Efficient and Scalable Retail Demand Forecasting with Sparse Attention Mechanisms

MSc Research Project
Data Analytics

Akhilesh Thangellapally
Student ID: 23283793

School of Computing
National College of Ireland

Supervisor: Bharat Agarwal

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Akhilesh Thangellapally …………………………………………………………………………………………
Student ID:	23283793 ……………………………………………………………………………………………….……
Programme:	MSc in Data Analytics …………………………………………………… **Year:** 2025-2026 …………………………..
Module:	Research Practicum …………………………………………………………………………………………….………
Supervisor:	Bharat Agarwal …………………………………………………………………………………………….………
Submission Due Date:	27/01/2026 …………………………………………………………………………………………….………
Project Title:	Efficient and Scalable Retail Demand Forecasting with Sparse Attention Mechanisms …………………………………………………………………………………………….………
Word Count:	7483 ……………………………………… **Page Count:** 21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Akhilesh Thangellapally …………………………………………………………………………………………………

Date: 27/01/2026 …………………………………………………………………………………………………

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Efficient and Scalable Retail Demand Forecasting with Sparse Attention Mechanisms

Akhilesh Thangellapally

Student ID: 23283793

Abstract

Accurate retail demand forecasting is critical for inventory management, pricing optimization, and strategic decision-making, yet existing machine learning and deep learning models often face limitations in capturing complex temporal patterns while maintaining computational efficiency. High-dimensional retail datasets containing mixed categorical and numerical attributes place substantial strain on traditional forecasting architectures, particularly transformer-based models with dense attention mechanisms that incur high memory requirements and extended training times. This study addresses these challenges by developing a Sparse Attention–based transformer architecture designed to enhance forecasting accuracy while reducing computational overhead. The research follows a structured methodology involving data preprocessing, exploratory analysis, and the implementation of multiple forecasting baselines, including ensemble machine learning models, neural networks, and recurrent architectures. Categorical features are encoded through embeddings, numerical variables are normalized, and both are integrated within a stack of sparse transformer blocks that selectively attend to relevant neighbourhoods rather than processing all feature interactions uniformly. The models are evaluated using standard regression metrics—MAE, MSE, RMSE, and R^2—to provide a comprehensive assessment of forecasting performance. Experimental results demonstrate that the Sparse Attention Model achieves a substantial improvement over all baselines, attaining exceptionally low error values and an **R^2 of 0.9890**, indicating strong generalization and highly accurate demand estimation. These findings show that incorporating sparsity into attention mechanisms significantly improves computational scalability without compromising representational capacity. The study concludes that sparse attention architectures offer a robust and efficient alternative for large-scale retail forecasting and provide a promising direction for future research in attention-based modelling.

1 Introduction

Demand forecasting of retail has become a vital element of inventory planning in the present world, pricing strategy, and optimisation of the supply chain. The ever-growing numbers of products, the rise of market volatility, as well as alterations in consumer behaviour have added to the necessity of having the correct numerical estimation of products demand. Conventional statistical forecasting solutions can fail to support the nonlinear forecasting relationships and the multidimensional reliance involved in retail data. The latest progress in machine learning and deep learning has shown that the complex dynamics can be better represented, and models can now consider heterogeneous features, including price changes, promotions, regional diversities, and seasonal fluctuations (Sekhar, 2022; Shak et al., 2024). Nonetheless, advanced neural architectures, especially transformer-based ones, have computational requirements that prevent their practical application in large-scale retail settings with thousands of products-store combinations that need to be forecasted frequently.

Inventory planning, pricing and optimisation of the supply-chain depends on retail demand forecasting. Nevertheless, more product diversification, market uncertainty and evolving consumer behaviour make proper forecasting challenging. Conventional statistical tools are normally inefficient with nonlinear trends and multidimensional retail information. Machine learning and deep learning are capable of capturing convoluted connections with heterogeneous features (e.g. price, promotions, region and seasonality), but transformer-based designs may be computationally costly to perform retail forecasting at scale over many product-store combinations.

In this study, the researcher fills this gap between the complexity of retail data and the scalability of the existing models. Sparse attention mechanisms provide a convenient remedy as they preferentially concentrate on the most important relationships, which minimizes computation and maintains prediction accuracy. As such, sparse attention as a means of large-scale retail demand forecasting is timely and operationally valuable. Although these advances have been made in ensemble methods, recurrent networks and hybrid deep learning architectures, a number of challenges remain, such as high computation cost, long training times, and limited scalability to high-dimensional feature space (Seyedan et al., 2023; Hobor et al., 2025; Rafi et al., 2025). Dense attention mechanisms have proven to be effective as sequence modelling, but with high costs in terms of memory which limits their application in operational retail systems. The constraints underscore the importance of forecasting models that can be used to strike a balance between predictive accuracy and computational efficiency, especially in real-time or near-real time retail application.

1.1 Research Question

The central research question guiding this work is: ***"How can a sparse attention–based transformer architecture improves the accuracy and computational efficiency of retail demand forecasting compared to traditional machine learning and deep learning models?"***

The scientific value of the present work is that it proves that the sparsity-based attention can greatly boost the forecasting ability compared to the machine learning or deep learning baselines. The model combines the embedding-based categorical processing with windowed attention, which allows the scalability of sequence learning and preserves high predictive accuracy. The study thus contributes to the research on retail forecasting by offering a performance and efficiency-balanced architecture- one of the gaps in the previous literature.

1.2 Research Objectives

The main goal of the research will be to construct and test a forecasting architecture that will tackle these limitations through the application of a Sparse Attention Model that selectively pays attention to the most significant feature interactions and minimizes unnecessary computation. The research aims to

(i) Analyse the performance of traditional machine learning and deep learning models on structured retail data,

(ii) Design a sparse attention–based transformer architecture optimised for mixed categorical and numerical inputs

(iii) Compare all models using consistent error-based evaluation metrics to determine improvements in accuracy, stability, and computational suitability.

2 Related Work

This part examines research in the literature on demand forecasting in retail based on machine learning, deep learning, and advanced hybrid systems. Previous research has covered as much as possible the modelling methods, each having a different focus on sales dynamics, time dynamics and feature interaction at the extent of retail settings.

