Retail Store wInventory Forecasting

Author 1 Name: Fathhur Rahaman Sams
Author 1 Gmail: frs14102@gmail.com

Author 2 Name: Mahir Faysal Haque Dipto
Author 2 Gmail: dipto.haq771@gmail.com

Author 3 Name: Sanjana Kazi Supti
Author 3 Gmail: sanjanakazisupti@gmail.com

Author 4 Name: Faisal Ahmed
Author 4 Gmail: faaisal8898@gmail.com

Author 5 Name: Md. Junaeid Ali
Author 5 Gmail:junaeidali2002@gmail.com

ABSTRACT

Abstract—Demand forecasting and dynamic pricing are two of the great enhancements for optimizing profitability. Inventory optimization has remarkable challenges in supply chain management, where inaccurate predictions cause substantial financial losses due to stockouts or overabundant inventory. The traditional method often struggles to achieve an accurate inventory level. To estimate early Inventory prediction and dynamic pricing, machine learning models can play a crucial role. In analyzing data related to retail stores, the machine learning model would be a good throne. This improves business outcomes and helps to grow overall productivity. Our hybrid ARIMA-LSTM model combines statistical time-series analysis with deep learning, which uplifted in capturing linear and nonlinear patterns. This approach defines an eye-catching improvement in forecasting accuracy compared to the traditional standalone single model approach. The gradient boosting model showed good results for inventory optimization and also achieved a high coefficient of determination. It reduced overstock, maintaining minimal stockouts. The implementation of XGBoost for dynamic pricing resulted in measurable revenue growth. Furthermore, explainable AI techniques such as SHAP (SHapley Additive exPlanations) and Lime(Local Interpretable Model Agnostic Explanation) were utilized to identify the influencing factors in the machine learning model. Hypertuning was also conducted using GridSearchCV and RandomizedSearchCV. The evaluation metrics include MSE, MAE, RMSE, and R^2 scores. In conclusion, it is well determined that Retail performance is affected by the demand patterns, the inventory levels, and competitor pricing. AI-powered analytics reveal insights that allow businesses to strategize with data and align operational activities with market and demand interests.

Keywords—Inventory Optimization, Demand Forecasting, Dynamic Pricing, ARIMA-LSTM, XGBoost, Random Forest, Elbow Method, K-Means, PCA, LightGBM, SHAP, LIME, Machine Learning, K-Fold, Randomized Search CV.

I. INTRODUCTION

Retail store inventory forecasting remains crucial for business efficiency because it ensures optimal inventory managers alongside satisfied customers while controlling operational costs effectively. Sihotang J. [1] concerned that having too large an inventory can result in high storage costs and excessive tied-up capital. The forecasting process determines necessary future inventory amounts through analysis of historical sales data, patterns of consumer activities, and general market climate. Neutral analysis of these factors enables businesses to determine future market demand better so they can properly stock their stores to meet customer needs and avoid both overstocking issues and stockout situations. Retailers can effectively predict forthcoming product requirements through proper forecasting to create strategic inventory decisions. The implementation of proper forecasting enables businesses to manage their inventory properly, which helps reduce waste expenses while improving profits, particularly for perishable goods. Products with high consumer demand lead to increased customer purchase frequency. Research in this paper explores critical aspects of retail inventory management with the main objective of enhancing forecasting capabilities. AI processes extensive multidimensional data types as it learns current changes in real-time weather conditions to protect retailer operations. Li M. et al. [2] apprise Dynamic pricing strategy as an effective tool to control demand and respond to market uncertainty has been used by many companies. The dataset analysis highlights three main tasks consisting of time series forecasting and inventory level optimization as well as dynamic pricing. Punia S. et al. [3] addressed that the time-series data-based demand forecasting methods can be classified into three categories: statistical methods, machine learning methods, and hybrid methods. Machine learning techniques, such as ARIMA, LSTM, XGBoost, and Random Forest, have demonstrated superior accuracy in retail demand forecasting, helping to improve inventory management and operational efficiency [1]. Accurate demand forecasting was obtained by implementing ARIMA along with LSTM and XGBoost and Random Forest and Linear Regression and Decision Tree and Gradient Boost and LightGBM. Our analysis incorporated the interpretability methods SHAP and LIME. Using AI for inventory management showcases its central function in optimizing decision-making along with stock efficiency and the prevention of losses in retail facilities.

RQ1: Does inventory level have a significant impact on the number of units sold in a retail setting?

RQ2: Is there a significant difference in demand forecasts across different regions?

RQ3: How does the accuracy of demand forecasting evolve and improve over time?

RQ4: Does the accuracy of demand forecasting improve over time?

RQ5: Does sales variance differ significantly across different stores?

Businesses can improve their forecasting models, ensuring both accuracy and favorable to change conditions. By the use of statistical methods, machine learning, and deep learning techniques, retailers can easily predict future demand more precisely. The success of these models relies on the quality of the data on which they are based. With the right forecasting methods and regular improvements, retailers can better predict demand, work more efficiently, and keep customers satisfied. In conclusion, retail store inventory forecasting is a dynamic and essential practice for businesses aiming to optimize stock levels, reduce costs, and enhance customer satisfaction. By using effective forecasting methods and making continuous improvements, retailers can better anticipate demand, enhance efficiency, and ultimately boost both operational success and customer satisfaction.

II. RELATED WORKS-(50-80)

In [4], Hasan et al. benchmarked forecasting models on historical sales data from Walmart to predict their future sales. The authors provided a comprehensive theoretical overview and analysis of the time series forecasting models. Then, we applied these models to the forecasting challenge dataset (M5 forecasting by Kaggle). Results suggest that the ARIMA model outperforms the Facebook Prophet and LightGBM models, but the LightGBM model achieves a huge computational gain for the large dataset with negligible compromise in the prediction accuracy. They used time-series data obtained from M5 forecasting Kaggle to predict future sales of Walmart products for the subsequent 28 days. They implemented the ARIMA and advanced models, including the Prophet model from Facebook and the LightGBM model from Microsoft. Further work includes developing an ensemble model using the benchmarked models and optimizing the performance to minimize RMSE further. LightGBM provided a nearly similar RMSE as ARIMA but considerably faster implementation time when considering the whole dataset.

Haque et al. [5] addressed the gap in retail demand forecasting by implementing macroeconomic variables such as the Consumer Price Index (CPI), Index of Consumer Sentiment (ICS), and unemployment rates into time series data alongside historical sales. Their research developed and compared regression and machine learning models to improve prediction accuracy, emphasizing the importance of these factors in consumer spending behavior. The results demonstrate that integrating these external economic indicators enhances the forecasting models' ability to predict retail demand more effectively. The authors also employed a range of regression and machine learning techniques to predict retail demand, integrating macroeconomic indicators with historical sales data. They compared the performance of these models to identify the most effective approach for incorporating external economic factors into forecasting. Their comparative analysis revealed that certain models, particularly those capable of handling multivariate time series, achieved higher accuracy in forecasting demand. Future research could explore the inclusion of additional macroeconomic variables or apply the forecasting models to different retail contexts to further validate their effectiveness.

In [6], Singhal et al. explored artificial intelligence (AI), specifically the Random Forest Regressor (RFR), to enhance demand forecasting, pricing, and inventory management in retail by analyzing factors such as competitor pricing, customer preferences, and inventory levels, achieving mean squared errors (MSE) of 8.15% for demand forecasting and 1.11% for price prediction. The authors employed the Random Forest Regressor (RFR) for both demand forecasting and price prediction, training the models on historical data with features including product prices, competitor prices, customer preferences, and seasonal trends for demand forecasting and competitor prices, inventory levels, and customer preferences for price prediction. They evaluated the models using mean squared error (MSE). A rule-based system was implemented for inventory management, using the demand forecasts to determine restocking needs based on predefined thresholds. Future studies could investigate the incorporation of more diverse datasets/ explore the model's applicability in different retail sectors to enhance its generalizability and effectiveness further. The paper discusses how traditional methods falter in dynamic retail contexts, a challenge the hybrid model overcomes by addressing both nonlinear and linear data patterns.

Falatouri et al. [7] explored predictive analytics for demand forecasting in retail supply chain management (SCM), where LSTM performed better for products with stable demand, while SARIMA was more effective for seasonal items. Additionally, SARIMAX, which includes external promotional factors, significantly improved forecasting accuracy for products influenced by promotions. The dataset comprises more than 37 months of actual retail sales data collected from an Austrian retailer. The results reveal that for salad, SARIMA achieved an MAPE of 13 and an RMSE of 1009 compared to LSTM's MAPE of 14 and RMSE of 949, while for tomato, LSTM recorded an MAPE of 44 and an RMSE of 1170 versus SARIMA's MAPE of 48 and RMSE of 1163. For potato, LSTM outperformed SARIMA with an MAPE of 15 and an RMSE of 1184 compared to SARIMA's MAPE of 25 and RMSE of 1626, whereas for cucumber, SARIMA was superior with a MAPE of 16 and RMSE of 2854 against LSTM's MAPE of 35 and RMSE of 4502. Incorporating external factors using the SARIMAX model improved forecasting accuracy significantly: salad's MAPE dropped from 13 to 6 (a 53% improvement) and tomato's from its baseline to a 54% improvement, while potato and cucumber saw improvements of 28% and 12.5%, respectively. Overall, these findings underscore that the effectiveness of forecasting models in retail SCM is highly dependent on product characteristics and the integration of external promotional factors.

In [8], Taparia et al. proposed a demand prediction model that considers historical demand data and the SKU price to forecast the demand. They evaluated the approach on 1000 SKUs, and the result showed that the Random Forest is the best-performing regressor algorithm with the lowest mean absolute percentage error (MAPE) of 8%. Furthermore, the hybrid model resulted in a lower inventory and lost sales cost with an MAPE of 7.74%. The dataset was collected from a retail store and contains weekly demand and price data for 1000 SKUs over 104 weeks, along with competitor and competitive commodity price data. The study evaluates five machine learning regression algorithms—Linear Regression (LR), Polynomial Regression (PR), Support Vector Regression (SVR), Decision Tree Regression (DTR), and Random Forest Regression (RF)—to predict demand for each SKU using historical demand and price data. They developed a hybrid model by selecting the best two performing algorithms for each SKU based on the lowest sum of absolute deviations, then aggregating their predictions into a new dataset fed into LR for final forecasting. The process involves training on 53-82 weeks of data, tuning the hybrid model on 83-97 weeks, and testing on 98-104 weeks, with inputs including optimized exponential smoothing predictions, trend-seasonality factors, and pricing variables.

