SvedbergOpen
DISSEMINATIONOF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

An Empirical Evaluation of Deep Learning and Machine Learning Methods towards Retail Sales Prediction

Riddhi J. Kotak[1], Jinal H. Tailor[2], Amit Lathigara[3], Krunal Vaghela[4]

[1]Research Scholar, PhD-Computer Engineering/IT, Gujarat Technological University, Ahmedabad. kotakriddhi@gmail.com
[2]MCA, S. S. Agrawal Institute of Management & Technology, Navsari, India jinal.tailorssa@gmail.com
[3]RK University, Rajkot, India. amit.lathigara@rku.ac.in
[4]Engineering & Technology, Marwadi University, Rajkot, India krunal.vaghela@marwadieducation.edu.in

Abstract

The retail sales forecasting problem is difficult due to the factors of seasonality, non-linearity of demand, the contextual variables which are heterogeneous and the time series dynamics which changes over time. This work compares the various statistical, machine-learning and deep-learning methods to forecast weekly retail sales empirically and from a time-aware perspective. The Sample Superstore consists of 10,194 transactions with 21 original attributes collected between 2019 and 2022. Observations per transaction were weekly aggregated and transformed into a leakage-aware representation of history with the lags of 1 week, 2 weeks, 4 weeks, 8 weeks, 13 weeks, 26 weeks, and 52 weeks. The models were assessed on the 80:20 chronological holdout in addition to the time-series validation, and were ARIMA, Random Forest, Ridge Regression, SGD Regressor, LSTM and Transformer. The test error of Ridge Regression was the lowest with MAE=5,009.71, RMSE=6,580.95, MAPE=37.17%, and R^2=0.249 while SGD Regressor and Random Forest were the second and third respectively. The weekly setting gave rise to higher errors for LSTM, ARIMA and Transformer models. The findings indicate that the more complex the architecture, the more difficult the task of forecasting is not necessarily the more accurate the forecasting. This study thus provides a repeatable comparison concentrating on the temporal representation, leakage control, the validation strategy, regularization, and the model capacity and available data.

Keywords

Machine learning, time series, sales prediction, random forest, Regressor, univariate approach, multi variant approach.

1. Introduction

In the context of artificial intelligence (AI), retail sales forecasting is a significant application that has increasingly become a major challenge for retailers given the amount of data that can be collected from transactional activity, including product, customer, store and sales data [1]. Forecasting can be useful in managing inventory, production planning, supply-chain operations, and strategic decision making [2]. Consumer behavior, seasonality, promotions, various sales channels, and data variability, however, introduce an element of difficulty for prediction [3, 4].

In the AI context, retail forecasting is a process of learning temporal patterns and non-linear relationships from the historical and contextual data. The purpose of this study is to compare univariate and multivariate forecasting methods in similar predictive framework. The univariate approach is mainly based on the previous sales pattern, while the multivariate approach uses other explanatory variables. To examine the impact of feature representation and learning mechanisms on forecasting, the Random Forest, Ridge Regressor and SGD Regressor are tested.

This research centres on the comparison of different types of forecasting univariate and multivariate forecasting Both kinds of time series forecast ought to segments: univariate forecasting as shown in Fig.1. below. To predict future phenomena a model needs to be taught based on data to identify recurrent patterns. A well-constructed forecast model can improve a supermarket's revenue greatly by making more informed decisions and by gaining insight of the customer behaviour.

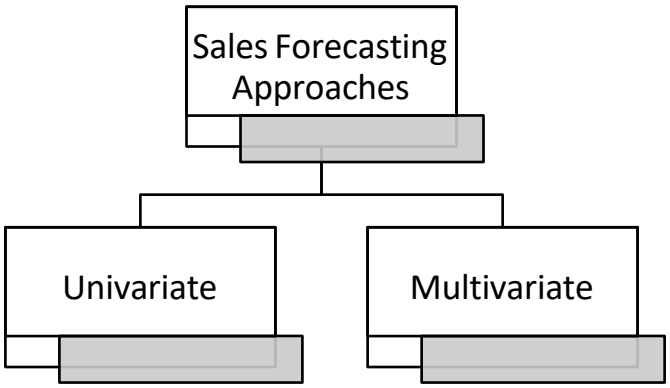


Fig 1. Sales forecasting approaches

Retail sales forecasting is a complex, temporal-dependent, nonlinear, explanatory variables with various types, noisy and seasonal, and changing consumer behavior predictive-learning problem, from an artificial intelligence point of view. In this regard, the aim of this research goes beyond the scope of a retail specific forecasting solution. The study explores the representation and exploitation of information in the temporal and contextual data, as it is presented through the different learning paradigms. Random forest is a nonlinear ensemble learning, Ridge regression is a regularized linear learning, and SGD Regressor is optimization-based incremental learning. They compare model structure, regularization, feature representation and optimization to gain insight on how it affects forecasting.
The methodological value of this study is the comparative study of univariate and multivariate forecasting in the same context of implementing artificial intelligence-based forecasting. The univariate setting examines the predictive information that comes mainly from the past sales behaviours, while the multivariate setting includes other variables to see if more predictive information is available. In this work these methods are not considered as being independent forecasting options, but as complementary representations of the same prediction problem.

The key contributions of this study are: (1) comparison of classical forecasting paradigms to machine-learning forecasting paradigms in the same retail forecasting environment; (2) analysis of the data from different perspectives to better understand the data's univariate and multivariate structure for influencing the temporal prediction task; (3) correlation and exploratory feature-selection analysis of the relationships between features; (4) validation of multidimensional models using MAE, RMSE, MAPE and R^2; and (5) examination of the trade-offs among model predictive accuracy, model interpretability and model computational complexity. The outcome of this is a methodology that can be applied to other time-dependent prediction problems with structured heterogeneous data that are also time dependent.

Research novelty and contribution: We are deviating from the usual way of comparing already existing forecasting algorithms based on their predictive performance; in this study, we are taking a different approach and look at retail forecasting with respect to the temporal representation of the data, the complexity of the model, the regularization, and the number of effective samples. The leakage-aware experimental framework is defined for statistical, machine-learning, ensemble, recurrent, and attention-based forecasting paradigms.

A leaky-experimental common framework is created for the statistical, machine-learning, ensemble, recurrent, and attention-based forecasting paradigms. Methodologically consistent comparison is provided by the multi-horizon historical lag features, chronological holdout evaluation, time-series cross-validation and real sequential representations. The analysis that follows shows that the more complicated the architecture the less accurate the forecasting in the limitation of the effective temporal sample.

Thus, the novelty of this work is not an alternative forecasting algorithm, but rather the methodical evidence provided in order to be able to replicate the results and understand what conditions lead to the success of different AI forecasting paradigms when applied to structured temporal retail data.