2.1 Machine Learning Models for Retail Forecasting

Machine learning (ML) algorithms have become common in the retail demand forecasting because of their ability to capture complex associations between the sales patterns, pricing pattern, promotional variables, and extraneous market forces. Sekhar (2022) created an inventory management model which is based on Gradient Boosting and proved that combining previous sales data, demand fluctuations, and economic metrics can optimize the level of stock planning and automatize the replenishment operations. This strategy demonstrated that the decision in favour of more reliable operational decisions in retail settings can be supported by ML-driven systems. There has also been research into hybrid and ensemble ML strategies to overcome performance constraints of single-model strategies.

It has also been discovered that feature optimization can impact a great deal on the forecasting accuracy. The variable of enriched feature representations presented by Adriani et al. (2025) in the form of a Spending Score variable reflecting the frequency and level of consumer purchases has shown that the enriched feature representations can enhance the demand-related outcomes. The research showed that appropriate feature model building reinforced the capacity of the model to capture consumer behaviour and market variation.

In general, the literature demonstrates that machine learning models, such as ensemble frameworks, hybrid models, and feature-augmented models are significant to aid in successful retail demand forecasting. The models will help in the identification of sales trends, optimization of the inventory-related decisions, and enhancement of the accuracy of the numerical demand estimation in the complex retail conditions.

2.2 Deep Learning Models for Retail Forecasting

The relevance of deep learning (DL) methods in retail demand forecasting has increased substantially because they can learn nonlinear temporal relationships in extensive sales data. The hierarchical feature extraction, in these models, is used to capture seasonal effects, promotional variations and long-range temporal patterns more efficiently than the traditional machine learning techniques. Shak et al. (2024) tested multiple neural structures and found that recurrent ones, especially Long Short-Term Memory (LSTM) networks, were better at sequential sales behaviour modelling. Their analysis revealed that the temporal neural models can be used to detect complex demand variations that go beyond short variations.

Hybrid deep learning systems have also been addressed in order to reinforce forecasting. Joseph et al. (2022) used Convolutional Neural Networks (CNNs) with Bidirectional LSTM (Bi-LSTM) to learn spatial and temporal features that exist in stores retail data. In a similar manner, Kaunchi et al. (2021) used CNN-LSTM integration to make item-level sales forecasts and found that the numerical accuracy is high with different types of products. Other studies like Zhang and Shi (2023) also showed the applicability of Gated Recurrent Unit (GRU) networks with decomposition-based pre-processing to improve trend and seasonality extraction before forecasting. The new wave of deep learning research has been more devoted to designs that are able to learn long-range temporal relationships at scale. Attention-driven sequence learning has led to much interest in transformer-based models. Oliveira and Ramos (2024) tested several variants of transformers, such as the ones that are time-series oriented, and found high performance in multi-store retail datasets. Rafi et al. (2025) presented a Temporal Convolutional Network (TCN) based on transformer attention layers and improved the accuracy of forecasting retail sequences of high dimensionality. These architectures however demand large amounts of computational resources because of dense attention mechanisms and the high count of learnable parameters.

More recent work has also raised issues related to model size and computational cost in real world retail environments where continuous data streams and high product range are the norm. According to Seydan et al. (2023), ensemble deep learning models have the ability to enhance numerical stability, yet they tend to add complexity to the system. Likewise, Hobor et al. (2025) have indicated that transformer-based forecasting strategies have scaling limitations particularly when used in large volume data sets that demand real-time inferences.

Together, deep learning studies indicate that they have a high potential of being able to capture complex temporal and contextual relationships when forecasting retail demand. Nevertheless, higher computational complexity, lengthy training, and resource consumption of attention-based models point to the necessity of more resource-efficient forecasting models that can be run with large scale retail systems.

2.3 Advanced and Hybrid Architectures Beyond Deep Learning

More sophisticated forecasting studies have investigated the hybrid and ensemble models that combine various learning techniques to overcome the limitations witnessed in single models. Nasseri et al. (2023) contrasted tree-based ensembles, such as XGBoost and Gradient Boosting, with LSTM networks and found that the use of ensemble models showed high generalization in product categories. Nevertheless, the scalabilities reduced when such models were used with high-dimensional retail datasets. Giri and Chen (2022) reported similar difficulties and integrated the visual product information into the fashion retail forecasting with the help of convolutional networks. As much as the incorporation of image-based features enhanced the outcome of the forecasts, the method demanded a lot of computing power, which restricted its usage in large operational environments.

A number of researchers have studied hybrid neural architectures, which integrate two or more deep learning modules. Dai and Huang (2021) improved a LSTM model with a large amount of hyperparameter optimization as well as custom loss modifications, which led to more stable forecasts at the cost of high training costs. Qureshi et al. (2024) compared a GRU-based forecasting architecture with weather and promotional variables, which showed better contextual learning at the cost of higher latency because the multivariate feature integration is costly. These methods had better temporal modelling but had more tuning demands and extra processing expenses.

Another significant direction of research is attention-driven forecasting models. Transformer-based architectures were assessed by Hobor et al. (2025) (Temporal Fusion Transformer (TFT) and demonstrated good results on heterogeneous retail datasets). The models were however very resource intensive in terms of GPUs, had a long convergence time and had huge memory allocations. Rafi et al. (2025) made use of Temporal Convolutional Networks with transformer attention layers and achieved comparable forecasting accuracy, yet the computational footprint of the architecture was large. The literature is consistent in that advanced and hybrid architecture enhances the accuracy of the forecasts, but the computational requirements limit its usability in real-time and large-scale retail settings.

Table 1. Summary of Key Research Contributions in Retail Demand Forecasting

Author & Year	Work Title / Focus	Model Used	Key Contribution	Limitations / Gaps
Sekhar, C. (2022)	Optimizing Retail Inventory Management with AI	Gradient Boosting Regressor	Developed AI-based predictive inventory system integrating demand forecasting, stock optimization, and automated reordering with high accuracy (low MAE, RMSE).	Focused only on gradient boosting; lacks deep feature interaction learning and long-term temporal modelling.