Bi et al. [9] proposed a novel demand-aware tensor factorization approach, termed ATLAS. When evaluated on eight datasets from the IRI database, spanning various product categories like blades, coffee, and milk, ATLAS outperformed benchmarks like SARIMA, VAR, LSTM, and other tensor-based methods (BPTF, libFM, TRMF, CPD) across various scenarios, including different prediction horizons (Appendix 11) and data densities (Appendix 14), with notable improvements in RMSE. RMSE improvements ranging from 10% to 30%. The ATLAS model employs tensor factorization to decompose sales data into latent factors. This regularization captures sales interactions, such as competition or collaboration, by penalizing deviations from expected covariance structures in the latent factors, as detailed in the derivation of the regularization function (fg). The process begins with an alternating least squares (ALS) approach for regularized SVD (Appendix 1), followed by integration with time-series models like SARIMA (Appendix 3) or LSTM (Appendix 4), where the tensor's completed values serve as inputs for final sales predictions. The data originates from the IRI database, consisting of weekly sales records for eight product categories (e.g., blades, coffee, deodorant, diapers, frozen pizza, milk, photography, toothpaste) across 1,560 grocery stores over 208 weeks, from January 2008 to December 2011. This dataset, detailed in Appendix 8, provides a rich basis for testing, with sales amounts per product per store per week showing significant variation (e.g., mean sales ranging from $5.88 for deodorant to $110.14 for milk). The authors suggest that future research could enhance ATLAS by incorporating additional external factors, such as economic indicators or more detailed promotion data, to further improve forecasting accuracy.

Kumar et al. [10] investigated how AI optimizes inventory levels and enhances demand forecasting. They examine various models, including linear regression, decision trees, random forests, support vector machines (SVMs), and neural networks (particularly deep learning models like LSTM), which are trained and validated using techniques like k-fold cross-validation, with neural networks showing superior accuracy. The dataset comprises historical sales records, market data, customer feedback, operational data, and external variables like weather conditions and seasonal events, collected to provide a comprehensive basis for AI analysis. Future enhancements may integrate Internet of Things (IoT) devices for real-time inventory tracking, blockchain for secure data sharing, and advanced machine learning algorithms to further refine AI-driven inventory management and demand forecasting. Neural networks achieved a Mean Absolute Error (MAE) of 0.35 and Root Mean Squared Error (RMSE) of 0.45 in demand forecasting. AI implementation reduced excess inventory by 20% and stockouts by 15%, driven by automation and real-time monitoring.

In [11], Baloian et al. presented a methodology that, unlike traditional approaches that rely on current sales levels, uses predicted future sales levels to enable retail managers to simulate various scenarios and set realistic sales goals. The methodology is structured into two primary components, with an additional optimization extension. The first one is the prediction and model selection process- An algorithm evaluates multiple machine learning models—Gradient Boosting Decision Trees (GBDT), Support Vector Regression (SVR), XGBoost Regression (XGBR), Random Forest, and Weighted Evidence Regression Model (WEVREG). Hyperparameter calibration is performed using an Exhaustive GridSearch, which tests the constrained ranges of hyperparameters across all stores. The second one is the "Sales Goal Setting Model."- Using the predicted indicators (foot traffic Ep, conversion rate Cp, and average transaction value Tp), the model calculates the necessary percentage changes (VE, VC, VT) to achieve a user-defined sales goal (Sg). Sales are modeled as $S=E\times C\times T$. The model uses historical coefficients of variation to assign weights (WE, WC, WT) that reflect how easily each indicator can change, solving a cubic polynomial equation to determine feasible adjustments. This allows managers to simulate scenarios and identify actionable targets. Specific details about the dataset's source are not provided. The authors propose two main directions for future research: 1. Enhancing the reliability and uniformity of the algorithm's predictions across different stores, 2. Investigating other performance indicators (beyond foot traffic, conversion rate, and average transaction value) that could improve the effectiveness of sales and operations planning for managers. The tool achieves a prediction accuracy exceeding 90%. In a comparison with human performance (HP) across 30 stores, the algorithm's average error was -4.44 %, outperforming HP's 7.73%. The algorithm showed slightly better consistency (standard deviation of 0.507 vs. 0.516 for HP) and excelled in more instances (7 vs. 4), particularly where marketing campaigns were underestimated by experts. However, human experts occasionally outperformed the algorithm due to store-specific knowledge.

Saha et al. [12] investigated the use of advanced deep learning techniques to predict sales demand for an American multinational retail company. Their research introduces two forecasting approaches utilizing Long Short-Term Memory (LSTM) and Light Gradient Boosting Machine (LightGBM) models. The dataset originates from an American multinational retail company and spans sales data from January 29, 2011, to June 19, 2016. A variant of Recurrent Neural Networks (RNNs) used for time-series forecasting. It employs memory cells with input, forget, and output gates to capture long-term dependencies in the sales data. A gradient boosting framework based on decision trees, optimized for large datasets. It uses leaf-splitting techniques for efficient node division and includes parameters (e.g., learning rate, tree depth) to mitigate overfitting. LightGBM: Weighted Root Mean Squared Scaled Error (WRMSSE): 0.6255, Root Mean Squared Error (RMSE): 4436, Mean Absolute Error (MAE): 2125.28, Weighted Mean Absolute Percentage Error (WMAPE): 0.1058. LSTM: WRMSSE: 0.7977, RMSE: 5024, MAE: 2379.06, WMAPE: 0.1545. LightGBM outperforms LSTM, delivering lower error rates and better predictions.

In [13], Hübner et al. identified cross-cutting store-related planning issues and developed a planning framework for OC operations. The dataset originates from a combination of insights gathered through industry interviews and a comprehensive review of over 40 recent publications, drawing on both academic literature and real-world retail practices. Methodologically, the study employs a multi-method approach that combines industry interviews and a systematic literature analysis to develop a conceptual planning framework without specifying a distinct model name.

The integrated results and discussion emphasize the need for advanced operational research models to optimize fulfillment networks, order assignment, and inventory policies in the evolving retail landscape.

Wang and Gu [14] focused on three major US retail firms—Walmart, Costco, and Kroger. Their study employs the Least Absolute Shrinkage and Selection Operator (Lasso) to identify key economic indicators like the Consumer Price Index (CPI) and regular wage, which significantly influence sales across all three firms. For sales forecasting, Support Vector Regression (SVR) outperforms in the training set, while Multivariate Adaptive Regression Splines (MARS) excels in the testing set. The Lotka-Volterra Model (LVM) reveals a mutualistic relationship among the firms, suggesting they benefit from each other's presence. Data Envelopment Analysis (DEA) indicates that Kroger operates less efficiently than Walmart and Costco, though it has the potential for greater sales growth under stable conditions. The paper combines statistical learning, machine learning, and operations research techniques to tackle the three core objectives: Sales Forecasting (Predictive Analytics): Lasso identifies critical economic indicators (e.g., CPI, regular wage) by shrinking less relevant predictors. Three models are compared: MARS, SVR, and Deep Neural Network (DNN) using RMSE, MAE, and MAPE. SVR achieves the lowest MAPE (e.g., 4.36% for Walmart, 8.48% for Kroger) in the training set, and MARS outperforms with MAPEs of 4.27% (Walmart), 10.22% (Costco), and 9.4% (Kroger) in the testing set. Market Analysis (Diagnostic Analytics): Lotka-Volterra Model (LVM). Performance Assessment (Prescriptive Analytics): Data Envelopment Analysis (DEA). Significant economic indicators are CPI, regular wage, PPI(WM specific- due to upstream sensitivity), oil prices(WM specific), and DJT(Costco & Kroger specific due to logistic needs).

Martins and Galegale [15] explored the vital role of sales forecasting in retail, addressing challenges like stock shortages and excess inventory due to inaccurate predictions. Their integrative literature review (ILR) compares traditional forecasting methods, such as time series models (e.g., Simple Exponential Smoothing and ARIMA), with machine learning algorithms such as Random Forest (RF), Support Vector Machines (SVM), and Gradient Tree Boosting (GTB). The analysis draws from diverse datasets, including proprietary sales records from supermarket chains (e.g., 1,510,563 transactions from IRI Marketing Research) and Japanese point-of-sale data, alongside public datasets like Kaggle's 8,523-entry collection, spanning millions of transactions. The study conducts an ILR using the PRISMA-P protocol to review nine recent publications on sales forecasting systematically. It evaluates traditional time series models against machine learning algorithms across various retail contexts, employing metrics like RMSE, MAPE, accuracy, and AUC. Machine learning outperforms traditional methods in accuracy, especially with exogenous and endogenous variables, and excels at identifying hidden demand patterns, though traditional methods may suffice in stable markets due to their simplicity and lower cost.

In [16], Best et al. developed a simulation model of a retail store that incorporates various error-prone processes based on input from retailers. They conducted a full factorial test design to analyze how different operational errors contribute to inventory discrepancies, with a particular focus on positive inventory record inaccuracies. The dataset was developed through collaboration with professionals from eight different retail companies and associations who participated in the researchers' practitioner workshop. They conducted a full factorial test design to analyze how different operational errors contribute to inventory discrepancies, with a particular focus on positive inventory record inaccuracies. Future research could focus on developing specific countermeasures for each type of positive inventory discrepancy identified in the study. Retailers need to adjust specific process parameters to avoid inventory records becoming inaccurate, with phantom products playing a significant role in retail stockouts.

Esrar et al. [17] examined how a large-scale retail chain can improve its handling of excess seasonal inventory using three common strategies: information sharing, visibility, and collaboration. Their research employed a case study method focusing on one retail chain at three key organizational levels: strategic (head office), warehouses, and retail stores. This study examines how a large-scale retail chain can improve its handling of excess seasonal inventory using three common strategies: information sharing, visibility, and collaboration. The research employed a case study method focusing on one retail chain at three key organizational levels: strategic (head office), warehouses, and retail stores. The dataset was collected through primary research involving semi-structured interviews with senior-level employees at the strategic level (head office), warehouses, and retail stores of a Canadian big-box retailer. Future research could explore multicultural aspects of inventory management in the Canadian context, which was noted as particularly interesting due to the country's diverse demographics. The Implementation of information sharing, visibility, and collaboration strategies significantly improved the retailer's seasonal inventory management, despite challenges related to human errors, forecasting limitations, and return merchandise handling.

