International Journal of Innovative Research and Scientific Studies, 8(6) 2025, pages: 2883-2892

IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

Data-driven forecasting of sales influenced by climate variability using deep learning

Siriwan Kajornkasirat[1*], Chayanin Limrattanabunchong[2], Nattaseth Sriklin[3]

[1,2,3]*Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani Campus, Surat Thani, Thailand.*

Corresponding author: Siriwan Kajornkasirat (*Email: siriwan.wo@psu.ac.th*)

Abstract

This study investigates the statistical and machine learning models to improve sales forecasting accuracy in the retail sector by incorporating both transactional and environmental variables. Motivated by the limitations of traditional data handling systems within a case study company, the research proposes an integrated, centralized information system that enhances data accessibility, reduces redundancy, and supports timely decision-making. Multiple forecasting approaches—including SARIMA, SARIMAX, LSTM, Ordinary Least Squares (OLS), and Poisson regression—were evaluated using historical sales and weather data. Results indicate that regression-based models (OLS and Poisson) outperformed time series and deep learning models in terms of model fit and predictive power, emphasizing the effectiveness of simpler, interpretable methods when relevant features are included. The study also demonstrates that weather conditions, such as humidity and temperature, exhibit moderate correlations with sales volume, though their direct predictive contribution is limited when used in isolation. This data-driven framework offers a scalable solution for retail operations, contributing to cost reduction—such as minimizing reliance on third-party business intelligence tools—and promoting sustainable competitive advantage.

Keywords: Data-driven, Deep learning, Retail analytics, Sales forecasting, Weather data.

DOI: 10.53894/ijirss.v8i6.10223
Funding: The work was supported by Prince of Songkla University, Surat Thani Campus and Graduate School, Prince of Songkla University, Thailand.
History: Received: 8 August 2025 / **Revised:** 10 September 2025 / **Accepted:** 12 September 2025 / **Published:** 25 September 2025

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Acknowledgements: The authors thank Seppo J. Karrila for constructive comments on this manuscript. This research was funded by the Prince of Songkla University, Surat Thani Campus.
Publisher: Innovative Research Publishing

1. Introduction

Nowadays, organizations face significant challenges in dealing with scattered and unorganized data, which hampers effective decision-making and can lead to inaccurate analysis. Data stored across multiple, unconnected sources

complicates the retrieval of crucial information and increases the likelihood of errors that negatively impact organizational performance. Additionally, the absence of a centralized data management system leads to increased time spent on manual processes, resulting in inefficiencies and missed opportunities for timely actions [1].

Machine Learning (ML) offers powerful solutions to these challenges by managing data from various sources, and performing tasks such as cleaning, transforming, and integrating the data to make it analysis-ready. This significantly reduces human error and saves time previously spent on manual processes. ML is widely adopted in marketing, especially for forecasting, due to its ability to process large volumes of data and analyze complex patterns more quickly and accurately than traditional methods. It enables reliable forecasting of trends such as consumer behavior, sales, or market demand. Moreover, ML systems can continuously learn from new data, leading to improved accuracy over time [2].

There is also a growing interest in applying predictive analytics (PA) within supply chain management (SCM), particularly in retail. Given the volatile nature of consumer demand, intense market competition, and the complexity of global supply networks, data-driven decision-making is becoming essential. Predictive analytics, especially in demand forecasting, has shown great potential in enhancing retail SCM operations. Accurate and sophisticated demand forecasting enables better decisions in inventory control, purchasing, and assortment planning, ensuring product availability at the point of sale [3].

Data visualization further enhances the understanding of business data by presenting it in a visual or graphical format, as illustrated in Figure 1. This approach facilitates the identification of patterns, correlations, and outliers, enabling quicker and more effective decision-making. Visualization helps reveal emerging market trends and business dynamics while also clarifying relationships between key parameters that influence organizational goals [4, 5].

The integration of Business Intelligence (BI) systems enables the transformation of raw data into actionable insights. These systems help executives understand trends and customer behavior, such as purchasing patterns and market demand, more clearly. BI supports strategic decision-making with accurate data, reduces operational risks, and improves resource management efficiency. With real-time access to information through various technologies and tools, businesses can respond quickly and confidently to changing circumstances [6].

Today, data is often organized and presented visually using charts and graphs. Dashboards, in particular, have become a popular tool for real-time data visualization. These dashboards are developed using software platforms and tools such as marketing analytics systems, data management platforms, and website monitoring tools, as shown in Figure 2 [7, 8]. By integrating these technologies and strategies, companies can effectively manage complex environments, improve decision-making, and drive sustainable growth. Additionally, reducing operational costs, such as those associated with Power BI subscriptions, further enhances financial efficiency and competitiveness.

In this study, we explore the relationships between variables affecting sales through correlation analysis. We compare the performance of three forecasting models: Seasonal Autoregressive Integrated Moving Average (SARIMA), SARIMA with eXogenous variables (SARIMAX), and Long Short-Term Memory (LSTM), in order to identify the most suitable model for sales prediction. The results contribute to the development of a dynamic dashboard that supports data-driven decision-making for the company.

