Leveraging Self-Attention in Transformer Models for Improved Retail Demand Forecasting

MSc Research Project
MSc in Data Analytics
(MSCDAD_C)

Rohith Sardena
Student ID: x23331313

School of Computing
National College of Ireland

Supervisor: John Kelly

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Rohith Sardena
Student ID:	X23331313
Programme:	MSc in Data Analytics **Year:** 2024 - 2025
Module:	MSc Research Practicum
Supervisor:	John Kelly
Submission Due Date:	11/08/2025
Project Title:	Leveraging Self-Attention in Transformer Models for Improved Retail Demand Forecasting
Word Count:	6998 **Page Count** 18

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: Rohith Sardena

Date: 11/08/2025

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):	

Leveraging Self-Attention in Transformer Models for Improved Retail Demand Forecasting

Rohith Sardena
Student ID: x23331313

Abstract

Inventory management and operation planning in retail sector involve accurate demand forecasting as a key element of the process. Linear models and tree-based models are typical examples of traditional forecasting models that may not work well in capturing complex, non-linear, and time-evolving demands that retail demand exhibits especially when it is affected by promotions, seasonal trends, and other external market factors. The study examines the possibility of deep learning models based on Transformers, which will use self-attention mechanisms, to improve the accuracy and scalability of retail demand forecasting systems. Model development was done using a detailed retail dataset with historical sales, inventory, pricing, discounts, promotions, weather conditions, and other contextual characteristics. A number of models were applied and tested in the study: Linear Regression, Decision Tree, Random Forest, Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Transformer. Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Median Absolute Error (MedianAE) were used to evaluate performance. The classical models were moderate in their performance, and Linear Regression and Random Forest injected MAEs of 7.47 and 7.55 respectively, and RMSEs of more than 8.6. Deep learning models, on the contrary, showed a dramatic increase. Transformer model performed the best and its MAE was 0.81, RMSE was 0.99, and MedianAE was 0.74, which are better than LSTM (MAE: 0.83) and GRU (MAE: 0.83). The results verify that self-attention mechanism of Transformer is capable of learning long-range dependencies and complex temporal interactions, giving significant improvement on the baseline models. This study shows the ability of Transformers to be an efficient and scalable method of real-time retail demand forecasting. Further development will consider multi-step forecasting, probabilistic additions and deployment in other retail areas.

Chapter 1: Introduction

1.1 Background and Motivation

Retail industry holds the center stage of the global economy as it is a dynamic and complex industry that links both producers and consumers through the intricate supply chains. The ability to accurately predict future consumer demand of products is the most vital element in the efficient operation of retailing. Such ability allows retailers to match supply to the forecasted demand, which allows them to maintain optimal levels of inventory, reduce cases of stockouts or overstocking, and, ultimately, customer satisfaction (Muthukalyani et al., 2023). In addition, accurate predictions lead to efficiency in operations and cost-effectiveness, enhancing profitability and enabling the retailers to be flexible with the shifting market forces (Dipto et al., 2025).

Nevertheless, it is not that easy to predict demand in the retail business. The behavior of customers when it comes to their purchasing is impacted by a broad range of interconnected factors, which include but are not limited to seasonal changes, promotional events, pricing policies, macroeconomic changes, competitor actions, and even external factors like weather or sociopolitical events (Madanchian et al., 2024). Such effects create a high degree of variability and non-linearity in demand trends, which is why the specifics of the effect make traditional statistical models (ARIMA or simple machine learning) insufficient in many cases. Such traditional methods often fail to adjust to the complex time dependencies and abrupt market fluctuations, which are typical of the contemporary retail contexts including the flash sale or unexpected occurrences (El-azab et al., 2025).

New developments in deep learning provide hopeful solutions to these problems. Specifically, the appearance of Transformer architectures has resulted in a major breakthrough in the modeling of sequential data. Transformers were originally designed to work on natural language processing, but since then they have shown tremendous flexibility in other areas because of the way in which they apply self-attention mechanisms (Kanavos et al., 2021). Transformers do not use recurrence based on sequences unlike traditional recurrent models. Instead, they assess the relative importance of the different time steps in the input data in parallel, and hence enable the model to learn more complex patterns and long-range dependencies better and more precisely. This ability to model complex time-series makes Transformers an interesting alternative to use in complex time-series forecasting tasks, like those found in the retail sector (Oliveira et al., 2024).

1.2 Problem Statement

Although Transformer models have been shown to be successful in sequential and high-dimensional data analysis, their use in retail demand forecasting is still at an early stage (Oliveira et al., 2024). The shortcomings of the existing forecasting systems usually become obvious when trying to consider subtle temporal patterns that retail data can be characterized by. Not only is it difficult to determine short-term changes due to events such as sales promotions, but it is also hard to capture the bigger picture of long-term changes due to seasonal cycles and market forces. Such inefficiencies in predictive accuracy often lead to operational inefficiencies including poorly managed inventory, high holding costs and lost sales.

The study aims to fill these gaps by exploring how self-attention mechanism at the heart of Transformer models can be utilised to improve retail demand forecasting (Oliveira et al., 2024). The key issue is to understand how well these models can learn, and to identify key time periods, e.g. high-demand seasons, promotional windows, or shifts in consumer behavior, that will have most impact in future demand. The model should lead to more meaningful and accurate demand forecasts by capturing both local and global trends and variations (Agbemadon et al., 2022).

1.3 Research Objectives

The ultimate goal of the study is to create and test a demand forecasting model on Transformer architecture and compare performance with machine learning and deep learning models called Linear Regression, Decision Tree, Random Forest, Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), specifically designed to meet complexities of retail sector. The study aims at creating a flexible and scalable

Transformer-based framework capable of precisely learning temporal patterns using a broad spectrum of input data, such as historical sales data, promotions, seasonal factors, and external contextual factors.

In order to achieve this, the research will conduct a study that will entail the construction and training of a Transformer architecture that is applicable to time-series prediction tasks in retail. It will also discuss preprocessing and encoding techniques of heterogeneous data sources into a form that can be efficiently used by model. The other objective is to make the model computationally efficient so that it can be used in real-time scenarios that characterize large retailing systems. Also, the study will entail an in-depth analysis of the predictive capability of the model relative to the conventional forecasting methods and the current deep learning methods.