In [18], Wellens et al.. demonstrated that simple implementations of tree-based machine learning methods can substantially outperform traditional statistical forecasting techniques in retail while remaining computationally efficient. The study validates this approach using a dataset of 4,523 products from a leading Belgian retailer with various explanatory variables, including promotions and national events. The researchers compared various tree-based machine learning methods with different complexity levels against popular statistical methods for forecast accuracy, bias, and inventory performance. They utilized Shapley values and, for validation purposes, applied their framework to both their private dataset and the M5 competition dataset, ensuring their insights were structural and not specific to one dataset. Future research could focus on strategies to increase the adoption rate of machine learning forecasting methods among traditional retailers who currently rely on simpler statistical techniques. A simple, off-the-shelf tree-based machine learning implementation can improve forecast accuracy by 11.48% compared to traditional methods while remaining computationally efficient.

Nasseri et al. [19] compared tree-based ensemble forecasting (specifically Extra Tree Regressors/ETRs) with Long Short-Term Memory (LSTM) networks for retail demand prediction. The dataset consisted of historical demand data from a prominent Austrian retailer, encompassing daily demand for over 330 products across three main categories (fruits, fresh meat, soft drinks), totaling 5.2 million records over 76 months (January 2016 to February 2022), supplemented with internal data (pricing, promotions) and external variables (calendar data, weather conditions, COVID-19 information). The researchers developed a comprehensive methodology that included data preparation, feature creation (including lagged demand features for previous days/weeks), feature scaling and encoding, feature selection, and model training. They employed two primary models: an Extra Tree Regressor ensemble model using a bagging approach with 300 trees and a deep learning model combining LSTM layers for temporal features and dense layers for non-temporal features. Both models were evaluated using four metrics (MAPE, MAE, RMSE, and R^2) on a test dataset covering the period from January 2021 to February 2022. Future research could compare tree-based ensembles and deep learning approaches based on different time horizons, focusing on short-term, medium-term, and long-term demand predictions. Tree-based ensemble models, particularly ETR, consistently outperformed LSTM in retail demand forecasting across all product categories, showing particular strength in predicting demand for fresh meat products while offering advantages in terms of lower data preparation requirements, better interpretability, and more efficient training.

In [20], Ganguly and Mukherjee focused on improving prediction accuracy by implementing and comparing various models, including Random Forest (RF), Gradient Boosting (GB), Support Vector Regression (SVR), and XGBoost. The dataset used for this research was sourced from Favorita Stores, a retail chain in Ecuador. The methodology included hyperparameter tuning through randomized search cross-validation to optimize key parameters such as the number of estimators, maximum depth, and minimum samples split for the RF model to minimize root mean squared error (RMSE). Future research should incorporate external variables and explore other advanced models like Long Short-Term Memory (LSTM) networks and hybrid models combining multiple algorithms for potentially better results. Through hyperparameter tuning with randomized search cross-validation, the optimized RF model achieved an R-squared value of 0.945, substantially outperforming both the initial RF model and traditional LR, which had an R-squared of 0.531.

Hossam et al. [21] addressed critical retail challenges, such as inefficient queue management, poor demand forecasting, and ineffective marketing, by introducing a smart retail analysis system that leverages advanced machine learning technologies. Their system integrates YOLO-V8 for real-time person detection and BOT-SORT for precise multi-object tracking alongside a GRU model for demand forecasting, achieving high accuracy and operational efficiency. Experimental results demonstrate that YOLO-V8-X delivers a mean average precision (mAP) of 0.93 for customer detection, while the GRU model enhances demand prediction with an R2-score improvement of up to 3.215% and an MAPE reduction of up to 29.31% compared to baseline models. This innovative approach significantly improves retailer operations and customer experience in dynamic retail environments.

In [22], Jain et al. evaluated the effectiveness of machine learning models for sales prediction in various industries by analyzing historical sales data and comparing the performance of different machine learning algorithms, such as decision trees, random forests, and neural networks. The researchers used historical sales data, including time-series data from retail stores spanning five years, to analyze patterns and trends in average transactions by day of week and sales across different product categories. The study explored various machine learning algorithms for sales prediction, including Linear Regression, Random Forest Regressor, and K-Neighbors Regressor. Linear Regression models the relationship between input variables and sales using an equation of the form $Y = mX + b$, allowing for predictions based on various factors. Future work aims to create an intuitive interface that enables firms to enter their sales data and generate sales projections. Results show that machine learning techniques can accurately predict sales with high accuracy, outperforming traditional statistical methods.

III. METHODOLOGY

We explore the whole dataset to understand the concept. Then, we analyze the attributes to ensure the distribution. After that, we converted the categorical data to numerical data by doing label encoding, identified the null values, and filled the null values with median. Also, the outliers were removed. We created some columns by doing feature engineering, which is necessary for our work to enhance outcome prediction. Subsequently, the null and alternative hypotheses were tested using statistical models and the results, and validated the hypotheses related to our research question. After that, we trained the machine learning models for three challenges (Time Series Demand Forecasting, Inventory Optimization, and Dynamic Pricing), analyzed the model's prediction, and evaluated the results. For visualizations, we used XA, I, which gave us a clear scenario for our study. At last, we did an unsupervised learning process for better prediction without model training and testing. Figure 1 illustrates our proposed workflow diagram.

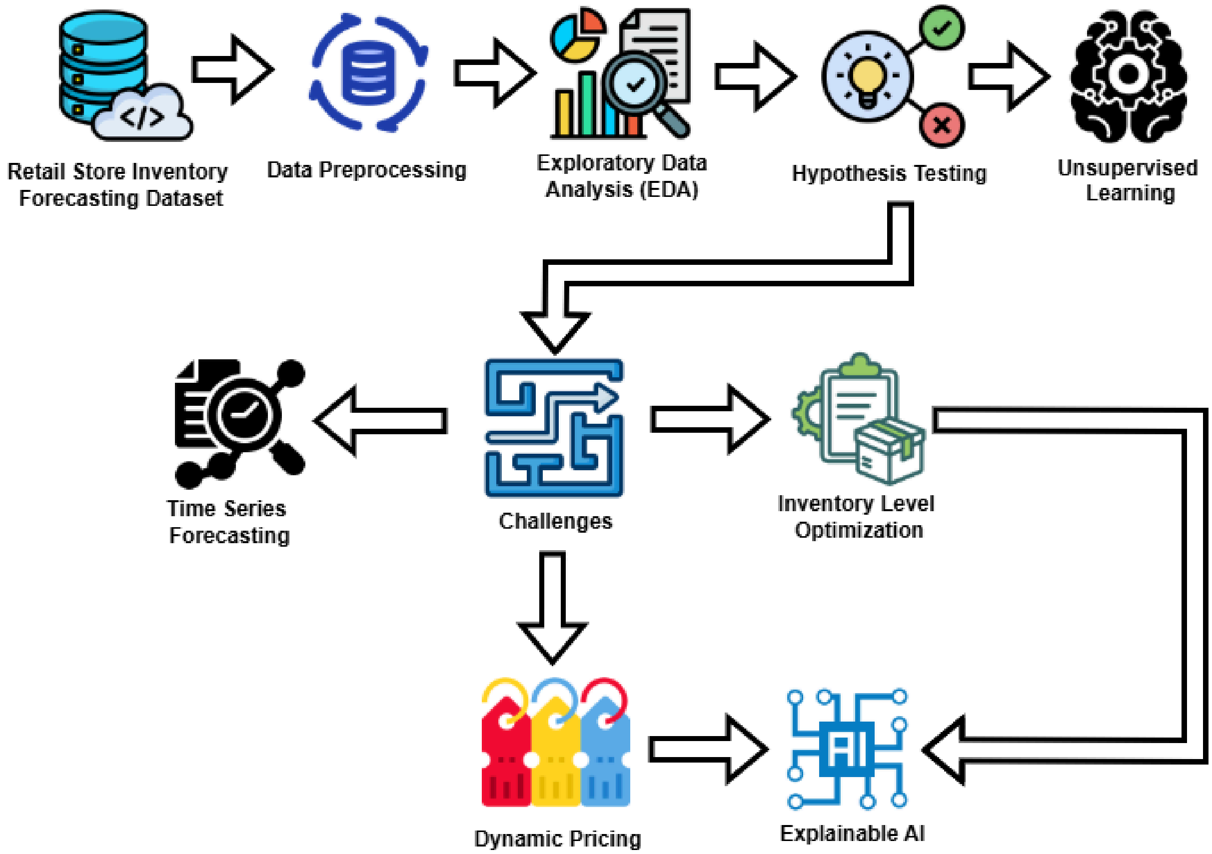

Fig.1 workflow diagram of our approach

A. DATA DESCRIPTION

For our research, we used a retail store inventory forecasting dataset from Kaggle [23] provided by Chauhan A. S. It's a synthetic dataset, yet the data is realistic for analyzing and forecasting retail store inventory demand. This data shows a good understanding of the retail store's inventory as well as the business matrix with forecast data. In addition, it gives us valuable insights for retail business research. The dataset consists of 15 attributes and 73000 data points.

All attributes act in the time duration of a day for a particular store. The dataset attributes include:

- Category: Represent the products in the store.
- Region: It consists of 4 areas, those are east, west, north and south
- Inventory level: The number of products available in store, ranging from 50 to 500
- Units sold: This represents in which quantities the product was sold per day. Ranging from 0 - 490
- Units ordered: The product ordered between 20-200 per day
- Demand forecast: It predicts the number of products a retail store will sell in a future period. Ranging from -9.99 to 519.
- Price: The price at which products are sold. Relays between 10 to 100

Table I shows the overall summary of the dataset. In this table, the Min and Max columns represent the minimum and maximum values remarked for each attribute, while the Mean and Standard Deviation columns provide the average value and the measure of variance for each attribute. Identifying time series demand forecasting, predicting the ideal price, and optimizing retail store inventory—this statistical analysis will be highly helpful.

TABLE I
Descriptive analysis of the dataset

Variables	Mean	Std	Min	Max
Inventory level	274.469877	129.949514	50	500
Units Sold	136.464870	108.919406	0	499
Units Ordered	110.004473	52.277448	20	200
Demand Forecast	141.494720	109.254076	-9.99	518.55
Price	55.135108	26.021945	10	100
Discount	10.009508	7.083746	0	20
Competitor Pricing	55.146077	26.191408	5.03	104.94

.