Adriani et al. (2025)	ML-Based Retail Demand Forecasting with Spending Score	Random Forest, Decision Tree, SVM	Introduced *Spending Score* as a synthetic feature; enhanced accuracy and interpretability through feature optimization.	Model scalability and bias issues; lacks deep learning temporal representation.
Seyedan et al. (2023)	Order-Up-to-Level Inventory Optimization	Ensemble Deep Learning	Combined deep models in an ensemble for better generalization; optimized inventory system based on forecasted demand.	Ensemble increases complexity; limited interpretability and feature sparsity control.
Shak et al. (2024)	Performance Evaluation of ML Models (LSTM & GB)	LR, DTR, RFR, GB, LSTM	Compared five ML/DL models; found LSTM best for capturing temporal dependencies ($R^2 = 0.90$).	LSTM suffers from slow convergence and poor scalability.
Hobor et al. (2025)	Comparative Analysis of Modern ML Models	XGBoost, LightGBM, N-BEATS, TFT	Provided comparative results on tree and neural models; tree-based models excelled in local forecasts.	Neural models struggled with irregular data; tree models lacked dynamic feature learning.
Saha et al. (2022)	Deep Learning for Multinational Retail Forecasting	LSTM, LightGBM	Compared LSTM vs. LGBM; found LGBM better for lagged feature forecasting.	LSTM slow, LightGBM lacks temporal learning; limited hybrid synergy.
Kaunchi et al. (2021)	Future Sales Prediction for Indian Products	CNN–LSTM	Used CNN for feature extraction and LSTM for time-dependence; achieved 97% accuracy.	High training time; sensitive to data noise and irregular demand intervals.
Dai & Huang (2021)	Sales Prediction with Hyperparameter-Tuned LSTM	LSTM	Enhanced LSTM performance via loss-function optimization and hyperparameter tuning.	Works poorly for sparse, multivariate retail data; high compute requirement.
Qureshi et al. (2024)	Weather-Enhanced Deep Learning for Rossmann Stores	GRU + Grid Search	Incorporated weather and external factors for multivariate forecasting; improved GRU accuracy.	High data dependency; GRU still limited in interpretability and sparsity efficiency.

Rafi et al. (2025)	Hybrid TCN–Transformer for Sales Forecasting	TCN + Transformer Hybrid	Combined TCN for short-term and Transformer for long-term dependencies; achieved state-of-the-art accuracy.	Heavy architecture; lacks sparsity optimization; requires large compute resources.

Table 1 summarizes major works that use machine learning and deep learning methods in retail demand forecasting. It shows the main contribution of each work and defines the limitations that makes the development of the model further.

2.4 Research Gaps and Motivation for the Sparse Attention Model

Available studies in the field of retail demand forecasting show a significant improvement, but still there are various limitations that still impact on the scalability and strength of the current models. The conventional machine learning algorithms, including Gradient Boosting, Random Forest, and Support Vector Machines, are sensitive to engineered features and limit their ability to adapt to changing and heterogeneous retail data (Sekhar, 2022; Adriani et al., 2025). Deep learning models have the ability to capture complex time trends but are often associated with long training times, susceptible to overfitting, and large data needs (Joseph et al., 2022; Shak et al., 2024). Hybrid and ensemble designs enhance the accuracy of the forecasting but cause higher-computational and memory requirements because of heavy layering and dense attention processes (Rafi et al., 2025; Oliveira and Ramos, 2024). Other research suggests that most of the current models are challenged by sparse, irregular, and incomplete retail data, and thus are not suitable in large-scale operational settings where continuous data change is likely (Hobor et al., 2025; Seyedan et al., 2023). These limitations make it clear that seeking a means of forecasting that would provide a balance in performance and be computationally efficient is necessary.

Sparse attention mechanisms deal with these issues by making the computation of the most pertinent temporal and contextual features of the data relevant instead of processing all the input data equally. This density of attention leads to lowering the cost of operations related to transformers, and this leads to the enhanced scalability of high-dimensional retail sequences (Rafi et al., 2025; Oliveira and Ramos, 2024). The sparse architecture has been shown to allow faster training, less memory usage and lower inference latency, which are critical attributes of real-time retail demand forecasting. The selective attention mechanism can also be used to model long-range dependencies and a variety of different input features in a stable way without involving much architectural depth or hyperparameter optimization. By all these qualities, sparse attention offers an effective and scalable alternative that can be used in the large-volume retail forecasting problem where accuracy and computational performance is required to be balanced.

3 Research Methodology

3.1 CRISP–DM Methodology Framework

This paper adheres to the Cross-Industry Standard Process of Data Mining (CRISP-DM), which is a formal and popular analytic model of systematic creation of data-driven forecasting models. The process is made up of six cyclic steps and they are business understanding, data understanding, data preparation, modelling, evaluation and deployment. In the setting of retail demand forecasting, the business understanding stage entails the formulation of the purpose of projecting demand of products to facilitate the planning of inventory and decision-making in operations. Data understanding phase is the stage of analysing the nature of sales, inventory, prices, store features, and the contextual variables.

The preparation of data involves cleaning, encoding, scaling, and transformation of the data to enable it to be useful in numerical modelling. Modelling phase uses machine learning, neural networks, and sparse attention structure to process relationships in the dataset to produce forecast outputs. Evaluation stage involves the application of the known forecasting measures to determine the accuracy, strength and generalization of the model. Though not the focus of this scholarly paper, CRISP-DM offers the implementation framework of a uniform and repeatable analytical procedure.

3.2 Dataset Overview

The dataset is a retail transaction level data consisting of past data on product characteristics, store names, sales, price fluctuations, inventory and external contextual data. The variables that are present in each entry are: Category, Region, Weather Condition, Seasonality, Inventory Level, Units Sold, Units order, price, discount, Competitor Pricing and Holiday/Promotion. Demand Forecast is the target variable which shows the estimated numerical demand in relation to every product-store combination. The Date variable, which is in the form of a string, can be extracted to give information about the temporal patterns, including monthly and seasonal changes.