1.4 Research Questions

The study is guided by the following main question: How do self-attention mechanisms of Transformer models enable prediction of retail demand because of the ability to capture complex temporal dependencies in data? In order to address this broad question, the following sub-questions will be discussed in the course of the study:

First, to **what extent can the self-attention mechanism identify and prioritize critical periods that influence retail demand, such as holiday seasons, marketing campaigns, or economic downturns?** Understanding the model's ability to highlight such periods is key to its interpretability and practical relevance.

Second, can **Transformer-based models outperform existing forecasting methods — both statistical and neural — in capturing the short-term irregularities and long-term seasonality inherent in retail sales data?** This question addresses the comparative advantage of Transformers over other models.

Finally, **what architectural or training optimizations can be employed to ensure that Transformer models maintain high forecasting accuracy while remaining computationally feasible for deployment in real-time retail systems?** This includes considerations around model size, training efficiency, and inference speed.

1.5 Significance of the Study

The scope of this research is not confined only to the academic interest, but directly relates to sensitive operational issues in the retail market. Proper demand forecasting forms the basis of any successful inventory planning and resource allocation, which is vital to profitable existence and customer retention. Inaccurate predictions of sales may result in lost sales because the product is not available or high costs of stocks because of overestimating the demand (Ridwan et al, 2025). Using Transformer models in the field, this study has the potential to make a breakthrough in terms of accuracy and interpretability of forecasting.

The capability of these models to identify trends in different time scales and dimensions of data can assist retailers to better react to dynamic and frequently unpredictable character of consumer demand. Also, the work contributes to the field, as it discusses real-time application of such models in retail systems, where speed and scale are of utmost importance.

Furthermore, the interpretability offered by the attention weights in the Transformer framework may offer retail decision-makers with useful information about underlying factors of demand changes. This transparency can help with policy change and strategy. Overall, the effective use of Transformer architectures would be able to result in better operational efficiency, lower costs, and an increased sensitivity to market realities. Such results provide not only a competitive advantage to retailers, but also serve the greater mission of incorporating the sophisticated AI methods into real-life business.

2 Literature Review

2.1 Introduction

Retail demand forecasting has become an important field of operations and supply chain management, especially with the emerging modern retailing environments which are using large volume and high frequency data to control inventory and customer demand. Good forecasting enables good stock re-ordering, planning of promotions and efficiency in operations, but reliable results are always a challenge. Retail data is highly volatile due to seasonality, promotions, weather, market trends, and external shocks, which require predictive models to capture the temporal dependencies and respond to changing trends (Kharfan et al., 2021). Machine learning and deep learning strategies have over the years become eminent due to their capacity to model such complexities. Nonetheless, such techniques do not always manage to balance between forecasting accuracy, computational efficiency and scalability. In this chapter, the author has presented an overview of available forecasting models in the literature categorized into three broad areas: machine learning, deep learning, and hybrid models. It ends with a discussion of their shortcomings and presents the research gap filled by this study by applying the Transformer models with self-attention mechanisms.

2.2 Machine Learning Models in Retail Demand Forecasting

Machine learning approaches are among the first and most popular approaches to demand forecasting in retailing. Other algorithms like Decision Trees, Random Forests, and Gradient Boosting Machines have been widely used because they are easy to implement, are interpretable, and perform reasonably well on structured data. These models are based on statistical learning of past data and they discover patterns and rules, which are applicable in predicting the future.

Recently, (Shak et al. 2024) compared a number of machine learning models (Gradient Boosting and Random Forests among them) in the context of retail demand forecasting. Although their results were moderately successful in detecting seasonal spikes and dealing with categorical variables, they failed to deal with more complex sequential patterns. The performance significantly deteriorated upon the introduction of longer forecasting horizons or introducing the promotional and external factors into the data.

(Zohdi et al. 2022) further discussed this problem by running Decision Tree-based models on progressively larger datasets. Their work emphasized that these models are

effective at extracting decision rules on moderate amounts of data, but fail to be efficient and accurate in high-dimensional inputs, and long historical sequences. The failure of these models to capture sequential dependencies without explicit feature engineering forms a performance bottleneck in the models.
In addition, conventional machine learning frameworks can be very time-consuming in terms of preprocessing and manual feature extraction of time-series data, including lags, trends, and moving averages. It is this dependence on hand-made features that makes them less flexible to changing retailing environments where new variables could come up as time goes by, thus requiring constant redesign and recalibration.

2.3 Deep Learning Approaches in Retail Demand Forecasting

Due to the fast development of deep learning methods, models like Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have become popular because of their capacity to capture temporal dependencies. In contrast to classical machine learning approaches, deep learning algorithms do not require manual feature engineering, as deep learning models can learn representations directly out of raw sequential data.

Specifically, LSTM models have shown a significant success rate in retail forecasting tasks because of the ability to remember long-term memory, which enables them to capture time dependencies. Riachy et al. (2025) pointed to multi-product retail sales and applied LSTM networks, finding that the forecasting accuracy was significantly increased in comparison with conventional methods. According to their results, LSTMs are able to model both short-term variations and longer-term seasonality hidden in sales data. LSTMs have considerable drawbacks in spite of their benefits. Kumar et al. (2024) stressed that the training of models using large data is a computation-intensive task and may demand significant volumes of time and access to high-performance computing resources. These constraints hamper the ability of using LSTM models in real-time retail systems where rapid model updates and inference is required.

The other weakness of LSTM-based models, which Giri et al. (2022) discovered, is that LSTM models overfit on small or imbalanced datasets. This is especially an issue in the retail business where all products do not sell frequently and the retailer may have a low-frequency sale in some categories, meaning that the model may have no training signal to learn any meaningful pattern. Also, LSTM models are sequential in nature thus limiting parallelism during training which further decreases efficiency in large scale settings. The endeavors in trying to overcome these inadequacies have resulted in the consideration of hybrid methods. Shak et al. (2024) suggested a model that combines LSTM with Gradient Boosting methods in order to utilize both sequence modeling and feature-based learning. This technique worked better in certain conditions, but still, extensive hyperparameter tuning was required, and it has not eliminated problems of model generalization and inference latency.