The remainder of this paper is organized as follows: Section 2 details the research methodology. Section 3 presents the results. Section 4 discusses the results obtained. Finally, Section 5 concludes the study and outlines directions for future research.

2. Materials and Methods

2.1. Dataset

The dataset utilized in this study was derived from the daily sales records of a retail company based in Surat Thani, Thailand. It specifically focuses on clothing items sold via a point-of-sale (POS) system, covering the period from October 1, 2020, to September 30, 2024. This dataset provides essential time-series information that is critical for analyzing sales performance and forecasting. The data include as variables daily sales figures, transaction dates, product categories, and revenue. Additionally, environmental factors such as temperature, humidity, dew point, and wind speed are incorporated to assess their influences on consumer behavior. Metrics related to latex purchasing and pricing are also included, offering insight into broader economic conditions that may indirectly affect sales. The dataset, stored in .csv format, serves as a valuable resource for time-series forecasting and supports strategic decision-making through detailed historical analysis (Figure 1).

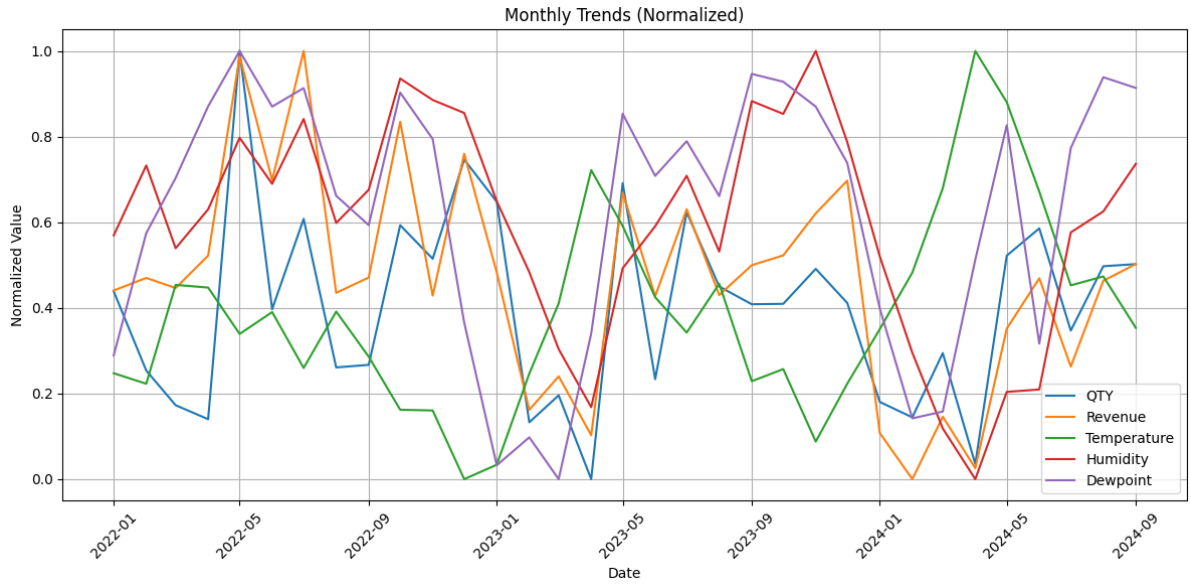

Figure 1.
Monthly Trends.

2.2. Data Preparation and Cleaning

Before conducting any analysis, the dataset underwent rigorous preparation and cleaning to ensure its quality and reliability. This process involved handling missing data, correcting entry errors, and standardizing the format of all variables. Outliers were identified and appropriately treated to minimize distortion in the results. Temporal data, in particular, was uniformly formatted to enable accurate time-series modeling. These preprocessing steps are critical to ensure that subsequent analyses and forecasts are both reliable and valid [1, 9].

2.3. Data Analysis

To derive actionable insights and construct accurate forecasting models, a structured data analysis pipeline was implemented. This process comprised three main stages: (1) correlation analysis to identify significant relationships between environmental and economic variables and sales performance; (2) forecasting models, where multiple modeling paradigms statistical, time-series, and deep learning were applied and tested under varying configurations of explanatory variables; and (3) model evaluation, which assessed the predictive accuracy and robustness of each model using standardized error metrics. The integration of these stages provided a comprehensive understanding of how external factors interact with consumer purchasing behavior over time and enabled the identification of optimal modeling strategies for sales prediction in retail environments.

2.3.1. Correlation Analysis

To thoroughly examine the impact of external factors on sales revenue, this study employed the Spearman rank correlation coefficient. As a non-parametric statistical method, Spearman's correlation is particularly suitable for time-series datasets that may not conform to the assumptions of normality, homoscedasticity, or linearity. Unlike the Pearson correlation, which is sensitive to outliers and requires a linear relationship between variables, Spearman's method assesses monotonic associations and remains robust under non-linear conditions, making it appropriate for real-world commercial data typically encountered in retail contexts.