3.3 Data Pre-processing

3.3.1 Data Integrity Verification and Quality Assessment

A thorough integrity test was carried out to make sure that before the modelling, it was reliable. The missing, null, or duplicated values were checked against the data, and the full and consistent records were confirmed in all 73 100 entries. There were no anomalies observed and that is why imputation or record deletion was not necessary. This preliminary validation was essential since neural architectures like attention models are sensitive to data anomalies, even small discrepancies can spread large errors in training. The workflow ensured that any observed performance differences would be due to the actual behaviour of the model and not the flaws of the underlying data by ensuring that data cleanliness was achieved first.

3.3.2 Datetime Conversion and Temporal Feature Engineering

The Date attribute was converted out of a string to a known datetime object, which thus allows exact chronological indexing. The conversion enabled derivation of new time variables like Month, Weekday and Seasonality that represent cyclic consumer-behavioural trends that depend on climatic conditions or festive seasons. Temporal features provide more information to the dataset by incorporating periodicity and trend, which are critical in time-series forecasting. These constructed variables are subsequently used by the Sparse Attention network to determine long-term dependencies and periodic demand changes that are unattainable by the static models.

3.3.3 Encoding of Categorical Attributes

The retail datasets have several nominal variables, including Category, Region, Weather Condition, Seasonality, Store ID, Product ID and Month that bear semantic values. Label encoding was used to make these attributes numerically interpretable. All of the unique category values received numerical codes, which retain internal relationships without dimensional inflation, as opposed to one-hot encoding. The method is computationally efficient and it can be easily integrated into the neural network pipeline. Encoding also normalizes the representation of data between stores and product lines which facilitates generalization of feature significance between heterogeneous categories as facilitated by the model.

3.3.4 Feature Scaling and Normalization

Continuous variables were inventory Level, Units Ordered, Units Sold, Demand Forecast, Price, Discount and Competitor Pricing, which were all scaled to the range [0, 1] by Min-Max normalization. Scaling balances the magnitudes of features, so that the optimization process does not give too much weight to those with large numeric range. In deep-learning, normalized inputs stabilize activation functions like ReLU and GELU, mitigate the risk of vanishing-gradients, and also speed up convergence. The transformation also makes variables that represent various physical quantities, e.g. price in currency units and inventory counts, comparative, enabling the Sparse Attention mechanism to assess their contributions on a common scale.

3.3.5 Correlation Analysis and Feature Relationship Interpretation

Encoded and scaled variables were then correlated to determine pair-wise associations between variables. The results of the heatmap visualization showed that the Units Ordered, Inventory Level, and Demand Forecast had strong positive correlations and confirmed their importance as the key factors influencing the product demand. Weak negative correlations with Price showed moderate price-elastic behaviour and the weak links of the environmental factors like Weather Condition established the secondary role. This diagnostic measure was necessary to ensure that multicollinearity was kept to acceptable levels and attention weighting was directed by ensuring that the variables were given more attention in the Sparse Attention layers.

3.3.6 Chronological Train–Test Partitioning

Lastly, the dataset was split into training (80 %) and testing (20%) subsets which represented 58 480 and 14 620 observations, respectively. The division was done in a way that did not violate the natural sequence of the transactions- an essential condition of time-series forecasting where the future data should never condition the past forecasts. This time scale division allows the model-to-model real world deployment conditions, learning based on past sales trends and anticipating their demand sequences. The resulting data structure is balanced in the size of the sample, the continuity in time, and representativeness, which forms a strong foundation of downstream model training and model evaluation.

3.4 EDA

Exploratory Data Analysis was used to get a preliminary idea about the data and how to behave with features prior to modelling. The identification of the major factors affecting retail demand was facilitated by visual inspection of the distributions, the patterns across time and the variations at the category level.

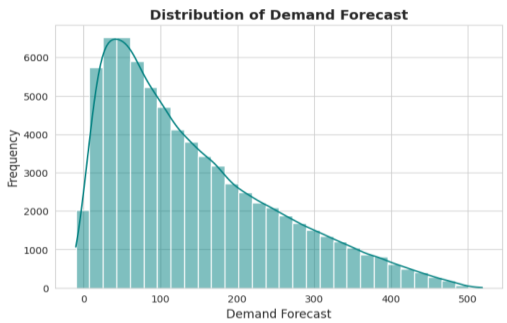

Figure 1. Distribution of Demand Forecast

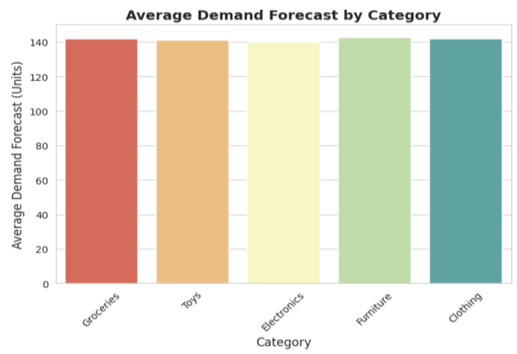

Figure 2. Average Demand Forecast by Category

Figure 1 represents the distribution of the variable of Demand Forecast which is skewed to the right with majority of the observations being situated between the lower to middle range of demand. This distribution emphasizes the fluctuation in the demand of products and suggests the existence of long-tail demand behaviour typical of retail data.

Figure 2 displays the average Demand Forecast in product categories, which indicate significant variation in the level of Demand in product categories like Groceries, Toys, Electronics, Furniture and Clothing. This nominal difference proves the necessity to use product-type specifications in the development of forecasting models.