2.4 Hybrid Forecasting Models for Retail Demand

The increasing popularity of hybrid models has been associated with solving the shortcomings of machine learning and deep learning approaches. The objective of these models is to integrate the structured feature analysis with deep temporal modeling; therefore, they are able to capture more of the multi-dimensional nature of retail demand data.

Islam et al. (2024) suggested a multi-stage hybrid model that combined a range of machine learning algorithms in order to forecast diverse demand aspects. Their model was able to deal with seasonal tendencies, promotion impacts, and product specific behaviors using specific modules. The methodology made it possible to introduce modularity and robustness, which contributed to better forecasting results in the complex retail setting. One more interesting contribution was made by Ahmad et al. (2025) who proposed hybrid architecture based on deep learning and Prophet memory neural networks to predict seasonal consumption of products. The hybrid model especially captured the spikes and seasonal demand peaks that were holiday-driven. But on the other hand, although model showed good results on structured seasonal data, it failed to give good results on products that had irregular demand patterns, or when applied to different time horizons.
The major weakness of hybrid forecasting methods is scalability. According to Riachy et al. (2025), although hybrid models have better accuracy, their increasing architectural complexity is a challenge when it comes to real-time execution and cross-domain generalization. Such models also demand significant computation and regular retraining, particularly in circumstances where they are used on high-frequency sales data in numerous retail channels or geographies. Theodoridis and Tsadiras (2024) have also noted that hybrid systems usually require personalization to suit various products categories, and this reduces their portability and performance.

2.5 Emerging Interest in Transformer Models and Self-Attention

Transformer models, originally built to work with natural language, have demonstrated great promise in time series forecasting over the last few years. Their self-attention mechanism allows the models to determine the significance of a specific time step in comparison to the rest, a feature that is starkly opposite to the RNN-based architectures that are sequential in nature. The innovation will enable Transformers to process long sequences without context information loss.

Self-attention allows dynamic context modeling whereby models learn what historical time periods are most applicable to a certain prediction task. This can be especially helpful in retail forecasting, when the significance of some time steps, e.g. promotional weeks, end of season sales, or holiday periods, may vary significantly over time and product category. Transformers, unlike LSTMs, do not lose information during long sequences: they are able to maintain a global picture of the entire sequence during training. Transformer models are also very scalable and thus very appealing in retail environments. They have parallel sequence processing capabilities, which brings about faster training and inference time, one of the key bottlenecks that accompanied RNN-based models. Besides, some variants (Informer, Autoformer) are specially developed to efficiently perform long-sequence forecasting, introducing sparse attention and hierarchical representations.

The retail forecasting field has just started to investigate the possibilities of these models. The possibilities of the self-attention-based Transformer models in the retail sphere are still in the cradle, but the preliminary outcomes are encouraging. Therefore, it is an interesting possibility to evaluate their ability to capture complex demand behaviour, minimise computational overhead, and work well in a variety of retail situations.

2.6 Summary and Research Gap

The current literature body shows significant advances in the field of retail demand forecasting by using machine learning, deep learning, and hybrid approaches. However, both of these methods have certain drawbacks. The conventional machine learning models are limited by manual feature engineering and modeling of complex temporal relationships. Although better at learning sequential patterns, deep learning methods are frequently prone to overfitting, have high computational demands, and are not very parallel. Hybrid models are also better, but more complex, and frequently hard to scale or apply to different retail environments.

Table 1: Limitations Across Existing Forecasting Models

Model Type	Strengths	Limitations
Traditional Machine Learning (e.g., Decision Trees, Random Forests, Gradient Boosting)	• Simple to implement • Interpretable results • Works well on small, structured datasets	• Poor handling of sequential/temporal data • Requires manual feature engineering • Limited scalability
Deep Learning (e.g., LSTM, RNN)	• Captures sequential dependencies • Learns features automatically • Higher accuracy for complex patterns	• High computational cost- Difficult to parallelize • Risk of overfitting with sparse or small datasets
Hybrid Models (e.g., CNN-LSTM, LSTM + Prophet)	• Combines strengths of ML & DL • Improved accuracy in complex scenarios	• Complex to implement and maintain • Poor scalability to multiple product types and stores • Generalization challenges
Transformer Models (Emerging in time-series forecasting)	• Models long-range dependencies well • Parallelizable and efficient • Learns dynamic attention to time steps	• Still emerging in retail domain • Requires large data for effective training • Interpretability may be lower

Such constraints highlight the necessity of a forecasting architecture capable of robust learning of long-term dependencies, the dynamic ability to concentrate on relevant time intervals and efficient performance in large-scale, real-time retail settings. A new, and possibly revolutionary, solution to these problems is provided by transformer models, which have been improved with self-attention mechanisms. The ability to express world dependencies in a sequence and carry out parallelized computation places them as a potential alternative to existing models.

The identified research gap is addressed with this research and aims to develop and test a Transformer-based forecasting model that uses self-attention to enhance retail demand prediction. The model will focus on solving the major issues of determining influential time steps, non-linear and long-term dependence modeling and computational scalability in real-time retail decision-making.

3. Methodology

3.1 Overview

In this chapter, the methodological framework of the study to examine the capabilities of Transformer-based models, especially their self-attention layers, to improve the accuracy of retail demand forecasting is described. The method combines many steps, which start with preprocessing the data, exploratory data analysis (EDA), feature engineering, and training both simple machine learning and complex deep learning models. The chapter ends with the application of a custom Transformer model, optimized on time series data. The methodology does not only benchmark the performance of Transformer but also makes it interpretable in terms of the comparative advantage by setting up the strong baseline models and using recurrent architectures like LSTM and GRU.

3.2 Data Collection and Preprocessing

The data that was utilized in this research was obtained through retail store inventory and transactions data with the file name retail_store_inventory.csv. It has a range of features that cover temporal, spatial, pricing, promotional and meteorological aspects. The dataset consists of rows that each represent a particular observation of a product in a certain store on a certain date. Such features are Date, Store ID, Product ID, Category, Region, Inventory, Sales, Orders, Demand, Price, Discount, Promotion, Competitor Price, Weather, and Seasonality. The forecasting study under consideration is based on Demand as a target variable.