In order to capture the temporal nature of consumer behavior, the analysis incorporated lagged effects of independent variables over a period of up to two months. This approach allowed for the examination of not only concurrent relationships (lag 0) but also the delayed influence of external variables at one-month (lag 1) and two-month (lag 2) spans. In retail environments, consumer responses to external stimuli such as shifts in weather patterns or commodity prices often manifest with a time delay. Therefore, investigating these lagged correlations provides a more realistic understanding of the causal dynamics that drive revenue fluctuations.

The variables selected for analysis were chosen based on both theoretical relevance and the consistency of their historical time-series records. These included daily average relative humidity (humidity_avg), expressed as a percentage; daily average ambient temperature (temperature_avg), measured in degrees Celsius; quantity of items sold (qty), expressed in units and representing direct sales volume; the average monthly latex price under the Free on Board (FOB) benchmark (latexPriceFOB), measured in USD per kilogram, serving as a proxy for macroeconomic or market conditions; and daily average dew point temperature (dewpoint_avg), in degrees Celsius, which reflects atmospheric moisture content and is known to influence consumer comfort and behavior.

Each variable was evaluated across the three lag periods by comparing its values to daily revenue, enabling the identification of both immediate and delayed correlations. The statistical significance of these relationships was assessed using p-values, with results interpreted according to standard significance thresholds: cases with $p < 0.001$ were considered highly significant**,** those with $p < 0.01$ were considered moderately significant, *and those with $p < 0.05$ were regarded as*

marginally significant. This rigorous examination provided empirical evidence of the strength and timing of relationships between external factors and revenue performance.

The findings from this correlation analysis served as a critical foundation for variable selection in the subsequent development of forecasting models. By identifying variables with statistically significant associations, whether immediate or delayed, the study ensured that only the most relevant, time-sensitive predictors were incorporated into predictive models. This strategy contributed not only to the improvement of forecast accuracy but also to the interpretability of the results, thereby enhancing the practical utility of the models for decision-making in retail operations.

2.3.2. Forecasting Models

To evaluate the impact of external environmental conditions on sales, a suite of forecasting models was specified using a combination of time series analysis, statistical regression, and deep learning methodologies. Each model was trained under varying configurations of explanatory variables and tested on historical monthly data. Meteorological variables (humidity, dewpoint, temperature) were sourced from publicly available climate records, while sales quantity and revenue data were extracted from internal business systems.

A total of seventeen model configurations were systematically developed across five methodological categories: SARIMA, SARIMAX, Long Short-Term Memory (LSTM), Ordinary Least Squares (OLS), and Poisson Regression.

2.3.2.1. SARIMA

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model served as a univariate baseline, relying solely on historical sales data without any external predictors. It is denoted as SARIMA(p,d,q)(P,D,Q)s, where parameters were selected based on AIC minimization.

2.3.2.2. SARIMAX

SARIMAX extends SARIMA by incorporating exogenous variables. In this study, variables were cumulatively added in the following order: humidity, dewpoint, temperature, and sales quantity (qty).

To account for temporal dependence, a first-order autoregressive term (AR(1)) was included via the model's internal AR structure:

$$Y_t = \emptyset_1 Y_{t-1} + \theta X_t + \varepsilon_t$$

where Y_{t-1} is the previous month's revenue, and X_t represents external variables.

2.3.2.3. LSTM

Long Short-Term Memory (LSTM) neural networks were implemented to capture non-linear and long-term dependencies in time-series data. LSTM models were tested with input combinations aligned with the SARIMAX models. First-order temporal information was naturally handled by the recurrent architecture of the network, which maintains memory of prior sequences.

2.3.2.4. Ordinary Least Squares (OLS)

OLS regression was applied to estimate the linear relationship between sales and the selected predictors. The basic model is specified as:

$$Y_t = \beta_0 + \beta_1 X_{\text{humidity}} + \beta_2 X_{\text{temperature}} + \beta_3 X_{\text{latexPriceFOB}} + \beta_4 X_{\text{dewpoint}} + \beta_5 X_{\text{qty}} + \varepsilon_t$$

A first-order OLS model was also included, incorporating the previous period's revenue as an autoregressive term:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 X_{\text{humidity}} + \beta_3 X_{\text{temperature}} + \beta_4 X_{\text{latexPriceFOB}} + \beta_5 X_{\text{dewpoint}} + \beta_6 X_{\text{qty}} + \varepsilon_t$$

This structure helps capture autocorrelation effects in daily or monthly sales.

2.3.2.5. Poisson Regression

Poisson regression was used due to the count-like nature of the sales quantity variable (qty). The log-linear model form is:

$$\ln(Y_t) = \beta_0 + \beta_1 X_{\text{humidity}} + \beta_2 X_{\text{temperature}} + \beta_3 X_{\text{latexPriceFOB}} + \beta_4 X_{\text{dewpoint}} + \beta_5 X_{\text{qty}}$$

The first-order Poisson regression model adjusted for autocorrelation is specified as:

$$\ln(Y_t) = \beta_0 + \beta_1 Y_{t-1} + \beta_2 X_{\text{humidity}} + \beta_3 X_{\text{temperature}} + \beta_4 X_{\text{latexPriceFOB}} + \beta_5 X_{\text{dewpoint}} + \beta_6 X_{\text{qty}}$$

where Y_{t-1} is the lagged sales quantity.