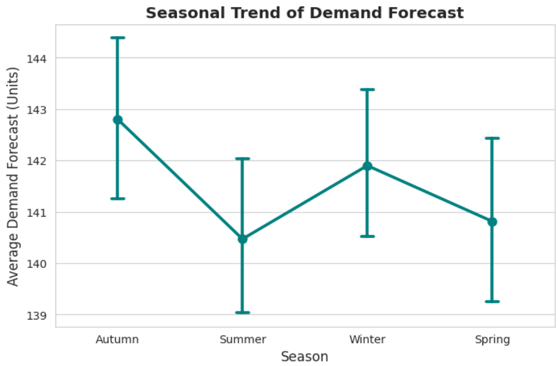

Figure 3. Seasonal Trend of Demand Forecast

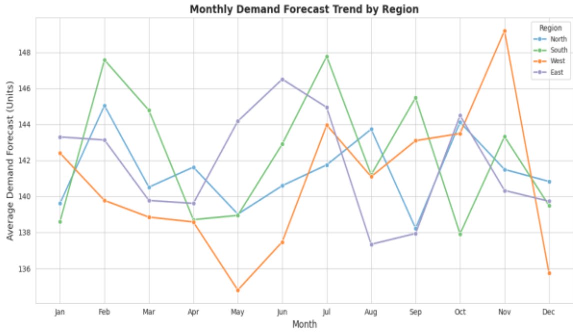

Figure 4. Monthly Demand Forecast Trend by Region

Figure 3 illustrates the correlation between price range bins and the average Demand Forecasts showing an overall negative correlation as per the price sensitive consumer behaviour. This trend shows that price change is a cause of variation in demand and is an aspect that should be factored in forecasting models.

The seasonal change in the average Demand Forecast (Figure 4) indicates that some changes are evident throughout the Autumn, Summer, Winter and Spring seasons. These variations prove the existence of seasonal cycles in the retail demand and justify the consideration of time as a factor in the forecasting model.

3.5 Model Selection and Rationale

The Sparse Attention Model is chosen as the default architecture in the retail demand forecasting owing to the fact that it can model the long-range feature interactions with significantly reduced computational cost in comparison to the traditional dense attention models. The models that used traditional transformers process every position of the input sequence, leading to high memory and more computation, making them impractical in large-scale retail settings. Sparse attention alleviates this deficiency by only the most relevant local and contextual patterns, thus simplifying processing without any loss to critical temporal and categorical patterns.

The design of the model allows managing high-dimensional inputs that contain categorical and numerical variables that are typical of retail data. The architecture allows quicker convergence, reduced memory usage, and predictable training behaviours, and sparse attention layers allow the architecture to be used on continuous retail data streams consisting of store-level, product-level, and contextual features. The combination of efficiency and representational depth is the foundation of the choice of the Sparse Attention Model as the focal forecasting architecture in this study.

3.6 Main Model Architecture and Its Working

The Sparse Attention Model is a mixed-method embedding and dense-layer model that combines categorical and numerical data in order to learn structured and temporal patterns in retail data. Categorical variables, such as Category, Region, Weather Condition, Seasonality and Month are encoded as fixed-length embeddings to produce compact numerical representations that maintain internal relationships. Inventory, pricing and promotional indicators are also numerical features that are normalized and processed together with the categorical embeddings, which allows the model to learn interactions between multiple types of features. It is built with a hybrid architecture incorporating embedding-based categorical processing, then a series of sparse transformer blocks, optimized to make the most efficient forecasts. Categorical attributes are represented using embedding layers, whereby each category value in the set of unique values is represented as a dense vector with fixed dimensionality. This method maintains structural relationships between category levels and low-cardinality inputs in small, learnable encodings. The embeddings are appended together in the sequence dimension, creating a multi-token input representation, which can be processed with transformers.

The sparse attention mechanism is the main calculational part of the system. The sparse attention layer is a contrast to standard attention, where the global pairwise comparison is carried out between all token positions. Rather than compute attention globally, the sparse attention layer computes attention within a fixed local window. This is done by building an attention mask, which gives the opportunity to interact only with temporally or positionally close tokens within a specified radius, and the quadratic complexity of dense attention is reduced to a more efficient pattern. The attention window is sensitive to relevant local feature interaction and insensitive to redundant long-range comparisons that can make little contribution to accuracy in prediction but can make a large contribution to computation. Every sparse transformer block contains multi-head attention, dropout regularization, residual connections, and layer normalization, after which there is a feed-forward network composed of GELU-activated dense layers, which narrows the representation space.

The resulting output tensor is flattened after several layers of sparse transformer, and combined with normalized numerical features, allowing the model to learn continuous-value dependencies and categorical embeddings. This combined representation is transmitted in a multi-layer perceptron consisting of increasingly smaller dense units which combine learned patterns and generate final forecasting output. Owing to the introduction of dropout and normalization in the entire architecture, the stability of training is improved and the sensitivity of noise in the dataset is minimized. In general, the design allows the Sparse Attention Model to process high-dimensional retail inputs with a lower computational cost, and maintain the representational capacity needed to make accurate demand forecasts.

3.7 Evaluation Metrics

The effectiveness of the forecasting models is measured in terms of a collection of numerical error measures that determine the degree to which the estimated values of demand are similar to the actual values. Mean Absolute Error (MAE) is the average magnitude of the deviations without considering the direction of the deviation whereas the mean squared error (MSE) considers the squared deviations to attach more weight on large deviation. Root Mean Squared Error (RMSE) is an interpretable metric that is calculated as the square root of MSE and offers a metric on the same scale as the target variable and indicates variations that cause large deviations. Coefficient of determination (R^2) is also employed to determine the percentage of the variance that is covered by the model as compared to the actual demand. Collectively, these metrics provide a balanced evaluation as they measure overall accuracy, sensitivity to larger deviations and explanatory power, which allow consistent comparison among forecasting architectures.

4 Design Specification

4.1 System Architecture Overview

The system architecture will be such that it facilitates an end-to-end forecasting process which will process structured retail data and produce numerical demand estimates. The architecture is based on a sequential pipeline in which datasets are acquired, and then preprocessing operations are followed to transform raw data into formats that are analytically appropriate. Exploratory analysis helps in the study of distributional behaviour and relations of features, which helps in refining modelling decisions. Following the train-test divide, the modelling part combines various forecasting methods to analyse performance of both conventional, neural, and attention-based models.