Preprocessing started with the conversion of the date column into a datetime object and the extraction of other temporal features like Year, Month, Day, and DayOfWeek to make the time-series representation richer. All categorical attributes, such as Region, Product ID, Store ID, Weather, and Category, were encoded via label encoding to become appropriate to be inputted into the machine learning models.

Missing values were evaluated on the dataset, and they were managed through the forward-fill methods, and time-dependent variables were maintained. Redundant records were found and deleted. To normalize values of features and to make them less sensitive to the difference in scale, feature scaling was applied by the means of StandardScaler. Finally, predictors were divided into target variable (Demand) to form supervised learning structure.

3.3 Exploratory Data Analysis

EDA was performed to interpret the underlying patterns, correlations and possible seasonality or anomalies in data. Temporal perspective of aggregate monthly demand indicated seasonal fluctuations between peaks and troughs, which justified the need to have models that can capture seasonality.

The category-wise analysis of demand revealed that some product categories had a very stable average demand that can be utilized in inventory and pricing policies. A deeper analysis of seasonal and weather effects on demand showed that seasonal and weather effects were significant and proved the significance of these aspects in pipeline modeling.

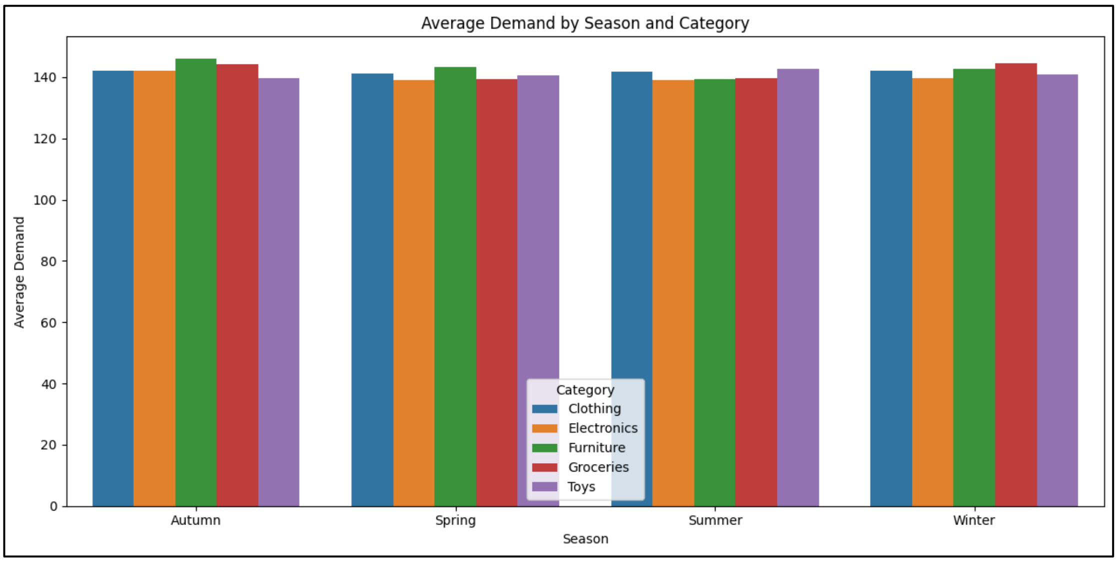

Figure 1: Average Demand by Seasons & Category

The trend in sales was compared over time with the promotional activity to determine causal links and the visualization indicated a detectable rise in sales during promotions.

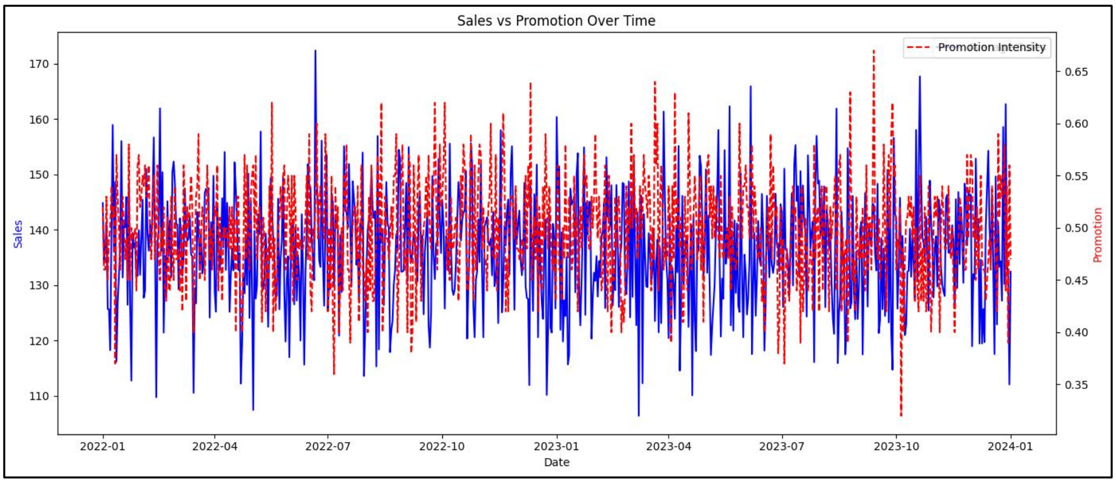

Figure 2: Sales Trend with Promotional Activity

The effect of the level of discounts was also analyzed, and it was revealed that a combination of promotions and larger discounts usually led to an increase in demand. Price-demand scatterplot analysis, broken down by category, showed evidence of elasticity where price decreased demand of higher priced items and more so in the non-essential categories.