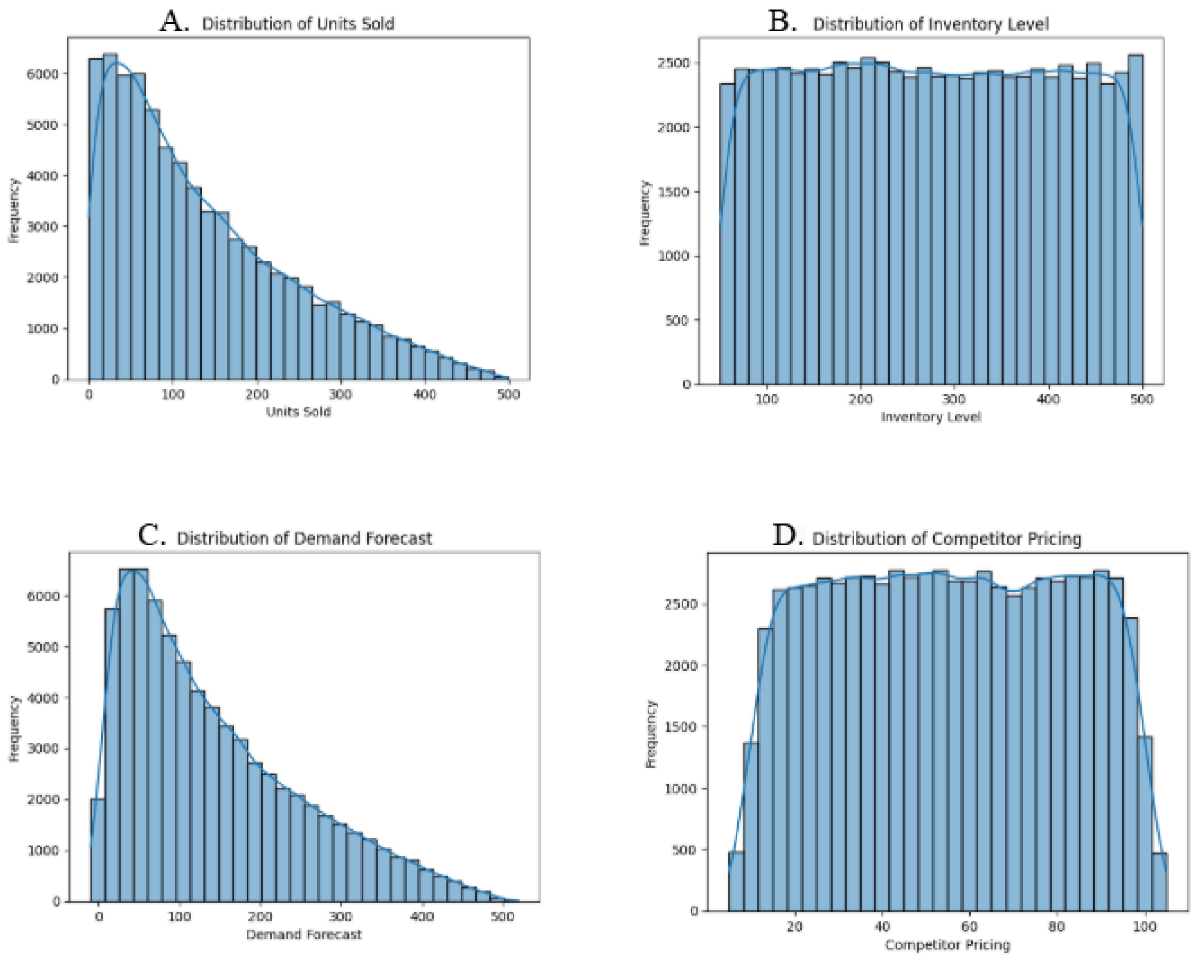

Fig 2. Distribution of Units Sold, Inventory Level, Demand Forecast, and Competitor Pricing

Inventory Level, Units Ordered, and Competitor Pricing are uniformly distributed, while Units Sold and Demand Forecast are left-skewed. The skew shows low sales or underestimations in forecasting. Stable stock levels and ordering patterns are evident from uniform distributions. Skewed data possibly requires transformations for better analysis.

B. HYPOTHESIS DEVELOPMENT

We formulated our research questions into hypotheses to analyze how the dataset's features impact retail store inventory forecasting by applying some statistical techniques to the working dataset. In our research question RQ1, we examine the relationship between inventory level and units sold to determine if stock availability impacts sales. Based on this, we formulate our first hypothesis H_{01}: There is no correlation between Inventory Level and Units Sold. On the other hand, the RQ2 focuses on the significant difference in demand forecast across different regions. In H_{02}, Demand Forecast does not differ across different regions. Additionally, RQ3 queries whether the accuracy of demand forecasting gets better or not over time. And the consequence regarding RQ3, H_{03} defines the accuracy of Demand Forecast does not improve over time. Moving on to RQ4, which explores the difference in sales variance across different stores. In the fourth null hypothesis, Sales variance is equal across different stores. Lastly, for RQ5, which examines the effect of competitor pricing on units sold across different regions, we proposed the following fifth hypothesis, which determines that the effect of competitor pricing on units sold does not vary across regions.

These hypotheses help us understand how inventory affects sales, whether forecasts differ by region or not, and if prediction accuracy improves over time, helps better retail decisions or not.

C. STATISTICAL TOOLS

Several statistical testing approaches are used to determine whether the null hypothesis should be rejected or failed to be rejected. For the first hypothesis (H1), Pearson's correlation coefficient was applied, which is a broadly used testing approach that measures the linear relationship between inventory level and units sold, and it ranges in value between -1 and +1. This testing shows that both variables are positively or negatively associated with each other by following the continuous and normal distribution. The correlation coefficient of 0.5902 indicates a moderate positive correlation between Inventory Level and Units Sold. The computed p-value was 0.0, lower than the conventional significance level ($\alpha = 0.05$). For the second hypothesis, we have used ANOVA (Analysis of variance). This test helps assess whether variations in demand forecasts are influenced by regional differences. Mostly, if the p-value associated with the F-statistic is less than 0.05, then we reject the null hypothesis, which indicates that all the means of the groups are not equal. We used Two-Way ANOVA instead of One-Way ANOVA when we wanted to analyze the effect of two independent variables (factors) on a dependent variable. We used Two-Way ANOVA to analyze if the effect of Competitor Pricing on Units Sold differs by Region. This test helps determine both the individual and combined impact of these two factors. Mainly, if the p-value corresponding to the interaction effect is less than 0.05, we reject the null hypothesis, which implies that the effect of Competitor Pricing on Units Sold varies significantly by Region. We used Mean Absolute Error (MAE) in hypothesis 4 to evaluate whether forecast accuracy improved over time by comparing the average absolute difference between actual sales (Units Sold) and forecasted demand in the first and last six months. A lower MAE in the later period would indicate improved accuracy. For the fifth hypothesis, Levene's Test was applied to estimate whether sales variance remained consistent across different Stores or not. This test is important for confirming the homogeneity of variance and indicates that differences in sales across stores are statistically notable.

D. UNSUPERVISED LEARNING

In this paper, based on the dataset, we applied unsupervised learning after cleaning and splitting the dataset. At first, we applied the Elbow method using K-Means for clustering. Here, Unsupervised learning is an approach where models analyze data without labeled outputs to find hidden patterns and also without structures or groupings. First of all, the Elbow Method is a technique used in clustering to find the optimal number of clusters by plotting the within-cluster sum of squares against the number of clusters. Here, we implemented different types of algorithms like K-Means Algorithm and PCA. K-Means Algorithm, a widely used clustering method that partitions data into k clusters by iteratively assigning points to the nearest centroid and updating the centroids. Moreover, Principal Component Analysis (PCA) is a dimensionality reduction technique that switches data into a set of uncorrelated principal components and captures the most significant variance, while K-Means works well for spherical clusters. PCA is often used before clustering to reduce high-dimensional data while keeping essential patterns.

– Elbow Method

The Elbow Method is a widely used technique in K-Means clustering to determine the optimal number of clusters for a dataset. It works by plotting the within-cluster sum of squared errors (WCSS) against different values of K and identifying the point at which the reduction in WCSS slows down significantly, forming an elbow shape in the graph. This elbow point represents the most appropriate number of clusters, ensuring a balance between accuracy and computational efficiency. In this study, the Elbow Method was applied by first preprocessing the dataset to remove inconsistencies and normalize values. The K-Means algorithm was then run with varying numbers of clusters, and the WCSS was calculated for each case. The results were plotted, allowing for a visual identification of the elbow point, which guided the selection of the optimal number of clusters. By using this data-driven approach, the clustering process avoids arbitrary decisions and ensures meaningful, interpretable outcomes.

– K-means clustering

K-means clustering is a widely used unsupervised machine learning algorithm for partitioning a dataset into a set of distinct clusters. The objective of K-means is to minimize the variance within each cluster while maximizing the variance between clusters. The algorithm operates by first initializing a set of K centroids (the number of clusters), usually chosen randomly. The data points are then assigned to the nearest centroid based on a distance metric, typically Euclidean distance. Once all points are assigned to clusters, the centroids are recalculated as the mean of the points within each cluster. This process of assignment and centroid update is repeated iteratively until convergence, which occurs when the centroids no longer change significantly or the maximum number of iterations is reached. K-means is particularly effective for large datasets and can be applied in various fields, such as market segmentation, image compression, and anomaly detection, though it has some limitations, such as sensitivity to the initial placement of centroids and the need for the user to specify the number of clusters in advance.

– Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction technique widely used in data analysis and machine learning. It transforms a dataset with potentially correlated features into a set of linearly uncorrelated variables called principal components. The main goal of PCA is to reduce the number of variables while retaining as much variance as possible, thus simplifying the dataset without losing significant information. This is achieved by identifying the directions (principal components) in which the data varies the most. Each component is a linear combination of the original features, and the first principal component captures the largest variance, followed by the second, and so on. PCA is often applied to preprocess data before performing tasks like regression, classification, or clustering. It helps improve computational efficiency, enhances model performance by reducing noise, and aids in better visualizing high-dimensional data.

