D.: Forecasting and stock control for intermittent demands. J. Oper. Res. Soc. **23**(3), 289–303 (1972). https://doi.org/10.1057/jors.1972.50
5. Davis, D.F., Mentzer, J.T.: Organizational factors in sales forecasting management. Int. J. Forecast. **23**(3), 475–495 (2007). https://doi.org/10.1016/j.ijforecast.2007.02.005
6. Engle, R.F.: Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica **50**(4), 987–1007 (1982). https://doi.org/10.2307/1912773
7. Ensafi, Y., et al.: Time-series forecasting of seasonal items sales using machine learning - a comparative analysis. Int. J. Inf. Manage. Data Insights **2**(1), 100058 (2022). https://doi.org/10.1016/j.jjimei.2022.100058
8. Favorita, C., et al.: Corporación Favorita Grocery Sales Forecasting (2017). https://kaggle.com/competitions/favorita-grocerysales-forecasting
9. Fildes, R., Makridakis, S.: Forecasting and loss functions. Int. J. Forecast. **4**(4), 545–550 (1988). https://doi.org/10.1016/0169-2070(88)90131-8
10. Fiorucci, J.A., et al.: Models for optimising the theta method and their relationship to state space models. Int. J. Forecast. **32**(4), 1151–1161 (2016). https://doi.org/10.1016/j.ijforecast.2016.02.005
11. Garza, F., et al.: StatsForecast: lightning fast forecasting with statistical and econometric models. PyCon Salt Lake City, Utah, US 2022 (2022). https://github.com/Nixtla/statsforecast
12. Haselbeck, F., et al.: Machine learning outperforms classical forecasting on horticultural sales predictions. Mach. Learn. Appl. **7**, 100239 (2022). https://doi.org/10.1016/j.mlwa.2021.100239
13. Holt, C.C.: Forecasting seasonals and trends by exponentially weighted moving averages. Int. J. Forecast. 20(1), 5–10 (2004). https://doi.org/10.1016/j.ijforecast.2003.09.015
14. Addison Howard et al. M5 Forecasting - Accuracy (2020). https://kaggle.com/competitions/m5-forecasting-accuracy
15. Hyndman, R.J.: A brief history of forecasting competitions. Int. J. Forecast. **36**(1), 7–14 (2020). https://doi.org/10.1016/j.ijforecast.2019.03.015
16. Hyndman, R.J., Athanasopoulos, G.: Forecasting: Principles and Practice. 3rd. Melbourne, Australia: OTexts (2021). https://OTexts.com/fpp3
17. Hyndman, R.J., Khandakar, Y.: Automatic time series forecasting: the forecast package for R. J. Stat. Softw. **27**(3), 1–22 (2008). https://doi.org/10.18637/jss.v027.i03
18. Kim, S., Kim, H.: A new metric of absolute percentage error for intermittent demand forecasts. Int. J. Forecast. **32**(3), 669–679 (2016). https://doi.org/10.1016/j.ijforecast.2015.12.003

19. Kosovan, O.: Fozzy group hack4retail competition overview: results, findings, and conclusions. Market Infrastruct. 67 (2022). https://doi.org/10.32843/infrastruct67-42
20. Kosovan, O., Datsko, M.: Interpretation of machine learning algorithms for decision-making in retail. Econ. Soc. **47** (2023). https://doi.org/10.32782/2524-0072/2023-47-47
21. Makridakis, S.: Accuracy measures: theoretical and practical concerns. Int. J. Forecast. **9**(4), 527– 529 (1993). https://doi.org/10.1016/0169-2070(93)90079-3
22. Makridakis, S., Spiliotis, E., Assimakopoulos, V.: The M5 competition: background, organization, and implementation. Int. J. Forecast. **38**(4), 1325–1336 (2022). https://doi.org/10.1016/j.ijforecast.2021.07.007
23. Marquez, J.: Time Series Analysis. James D. Hamilton, 1994, (Princeton University Press, Princeton, NJ), 799 pp., US$55.00, ISBN 0-691-04289-6". In: International Journal of Forecasting 11.3 (1995), pp. 494–495. https://ideas.repec.org/a/eee/intfor/v11y1995i3p494-495.html
24. Nikolopoulos, K., et al.: An aggregate-disaggregate intermittent demand approach (ADIDA) to forecasting: an empirical proposition and analysis. J. Oper. Res. Soc. **62**(3), 544–554 (2011). https://doi.org/10.1057/jors.2010.32
25. Teunter, R.H., Syntetos, A.A., Zied Babai, M.: Intermittent demand: linking forecasting to inventory obsolescence. Eur. J. Oper. Res. **214**(3), 606–615 (2011). https://doi.org/10.1016/j.ejor.2011.05.018
26. Vallés-Pérez, I., et al.: Approaching sales forecasting using recurrent neural networks and transformers. Exp. Syst. Appl. **201**, 116993 (2022). https://doi.org/10.1016/j.eswa.2022.116993