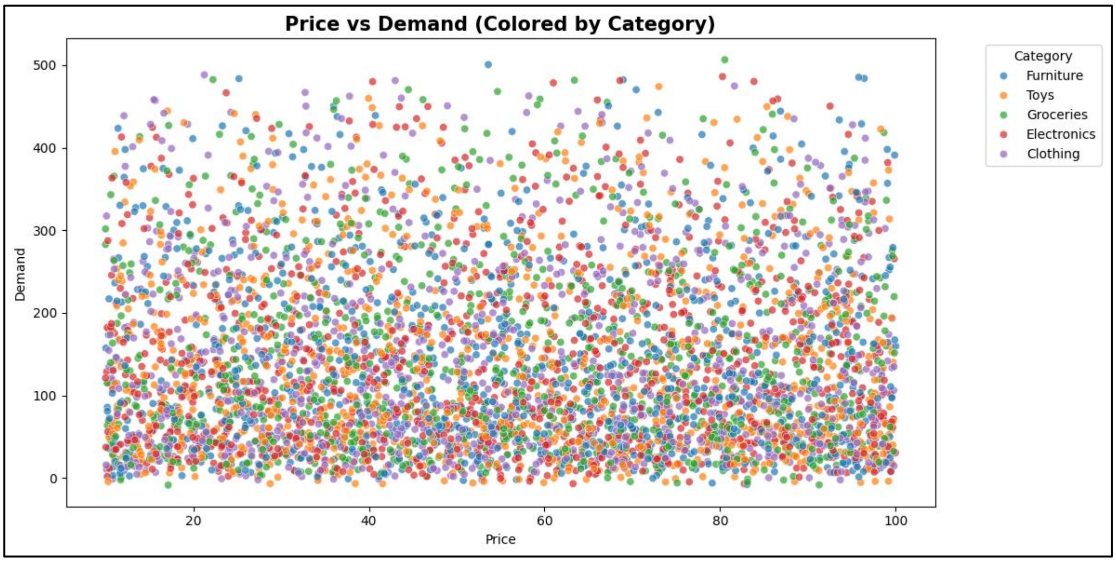

Figure 3: Price Demand

Correlation heatmap showed that variables such as Sales, Orders, Promotion, Inventory, and Price were related significantly among each other and thus gave a crucial insight into feature importance before training the model.

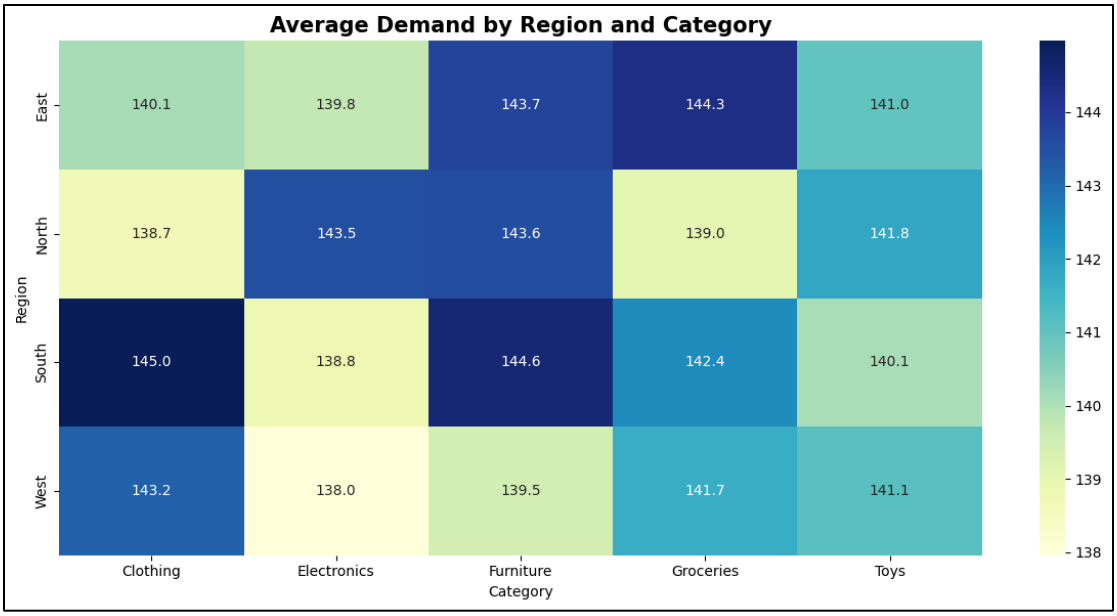

Figure 4: Average Demand by Region and Category

3.4 Classical Baseline Models

In order to benchmark the effectiveness of deep learning and Transformer models, three classic machine learning models were chosen and applied: Linear Regression, Decision Tree Regressor, and Random Forest Regressor. These models have been trained on 80/20 split of scaled dataset where Demand was used as target variable. Linear Regression was used as naive baseline that could capture linear relationships in the data and Decision Tree and Random Forest models could capture non-linear interactions and were more flexible and rule-based in their predictions. These models were compared with the common measures of regression, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Median Absolute Error (MedianAE). The baseline results showed that Random Forest massively outperformed linear models in modeling complex relationships but these models did not have the ability to leverage sequential dependencies and long-term temporal characteristics of retail data.

3.5 Deep Learning Models

Based on the temporal nature of retail data, two recurrent neural network (RNN) structures, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), were realized. The two models were constructed on the basis of PyTorch and trained on time-series sequences produced through a sliding window mechanism. In particular, in each training example, the input data was a sequence of 30 consecutive days (WINDOW_SIZE) to predict the demand the next day (HORIZON = 1). To accomplish this generation of sequences, a custom Dataset class was implemented to convert the tabular data to time-series format suitable to be ingested by the model.