This model configuration allows for estimating the immediate and cumulative effects of predictors while accounting for the influence of prior sales.

Here, the variables X_{humidity}, $X_{\text{temperature}}$, $X_{\text{latexPriceFOB}}$, X_{dewpoint} and X_{qty} stand for Humidity, temperature, latex price under the FOB benchmark, dewpoint, and sales quantity, respectively.

2.3.3. Model Evaluation

In order to select the most appropriate forecasting model for predicting sales influenced by external environmental factors, a systematic model evaluation procedure was conducted. The study adopted a structured approach beginning with model specification, followed by training, validation, and performance comparison using well-established statistical metrics. The aim was to assess the predictive accuracy and explanatory power of various model types under different configurations of independent variables.

The evaluation involved four key metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). These metrics were selected for their capacity to capture different aspects of model performance.

- MAE represents the average of absolute differences between actual and predicted values, offering an intuitive measure of prediction error.
- MSE calculates the mean of squared differences, penalizing larger errors more heavily.
- RMSE, the square root of MSE, retains the original unit of measurement and provides a more interpretable metric.
- R^2 indicates the proportion of variance in the dependent variable explained by the model. A value of 1 represents perfect prediction, while values closer to 0 or negative suggest poor model performance [10].

The evaluation metrics are defined as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (1)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (2)$$

$$RMSE = \sqrt{MSE} \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (4)$$

Where:

n is the number of samples
y_i is the actual value
$\hat{y}_i$ is the predicted value
$\bar{y}$ is the average of actual values

3. Results

3.1. Correlation Analysis

The Spearman correlation analysis between environmental and sales variables, incorporating lagged effects, is summarized in Table 1. The results indicate that average daily relative humidity (humidity) exhibits a strong and highly significant positive correlation with sales revenue at lag 0 (r = 0.761, $p < 0.001$). This positive association persists at a moderate level with a one-month lag (lag 1; r = 0.444, $p < 0.05$), but diminishes and becomes non-significant by the two-month lag (lag 2; r = 0.200). These findings suggest that humidity exerts an immediate and short-term delayed influence on revenue performance.

Table 1.
Spearman correlation analysis of environmental and sales variables with lagged effects.

Variable	Lag 0	Lag 1	Lag 2
Humidity	0.761***	0.444*	0.200
Temperature	-0.530**	-0.124	0.106
qty	0.684***	0.032	0.103
latexPriceFOB	-0.161	-0.161	-0.011
dewpoint	0.616***	0.557***	0.474**

Note: Spearman Test: Spearman's correlation coefficient; Lag: time lag (months). *, **, *** Significant at the 0.05, 0.01, and 0.001 level (two-tailed).

Average daily temperature (temperature) shows a significant negative correlation with revenue at lag 0 (r = -0.530, $p < 0.01$), indicating that higher temperatures are associated with lower sales on the same day. However, the correlation weakens and becomes statistically insignificant at lag 1 and lag 2, implying that the temperature's impact on sales is mostly immediate rather than delayed.

The quantity of items sold (qty) demonstrates a strong and highly significant positive correlation with revenue at lag 0 ($r = 0.684$, $p < 0.001$), reflecting the direct effect of sales volume on revenue. No significant associations are observed at lag 1 and lag 2, which aligns with the expectation that sales quantity is contemporaneously linked to revenue.

The average monthly latex price under the FOB benchmark (latexPriceFOB) exhibits weak negative correlations across all lag periods, none of which reach statistical significance (lag 0: $r = -0.161$; lag 1: $r = -0.161$; lag 2: $r = -0.011$). This suggests that fluctuations in latex price have limited direct influence on daily sales revenue within the examined timeframe.

Lastly, the average daily dew point temperature (dewpoint) shows a strong and highly significant positive correlation with revenue at lag 0 ($r = 0.616$, $p < 0.001$) and lag 1 ($r = 0.557$, $p < 0.001$), with a moderate yet significant correlation persisting at lag 2 ($r = 0.474$, $p < 0.01$). This indicates that dewpoint has both immediate and sustained effects on sales performance, possibly reflecting its impact on consumer comfort and behavior.

To supplement the statistical findings, a correlation heat map was constructed and is presented as Figure 2, visualizing the strength and direction of relationships between sales revenue and relevant variables. The analysis is based on aggregated daily sales data from a retail store spanning the years 2020 to 2024. The correlation coefficients depicted in the heat map reinforce the findings from the Spearman analysis, revealing that quantity sold**,** humidity**,** and temperature are the variables most strongly associated with sales revenue, with correlation coefficients of 0.80**,** 0.49**,** and -0.45**,** respectively. The strong positive correlation between revenue and quantity sold emphasizes the dominant role of sales volume in driving revenue. Meanwhile, humidity shows a moderate positive association, suggesting its facilitating role in sales, whereas temperature exhibits a negative relationship, indicating potential inhibitory effects on consumer purchasing behavior during hotter days. These patterns, captured visually in Figure 2, further underscore the climatic sensitivity of consumer demand and provide valuable guidance for the integration of weather-driven variables in sales forecasting models.