2. A review of the literature

A corpus of research on the approach and framework of sales forecasting for retail products is reviewed in this section. The pursuit of survival strategies, or to flourish in the competitive market environment is being pursued now by every organisation. In this section, some of the relevant literature is cited.
Retail Sales forecasting lies at the heart of supply chain and inventory management and there have been several alternative methods to enhance its accuracy. To improve predictive models, different statistical and

machine learning methods have been recently tested on different types of datasets. A summary of the contrasting evaluation of the literature review is provided in the table below. 1.

Before we discuss machine learning, let us first describe traditional forecasting techniques used within retail:

Author Name	Paper Published	year	Dataset	Methodology
Koh Ya Wen1,*, Minnu Helen Joseph2 and V. Sivakumar3	EAI	2024	Bigmart Sale dataset	Random Forest (without tuning) + Multivariate Approach
H. Chan, M.I.M,Wahab	Supply Chain Analytics 5 (2024), Elsevier	2024	massive retailer with multiple locations within Canada	LASSO with Poisson regression , linear regression with Poisson regression , XGB , SVM ,MLP, GBM, and.
Hitesh S.M1, Yukthi A2, Prof.Ramya B.N3	International Journal of Novel Research and Development	2024	efashion store	Gradient boosting, random forest regression, and linear regression
Taha Falatouri 1,2, Mehran Nasseri 1,2 , Farzaneh Darbanian 1,2,and Patrick Brandtner 1,2,*	MDPI	2023	Supermarket Dataset	extra tree regressors (ETRs), LSTM networks
P. Guru1, J. Sathyapriya2*, K. V. R. Rajandran3, J. Bhuvaneswari4, C. Parimala5 [1]	International Journal of intelligent systems and applications in engineering	2023	Bigmart Sale dataset	Random forest algorithm + Multivariate Approach
Sebastian Maier b,a, Yara Kayyali Elalem a,∗, Ralf W. Seifert a,c	International Journal of Forecasting, Elsevier	2022	Dell data set	gated recurrent units, Long short-term memory, convolutional neural networks and are examples of deep neural networks (DNNs).
Yu-Wei Gu and Chih-Hsuan Wang *	MDPI	2022	Walmart, Costco, and Kroger Data set	the Lotka–Volterra model (LVM), support vector regression (SVR), multivariate adaptive regression splines (MARS),and data envelopment analysis (DEA)

TABLE 1. COMPARATIVE STUDY OF LITERATURE REVIEW

Koh Ya Wen et al., (2012) [1] used the Bigmart Sale dataset in their work. (2024) employed a multivariate method, Random Forest (with no tunning). Demonstrating contrary results to the traditional statistical model used in prediction, however, was their finding that better-than-fair bi- scillation forecasting inferences can be achieved using ensemble type "learning" methods. Similarly to the above, Guru et. (2023) have also demonstrated that ensemble learning is very effective, even for sales forecasting, using the Bigmart Sale dataset and the Random Forest algorithm with the multivariate method.

In Hitesh S.M. et al. (2024) [3], the focus is on sales forecasting for a large Canadian outlet. Multiple regression approaches were used, such as multilayer perceptron's (MLP), support vector machines (SVM), gradient boosting machines (GBM), linear regression, Poisson regression, LASSO with Poisson regression, gradient boosting machines (GBM) and XGBoost gradient boosting (XGB). They found that more complex machine learning models like XGB and MLP outperform classical models like regression when dealing with large data sets from large retailers.

There are three frameworks of interval series: ARIMA, Simple Exponential, and Holt-Winters. Smoothing schemas are the traditional statistical techniques which predict future sales level from historical data. While they can work with linear patterns, they can't cope with the complexity of multidimensional and complicated data in retail applications.

For Prediction Regression model can be applicable to Utilize the relationship between independent and dependent variables, also regression models have been vastly used for polynomial and linear data for prediction. For large dataset regression model can handle it in effective way for linear, non-linear data.

2.1 Research Gap in AI-Based Retail Forecasting

Existing studies demonstrate increasing adoption of both classical machine-learning and deep-learning techniques for retail forecasting. Tree-based ensemble algorithms such as Random Forest and XGBoost are particularly effective for structured tabular data because they can capture nonlinear interactions among explanatory variables. Deep-learning architectures such as LSTM, GRU and CNN, in contrast, provide mechanisms for automatically learning complex temporal representations and long-range dependencies. However, their increased representational capability is generally accompanied by greater computational requirements, larger training-data requirements and reduced interpretability.

The reviewed literature also indicates that many studies concentrate either on a particular forecasting algorithm or on a single data representation. Consequently, comparatively less attention is given to understanding how the transition from univariate temporal representation to multivariate feature representation changes the effectiveness of different AI learning mechanisms. Furthermore, predictive accuracy alone does not provide a complete basis for AI model selection; interpretability, computational efficiency, scalability and robustness are also important considerations for practical decision-support systems.

This study addresses this gap by examining forecasting approaches from an AI model-selection perspective. It considers differences among regularized regression, optimization-based learning, ensemble learning and temporal/deep-learning approaches while examining the role of univariate and multivariate information representations. The resulting analysis is intended to identify not merely an accurate forecasting technique, but the conditions under which different AI paradigms are appropriate for structured temporal prediction.

Recent studies have shown that machine-learning and deep-learning techniques have the potential of improving retail forecasting, such as tree-based ensembles and LSTM-based networks [2]–[5]. But, there has been methodological shortcomings in the current literature. First, the algorithms are often tested with varying data sets, the algorithms' representations of the data, the length of forecast, or the manner in which it is validated, so that it is hard to isolate the impact of the learning mechanism itself.

Third, deep-learning architectures have a higher level of representational capacity, and comparatively little work has focused on how performance of these architectures' changes when the number of temporal observations that are effective for representing the problem is reduced by aggregation and/or a long historical lag requirement. Thirdly, incorrect random validation will lead to optimistic forecasts in time-dependent forecasting problems, highlighting the importance of chronological validation [11].

Furthermore, extensive forecast experiments have shown that the use of more complex models is not always associated with better predictions [12; 13]. Hence, there is no existing study that can be used to control and compare the temporal representation, effective sample size, regularisation, and complexity of the models together. To fill this gap, this study compares these six methods in terms of their weekly forecasting performance, evaluated under a unified framework, with multiple lag features, chronological holdout testing, and time-series cross-validation. The contribution in this paper is not only about the accuracy of a given forecasting model, it is about the conditions under which it generalizes well, when using different learning paradigms, such as AI.

3. Machine Learning Methods for predicting retail sales

In this section will provide a concise detail about machine learning models Random Forest, Decision tree, SVM (Support Vector Machine) , ANN (Neural Network), Deep Learning Approaches, RNN (Recurrent Neural Network), Facebook prophet model, – CNN, Hybrid Models, with its methodologies, parameters and functionality, performance outcomes for retail products sales forecasting how it can be effective. Comparative performance of the retail products sales forecasting models is also explored and showcased and results of effectiveness of these models in capturing demand patterns, trends, seasonality and nonlinearity with used super markets Dataset are explored and highlighted.