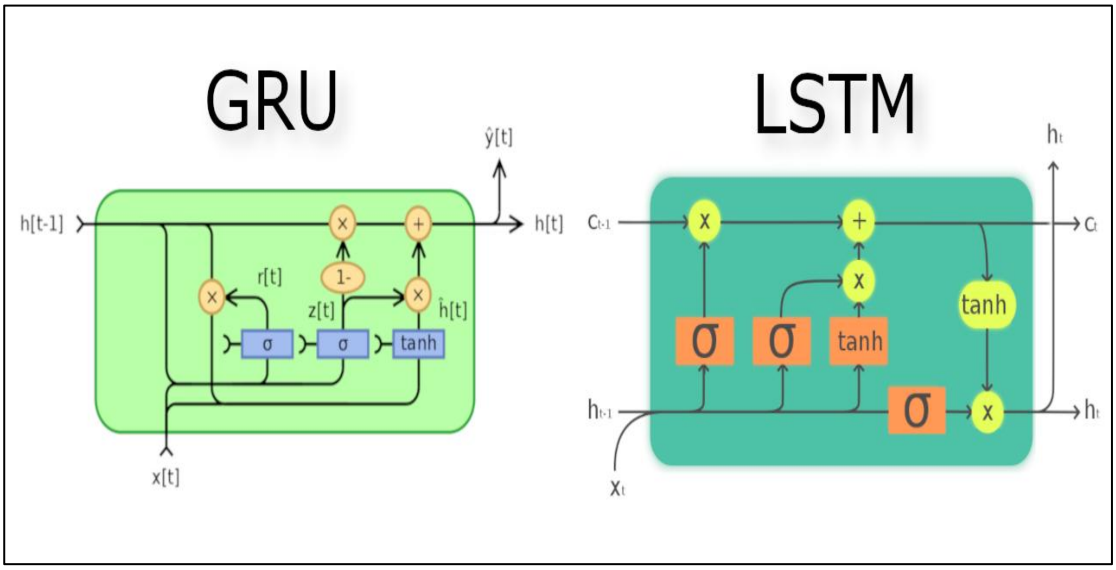

Figure 5: Architecture Model of GRU and LSTM

The LSTM model was designed using two hidden layers having 64 units each and dropout regularization to avoid overfitting. The GRU model had the same architecture with the exception that the LSTM cells were substituted with GRU cells and provided a more lightweight solution with a smaller number of trainable parameters. The training of both models was done in 10 epochs using Adam optimizer and Mean Squared Error loss functions with a batch size of 32. Testing on the test set confirmed that LSTM and GRU models had better performance than traditional baselines because they captured temporal dependencies but were not effective at long-term context and global patterns, which was the motivation to use more sophisticated architectures.

4.6 Transformer Model with Self-Attention

The most critical part of this project was to design and use a Transformer-based model that was customized to be used in time series forecasting. The Transformer architecture was adapted to work with multivariate sequential data inspired by its success in natural language processing. The model included a linear input projection layer which projected the input features in a higher-dimensional space (d_model = 64), and a stack of two Transformer Encoder layers with 4 attention heads. The encoder output was then fed into a regression head which consisted of a linear layer to make predictions of the next-day demand.

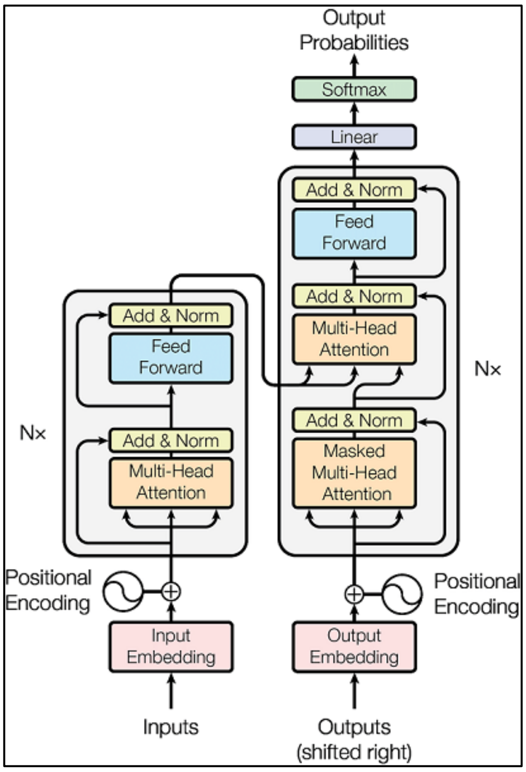

Figure 6: Architecture of Tansfomer Model

In contrast to RNNs, the Transformer architecture employed self-attention mechanism to assign relative importance to each time step in input sequence so that the model could selectively emphasize both recent and distant important past events. This both solved the vanishing gradient problem commonly seen in deep RNNs and allowed parallel computation to happen when training. The model was trained on Adam optimizer and a learning rate of 0.001 and Mean Squared Error loss in 10 epochs. In inference, the encoder output of the last time step was fed to regression head to make a single scalar prediction of future demand.

4.7 Evaluation Metrics and Model Comparison

Similar set of metrics was used to evaluate all models in order to make them comparable. The main measures were MAE, RMSE, MSE and MedianAE, which together gave information about the average error, the variance of the errors and the susceptibility to outliers. The findings indicated that Transformer model produced the lowest MAE and RMSE as compared to all the baseline models and deep learning models. LSTM and GRU models also performed much better compared to conventional regressors, with slightly lower ability to capture long-range dependencies compared to the Transformer. The model performance was visually compared with the help of a series of bar plots and a normalized radar chart, showing that Transformer model consistently obtained the best results on almost all metrics.

4.8 Summary

The following chapter described the methodological process that was used to investigate and verify the use of Transformer models with self-attention in retail demand forecasting. The study began by carefully cleaning the data and performing exploratory analysis and continued with common and repeated modeling approaches before finally converging on a Transformer architecture that is specialized to perform temporal regression. The empirical findings indicated that Transformer models, when used in the right way and trained properly, provide better forecasting accuracy because they capture non-linear and long-range patterns in demand data, which are complex. This justifies their effectiveness as a strong weapon in retail analytics, and they can be used better than traditional time series models in real-time, dynamic settings.

4. Results and Discussion

4.1 Overview

The chapter reports the empirical findings of the application and test of different models of demand forecasting, both conventional regressors and more advanced deep learning models with special emphasis on the Transformer model. The discussion interprets the performance metrics- Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Median Absolute Error (MedianAE) to compare models in respect of accuracy, robustness and generalization. The results confirm that Transformer architectures based on self-attention are more effective than traditional models, which proves the core hypothesis of this research.