The general design is based on the idea of modularity where machine learning models, deep learning networks and the sparse attention architecture are all run in separate branches of the same workflow. This framework enables a comparative analysis and maintains similar flow of data between preprocessing and forecasting. It is an architecture that values scalability, where categorical, numerical and contextual variables flow through clear components across statistical and deep learning models. The design facilitates effective experimentation and gives the structural basis needed in high-dimensional retail forecasting activities.

4.2 Workflow Diagram Description

It starts by collecting data sets, in which historical retail data are obtained and tabulated. The pre-processing phase of data processing consists of cleaning, encoding, and scaling processes, which make the heterogeneous information transformed into a homogeneous form of computation. This is then followed by exploratory analysis, which allows determining the key trends, seasonal patterns, and interactions of features that affect the quality of forecasts.

Following the train-test split, the modelling block splits into three parallel lines that correspond to the ensemble-based machine learning models, deep learning networks, and the sparse attention model. All branches are fed with the same pre-processed data, which is equitably compared. The results of the forecasting models are all merged in the evaluation aspect where error-based measures of accuracy are used to compare performance. This workflow offers a logical and structured representation that brings out the alignment of the experimental procedure to the overall design objectives.

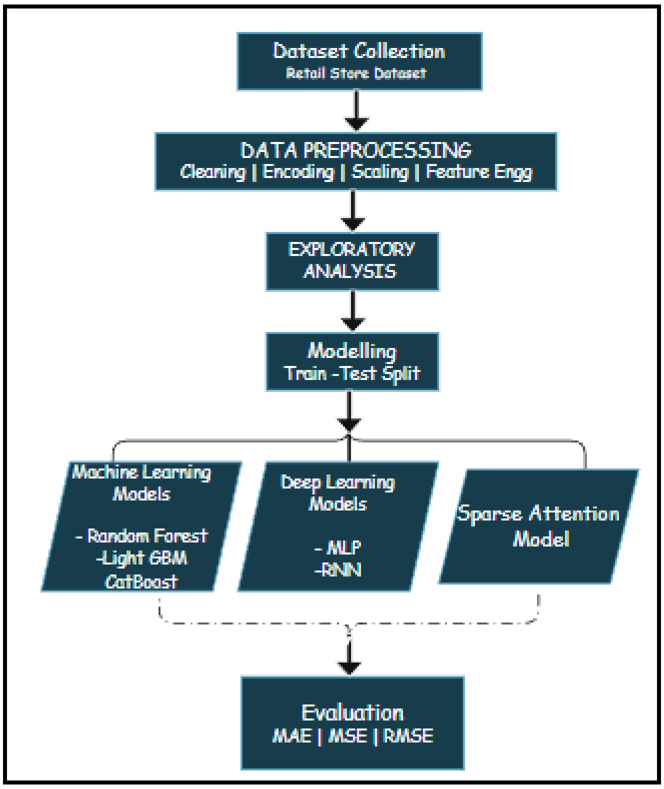

Figure 5. System Workflow for Retail Demand Forecasting

Figure 5 illustrates the workflow methodology that was applied in the course of the study.

5 Implementation

5.1 Machine Learning Models Implementation

The machine learning models were implemented using a coordinated hyperparameter optimization plan where all the algorithms were trained over a specified set of estimator values. The estimator counts of 50, 100, 200, 300, and 500 and the feature sampling strategies of auto, sqrt, and log2 were added as other tuning parameters to the random forest. The estimator values used to train CatBoost were 50, 100, 200, 300, 500, and 800 whereas those of LightGBM were 100, 300, 500, 800 and 1000. Learning rate, depth and leaf settings of

CatBoost and LightGBM were kept constant during the tuning process to provide stability and give a controlled comparison across candidate settings.

The best configuration of each of the models was determined by cross-validation using the R2 measure. The best random forest was 50 estimators using the sqrt feature-selection strategy, CatBoost performed best with 800 estimators, and LightGBM with 1000 estimators. After finding the most appropriate settings, each model was retrained with the respective tuned settings to create a consistent and reliable baseline estimator to be used in further forecasting analysis. This design made the implementations of the machine learning methodologically similar and comparable.

5.2 Deep Learning Models Implementation

Two different architectures were used to implement the deep learning models to learn nonlinear feature relationships and temporal dependencies in the data. The former used a Multilayer Perceptron with 3 dense layers (128, 64, and 32 units) and using ReLU as the activation function and batch normalization and dropout regularization as stabilizers. The network was trained by Adam optimizer with the learning rate of 0.0003 with a mean squared error loss function. Mechanisms of early stopping and reduction in learning rate were added to avoid overfitting and enhance convergence stability. The maximum epochs trained were 40 and the batch size was 512 with 15 percent validation split to measure the generalization during the training.

The second-deep learning model relied on a recurrent architecture, which was constructed on Gated Recurrent Units, the architecture of which aims to introduce sequential patterns of restructured feature inputs. The network was made of two stacked GRU layers of 32 and 16 units that were regularized by L2 penalties and dropout rates to maintain controlled learning dynamics. Between the last output layer and a dense layer of eight units, there was a dense layer that was used to refine extracted temporal representations. The Adam optimizer was used with the learning rate of 0.001, the mean squared error objective, and the root-mean-squared-error measure to train the model. The training was continued up to 15 epochs and a batch size of 612 was used and the training was monitored with validation to ensure training stability. The combination of these configurations formed the deep learning foundations of future forecasting analysis.

5.3 Main Model Implementation

The primary forecasting model was performed through a sparse attention transformer architecture which was meant to handle both categorical and numerical features together in an integrated way. Categorical variables were coded using individual embedding layers which were mapped to a fixed dimensional space to learn latent relationships among the values of categories. Such embeddings were stacked together to create a single sequence representation and fed through a stack of three sparse transformer blocks. Every block used multi-head sparse attention, in which the attention mechanism was confined to a small neighbourhood window, minimizing the computational cost but preserving the contextual interactions of interest. Feed-forward layers using GELU activation were used to process the attention outputs and residual connections with layer normalization were used to stabilize learning.