3.1 Random Forest models and decision trees

This is an ensemble approach called Random Forest, which does multiple decision tree forecasting to provide greater forecasting resistance and accuracy. This approach is useful in cases where the relationship between attributes is not linear and is complicated, and when the data are high-dimensional. Random

Forest is extremely useful for the multivariate retail sales prediction problem, because it can adapt itself and has an inbuilt ability to adapt to overloading. But with extremely large amounts of data, they can be hard to compute and take up a long training time.

XGBoost: Due to XGBoost remarkable predictive performance and high computational efficiency, Extreme Gradient Boosting (XGBoost), an effective tree-based ensemble technique is nowadays extensively utilized in sales forecasting applications. It is particularly effective for situations where there are a large number of external influences affecting sales. While XGBoost has many benefits, it has also a number of hyperparameters that need to be carefully tuned, including learning rate and tree depth, and which can make it more difficult for a practitioner with less experience to implement.

In terms of artificial intelligence, ensemble methods provide a way of learning from complex data through the use of multiple weak or diverse predictive structures that are connected to a more powerful decision-making mechanism. Random Forest builds a number of decision trees with randomly selected samples and feature subsets and combines their results to lessen the reliance on the fitted single decision tree, making it more robust against variance. This property allows it to work for diverse multivariate data with nonlinear relationships and interactions between features.

Boosting is another ensemble-learning technique that generates successive learners to correct the mistakes by previous learners. With the iterative correction, more complicated predictive relationships can be expressed. These properties show the potential of ensemble AI systems to learn interaction in high dimensional structured data without an explicitly specified mathematical relationship between all of the predictors and the target variable.

3.2 Networks of Neural Systems

Artificial Neural Networks (ANNs) are computational models created to extract intricate patterns from data Inspired by the framework and functioning of the human brain. They are able to catch non-linear relationships in sales data, even when the underlying aspects are random and not well understood. Nonetheless, ANNs are a 'black box' model and provide little insight into the variables that influence the results.

RNNs are especially helpful for forecasting time series, although they can recall past inputs, RNNs can interpret sequential data. a form of RNN, LSTM networks, are frequently used to predicting trends in retail sales data, because of its ability to capture temporal dependencies.

3.3 Facebook Prophet

Facebook Prophet: Recurrent neural networks (RNNs) can learn to recognize events over time, remembering what it has previously seen, so they're definitely useful for figuring out variation over time.

Prophet is a time-series forecasting model designed for enterprise data, which was created by Facebook. It's hugely adaptable, accepting seasonality, vacations, and tendency switches in consumer expenditure, rendering it a sound applicant for solitary forecasting situations. Non-specialists often find Prophet easier to use and more easily understood.

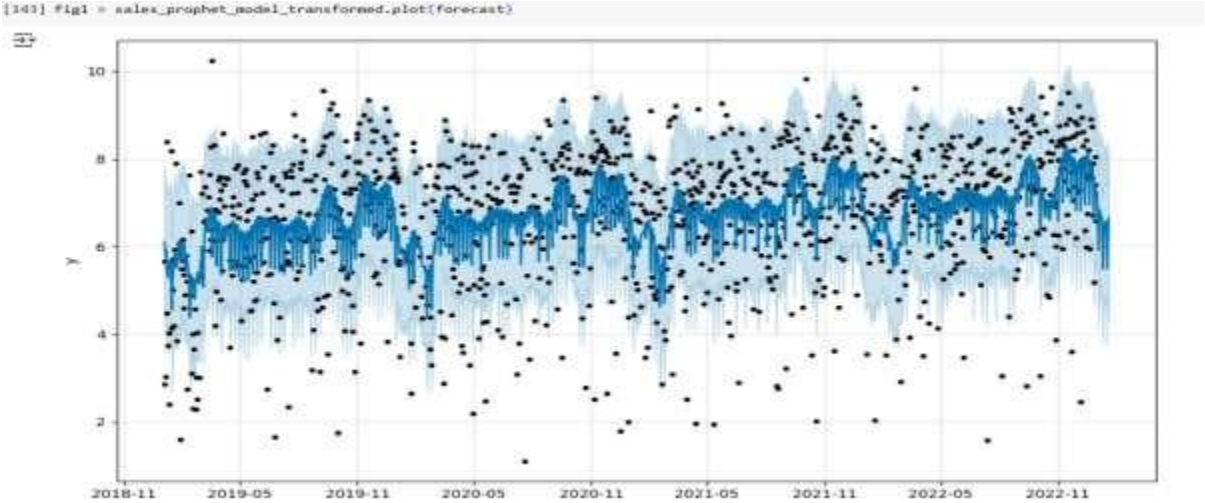


Fig 2. Facebook Prophet Model

In Fig.2. are the illustration assists in evaluating the accuracy of the model and highlighting areas in which sales at supermarkets forecasting requires to be enhanced.

3.4 SVMs, or support vector machines

Traditionally used for classification, SVM for Support Vector Regression can also be improved for regression issues related to sales forecasting. Small to medium-sized datasets with non-linear relationships can be handled well with SVR..

3.5 Deep Learning Approaches

CNNs (Convolutional Neural Networks) : formally used in image processing, CNNs are use for time series sales forecasting by extracting important features from raw sales data.

Hybrid Models: By integrating machine learning models alongside additional approaches to deep learning or conventional methods, hybrid models have been demonstrated to increase prediction accuracy by utilizing the advantages of several different strategies.

4. Data Preparation and Feature Engineering

The quality of the data that machine learning models use determines how well they perform. Important facets of data preparation for retail sales forecasting are covered in this section.
Order Date, Order ID, Ship Mode Ship Date, Customer Name, Customer ID, Segment, Country/Region, City, State/Province, Postal Code, Region, Product ID, Category, Sub-Category, Product Name, Sales, Quantity, Discount, and Profit are among the 10195 attributes for sales data that we used from the Sample Superstore data set from 2019 to 2022.
As shown in Fig.3. First, do analysis of data by applying few methods and command, identify Unique attributes from dataset and we are getting below mentioned values.



Fig 3. Data Preparation

then check for Null entity in dataset but no entity in data set contains Null value so it will help more for our work, then check entities data type it will help for further work and commands.
Next Apply Command to check Sales, Profit, Quantity values In Box Plot graph as shown below in Fig.4 Boxplot Graph for Sales and Fig.5. Box plot graph for Quantity.