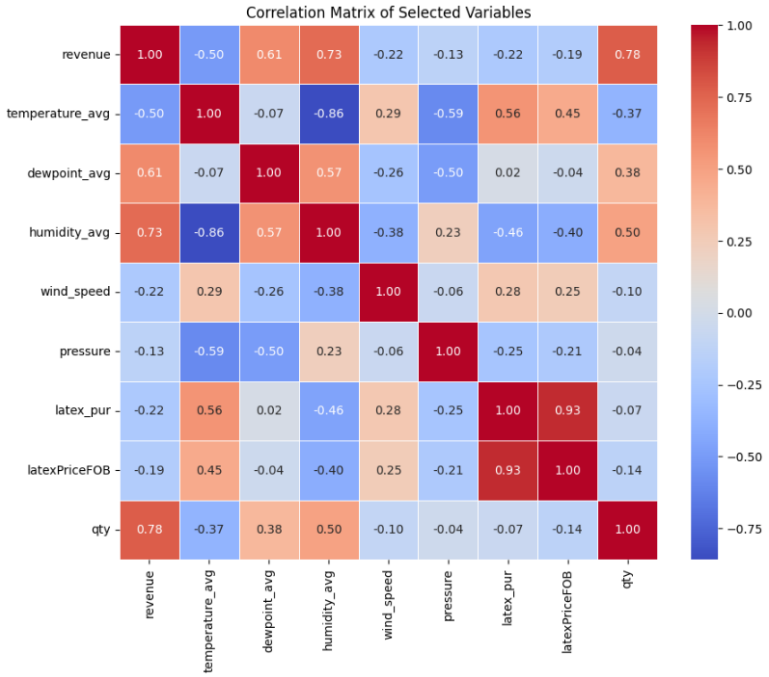

Figure 2.
Correlation matrix heatmap.

Overall, the analysis highlights that certain environmental variables such as humidity and dewpoint have both immediate and lagged associations with sales revenue, indicating a sustained influence over time. In contrast, variables like temperature primarily affect revenue contemporaneously, with diminishing effects in subsequent months. Additionally, the quantity of items sold (qty) exhibits a strong and highly significant immediate correlation with revenue, as expected, since it directly represents the sales volume on a given day. However, this relationship does not persist in the lagged periods, suggesting that qty reflects current consumer demand rather than serving as a leading indicator. These temporal patterns underscore the importance of considering both the timing and nature of variable effects when modeling and forecasting sales with external environmental and behavioral factors.

3.2. Model Evaluation

To rigorously assess the influence of environmental variables on sales forecasting, a comprehensive model evaluation was conducted across multiple modeling paradigms: univariate and multivariate time series analysis (SARIMA, SARIMAX), deep learning (LSTM), and statistical regression approaches (Ordinary Least Squares and Poisson regression). The evaluation aimed to identify both the most accurate and the most interpretable model for capturing the complex relationships between meteorological factors and sales quantity.

Table 2 presents the performance of all model variants using three key statistical indicators: the coefficient of determination (R^2), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). These metrics jointly evaluate model fit, parsimony, and explanatory power. The univariate SARIMA model, using only historical sales data, served as a benchmark and achieved an R^2 of merely 0.027, highlighting its limited predictive capability when external variables are excluded.

Table 2.
Performance comparison of time series models based on R^2, AIC, and BIC.

Type	Variables	R^2	AIC	BIC
SARIMA	Only y	0.027	353.348	356.173
SARIMAX	Humidity	-0.383	344.827	346.521
SARIMAX	Humidity, dewpoint	-0.094	345.472	347.732
SARIMAX	Humidity, dewpoint, temperature	-0.304	347.237	350.061
SARIMAX	Humidity, dewpoint, temperature, qty	0.786	343.903	347.293
LSTM	Humidity	0.399	-	-
LSTM	Humidity, dewpoint	0.234	-	-
LSTM	Humidity, dewpoint, temperature	-1.382	-	-
LSTM	Humidity, dewpoint, temperature, qty	-0.718	-	-
OLS	Humidity	0.535	859.045	862.038
OLS	Humidity, dewpoint	0.589	856.952	861.441
OLS	Humidity, dewpoint, temperature	0.593	858.701	864.687
OLS	Humidity, dewpoint, temperature, qty	0.793	838.352	845.834
Poisson	Humidity	0.536	3.354×10^5	3.354×10^5
Poisson	Humidity, dewpoint	0.588	2.960×10^5	2.960×10^5
Poisson	Humidity, dewpoint, temperature	0.588	2.956×10^5	2.956×10^5
Poisson	Humidity, dewpoint, temperature, qty	0.791	1.509×10^5	1.509×10^5

As progressively more environmental variables were incorporated into the SARIMAX framework, the model's performance improved substantially. The full SARIMAX specification, which includes humidity, dewpoint, temperature, and sales quantity (qty), achieved a markedly higher R^2 value of 0.786, along with the lowest AIC (343.903) and BIC (347.293) among time series models. This confirms that exogenous climatic variables significantly enhance the forecasting power of autoregressive models.