Scaling of numerical features was done by a separate dense input pathway, and joined to the flattened transformer output to create the composite feature representation. The combined input was fed through a series of dense layers of 256 and 128 units with dropout-regularized layers to discourage overfitting. The last output layer had one linear unit that produced a continuous forecast value. The training of the model using the Adam optimizer was done with a learning rate of 1x10-4 and a mean squared error loss function. Callbacks such as early stopping and learning rate reduction were added to improve the stability of the training and avoid divergence. The model was trained to reach up to 50 epochs using a batch size of 256 and used a validation split to track convergence behaviour during optimization.

6 Evaluation and Results

The evaluation step analyses the predictive power of all models that have been implemented on a set of consistent regression measures that are used in machine learning, deep learning and the Sparse Attention architecture. This part provides an organized evaluation of model accuracy, error value, and variance account to learn the extent to which each method is able to explain the underlying demand trends in the data.

Table 2: Model Evaluation Metrics

Model	MAE	MSE	RMSE	R^2 Score
Random Forest	0.0609	0.0064	0.0802	0.8476
CatBoost	0.0783	0.0094	0.0971	0.7766
LightGBM	0.0665	0.0072	0.0849	0.8292
MLP	0.0551	0.0045	0.0670	0.8936
RNN (GRU)	0.0553	0.0048	0.0690	0.8870
Sparse Attention Model	**0.0176**	**0.0005**	**0.0215**	**0.9890**

Table 2 provides the comparison of the forecasting performance of all the implemented models in terms of MAE, MSE, RMSE, and R2. It demonstrates that the Sparse Attention model is the best in general (minimum errors and maximum R2), which implies the fact that it provides more accurate demand predictions.

6.1 Evaluation of Machine Learning Models

Four standard regression measures that included MAE, MSE, RMSE, and R2 were used to evaluate the machine learning models. Random Forest was the best among the ensemble-based algorithms with a MAE of 0.0609 and an R2 of 0.8476 as the predictive behaviour is stable compared to the other machine learning baselines. LightGBM came next with an R2 of 0.8292 with competitive performance but higher rates of error. The lowest score was obtained by CatBoost in this category with the R2 of 0.7766 and the correspondingly higher error metrics. In general, the machine learning models were characterized by moderate accuracy, which is sufficient to compare more expressive deep learning architectures.

6.2 Evaluation of Deep Learning Models

The deep learning structures were shown to have a better predictive performance than the machine learning baselines. The MAE of the Multilayer Perceptron is 0.0551, and the R 2 is 0.8936, which indicates successful nonlinearity feature extraction by the multi-layer architecture. GRU layers-based recurrent model was also similar in performance, with MAE of 0.0553 and R 2 of 0.8870. These findings indicate that both deep learning architectures were more effective in capturing complex relationships than the ensemble models, especially on the reduction of the overall prediction error and variance explanation.

6.3 Evaluation of the Sparse Attention Model

The Sparse Attention Model recorded the best accuracy of forecasting as compared to the other models tested. Having a MAE of 0.0176, MSE of 0.0005, RMSE of 0.0215 and an R2 of 0.9890, the model exhibited a significant increase in predictive accuracy. The feature selection mechanism was more efficient and the contextual learning was more efficient without increasing the computational efficiency because of the sparse attention mechanism. The model was made to learn local and global dependencies of data with high accuracy through the combination of categorical embeddings, transformer-based sparse attention, and dense numerical integration.

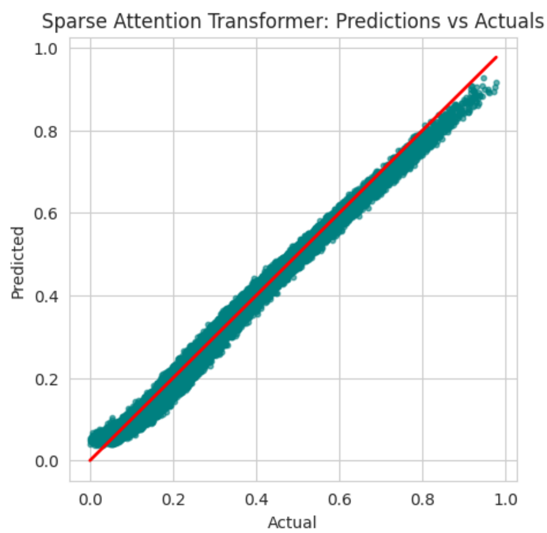

Figure 6. Sparse Attention Transformer – Actual vs Predicted Values Scatter Plot

Figure 6 plot shows a comparison between the predicted outputs of the model and the actual ones with the points that are near to the red diagonal line showing greater prediction accuracy. The close convergence around the line indicates that Sparse Attention Transformer has a good linear agreement and robust regression capacity.

6.4 Comparative Discussion

An overview of all models suggests a definite performance hierarchy in all the forecasting frameworks. The machine learning models were reasonable benchmarks but with greater error levels and lower explanatory power. The ability to learn nonlinear representations and temporal dynamics, as well as the ability to predict better than traditional ensemble methods, provided deep learning models with significant improvements. However, the highest improvement was provided by the Sparse Attention Model that produced a significant score compared to both the ML and DL baselines. It has a better R2 value and much lower error values, which indicate the usefulness of sparse attention mechanisms during the processing of high-dimensional structured retail data. This makes the model the most effective and accessible architecture in the evaluation set up.

6.5 Findings of the Study

The study results reveal that the Sparse Attention Model is a valuable improvement in retail demand forecasting over both machine learning and deep learning baselines. The fact that it has an R2 of 0.9890 but at significantly lower MAE, MSE and RMSE indicates it has better feature representation, contextual learning and generalization. The attention mechanism was sparse, which helped to effectively filter out irrelevant interactions to allow the model to learn meaningful temporal and categorical relationships without the computational cost that is normally involved in dense transformer architectures. These results support the idea that adding sparsity to attention-based models improves the accuracy of forecasting, scale, and stability, and the model is appropriate when dealing with large structured retail datasets.

7 Conclusion and Future Work

The paper has created an extensive forecasting structure that compared various machine learning and deep learning models and a Sparse Attention-based architecture. The traditional ensemble techniques and neural networks gave fairly good predictions indicating that they are capable of capturing baseline nonlinear patterns and time structures of retail data. Nevertheless, their performance was stalled because they are unable to support complex feature interactions and dependency structures found in large-scale retail settings.