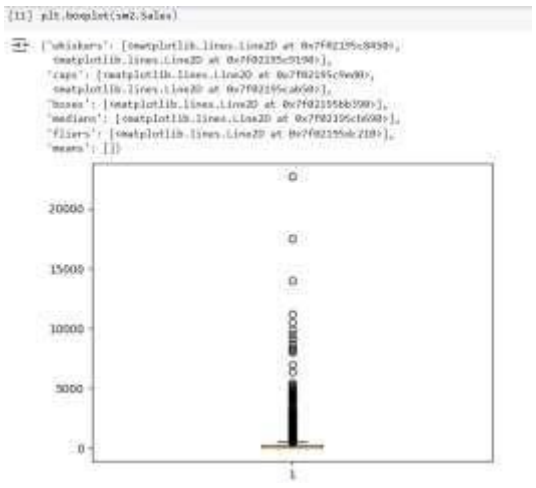


Fig 4. Boxplot Graph for Sales

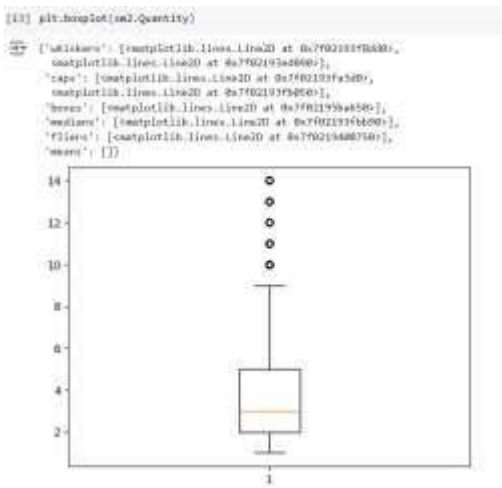


Fig 5. Box plot graph for Quantity

Then we start working on Time series Data in which because dataset is of Supermarket sales so we decided to work in Weekley time series forecasting because it Captures Shopping Patterns, Reduces Noise, Matches Business Decisions, Balances Granularity and Accuracy, Handles Missing Data Better, Improves Computational Efficiency, Enhances Seasonality Detection, Eases Comparison Across Years,through as shown in Fig. 6. Weekly sales outcome , Fig. 7. Weekly Quantity outcome.

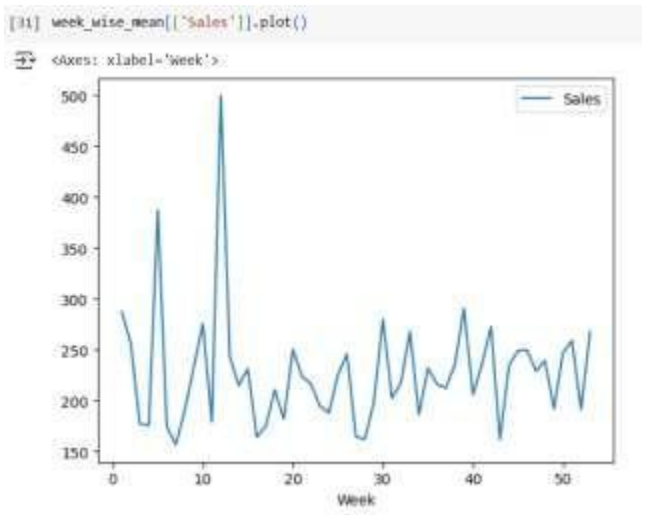


Fig. 6. Weekly sales Analysis

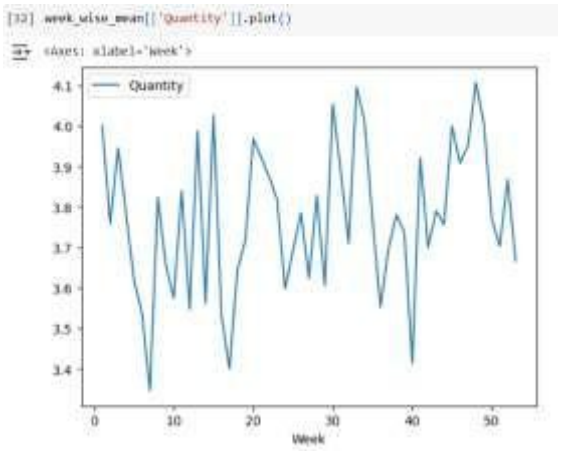


Fig. 7. Weekly Quantity Analysis

The second correlation heatmap in Python was created with Seaborn as illustrated in Fig. 8. Using Seaborn to create a heatmap graph. The relationship between several numerical parameters in the supermarket sales data is plotted.

As Sales and Discount are negatively correlated, a higher discount might not always result in a higher total sales revenue. Other factors affect the revenue more than the quantity sold, as Sales and Quantity are weakly correlated.Time-related variables (Week, Month, Season) are highly correlated and one of these variables could be redundant in forecasting models.



Fig. 8. Heatmap Graph Using Seaborn

The work is then continued with the use of the methods for handling incomplete data namely the Attribution and Interpolation methods in dealing with gaps in sales data.

Forecasting accuracy can be significantly affected by feature selection, which includes the process of choosing the most pertinent features (such as promotions, holidays, and weather).

The literature review covers a wide variety of forecasting models such as ARIMA/SARIMA, Prophet, Support Vector Regression, XGBoost, ANN, RNN, LSTM, CNN, Transformer and hybrid models to provide the general background of the methods employed in retail forecasting. These literature-based methods, however, should not be considered as methods that have been tested in the present experiment. The final experimental comparison is limited to only the models ARIMA, Random Forest, Ridge Regression, SGD Regressor, LSTM and Transformer. The six models were chosen to consider the three main categories of forecasting paradigms: statistical time-series modelling (ARIMA), nonlinear ensemble learning (Random Forest), regularized linear learning (Ridge Regression), optimization-based incremental learning (SGD Regressor), recurrent deep learning (LSTM), and attention-based deep learning (Transformer). The results of the experiments and the conclusions drawn with respect to these models are only those for the aforementioned six models.

Model	MAE	RMSE	MAPE (%)	R^2	Correlation
Ridge Regression	5,009.71	6,580.95	37.17	0.249	0.530
SGD Regressor	5,167.27	6,777.56	38.95	0.204	0.491
Random Forest	5,620.14	6,933.79	41.58	0.167	0.468
LSTM	5,836.60	7,570.90	38.48	0.006	0.414
ARIMA (1,1,1)	6,026.97	7,578.76	44.27	0.004	0.602
Transformer	6,155.99	7,737.75	42.11	−0.038	0.394

Table.2. Forecasting Models

4.1 Experimental Methodology

The dataset consists of a single paragraph. There is only one paragraph in the dataset. The data set studied is the Sample Superstore transactions data set with 10,194 transactions and 21 original attributes that spans the time period from 3 January 2019 to 30 December 2022. The variables are Order information, Shipping information, Customer information, Geographical information, Product hierarchy, Sales, Quantity, Discount and Profit. Sales is regarded as the main forecasting objective.