E. SUPERVISED LEARNING

I. Time-series forecasting: ARIMA-LSTM hybrid (architecture diagram):

The ARIMA+LSTM hybrid model combines the strengths of both ARIMA (AutoRegressive Integrated Moving Average) and LSTM (Long Short-Term Memory) networks to improve time series forecasting. ARIMA is a classical statistical model that excels in capturing linear patterns in time series data, such as trends, seasonality, and noise. However, it may struggle with nonlinear relationships. To address this limitation, the residuals (errors) from the ARIMA model, which represent the part of the data not explained by ARIMA, are fed into an LSTM network. LSTM, a type of recurrent neural network, is well-suited to model complex, nonlinear dependencies in time series data. The hybrid approach first uses ARIMA to predict the linear components of the time series, and then applies the LSTM to model and predict the nonlinear residuals. The final forecast is obtained by combining the ARIMA predictions with the LSTM-predicted residuals, resulting in a more accurate and robust model for time series forecasting. This methodology leverages the complementary strengths of both models, making it particularly effective for datasets that exhibit both linear and nonlinear characteristics.

II. Inventory optimisation: Tree-based Regressors:

– Random Forest Regressor:

The Random Forest Regressor is an ensemble learning method used for regression tasks, leveraging multiple decision trees to improve predictive accuracy and reduce overfitting. In this study, the Random Forest Regressor was employed to model complex relationships within the dataset by constructing a collection of decision trees, each trained on a random subset of the data. The final prediction was obtained by averaging the outputs of all trees, ensuring robustness and stability in the results. This approach mitigates the limitations of individual decision trees, such as high variance, by aggregating multiple models to enhance generalization. The algorithm was implemented by selecting a suitable number of trees and tuning hyperparameters such as the maximum depth and minimum samples per split to optimize performance. By utilizing this method, the study achieved reliable and interpretable regression results, making it a suitable choice for handling nonlinear dependencies and high-dimensional data.

– Gradient Boosting:

Gradient Boosting is a powerful ensemble machine learning technique used for both regression and classification tasks, which builds models in a sequential manner to correct the errors made by previous models. In this approach, each new model is trained to predict the residuals—or the differences—between the actual values and the predictions of the previous models, effectively minimizing the loss function over iterations. The final prediction is made by aggregating the outputs of all individual models, typically decision trees, which collectively form a strong predictive model. Gradient Boosting is known for its high accuracy and ability to handle complex, non-linear relationships in data. Its flexibility in choosing loss functions and regularization techniques makes it suitable for applications where precision and performance are critical. In this study, Gradient Boosting was applied to capture intricate patterns in inventory dynamics and sales behavior, offering a robust solution for forecasting and decision-making.

– XGBoost:

XGBoost (Extreme Gradient Boosting) is a powerful and efficient machine learning algorithm based on the gradient boosting framework, widely used for supervised learning tasks like regression and classification. It improves upon traditional gradient boosting by incorporating several optimizations, including parallelization, regularization, and tree pruning, which enhance performance and reduce overfitting. XGBoost constructs an ensemble of decision trees in a sequential manner, where each subsequent tree corrects the errors of the previous one. This approach leads to high predictive accuracy, making it particularly suitable for large datasets with complex relationships. XGBoost also allows for custom objective functions and evaluation criteria, providing flexibility in handling a wide range of problems. Due to its speed and scalability, it has become a popular choice in data science competitions and real-world applications.

– Linear Regression:

Linear Regression is a fundamental statistical and machine learning technique used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. In its simplest form, known as Simple Linear Regression, it models the relationship between two variables by fitting a straight line, where the slope and intercept are determined to minimize the error between predicted and actual values. In cases involving multiple predictors, Multiple Linear Regression is used. This method assumes a linear relationship among the variables, constant variance of the errors, and independence of observations. In this study, Linear Regression was utilized to establish a baseline model for predicting target values based on historical data features, allowing for the comparison of its performance with more complex machine learning algorithms.

– Decision Tree:

A Decision Tree is a supervised machine learning algorithm used for both classification and regression tasks. It operates by recursively splitting the dataset into subsets based on feature values, aiming to achieve the most homogenous subsets with respect to the target variable. At each internal node, a decision is made based on the value of a feature that maximizes the homogeneity of the resulting partitions, typically using criteria like Gini impurity or Information Gain. The process continues until a stopping criterion is met, such as a maximum tree depth or a minimum number of samples per leaf. The resulting structure resembles a tree, with each branch representing a decision rule and each leaf node representing an outcome or prediction. Decision Trees are easy to interpret, as they provide a clear path from input features to the final decision, making them a popular choice in applications where interpretability is important. However, they are prone to overfitting, particularly with very deep trees, and may require techniques like pruning or ensemble methods, such as Random Forests, to improve their generalization ability.

III. Dynamic Pricing

In this part we used the same models except one (LightGBM) and in each model we used K-Fold cross validation and Randomized Search CV to improve the accuracy of prediction.

– LightGBM:

LightGBM (Light Gradient Boosting Machine) is a highly efficient and scalable gradient boosting framework developed by Microsoft. It is designed for large datasets and is particularly effective for tasks such as classification, regression, and ranking. LightGBM uses a novel tree learning algorithm that grows trees leaf-wise rather than level-wise, which results in a more accurate model and faster training times. This technique allows it to handle large-scale data with better computational efficiency compared to traditional gradient boosting methods. Additionally, LightGBM supports categorical features directly, reducing the need for pre-processing and providing a more streamlined modeling pipeline. It also includes several advanced techniques such as histogram-based learning and parallel and GPU learning, which further enhance its speed and scalability. LightGBM has gained widespread use in machine learning competitions and real-world applications due to its speed, accuracy, and ability to handle a variety of data types effectively.

– K-Fold Cross-Validation:

K-Fold Cross-Validation is a statistical method used for evaluating the performance and robustness of machine learning models. In this technique, the dataset is divided into 'K' equal-sized subsets or "folds." For each iteration, one fold is held out as the validation set, while the remaining K-1 folds are used for training the model. This process is repeated K times, with each fold serving as the validation set once. The results from each iteration are then averaged to provide a more reliable estimate of model performance, reducing the risk of overfitting and providing a comprehensive view of how the model generalizes to unseen data. K-Fold Cross-Validation is commonly used for both model selection and hyperparameter tuning, as it helps ensure that the model performs well across different subsets of the data.

– RandomizedSearchCV:

Randomized Search Cross-Validation (RandomizedSearchCV) is a method used to find the best hyperparameters for a machine learning model. Unlike traditional grid search, which exhaustively tests all possible combinations of hyperparameters, RandomizedSearchCV performs a random search over a specified hyperparameter space. It randomly selects a fixed number of hyperparameter combinations from the search space, evaluating each configuration's performance on a given task. This technique allows for a more efficient search, especially when the search space is large, as it does not need to evaluate every possible combination. RandomizedSearchCV is particularly useful when dealing with complex models and large datasets, as it can significantly reduce computational costs while still providing a good approximation of the optimal hyperparameters. This method is implemented in libraries such as

Scikit-learn and is often used in conjunction with cross-validation to assess the performance of each set of hyperparameters. By selecting the combination that maximizes the model's performance, RandomizedSearchCV helps in improving the generalization and accuracy of machine learning models.

F. EXPLAINABLE AI TECHNIQUE

Explainable AI (XAI) is a technique that helps to understand and explain the machine learning model's decisions. Two widely used methods for explaining models are SHAP (SHapley Additive exPlanations) summary plots and LIME (Local Interpretable Model-agnostic Explanations) tabular plots.

A SHAP summary plot gives an overall view of how each feature contributes to model predictions. It visualizes the impact and direction of each feature's performance for multiple predictions. It finds the main features behind a model prediction and decides it. This is a wonderful way to visualize and understand models, ensure fair predictions, and understand feature importance.

On the other hand, a LIME plot provides insights into individual predictions by locally similar to the complex model with the simple explainable linear model. LIME visualizes the essential feature in a specific decision by changing the input data. It observes changes in model predictions and makes it useful for the instance level.

IV. RESULTS

Before starting the experiment, we calculated any missing values in the dataset. After confirming that there were no missing values, we removed outliers using the IQR method and capping. To train the model, we first applied hypothesis testing to the dataset. Firstly, Pearson's correlation test was conducted to examine the relationship between Inventory Level and Units Sold. The null hypothesis(H_0) stated that there is no significant correlation between the two variables. The results showed a Pearson correlation coefficient of 0.590, indicating a moderate positive correlation. The p-value was 0, which is less than 0.05, meaning that the result is statistically significant. Thus, we reject the null hypothesis. This means there is a significant relationship between Inventory Level and Units Sold.

Similarly, an ANOVA test was performed to determine whether Demand Forecast significantly differs across different regions. By the testing we get,

TABLE II

ONE-WAY ANOVA

Source	**Sum of Squares (SS)**	**Degrees of Freedom (df)**	**F-Statistic**	**P-value**
Regions	35,653.58	3	0.601	0.614

From Table.II, the p-value obtained from the One-Way ANOVA test is 0.614, which is greater than the significance level of 0.05. When the p-value is greater than the common significance level (0.05), it suggests that there is not enough evidence to reject the null hypothesis (H_{01}), indicating that there is no significant difference in Demand Forecast across regions. In addition, a Two-Way ANOVA test was conducted to analyze whether the effect of Competitor Pricing on Units Sold varies across regions and the test results are given in Table III.

TABLE III

TWO-WAY ANOVA

Source	**Sum of Squares(SS)**	**Degrees of Freedom(df)**	**F-statistic**	**P-value**
Regions	35,653.58	3	0.502	0.493
Competitor Pricing	14,131,430.00	959	0.623	0.895
Interaction	28,766,850.00	2,877	0.422	0.993

Table III shows the p-value is much higher than the significance level of 0.05. This means the null hypothesis can’t be rejected, suggesting that Competitor Pricing does not significantly impact Units Sold differently across regions. Furthermore, the accuracy of the Demand Forecast over time for H_{04} was evaluated using Mean Absolute Error (MAE). The MAE for the first half of the data was 8.29, while for the second half, it was 8.32. Since there was no significant improvement in forecast accuracy, we accepted the null hypothesis. Lastly, Levene's test was performed to check if sales variance is equal across different stores. The test statistic was 1.634, and the p-value was 0.163, which is greater than 0.05. This led to the failure to reject the null hypothesis, indicating that sales variance is consistent across stores. Furthermore, the Table IV represents the overall overview of this hypothesis testing, values and it's final outcomes which indicates the acceptance of most of the hypotheses apart from the Hypothesis-1. Only one hypothesis was rejected (H_{01}), showing a significant correlation between inventory level and units sold and the others were accepted, indicating no major regional or temporal variation in forecast accuracy, demand, or competitor effects.