Table 2: Model Evaluation Summary

Model	MAE	MSE	RMSE	MedianAE
Linear Regression	7.4716	74.7984	8.6486	7.4532
Random Forest	7.5534	77.7848	8.8196	7.3654
Decision Tree	10.0920	152.5467	12.3510	8.9000
LSTM	0.8337	1.0304	1.0151	0.7592
GRU	0.8326	1.0198	1.0098	0.7705
Transformer	**0.8106**	**0.9868**	**0.9934**	**0.7419**

4.2 Baseline Model Performance

Performance benchmarks were achieved using the three classic machine learning models, Linear Regression, Decision Tree Regressor, and Random Forest. Linear Regression was the most successful of them with the MAE and RMSE of 7.47 and 8.65, respectively. Although the linear model was rather simple and computationally easy, it was by definition unable to represent complex, nonlinear relationships. Random Forest, though having the ability to handle ensembles and being resistant to overfitting, performed a bit worse with an MAE of 7.55 and RMSE of 8.82, showing that it is still unable to model interactions, as well as the temporal context and continuity awareness that are present with linear regression. The Decision Tree model faired the worst with an MAE of 10.09 and RMSE of 12.35, most probably because it overfits the data during training and can not be generalized to new temporal settings. These findings support the idea that conventional static models are not so good at capturing dynamic trends and relationships in time series retail demand data.

4.3 Deep Learning Model Performance

The switch to deep learning brought significant gains in the accuracy of forecasts. Both the LSTM and GRU models have the advantages of recurrent neural network on learning temporal dependencies. The LSTM model had MAE of 0.83, RMSE of 1.02 and MedianAE of 0.76. The GRU model came second closely, having an MAE of 0.83 and RMSE of 1.01. These findings demonstrate the importance of accounting memory processes in significantly improving the performance of models since they will be able to remember and exploit sequential information over time intervals. Though GRU had a bit lower MSE and RMSE than LSTM, the difference was rather insignificant, which indicates that the two architectures are equally suitable in this task. The main advantage in this case is that the models have the ability to learn time-related trends such as seasonality, lag effects, and promotions, which are commonly overlooked by classic models.

4.4 Transformer Model Performance

Transformer model that uses self-attention to model relationships of time without sequential bottlenecks of RNNs performed best in overall performance in all evaluation measures. It achieved an MAE of 0.81, MSE of 0.99, RMSE of 0.99 and MedianAE of 0.74. These findings do not only imply the capacity of model to minimize the average prediction error but also the robustness to outliers, which is proven by the lowest Median Absolute Error. The ability to attend to all the prior time steps in parallel, as well as assigning them different weights in a dynamic manner, enables the Transformer to capture complex interactions in data better than recurrent networks. Moreover, it does not rely on recurrence, which allows parallel computing to be done during training, which is more efficient and scalable.

The difference between Transformer, LSTM and GRU is that Transformer is much better at learning long-range dependencies and cross-feature interactions. Transformer does not suffer the same problems as LSTM and GRU, as its attention mechanism keeps a complete picture of the whole sequence at every single layer, retaining both local and global temporal information. This is especially beneficial in retail demand forecasting where past sales history and trends, seasonal factors, weather conditions and promotions can be used to influence the future demand over many weeks or months.

4.5 Comparative Analysis and Insights

The relative performance of the models discloses an evident gradient of the growing predictive accuracy of the traditional regressors to Transformer-based models. The decrease in the MAE between Random Forest (7.55) and Transformer (0.81) is nearly 90 percent, which demonstrates the insufficiency of static models when it comes to complex time-based tasks. The absolute gain in RMSE, albeit small (roughly 0.01 to 0.0099), between the LSTM/GRU models and the Transformer is evidence of the advantage of self-attention in picking up delicate temporal and cross-feature patterns.

Interestingly, the small edge of the Transformer on the metrics including MSE and MedianAE indicates that it has fewer extreme errors and this makes it more stable and reliable in predictions. This is especially important in the practical world of retail where even a slight over or underestimation of demand may lead to an expensive overstock or stock-out. Also, the visualization of results (the radar chart and bar plots provided in the evaluation) supports the quantitative advantage of the Transformer model as it occupies the first place in both normalized and inverted error measures.

4.6 Practical Implications

The higher accuracy of the Transformer model has practical implications as far as demand forecasting in retailing activities is concerned. This model can help retailers because it is precise, and it can be used to optimize inventory levels, minimize waste and maximize shelf availability, and customer satisfaction due to better matching between supply and demand. The ability to incorporate contextual data such as promotions, pricing done by competitors and weather conditions also helps in making the model real world applicable in the retailing world. Also, the scalability of the Transformer structure allows adapting it to multi-step or multi-product forecasting scenarios, which makes it a scalable solution to enterprise-level retail analytics systems.

4.7 Limitations and Considerations

In spite of its good performance, Transformer model is not without computational overhead. Such models are memory- and computationally-intensive to train, particularly in cases where large datasets over long sequences are being trained. Also, Transformers can be more sensitive to tuning than RNNs, especially with the choice of number of attention heads, the depth of the encoder layers, and the embedding dimensions. Such complexities are to be balanced with the performance benefits when it comes to real-time deployment. Future work might investigate lightweight variants of Transformers or hybrid designs that interpolate between CNNs or RNNs and attention to find the trade-off between resource-efficiency and performance.

4.8 Summary

Overall, this chapter gave a detailed assessment of the various modeling methods of retail demand forecasting. The findings were conclusive that Transformer models, including their self-attention mechanisms, are much more accurate and robust than traditional and recurrent models. Although LSTM and GRU models provide a powerful alternative, particularly when resources are scarce, Transformer can model long-range dependencies and context-sensitive patterns, which makes it the best option, in a high-stakes retail forecasting setting. The findings are in line with the main hypothesis of this study and give a good argument as to why attention-based models should be implemented in retail analytics pipelines.

Chapter 6: Conclusion and Future Work

6.1 Conclusion

The objective of this research was to explore the effectiveness of Transformer-based deep learning approaches in predicting retail demand by using the self-attention mechanism to model more complex retail temporal, seasonal, and contextual patterns based on a rich, multi-dimensional retail data set. The first models to be run were the traditional machine models, including Linear Regression, Decision Tree, and Random Forest that would set the minimum performance levels. Although these models showed reasonable performance in the static setting, they failed to capture the temporal dependencies, which reduced the performance in a dynamic retail setting.