Using Sample Superstore, all experiments were performed with a transactional dataset that includes 10,194 transactions and 21 original variables from 3 January 2019 to 30 December 2022. Each row is a single retail transaction/order-product record not a time-series observation.
These 21 variables are: Row ID, Order ID, Order Date, Ship Date, Ship Mode, Customer ID, Customer Name, Segment, Country/Region, City, State/Province, Postal Code, Region, Product ID, Category, Sub-Category, Product Name, Sales, Quantity, Discount, and Profit. The variables have been divided into five main groups: (i) order and shipping data, (ii) customer data, (iii) geographical data, (iv) product data and (v) numerical data of the outcomes of the transactions.
The choice of the forecasting target was made to be Sales. The values for the original transaction-level Sales vary considerably from about 0.444 to 22,638.48 with a mean value of about 228.23 and a standard deviation of about 619.91. The experimental pipeline did not have any missing data in the variables used.
The original data is a transactional data set and not a set of observations taken at regular intervals of time, so Order Date was converted to a datetime and the data were sorted chronologically. Observations were then aggregated to a weekly level with the weekly Sales being the forecast target. Then historical lags of 1, 2, 4, 8, 13, 26 and 52 weeks were created to represent short-, medium- and annual-time behaviors, respectively.
Sales and lagged historical data, along with some contextual data such as Profit, Discount, and Quantity was selected for use in the multivariate forecasting representation. No data leakage occurred because only the values prior to the origin of the forecasts were used. In order to have enough historical data, the maximum 52 weeks lag was applied and observations with less historical data were removed, leaving 157 complete weekly observations for the machine-learning experiment based on lag. The first 125 observations were used to develop the model and the last 32 observations were kept as the unseen chronological test set.

4.1.1 Feature Engineering

Lagged features were created at 1, 2, 4, 8, 13, 26 and 52 weeks in order to represent the historical demand information in the model without future data leakage. These lags are the historical behaviour in real time, short term, quarterly, half-yearly and annually.

Historical lag information was created for Sales, Quantity, Discount and Profit. Another form of temporal seasonality was also accounted for using sine and cosine transformations of week number. Only historical data prior to the forecast origin were used as predictors. The variables such as Quantity, Discount, Profit, and Sales data that were used were all contemporaneous with the week being predicted, so they would not necessarily be present at the forecast time and could contain information leakage.

Observations with less than 52 weeks of historical data were eliminated since a maximum lag of 52 weeks was used. This led to 157 complete observations for the week based machine-learning experiment.

4.1.2 Forecasting Horizon

This study uses a one-week-ahead forecasting period. A forecast for each origin is made using information up to the previous week to forecast the aggregate Sales for the next week. This horizon was chosen since short-term forecasts (weekly) are directly used for inventory planning, replenishment and operational decisions for retail stores.

The forecasting horizon is not changed in the comparative experiment to ensure that the performance differences are due to the model behaviour and not difference in the forecasting horizon.

4.1.3 Chronological Train-Test Splitting

Since the observations are a time series, random splitting of the train-test was avoided. The observations were carried out in chronological order following lag construction. The first 125 weekly observations (80%) were used to develop the model, and the last 32 observations (20%) were kept as an unseen temporal test set.
The "held out" test period is roughly the same time as 27 May 2022–30 December 2022. The observations made during the tests were not utilized for feature scaling, hyperparameter optimization, model selection or training.

In all cases of hyperparameter optimization using machine learning, five-fold TimeSeriesSplit was used only during the training time. So each validation segment itself will take place chronologically after the training observations it corresponds, which means that each validation segment is less prone to lookahead bias and can be used to assess the validity of the forecasts in a more realistic way [11].

4.1.4 Data Scaling

Feature scaling was used on just models sensitive to the magnitude of the features. The training observations were used to fit the scaling parameters and then the same parameters were applied to the validation and test observations. This is done so that statistical data collected in the future test period does not affect the modelling process.

The SGD based learning was standardized and the deep-learning input variables were normalized using only the learning data parameters. Unlike most other machine learning models, feature scaling is not required in Random Forest since splitting is done in trees is monotonically scaled independent.

The model configuration and hyperparameters are provided in this section.

Model	Final configuration
ARIMA	ARIMA(1,1,1); univariate weekly Sales
Random Forest	n_estimators=200, max_depth=10, max_features=0.7, min_samples_leaf=2, random_state=41
Ridge Regression	alpha=100
SGD Regressor	alpha=0.1, L2 penalty, standardized predictors
LSTM	64 hidden units, Adam optimizer, learning rate = 0.0005, chronological sequential input
Transformer	d_model=64, 4 attention heads, feed-forward dimension = 128, 2 encoder layers, dropout = 0.1, 26-week sequence

Table.3. the model configuration and hyperparameters

The hyperparameters were only tuned on the training set. Random Forest, Ridge Regression and SGD Regressor were chosen for their evaluation of candidate parameter settings via time-series cross-validation. Final parameter values were chosen based on the parameters' abilities to validated forecast.

The architecture of the LSTM was 64 units, trained with Adam optimizer with a learning rate of 0.0005. No shuffling of observations was carried out and training was not randomly shuffled.

Temporal windows of 4, 8, 12 and 26 weeks were examined for the Transformer. The selection of sequence length was based on chronological rather than final test performance. The model dimension was set to 64, the number of attention heads to 4, the number of encoder layers to 2, the feed-forward dimension to 128 and dropout rate to 0.1, with the selected architecture, input sequence of 26 weeks, and model architecture of 64.

4.1.5 ARIMA Configuration and Stationarity Analysis

In this step, you configure the ARIMA model and analyze its stationarity.In this step, you set up the ARIMA model and check for stationarity.

The weekly training series was first investigated with Augmented Dickey-Fuller (ADF) test before the statistical modelling. The ADF statistic was around −7.046 and the P value was less than 0.001, indicating that there is evidence that the training series investigated does not have a unit root.

The statistical baseline used was ARIMA(1,1,1). This result is interpreted with great caution due to the stationarity diagnostic, and is noted as a robustness extension for differencing that would be desirable, namely the \(d=0\) and seasonal specifications.

4.1.6 Evaluation Metrics

The accuracy of the forecast was assessed by the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), coefficient of determination (R^2) and Pearson correlation. Average absolute forecast deviation (MAE), Root mean square error (RMSE), Mean absolute percentage error (MAPE), and the amount of variance explained (R^2) give more weight to larger errors, while correlation assesses the correspondence between the actual and forecasted temporal movements.
All models reported were evaluated and investigated under the same unobserved test period, to ensure comparability.

The relationships between several numerical parameters in the supermarket sales dataset are visualized using a pair plot graph (scatterplot matrix) made with Seaborn's sns.pairplot() as shown in Fig 9. Pair plot graph.