LSTM models displayed mixed and often inconsistent performance. While the model using only humidity achieved a moderate R^2 of 0.399, the inclusion of additional variables led to deteriorating results, with the full specification yielding a negative R^2 (–0.718). This indicates overfitting and poor generalization, potentially due to the limited size of the dataset and the complexity of nonlinear interactions. Furthermore, the LSTM's inability to outperform simpler statistical models makes its applicability dubious in this context, especially when data availability and interpretability are critical.

Regression-based models, namely OLS and Poisson regression, demonstrated robust performance, particularly in their most comprehensive specifications. The full OLS model, integrating all four predictors, yielded the highest R^2 across all tested configurations (0.793) and relatively low mean absolute error (MAE). In contrast, Poisson regression achieved a comparable R^2 of 0.791, but excelled in terms of AIC (1.509×10^5), indicating greater statistical efficiency in handling the skewed or count-like nature of sales data.

To further investigate model accuracy, Table 3 compares the best-performing model from each method based on the error metrics MAE, MSE, RMSE, and R^2. The results reinforce the superiority of the full OLS and SARIMAX models, both of which exhibit high explanatory power and acceptable error levels. Notably, the OLS model, despite higher MSE and RMSE due to the magnitude of revenue values, provides a straightforward and interpretable mathematical relationship between inputs and output.

Table 3.
Comparison of forecasting model performance.

Model	MAE	MSE	RMSE	R^2 Score
SARIMA (Only y)	0.170	0.030	0.180	0.027
SARIMAX (humidity, dewpoint, temperature, qty)	0.180	0.050	0.230	0.786
LSTM (humidity, dewpoint, temperature, qty)	0.170	0.050	0.220	-0.718
Poisson Regression (humidity, dewpoint, temperature, qty)	5.628×10^4	1.204×10^7	3.470×10^3	0.791
OLS (humidity, dewpoint, temperature, qty)	**5.57×10^4**	**4.667×10^9**	**6.831×10^4**	**0.793**

The derived OLS regression equation is as follows:

$$\text{OLS } Y_{\text{Revenue}} = -2.433 \times 10^{6} + 3.403 \times 10^{4}\, x_{humidity} - 3.928 \times 10^{4}\, x_{dewpoint} + 1.028 \times 10^{5}\, x_{temperature} + 2.490 \times 10^{1}\, x_{qty}$$

This formulation offers key insights into the relative influence of each independent variable. Temperature exhibits the strongest positive association with sales revenue, followed by humidity, while dewpoint has a negative effect. The positive coefficient for quantity confirms its autoregressive contribution to revenue prediction.

To visualize the predictive capability of the full OLS model, Figure 3 presents a time-series plot comparing actual revenue against forecasted revenue over the study period. The OLS model was trained using 80% of the data and tested on the remaining 20%. The training and forecasting process was implemented using Python's statsmodels and matplotlib libraries. The forecasting plot (generated using the code shown below) clearly illustrates the model's ability to track revenue trends in the test period.

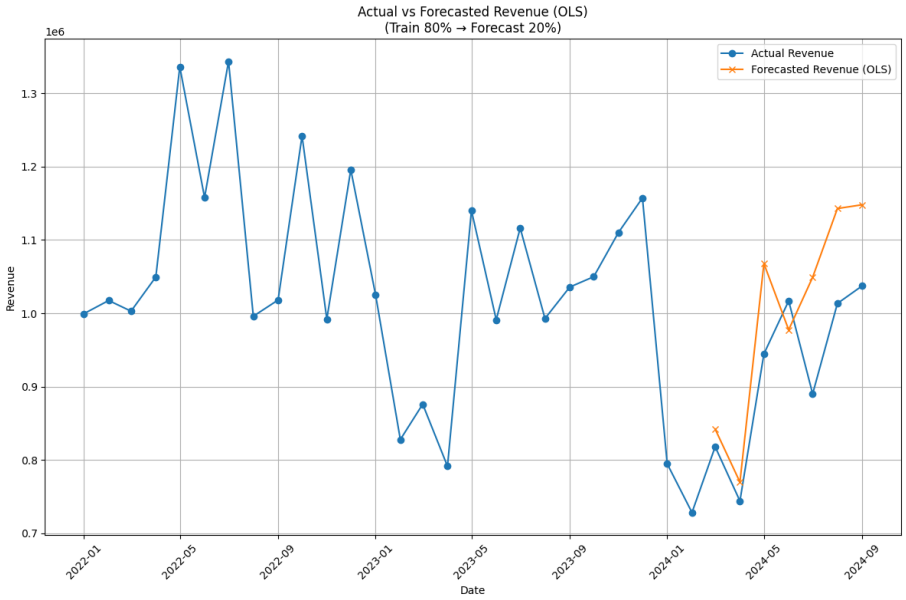

Figure 3.
Comparison of actual and forecasted monthly revenue using OLS regression model

In summary, the results demonstrate that traditional statistical approaches, particularly OLS and SARIMAX, outperform more complex deep learning models such as LSTM in this domain. Their superior performance can be attributed to better alignment with the structure and scale of the data. Furthermore, the OLS model stands out as the most interpretable and statistically robust, making it a suitable choice for practical implementation in business intelligence systems aimed at incorporating environmental intelligence into sales forecasting frameworks.