TABLE IV

Hypothesis Table

Hypothesis Number	**Null Hypothesis (H0)**	**Testing Tool**	**Testing Score**	**Final Result(Accept/ Reject H0)**
Hypothesis-1 (Pearson Correlation)	No significant correlation between Inventory Level and Units Sold	Pearson Correlation	Pearson Correlation Coefficient: 0.5901	Reject
Hypothesis-2 (ANOVA Test)	No significant difference in Demand Forecast across regions	ANOVA test	F- Statistic: 0.601	Accept

Hypothesis-3 (Two-Way ANOVA)	The effect of Competitor Pricing on Units Sold does not vary across Regions	Two Way ANOVA Test	F- Statistic: 0.8855	Accept
Hypothesis-4 (MAE Test)	The accuracy of Demand Forecast does not improve over time	Mean Absolute Error (MAE)	MAE (First Half): 8.2898 MAE (Second Half): 8.3229	Accept
Hypothesis-5 **(Levene's Test)**	Sales variance is equal across different Stores	Levene's Test	Levene's Test Statistic: 1.63378	Accept

After the statistical tools were performed for every null hypothesis, we applied Unsupervised Learning in a way that uncovers the hidden patterns or groupings within the features. Before that, we preprocessed the dataset and then used the Elbow method with the K-Means algorithm for clustering and finding the anomaly and using PCA visualization.

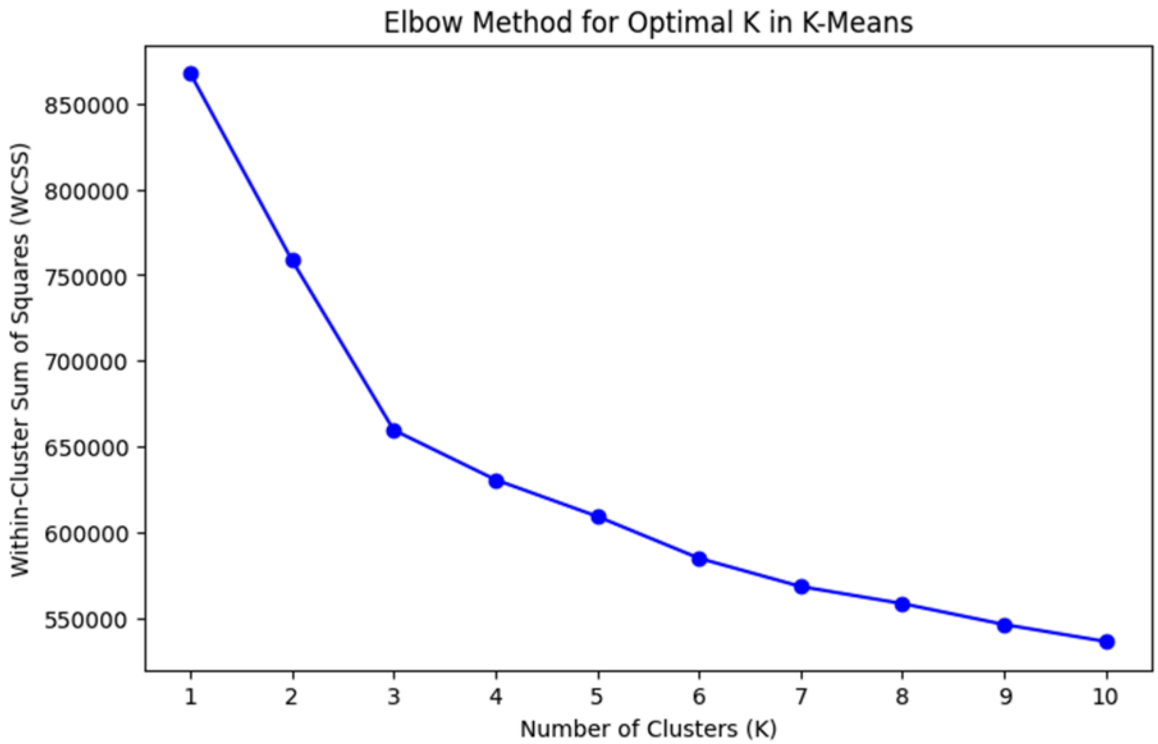

Fig 3. Elbow Method for Optimal K in K-means

By the visualization of K-Means shown above, the Elbow Method plot in Fig 3 shows the Within-Cluster Sum of Squares (WCSS) decreasing as the number of clusters (k) increases. The rate of decrease slows down after k=4, suggesting this is the optimal number of clusters where adding more clusters does not significantly reduce WCSS. In addition, we found that for the number of 3 clusters, the data points are well-separated into three distinct clusters, as shown in Fig. 4 and the Silhouette Score (K-Means) is 0.1249 and the Davies-Bouldin Index (K-Means) is 2.1846.

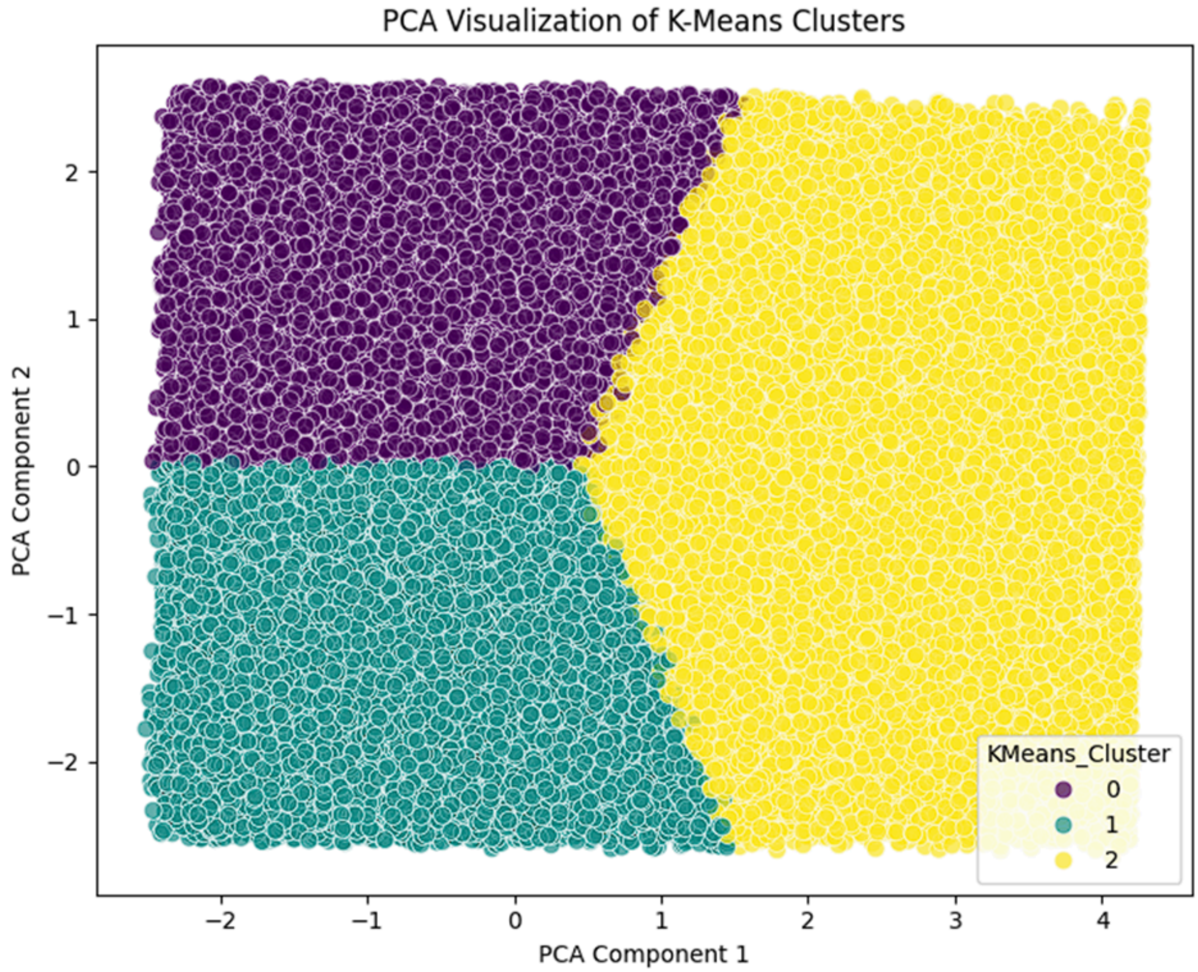

Fig 4. PCA visualization of K-Means Clusters

After the visualization of unsupervised learning (clustering), we analyze three different approaches as challenges. For each challenge, we added new additional features, used Label Encoder to convert the categorical data to numeric data, Min-Max scaler for feature scaling in order to convert the variables in range of 0 to 1 and a correlation heatmap, shown in Fig.5 to identify which pairs of attributes have the highest correlation and how well they correlate with each other for each analysis. Here, the color density indicates the strength of the correlation: the denser the color, the stronger the correlation between attributes.