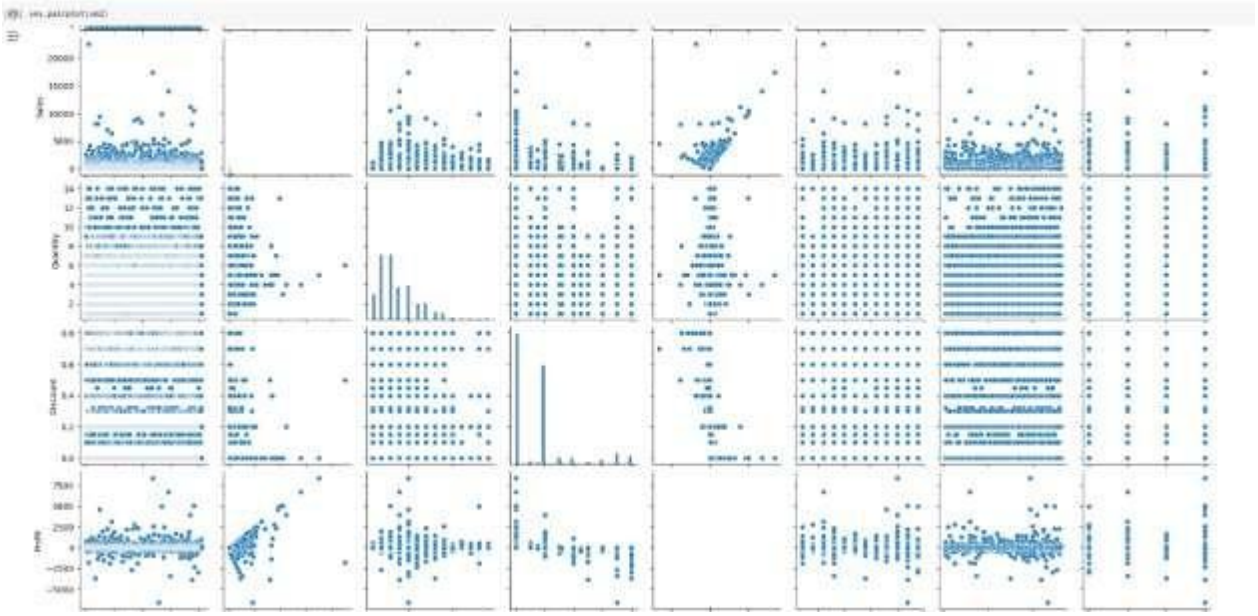


Fig 9. Pair plot graph

Sales, profit, quantity, and discount are a few of the significant numerical variables in the supermarket sales dataset that are visually represented by this pair plot. While the scatter plots show pairwise correlations, the diagonal plots show the histograms of individual variables.

A significant finding is that here is a moderate beneficial relationship between sales and profit, indicating that bigger numbers of sales do not consistently translate towards higher profit margins. Rising expenses for operations and aggressive promotional tactics are two contributing factors to this. Additionally, the marketing-discount correlation implies that there could be an adverse relationship because higher discounts rarely translate towards comparable spikes in earnings from sales. The finding that rising selling

tactics typically reduce profit margins and adversely impact total profitability is reinforced by the profit-discount research, it clearly demonstrates an overall decline.

Moreover, level lines demonstrate that reductions are utilized in specific amounts instead of steadily in the amount relative discount formulation. More selling does not always mean more money, according to a thought-provoking examination of the money vs. amount story. This may be due to market variables, such as pricing strategies, product lines or seasonal fluctuations in demand. These findings demonstrate how essential it is to correctly consider discounting strategies along with external variables into consideration when forecasting supermarket revenue and yield.
Data transformation: For instance, to produce retail statistics for machine learning algorithms, it is necessary to adjust, scale, and encrypt fraudulent information. Data must assess its findings of a retail sales prediction model.

4.2 Weekly Temporal Aggregation

This resulted in a weekly aggregation level being chosen as the granularity for forecasting – which was balanced with the focus on maintaining the information in time, reducing noise and computational efficiency. Temporal patterns can be both significant and highly variable at the transaction level or at a daily level, but can be masked by too high aggregation. A weekly level of granularity is an intermediate level of granularity that provides sufficient detail to capture the recurring demand behaviour without being overwhelmed by a lot of transactions.

This same aggregation also yields a consistent sequence of time to model with the AI for lag-based features and checking sequential dependencies. Weekly forecasting is not too coarse to be useful for inventory replenishment, but not too fine to be useful for short-term planning, and is sufficiently fine to be useful in detecting changes in demand. Most importantly, all the forecasting models compared have been run at the same time resolution, which is important for comparing the performance of the models and isolating differences in performance in terms of time aggregation and approach to learning.

4.3 Hyperparameter Optimization in AI Model Development

Hyperparameter optimization is a crucial part of AI model development because not only does the algorithm play a pivotal role, but the performance of the model can also be influenced by the hyperparameters used. Parameters that specify the complexity of the model, regularization and learning behaviour can influence the generalization and computational needs of the model. Random Forest has several important parameters such as number of trees, maximum depth of trees, minimum number of samples needed to split and number of features to consider at each split. The main parameters of the Ridge Regression are the regularization parameter, while the SGD Regressor is affected by the learning rate, the strength of the regularization and the number of iterations used for optimization.

Hyperparameter tuning needs to be done on the training set only, to prevent information leakage from the test set. In case of time series forecasting, the temporal order should remain intact also in the validation process, rather than using an arbitrary random shuffle. This allows for model configurations which are able to generalize to future observations, and not just memorize historical observations.

4.4 Hyperparameter Optimization in AI Model Development

Hyperparameter optimization is a crucial part of AI model development because not only does the algorithm play a pivotal role, but the performance of the model can also be influenced by the hyperparameters used. Parameters that specify the complexity of the model, regularization and learning behaviour can influence the generalization and computational needs of the model. Random Forest has several important parameters such as number of trees, maximum depth of trees, minimum number of samples needed to split and number of features to consider at each split. The main parameters of the Ridge Regression are the regularization parameter, while the SGD Regressor is affected by the learning rate, the strength of the regularization and the number of iterations used for optimization.
Hyperparameter tuning needs to be done on the training set only, to prevent information leakage from the test set. In case of time series forecasting, the temporal order should remain intact also in the validation process, rather than using an arbitrary random shuffle. This allows for model configurations which are able to generalize to future observations, and not just memorize historical observations.

4.5 A technical analysis of the performance of forecasting

A technical analysis of the performance of forecasting. The forecasting performance of the six models has been summarized in Table X for a test period that was never seen during the forecasting process. Ridge Regression produced the lowest forecasting errors, with MAE = 5,009.71, RMSE = 6,580.95, MAPE = 37.17%, and R^2 = 0.249. The second lowest RMSE was obtained by SGD Regressor with the value of 6,777.56

while the next lowest RMSE value was 6,933.79 by Random Forest. LSTM, ARIMA (1,1,1), and Transformer produced RMSE values of 7,570.90, 7,578.76, and 7,737.75, respectively. It will be useful to understand these differences by analysing their interplay with the model complexity, temporal representation, regularization and effective sample size.

4.5.1 Ridge Regression

The lowest RMSE (6,580.95) was obtained using Ridge Regression. It has the advantage that it can regularize the coefficients when the lag features are correlated [7]. The result indicated that the engineered temporal features contained a significant linear pattern from the weekly sales data, which makes it unnecessary to use more intricate nonlinear models.

4.5.2.SGD Regressor

SGD Regressor was able to get RMSE = 6,777.56 which was very close to Ridge Regression. Both have regularized linear representations, while SGD is iterative and is more prone to the sensitivity of the features and convergence and scaling parameters [8]. Its slightly higher error means that optimisation in this relatively small weekly data-set isn't clearly evident.