4. Discussion

The empirical results of this study underscore the effectiveness of statistical and machine learning techniques in enhancing retail sales forecasting, particularly when environmental and transactional variables are considered concurrently. Among the evaluated models, traditional regression techniques, namely Ordinary Least Squares (OLS) and Poisson regression, consistently outperformed more complex models such as SARIMA and LSTM in terms of goodness-of-fit and model selection criteria. The higher R^2 values alongside more favorable AIC and BIC scores observed for the regression-based models suggest that, in certain contexts, model simplicity and interpretability can outweigh algorithmic complexity when paired with relevant feature selection. This finding echoes Ramos and Oliveira [11] conclusion that robust forecasting often hinges on carefully chosen covariates rather than model complexity alone.

The analysis identified sales quantity as the most significant determinant of revenue, with a strong positive coefficient in the OLS model. This result aligns with theoretical expectations and practical retail operations, where volume is a direct driver of income. While environmental factors, specifically humidity and temperature, were found to influence revenue to some extent, their statistical significance was limited within the regression framework. Nevertheless, Spearman rank correlation analysis revealed lagged effects of these meteorological variables on sales, supporting the notion that external environmental conditions can exert indirect or delayed influence on consumer purchasing behavior.

These observations resonate with prior empirical studies. Roth Tran [12] found that high humidity tends to increase retail activity, whereas elevated temperatures tend to suppress it, particularly in open-air or tropical market settings. Similar effects were observed in the present study, where humidity exhibited a positive correlation with sales while temperature was negatively associated. This consistency reinforces the broader theoretical understanding of weather-dependent consumer behavior.

Despite its potential, the LSTM model demonstrated relatively poor performance in this study. While LSTM networks are designed to capture sequential dependencies and nonlinear dynamics, limitations such as insufficient input diversity, short time-series length, or suboptimal hyperparameter tuning, may have hindered their predictive capacity. Nevertheless, prior research has highlighted LSTM's strengths in complex forecasting environments. For instance, Hurtado-Mora, et al. [13] and Shak, et al. [14] reported improved accuracy using LSTM in structured retail and supply chain settings when supported by extensive feature engineering and optimization. This suggests that while the current implementation was suboptimal, the architecture holds promise for future applications if adequately refined.

The implications of these findings extend beyond methodological performance. From a practical perspective, accurate revenue forecasting enables businesses to engage in more informed decision-making, including inventory planning, resource allocation, and promotional campaign management. By integrating environmental factors, retailers can anticipate demand shifts driven by seasonal or weather-related influences, increasing agility and responsiveness. Visualization through interactive dashboards further enhances managerial capability, allowing real-time monitoring and proactive adjustment to external changes.

Moreover, the methodological framework applied in this study contributes to ongoing efforts across multiple sectors. Similar forecasting approaches have been applied successfully in retail furniture sales [15] e-commerce platforms using multimodal inputs [16] and tourism demand prediction via mixed-frequency data [17]. These cross-domain applications affirm the versatility of statistical and machine learning tools in capturing complex demand dynamics when paired with rich data environments.

In summary, this study highlights that well-specified and interpretable regression models remain powerful tools for forecasting when enriched with domain-specific variables. While advanced neural models like LSTM show theoretical appeal, their performance is contingent upon data richness and tuning precision. Future work may explore hybrid architectures that combine the interpretability of statistical models with the nonlinear learning capacity of deep networks, potentially offering more robust and generalizable forecasting solutions across varied business contexts.

5. Conclusion

This study evaluated the performance of multiple forecasting models SARIMA, SARIMAX, LSTM, Ordinary Least Squares (OLS), and Poisson regression in predicting daily retail sales in relation to both transactional and weather-related variables. Among the models tested, traditional regression methods (OLS and Poisson) consistently demonstrated superior predictive performance, as reflected by higher R^2 values and lower error metrics (MAE, MSE, RMSE), compared to more complex time series and deep learning models.

The SARIMA model, designed to capture seasonality within sales data, yielded poor performance with an R^2 score of –0.09, indicating limited explanatory capacity when applied as a univariate model. The SARIMAX model, which incorporated the exogenous variables temperature, humidity, and wind speed, slightly improved the model fit, but still produced a negative R^2 value (–0.66), suggesting that the weather variables alone were insufficient for accurate sales prediction. The LSTM model, while theoretically well-suited to capturing non-linear temporal dependencies, also underperformed in this context (R^2 = –0.61). Its suboptimal results likely stem from limited feature diversity, data sparsity, or the need for further tuning.