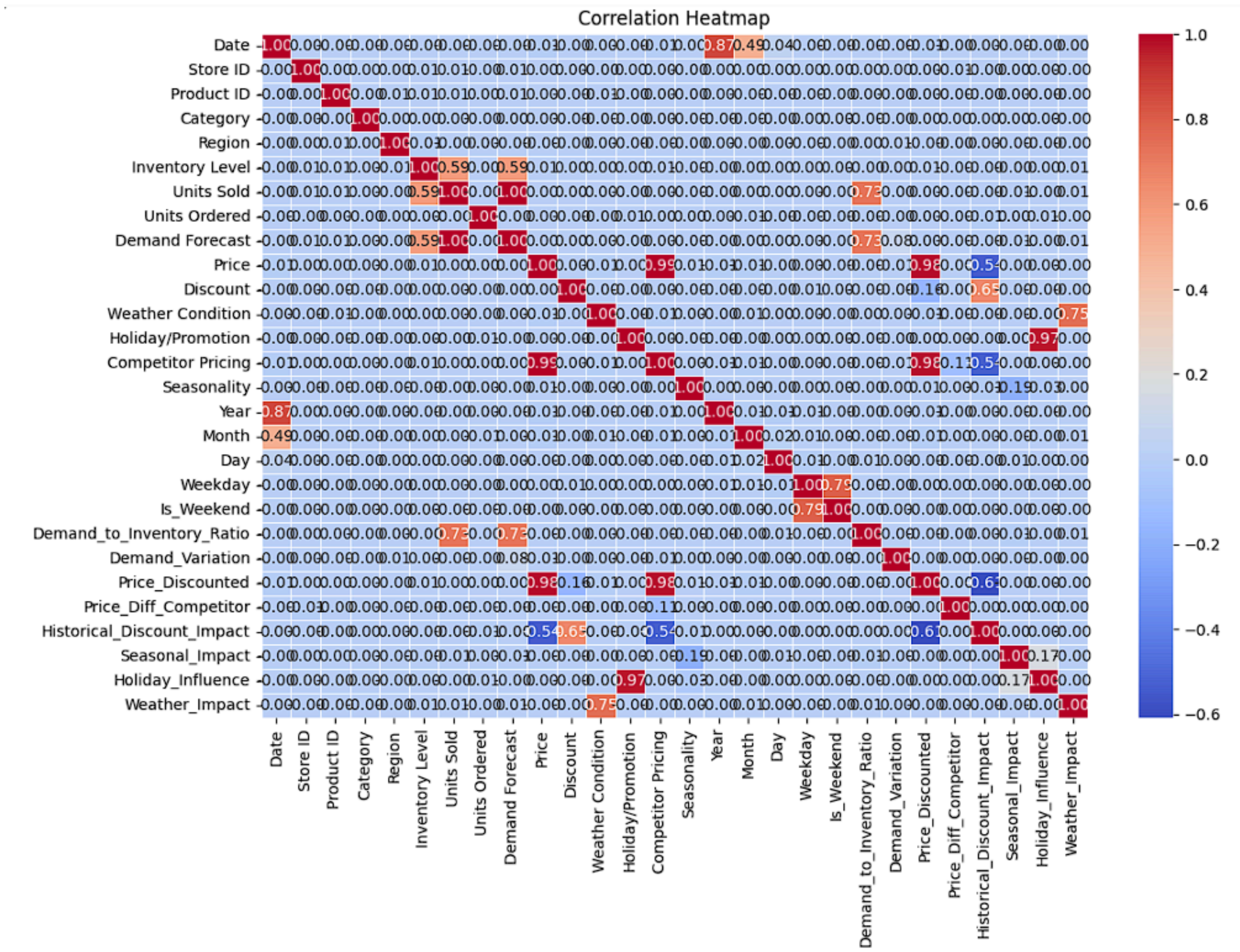

Fig. 5 (Correlation Heatmap)

After that, the train test split is performed by 80% (for model training purposes) and 20% (to test the model) individually for each challenge. By this correlation heatmap in Fig.5, we find that we should take the approach of regression analysis for this paper. For the first challenge (Time series demand forecasting), we applied a model that is a hybrid of ARIMA (linear machine learning model) and LSTM (non-linear deep learning model). First, it trains an ARIMA model on the training data and generates predictions. The residuals (errors) from the ARIMA model are then used as training data for an LSTM model, which learns patterns in the residuals. The LSTM is trained using a PyTorch-based neural network, and its predictions are added to the ARIMA forecasts to improve accuracy. Thus, we get the results from Table V that shows the Mean Absolute Error (MAE) is 0.0218, Mean Square Error (MSE) is 0.0007, RMSE is 0.0267 and the R square is 0.9884, which gives us the overview how well my model fits the data in this analysis.

TABLE V

Machine learning Model Performance (Time Series Forecasting)

Model Name	Mean Absolute Error (MAE)	Mean Square Error (MSE)	Root Mean Square Error (RMSE)	R Square Score
Hybrid ARIMA-LSTM	0.0218	0.0007	0.0267	0.9884

The challenge also visualized the time series forecasting for the Units sold around time in Fig.6. This plot shows Units Sold Over Time from early 2022 through early 2024 where X-axis contains the date from Jan 2022 to Jan 2024 (daily or near-daily data) and Y-axis represents Units Solds that range from about 100 to just under 200. Here, the blue shading suggests a confidence interval or moving range, which also fluctuates with the data. From Fig.6, we can see no strong upward or downward trend visible over the two years and the seasonality is subtle which has no clearly repeating patterns by month or quarter.

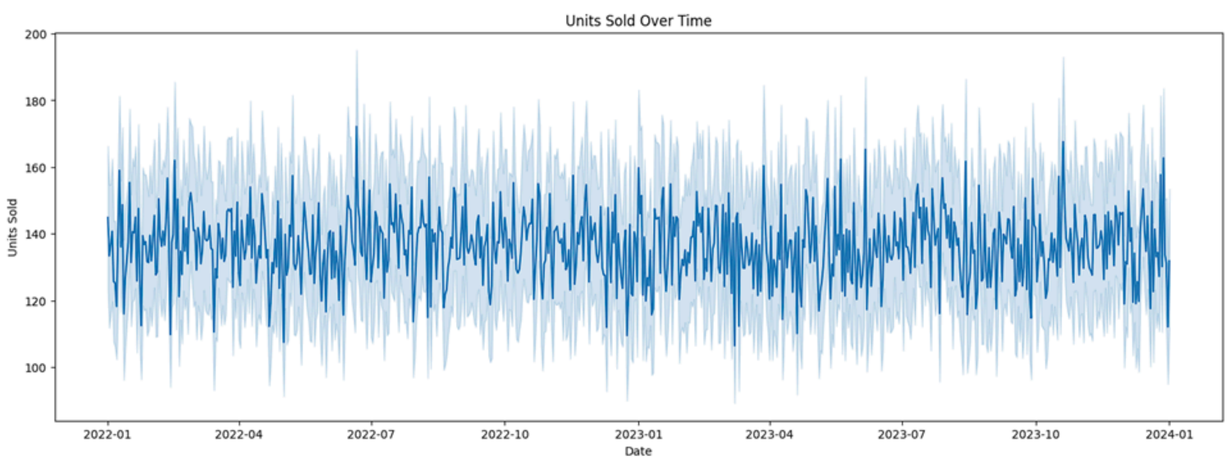

Fig 6. Time Series Forecasting Units Sold Over Time

On the other hand, for the second challenge (Inventory optimization), we used five different models, such as Linear Regression, Decision Tree, Random Forest Regressor, Gradient Boosting, and XGBoost, shown in Table VI. From that table we can see that the model Decision tree for this analogy has a R square value of 0.988 and generates Mean Square Error of 141.085. But, the rest goes with the R square value of 0.994 where we find Gradient Boosting as the best fitted model. Even so, the Gradient Boosting model performs well in this analysis with an MSE of 8.477 and R square of 0.994, we choose Random Forest Regressor as the best-suited model to explain the analysis with Explainable AI. Because it is a tree based model which gives the most optimum and accurate explanation through the Explainable AI techniques.

TABLE VI

Machine learning Model Performance (Inventory Optimization)

Model Name	**Mean Absolute Error (MAE)**	**Mean Square Error (MSE)**	**Root Mean Square Error (RMSE)**	**R Square Score**
Linear Regression	**7.472**	**74.664**	8.641	0.994
Decision Tree	**9.616**	**141.085**	11.878	0.988
Random Forest Regressor	**7.298**	**73.408**	8.568	0.994
Gradient Boosting	**7.279**	**71.860**	8.477	0.994
XGBoost	**7.287**	**73.063**	8.548	0.994

Afterwards we have applied the Explainable AI. The Fig.7 shows that the feature Demand Forecast has the highest impact, followed by Inventory Level, indicating that these two factors play a crucial role in our analysis. Other variables like Sales RollingMean 7 and Sales Lag 1 have minimal influence, while factors such as Competitor Pricing, Price, and Discount show almost no impact.

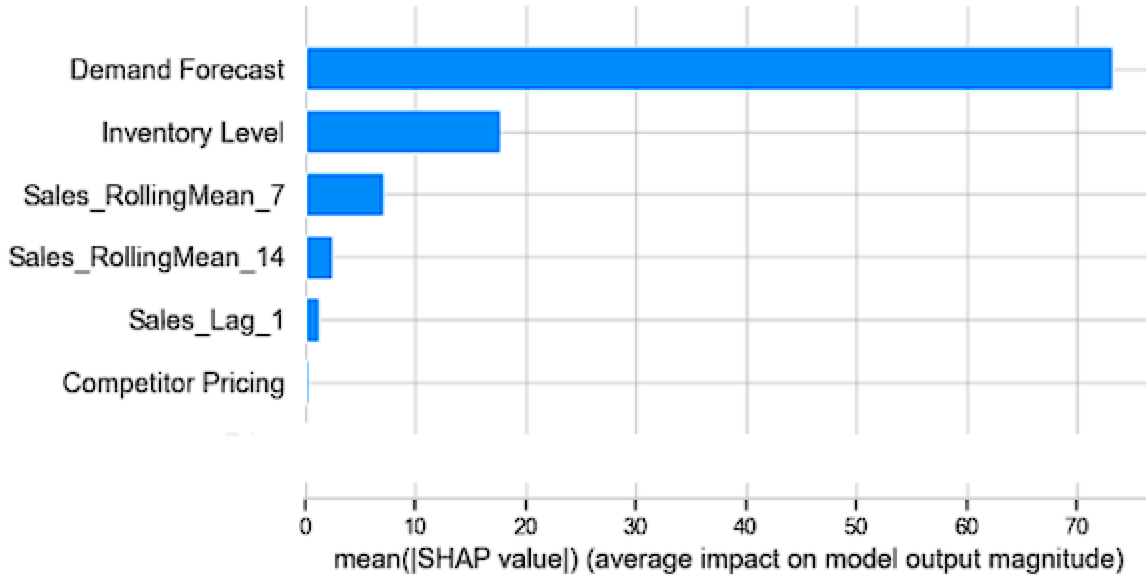

Fig 7. SHAP Summary Plot (Inventory Optimization) Using Random Forest

On the other hand, the LIME (Local Interpretable Model-agnostic Explanations) visualization in Fig.8 explains the impact of different features on a single prediction. The red bars indicate Demand Forecast, Inventory Level, Sales RollingMean 7, Units Ordered, and DayOfWeek that negatively affect the predicted value, while the green bars represent Discount, Price, and Sales Lag 7 that contribute positively. The most influential factor is Demand Forecast, which significantly decreases the prediction when it falls between 53.76 and 113.02, followed by Inventory Level and Sales RollingMean 7. On the other hand, factors like Discount and Price have minimal but positive contributions, suggesting they slightly increase the predicted value. Those factors lead the prediction to 55.71.