4.5.3 Random Forest

Random Forest obtained RMSE = 6,933.79 and R^2 = 0.167. It has nonlinear modelling capabilities which can be used to model complex feature interactions [6] but after the construction of lags 125 weekly observations were left. This small sample size could thus lead to other biases in exploiting this added flexibility, revealing the bias–variance trade-off.

4.5.4 LSTM

LSTM achieved RMSE = 7,570.90 and R^2 = 0.006. The goal of the proposed models was to account for sequential dependencies [9, 16] but weekly aggregation resulted in a significant reduction of the number of training sequences. This increased error rate indicates that there was not enough temporal sampling to take full advantage of the increased complexity of the model.

4.5.5 Transformer

Transformer produced RMSE = 7,737.75 and R^2 = −0.038. Although it is based on real temporal sequences, given the high representational capacity, there is a need for sufficient sequential data [10, 14]. The result shows that there is no general disadvantage of Transformer-based forecasting, and that the model is either too complex or the sample size is too small.

4.5.6 ARIMA

ARIMA (1,1,1) achieved RMSE = 7,578.76, MAPE = 44.27%, and R^2 = 0.004, despite the highest correlation (0.602). This indicates it was able to record some of the timing, but not the top of the forecast size. A first order differencing might not have been necessary due to the ADF test ($p < 0.001$) indicating stationarity and consideration of ARMA (d=0) and also of seasonal SARIMA models should be considered in future analysis.

5. Model Evaluation and Performance Measures

To ensure the precision and dependability of a retail sales prediction model, its performance needs to be examined. The capacity of the model to forecast future revenues will be assessed by employing several types of numerical and error-based actions. The mean standard error, root mean squared error, and mean percentage error are all commonly employed measurements for evaluating performance.

5.1 Mean Absolute Error (MAE) :

MAE is a simple and easy to comprehend metric that measures the average absolute value of the differences between actual and predicted sales. It provides an error amount in the same unit as the target variable. However, MAE does not penalize large forecasting errors as much as other measures :

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

where:

- y_i = actual sales value
- $\hat{y}_i$ = predicted sales value
- n = total number of observations

5.2. Mean Squared Error (MSE) :

MSE punishes over- and under-estimations disproportionately by squaring the deviations. It is more sensitive to outliers than MAE and thus often a better overall indicator of predictive power. Taking the square root converts the final output back into the original measurement scale of the dependent variable for simpler analysis:

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

Better model performance has been demonstrated by lower MSE values.

5.3. Root Mean Squared Error (RMSE)

RMSE indicate better model fit to past observations and greater proficiency in predicting future sales figures. Together, these error metrics provide meaningful insights into a forecasting model's accuracy and how well it limits deviations from real market movements:

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

When large errors need to be criticized with greater severity, RMSE is beneficial.

5.4. Mean Absolute Percentage Error (MAPE)

As MAPE estimates the percentage of error among actual and expected sales, it can be useful for comparing models across various data sets:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

Lower MAPE values indicate improved predicting accuracy.

5.5. R-Squared (R^2) Score

R^2 measures how well the independent variables explain the variance in sales:

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

where $\bar{y}$ is the mean of actual sales values. An R^2 value closer to **1** indicates a better fit.

To prevent overfitting evaluation through many information cross validation approximates needs to work properly.

Efficiency for retail products sales prediction must be evaluated properly and parameter used to achieve it needs to measure carefully. MAPE and R2 implied for Model fit and the non-negative size of difference between true value and measured value to evaluate MSE, RMSE and MAE will be use. Cumulatively this above mentioned all parameters used for better performance forecasting.

The comparison of Deep Learning. An additional deep-learning model, an LSTM model, was used to check if improvement in model complexity gives measurable forecasting advantages. LSTM was chosen since it has a gated recurrent structure that is better suited to learn the temporal dependency among sequential observations. In order to facilitate a consistent comparison, all models were evaluated with the same temporally ordered evaluation framework and performance measures.

The models have different accuracies and different computational complexity as indicated in Table X. After running LSTM, compare the results to the following sentences: The results show that the choice of an AI forecasting method should not only be based on model complexity, but also on the link between predictive performance and data characteristics/computational requirements.

Cross Validation Metrics Comparison of machine learning models with traditional statistical models Model – My research work, implemented in google colabs using Jupyter notebook python libraries using above mentioned database.

Table. 2. Forecasting performance across models

Model	MAE	RMSE	MAPE (%)	R^2	Correlation
Ridge Regression	5,009.71	6,580.95	37.17	0.249	0.530
SGD Regressor	5,167.27	6,777.56	38.95	0.204	0.491
Random Forest	5,620.14	6,933.79	41.58	0.167	0.468
LSTM	5,836.60	7,570.90	38.48	0.006	0.414
ARIMA (1,1,1)	6,026.97	7,578.76	44.27	0.004	0.602
Transformer	6,155.99	7,737.75	42.11	−0.038	0.394

The new results confirm that the task of forecasting did not benefit from the complexity of the model in all cases. Ridge Regression produced an MAE of 5,009.71, RMSE of 6,580.95, MAPE of 37.17%, and R^2 of 0.249. The next lowest RMSE was given by SGD Regressor with RMSE value of 6,777.56 and followed by Random Forest with RMSE value of 6,933.79.

Overall, Ridge Regression shows promise in its ability to generalize a linear representation in the presence of correlated temporal information in several lagged variables. L2 regularization puts bounds on the magnitude of the coefficients and may alleviate the instability that can be induced by correlated historical predictors.
Random Forest has higher nonlinear flexibility, but the number of useful training samples after constructing 52-week lags was 125. The relatively few timesteps used in the temporal sample limit the information which can be used for learning complex interactions between the trees. Hence, there was no correlation between extra nonlinear capacity and reduced out-of-sample error.
The results show that the complexity of the models has to be understood in terms of the amount and structure of the available temporal information.

6. The comparison of Deep Learning

An additional deep-learning model, an LSTM model, was used to check if improvement in model complexity gives measurable forecasting advantages. LSTM was chosen since it has a gated recurrent structure that is better suited to learn the temporal dependency among sequential observations. In order to facilitate a consistent comparison, all models were evaluated with the same temporally ordered evaluation framework and performance measures.

The models have different accuracies and different computational complexity as indicated in Table 2. After running LSTM, compare the results to the following sentences: The results show that the choice of an AI forecasting method should not only be based on model complexity, but also on the link between predictive perf6. Comparative analysis of models.

Retail sales forecasting helps in Inventory planning, Demand management and reducing waste. Depending on the characteristics of the data, the accuracy required and computational resources, different statistical, machine learning (ML) and deep learning (DL) models are commonly used.

6.1 Univariate Forecasting

Univariate analysis deals with predicting a future value of one variable, based on historical values of the same variable and accounts for such trends as seasonality and trend. Some of the popular techniques used are ARIMA/SARIMA model, ETS model, moving averages model, and Prophet. Such models are easy to understand and explain, but are not very good at predicting outcomes when other factors have a significant impact on demand.