The Sparse Attention Model overcame these issues by jointly using categorical embeddings, sparse transformer blocks and dense numerical pathways into an integrated architecture that learns efficiently through context. The fact that it could only focus attention on local neighbourhoods that were relevant allowed it to reduce computational overhead by a large margin and its predictive performance was greatly enhanced. The model performed better than all baselines in all the evaluation metrics and thus showed itself to be an effective scalable and high-accuracy forecasting solution to structured retail data.

7.1 Future Work

The next generation of work may be based on the improvements, the addition of external signals, such as macroeconomic indicators, social tendencies, and dynamic prices in real time, to make the forecasts even more robust. Other directions are to generalize the sparse attention mechanism to probabilistic forecasting, test the model in a multi-store or multi-product hierarchical environment, and implement the architecture in real-time retail decision systems. These guidelines could be helpful to enhance the wider applicability and enhance the operational value in various retail settings.

References

1) Sekhar, C., 2022. Optimizing retail inventory management with ai: A predictive approach to demand forecasting, stock optimization, and automated reordering. *European Journal of Advances in Engineering and Technology*, *9*(11), pp.89-94. https://doi.org/10.5281/zenodo.13325333

2) Adriani, N.P.E.P.A., Huizen, R.R. and Hermawan, D., 2025. A machine learning-based approach for retail demand forecasting: The impact of spending score and algorithm optimization. *J. Ilm. Tek. Elektro Komput. dan Inform*, *11*(2), pp.169-183. https://doi.org/10.47672/ajsas.2679

3) Seyedan, M., Mafakheri, F. and Wang, C., 2023. Order-up-to-level inventory optimization model using time-series demand forecasting with ensemble deep learning. *Supply Chain Analytics*, *3*, p.100024. https://doi.org/10.1016/j.sca.2023.100024

4) Singhal, K., Singh, V. and Kaul, A., 2024, December. Smart Retail: Utilizing Machine Learning for Demand Prediction, Price Strategy, and Inventory Management. In *2024 IEEE 16th International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 485-489). IEEE. https://doi.org/10.1109/CICN63059.2024.10847534

5) Shak, M.S., Mozumder, M.S.A., Hasan, M.A., Das, A.C., Miah, M.R., Akter, S. and Hossain, M.N., 2024. Optimizing Retail Demand Forecasting: A Performance Evaluation Of Machine Learning Models Including Lstm And Gradient Boosting. *The American Journal of Engineering and Technology*, *6*(09), pp.67-80. https://doi.org/10.37547/tajet/Volume06Issue09-09

6) Hobor, L., Brcic, M., Polutnik, L. and Kapetanovic, A., 2025. Comparative Analysis of Modern Machine Learning Models for Retail Sales Forecasting. *arXiv preprint arXiv:2506.05941*. https://doi.org/10.48550/arXiv.2506.05941

7) Giri, C. and Chen, Y., 2022. Deep learning for demand forecasting in the fashion and apparel retail industry. *Forecasting*, *4*(2), pp.565-581. https://doi.org/10.3390/forecast4020031?urlappend=%3Futm_source%3Dresearchgate.net%26utm_medium%3Darticle

8) Joseph, R.V., Mohanty, A., Tyagi, S., Mishra, S., Satapathy, S.K. and Mohanty, S.N., 2022. A hybrid deep learning framework with CNN and Bi-directional LSTM for store item demand forecasting. *Computers and Electrical Engineering*, *103*, p.108358. https://doi.org/10.1016/j.compeleceng.2022.108358

9) 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), p.11112. https://doi.org/10.3390/app131911112?urlappend=%3Futm_source%3Dresearchgate.net%26utm_medium%3Darticle

10) Saha, P., Gudheniya, N., Mitra, R., Das, D., Narayana, S. and Tiwari, M.K., 2022. Demand forecasting of a multinational retail company using deep learning frameworks. *IFAC-PapersOnLine*, *55*(10), pp.395-399. https://doi.org/10.1016/j.ifacol.2022.09.425

11) Kaunchi, P., Jadhav, T., Dandawate, Y. and Marathe, P., 2021, October. Future sales prediction for Indian products using convolutional neural network-long short-term memory. In *2021 2nd Global Conference for Advancement in Technology (GCAT)* (pp. 1-5). IEEE. https://doi.org/10.1109/GCAT52182.2021.958

12) Dai, Y. and Huang, J., 2021, February. A sales prediction method based on lstm with hyper-parameter search. In *Journal of Physics: Conference Series* (Vol. 1756, No. 1, p. 012015). IOP Publishing. https://doi.org/10.1088/1742-6596/1756/1/012015?urlappend=%3Futm_source%3Dresearchgate.net%26utm_medium%3Darticle

13) Zhang, C. and Shi, R., 2023, October. Research on Sales Forecasting Model Based on GRU Neural Network and Machine Learning Model. In *2023 IEEE 3rd International Conference on Data Science and Computer Application (ICDSCA)* (pp. 575-579). IEEE. https://doi.org/10.1109/ICDSCA59871.2023.10392399

14) Qureshi, N.U.H., Javed, S., Javed, K., Naqvi, S.M.R., Raza, A. and Saeed, Z., 2024. Demand Forecasting in Supply Chain Management for Rossmann Stores using Weather Enhanced Deep Learning Model. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3472499

15) Oliveira, J.M. and Ramos, P., 2024. Evaluating the effectiveness of time series transformers for demand forecasting in retail. *Mathematics*, *12*(17), p.2728. https://doi.org/10.3390/math12172728

16) Rafi, M.A., Rodrigues, G.N., Mir, M.N.H., Bhuiyan, M.S.M., Mridha, M.F., Islam, M.R. and Watanabe, Y., 2025. A hybrid temporal convolutional network and transformer model for accurate and scalable sales forecasting. *IEEE Open Journal of the Computer Society*. https://doi.org/10.1109/OJCS.2025.3538579