Conversely, the OLS model, particularly when incorporating both weather and transactional variables, showed the best model fit, indicating that interpretable linear models with relevant features remain powerful tools in retail forecasting. This outcome aligns with previous research and reinforces the notion that a simple model, when combined with domain-relevant variables, can often outperform more complex algorithms in certain real-world applications.

Future work should explore the integration of additional predictors beyond basic weather indicators such as promotional activity, customer foot traffic, economic indicators, and localized event data, to enrich model inputs and improve forecasting accuracy. In the case of LSTM and other deep learning models, further refinement through hyperparameter tuning, feature engineering, and use of larger datasets may unlock their full predictive potential. Additionally, hybrid models that combine the strengths of traditional statistical techniques and neural network architectures may offer a promising path forward.

Ultimately, accurate and interpretable forecasting models can provide valuable support for decision-making in retail operations, helping businesses adapt more effectively to both predictable trends and external environmental variability.

References

[1] T. Kongthanasuwan, N. Sriwiboon, B. Horbanluekit, W. Laesanklang, and T. Krityakierne, "Market analysis with business intelligence system for marketing planning," *Information*, vol. 14, no. 2, p. 116, 2023. https://doi.org/10.3390/info14020116

[2] V. Duarte, S. Zuniga-Jara, and S. Contreras, "Machine learning and marketing: A systematic literature review," *IEEE Access*, vol. 10, pp. 93273-93288, 2022. https://doi.org/10.1109/ACCESS.2022.3202896

[3] T. Falatouri, F. Darbanian, P. Brandtner, and C. Udokwu, "Predictive analytics for demand forecasting–a comparison of SARIMA and LSTM in retail SCM," *Procedia Computer Science*, vol. 200, pp. 993-1003, 2022. https://doi.org/10.1016/j.procs.2022.01.298

[4] D. J. Beltran, Y. Kangleon, A. K. Balan, and J. De Goma, "Credit card sales performance dashboard," in *Proceedings of the International Conference on Industrial Engineering and Operations Management*, 2021.

[5] N. Patel, "How to use data visualization in your content to increase readers and leads," 2023. https://neilpatel.com/blog/data-visualization/

[6] K. K. Halim, S. Halim, and Felecia, "Business intelligence for designing restaurant marketing strategy: A case study," *Procedia Computer Science*, vol. 161, pp. 615-622, 2019. https://doi.org/10.1016/j.procs.2019.11.164

[7] 1stCraft Team, "Data visualization," 2020. https://1stcraft.com/what-is-data-visualization/
[8] Geckoboard, "Digital dashboard examples," 2024. https://www.geckoboard.com/dashboard-examples/executive/digital-dashboard/
[9] J. P. Bharadiya, "The role of machine learning in transforming business intelligence," *International Journal of Computing and Artificial Intelligence*, vol. 4, no. 1, pp. 16-24, 2023. https://doi.org/10.33545/27076571.2023.v4.i1a.60
[10] W. Li and K. E. Law, "Deep learning models for time series forecasting: A review," *IEEE Access*, vol. 12, pp. 92306-92327, 2024. https://doi.org/10.1109/ACCESS.2024.3422528
[11] P. Ramos and J. M. Oliveira, "Robust sales forecasting using deep learning with static and dynamic covariates," *Applied System Innovation*, vol. 6, no. 5, p. 85, 2023. https://doi.org/10.3390/asi6050085
[12] B. Roth Tran, "Sellin'in the rain: Weather, climate, and retail sales," *Management Science*, vol. 69, no. 12, pp. 7423-7447, 2023. https://doi.org/10.1287/mnsc.2023.4799
[13] H. A. Hurtado-Mora, A. H. García-Ruiz, R. Pichardo-Ramírez, L. J. González-del-Ángel, and L. A. Herrera-Barajas, "Sales forecasting with LSTM, custom loss function, and hyperparameter optimization: A case study," *Applied Sciences*, vol. 14, no. 21, p. 9957, 2024. https://doi.org/10.3390/app14219957
[14] M. S. Shak *et al.*, "Optimizing retail demand forecasting: A performance evaluation of machine learning models including Lstm and gradient boosting," *The American Journal of Engineering and Technology*, vol. 6, no. 09, pp. 67-80, 2024. https://doi.org/10.37547/tajet/Volume06Issue09-09
[15] M. N. İnce and Ç. Taşdemir, "Forecasting retail sales for furniture and furnishing items through the employment of multiple linear regression and holt–winters models," *Systems*, vol. 12, no. 6, p. 219, 2024. https://doi.org/10.3390/systems12060219
[16] C. Li, "Commodity demand forecasting based on multimodal data and recurrent neural networks for E-commerce platforms," *Intelligent Systems with Applications*, vol. 22, p. 200364, 2024. https://doi.org/10.1016/j.iswa.2024.200364
[17] M. Hu, M. Li, Y. Chen, and H. Liu, "Tourism forecasting by mixed-frequency machine learning," *Tourism Management*, vol. 106, p. 105004, 2025. https://doi.org/10.1016/j.tourman.2024.105004