6.2 Multivariate Forecasting

Multivariate forecasting uses historical sales data as well as other data such as price, promotions, holidays, inventory and weather. Random Forest, Gradient Boosting, SVR, RNN and LSTM are some of the methods that can capture complex and nonlinear demand. These models can give better predictions, but take more data, engineering and computational power.

Therefore, the comparison of Univariate and Multivariate helps us to find appropriate models for AI based Retail Forecasting and Automated Model Selection.ormance and data characteristics/computational requirements.

6.3 Comparative Analysis in Research:

The comparison of two Univariant and Multivariant approach foreground competency and constraints of both methods.as univariate approach require minimum data and illustratable univariant model are used extensively as baseline forecasting strategies in retail sales forecasting. It apprehends Seasonality, trends and pattern to forecast metric in different domains. Where as multivariant approach to retail sales forecasting combine historical sales data accompanying multiple external variables to enhance accuracy and resilience of demand forecasting. Along with past sales, this method typically adds features such as promotional seasonality indicators, intensity, discounts, holidays, pricing, inventory levels, weather conditions, and macroeconomic variables, there by skilfully crafted complexity of consumers spending patterns.

Univariant approach best suited for single variable based on historical data pattern and multivariant approach better while considering multiple variables from retail sales data.

7. Discussion and Limitations

7.1 Discussion:

The results show that the performance of the forecasts is not only dependent on the complexity of the algorithm, but on the compatibility of the model with the data. The lowest RMSE value obtained was with Ridge Regression (6,580.95), followed by SGD (6,777.56) and Random Forest (6,933.79), with LSTM, ARIMA and Transformer having higher errors.

L2 regularization and correlated lag features helped to improve the performance of Ridge, while the small number of effective samples (125) limited the benefits of nonlinear and deep-learning techniques. The higher representational capacity of LSTM and Transformer may have necessitated more sequential data for them to make use of.

ARIMA was able to pick up some temporal movement (correlation = 0.602) but it had a fairly high error of forecast. The result of the ADF test showed that the data were stationary ($p < 0.001$), so other alternative models such as ARMA models and SARIMA models should also be explored.

7.2 Limitation

There are a number of limitations to the study. It is limited in generalizability as it is based on one four-year retail data set. The weekly aggregation and the construction of 52-week lag, reduced the effective sample size to 157, of which 125 were used for modeling, possibly impacting deep-learning models. Weather, economic, promotions and competitor pricing were not available.

Also, broader searches for hyperparameters may also yield different LSTM and Transformer results. The ARIMA(1,1,1) is just one possible statistical model; other possible ARMA/SARIMA models need to be considered. Lastly, the study only accounts for the 1-week-ahead forecasting and results might vary with other forecasting periods.

Overall, the more complex models did not necessarily perform better with respect to forecasting. Therefore, data availability, the temporal structure, regularization, the interpretability, computational cost, and predictive accuracy need to be taken into account for model selection.

8. Conclusion

In this study, the authors have been able to present a controlled comparison where the ARIMA method was compared with the other five models mentioned above; the models Random Forest, Ridge Regression, SGD Regressor, LSTM, and Transformer, for forecasting the retail sales values for the week. The experimental setup featured leakage-aware historical features, separation of trains per time, time-series validation, training-only preprocessing, and a common unseen period for testing, to ensure a fair assessment between different forecasting paradigms.

Ridge Regression performed with the minimum forecasting error, and its results were: MAE = 5,009.71, RMSE = 6,580.95, MAPE = 37.17%, R^2 = 0.249. But, the main conclusion is not that Ridge Regression is always better than more sophisticated forecasting methods. Instead, the results indicate that the higher the architecture, the less likely it is to produce better out-of-sample forecasting accuracy if the temporal sample size is small.

The results suggest that when the historical lag variables have predictive information that is correlated in time, temporal representation and regularization can be significant. Random Forest, LSTM and Transformer, on the other hand, need abundant data to provide accurate estimates of nonlinear and sequential relationships. The ARIMA results also highlight the importance of taking care in the specification of statistical models and their stationarity properties when making comparisons in performance.

In a practical sense, a retail forecasting system should choose models based on their proven forecasting accuracy, on the availability of data from the past, on their interpretability, on the complexity of the computing time required and on the complexity of deployment, rather than based on the sophistication of the algorithm used.

The proposed framework needs to be tested on multiple independent retail datasets, with various time-series forecasting horizons, rolling-origin validation, more comprehensive external variables, different ARMA/SARIMA specifications and more extensive deep-learning hyperparameter optimization. More temporal data should also be explored and multiple related product or store series to evaluate if larger sequential data sets result in better performance of LSTM and Transformer architectures relative to each other.

9. References

1. [1] Wen, K.Y.; Joseph, M.H.; and Sivakumar, V. (2024). Big Mart sales prediction using machine learning. EAI Endorsed Transactions on Internet of Things, 10.
2. [2] Chan, H.; and Wahab, M.I.M. (2024). A machine learning framework for predicting weather impact on retail sales. Supply Chain Analytics, 5, 100058.
3. [3] Nasseri, M.; Falatouri, T.; Brandtner, P.; and Darbanian, F. (2023). Applying machine learning in retail demand prediction—A comparison of tree-based ensembles and long short-term memory-based deep learning. Applied Sciences, 13(19), 11112.
4. [4] Elalem, Y.K.; Maier, S.; and Seifert, R.W. (2023). A machine learning-based framework for forecasting sales of new products with short life cycles using deep neural networks. International Journal of Forecasting, 39(4), 1874–1894.
5. [5] Wang, C.H.; and Gu, Y.W. (2022). Sales forecasting, market analysis, and performance assessment for US retail firms: A business analytics perspective. Applied Sciences, 12(17), 8480.
6. [6] Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.
7. [7] Hoerl, A.E.; and Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55–67.
8. [8] Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. Proceedings of COMPSTAT 2010, 177–186.
9. [9] Hochreiter, S.; and Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.
10. [10] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; and Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.
11. [11] Bergmeir, C.; Hyndman, R.J.; and Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. Computational Statistics & Data Analysis, 120, 70–83.

12. [12] Makridakis, S.; Spiliotis, E.; and Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. PLoS ONE, 13(3), e0194889.
13. [13] Makridakis, S.; Spiliotis, E.; and Assimakopoulos, V. (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. International Journal of Forecasting, 36(1), 54–74.
14. [14] Lim, B.; Arik, S.O.; Loeff, N.; and Pfister, T. (2021). Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4), 1748–1764.
15. [15] Box, G.E.P.; Jenkins, G.M.; Reinsel, G.C.; and Ljung, G.M. (2015). Time Series Analysis: Forecasting and Control, 5th ed. John Wiley & Sons.
16. [16] Hewamalage, H.; Bergmeir, C.; and Bandara, K. (2021). Recurrent neural networks for time series forecasting: Current status and future directions. International Journal of Forecasting, 37(1), 388–427.
17. [17] Kingma, D.P.; and Ba, J. (2015). Adam: A method for stochastic optimization. Proceedings of the 3rd International Conference on Learning Representations (ICLR).
18. [18] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.