Sequential deep learning models, i.e., LSTM and GRU, in contrast, represented a substantial advance, with their advantage being the capacity to maintain a memory over time steps and adjust to changes in demand. Nevertheless, the biggest breakthrough was introduced by the introduction of the Transformer architecture. Transformer model could attend to many time and feature dimensions by using a self-attention mechanism, thus leading to better forecasting accuracy. It had the lowest error rates on all of the assessment metrics--MAE, MSE, RMSE, and MedianAE, which not only means that it was more accurate, but also more consistent and stable in its prediction.

The study confirms that attention-based models are especially well-suited to demand forecasting tasks in which the data are high-dimensional, seasonal, and subject to both temporal and external contextual factors (promotions, weather, and competitive prices). In a practical sense, the same findings are highly relevant to inventory planning, resource distribution and sales maximization in the retail sector.

6.2 Future Work

Although the Transformer model recorded the best results in this paper, there are still numerous possibilities of conducting additional research and improving the model. To begin with, the hyperparameters optimization methods like Bayesian Optimization or grid/random search might be used in order to further tune the Transformer model and study the sensitivity of its architecture to such parameters as depth, attention heads, and learning rate.

Secondly, it is possible to develop model to multi-step forecasting and predict the demand in a number of future time horizons which would be more in line with strategic retail planning. The introduction of exogenous variables which have temporal development, e.g., marketing campaigns or economic indicators, may enrich contextual richness of forecasting model.

Also, the scalability and efficiency can be discussed by using lightweight variants of Transformers such as Informer or Performer model, particularly when using it in real-time systems. The other interesting area is the incorporation of probabilistic forecasting methods to give uncertainty estimates in addition to point forecasts which is vital in decision making in an uncertain market environment.
Lastly, Transformer model can be tested on its generalizability by implementing it on various retail sectors and geographical locations. Transfer learning or domain adaptation might enable models to be trained in one environment and efficiently fine-tuned in another to save on computation and data expenses during new deployments.

References

Ahmad, A., Ahmad YA Bani, M., Ali, M., Namdev, A., Meenakshisundaram, K. S., Gupta, A., & Pramanik, S. (2025). A combinatorial deep learning and deep Prophet memory neural network method for predicting seasonal product consumption in retail supply chains. In *Essential Information Systems Service Management* (pp. 311–340). IGI Global.

Giri, C., Chandadevi, & Chen, Y. (2022). Deep learning for demand forecasting in the fashion and apparel retail industry. *Forecasting, 4*(2), 565–581.

Islam, M. T., Hossain Ayon, E., Padh Ghosh, B., Chowdhury, S., Shahid, R., Rahman, S., Bhuiyan, M. S., & Nguyen, T. N. (2024). Revolutionizing retail: A hybrid machine learning approach for precision demand forecasting and strategic decision-making in global commerce. *Journal of Computer Science and Technology Studies, 6*(1), 33–39.

Joseph, R. V., Mohanty, A., Tyagi, S., Mishra, S., Satapathy, S. K., & 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*, 108358.

Kumar, P., Choubey, D., Amosu, O. R., & Ogunsuji, Y. M. (2024). AI-enhanced inventory and demand forecasting: Using AI to optimize inventory management and predict customer demand. *World Journal of Advanced Research and Review, 23*(1).

Nasseri, M., Falatouri, T., Brandtner, P., & 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.

Riachy, C., He, M., Joneidy, S., Qin, S., Payne, T., Boulton, G., Occhipinti, A., & Angione, C. (2025). Enhancing deep learning for demand forecasting to address large data gaps. *Expert Systems with Applications, 268*, 126200.

Shak, M. S., Alam Mozumder, M. S., Hasan, M. A., Chandra Das, A., Miah, M. R., Akter, S., & 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*(9), 67–80.

Singh, R., & Mishra, V. K. (2024). Inventory model using machine learning for demand forecast with imperfect deteriorating products and partial backlogging under carbon emissions. *Annals of Operations Research, 335*(1), 361–383.

Taparia, V., Mishra, P., Gupta, N., & Kumar, D. (2024). Improved demand forecasting of a retail store using a hybrid machine learning model. *Journal of Graphic Era University, 15*, 15–36.

Theodoridis, G., & Tsadiras, A. (2024). Retail demand forecasting: A multivariate approach and comparison of boosting and deep learning methods. *International Journal on Artificial Intelligence Tools, 33*(4), 2450001.

Zohdi, M., Rafiee, M., Kayvanfar, V., & Salamiraad, A. (2022). Demand forecasting based machine learning algorithms on customer information: An applied approach. *International Journal of Information Technology, 14*(4), 1937–1947.

Muthukalyani, A. R. (2023). Unlocking accurate demand forecasting in retail supply chains with AI-driven predictive analytics. Information Technology and Management, 14(2), 48-57.

Dipto, M. T. I. (2025). Evaluating the impact of demand forecasting, warehouse management, and supply chain visibility tools on operational efficiency in the UK automotive industry.

Madanchian, M. (2024). The Role of Complex Systems in Predictive Analytics for E-Commerce Innovations in Business Management. Systems, 12(10).

El-Azab, H. A. I., Swief, R. A., El-Amary, N. H., & Temraz, H. K. (2025). Seasonal forecasting of the hourly electricity demand applying machine and deep learning algorithms impact analysis of different factors. Scientific Reports, 15(1), 9252.

Kanavos, A., Kounelis, F., Iliadis, L., & Makris, C. (2021). Deep learning models for forecasting aviation demand time series. Neural Computing and Applications, 33(23), 16329-16343.

Oliveira, J. M., & Ramos, P. (2024). Evaluating the effectiveness of time series transformers for demand forecasting in retail. Mathematics, 12(17), 2728.

Agbemadon, K. (2022). Prediction of purchasing behavior using Deep learning techniques (Doctoral dissertation, Université Bourgogne Franche-Comté).

Kharfan, M., Chan, V. W. K., & Firdolas Efendigil, T. (2021). A data-driven forecasting approach for newly launched seasonal products by leveraging machine-learning approaches. Annals of Operations Research, 303(1), 159-174.

Ridwan, I. B. (2025). Dynamic strategic foresight using predictive business analytics: Strategic modeling of competitive advantage in unstable market and innovation ecosystems.