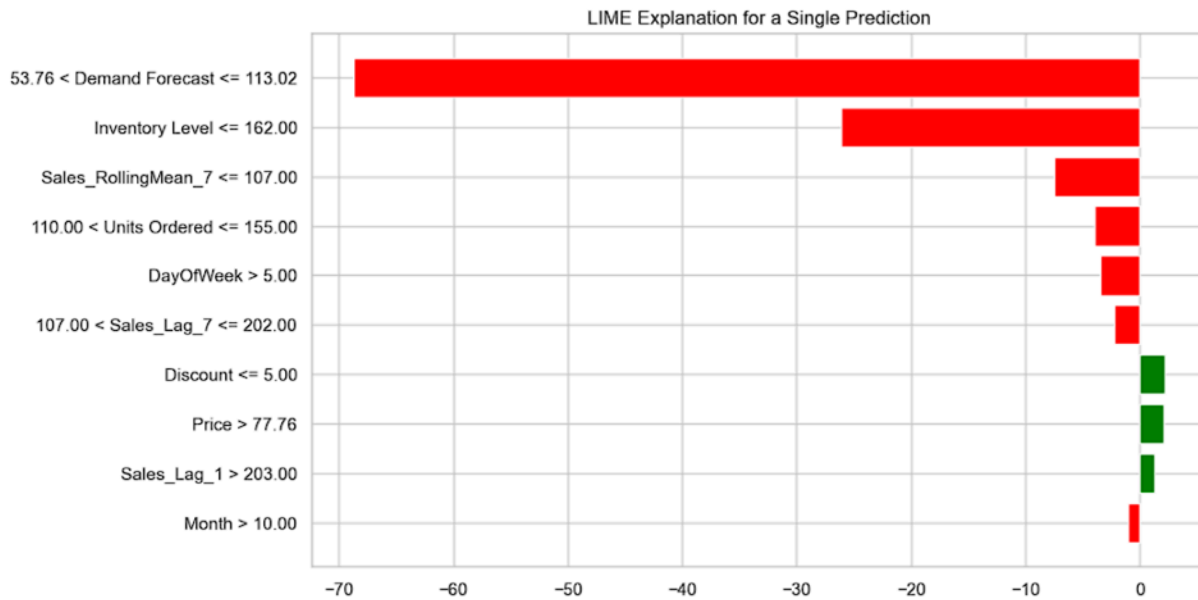

Fig 8. LIME Summary Plot (Inventory Optimization) Using Random Forest

The LIME summary table in Table 6 gives an overview of the predicted values which take part in this prediction both positive and negative.

<u>TABLE VII</u>

LIME SUMMARY TABLE

Feature	**Values**
Demand Forecast	60.06
Sales RollingMean 7	95.00
Sales RollingMeans 7	95.71
Units Ordered	127.00
DayOfWeek	6.00
Sales Lag 7	110.00
Discount	5.00
Price	80.00

Sales Lag 1	207.00
Month	11.00

Lastly, for challenge three (Dynamic Pricing), we applied K-Fold cross-validation with 5 folds to thoroughly validate the models accuracy and Randomized Search CV for hyperparameter tuning to search the best parameters for the model to analyze our prediction precisely to each model and visualized the loss curve for the best model. The results from each model are shown in TABLE 7.

TABLE VIII

Machine learning Model Performance (Dynamic Pricing)

Model Name	**Mean Absolute Error (MAE)**	**Mean Square Error (MSE)**	**Root Mean Square Error (RMSE)**	**R Square Score**
Linear Regression	**0.0008**	**0.0000**	**0.0026**	**0.9999**
Decision Tree	**0.0013**	**0.0000**	**0.0020**	**1.000**
Random Forest	**0.0004**	**0.0000**	**0.0009**	**1.0000**
Gradient Boosting	**0.0010**	**0.0000**	**0.0014**	**1.0000**
XGBoost	**0.0013**	**0.0000**	**0.0018**	**1.0000**
LightGBM	**0.0015**	**0.0000**	**0.0019**	**1.0000**

From these results of Table 7, we can see that apart from Linear Regression, other five models (Decision Tree, Random Forest, Gradient Boosting, XGBoost, LightGBM) fit the data accurately with the R square value of 1.00. From these findings, even though Random Forest has the lowest RMSE value of 0.0009, we choose the XGBoost model which seems faster than the other models to apply Explainable AI for the faster prediction with the visualization of the learning curve in Fig.9.

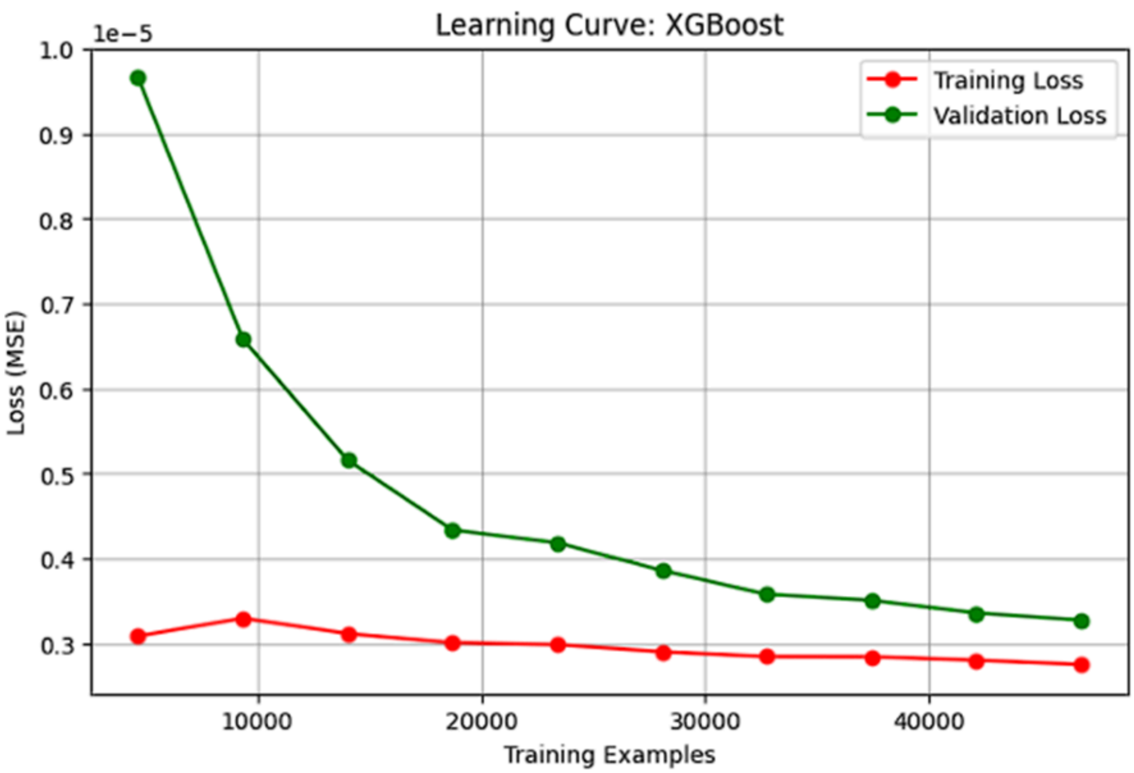

Fig 9. Learning Curve of XGBoost

The Fig.9 plot of the learning curve of XGBoost states the plot of loss (MSE) on the Y-axis against the training example on the X-axis. Here, the red line shows the training loss which interprets that the training loss is quite low and slightly decreases as the number of training examples increases and starts around 0.0000032 and drops to about 0.0000028. On the other hand, the green line shows the validation loss which interprets that The validation loss is higher than the training loss, especially at smaller training sizes and it decreases significantly as the training data increases from ~0.0000095 down to ~0.0000034. At first, Fig.9 shows a large gap between the training loss and validation loss which indicate the overfitting. Then, after more data is added, the gap narrows down and the validation improves indicating the model generalized better. Thus, we can conclude that the XGBoost model is learning effectively.

After the verification with the learning curve we applied SHAP LIME as the Explainable AI techniques. This SHAP summary plot displayed in Fig.10 proves that Competitor Pricing has the highest influence, indicating that changes in competitor pricing significantly affect the predicted outcome. Price Discounted and Price Diff Competitor also have noticeable effects, suggesting that pricing strategies play a crucial role in the prediction of the model. Other features like Inventory Level, Demand Forecast, and Weather Condition have a minimal impact.

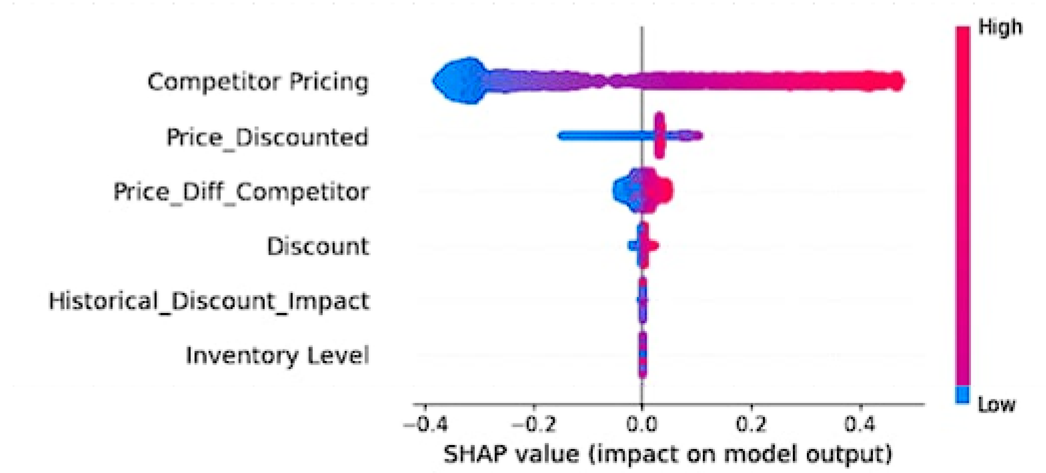

Fig 10. SHAP Summary Plot (Dynamic Pricing) Using XGBOOST

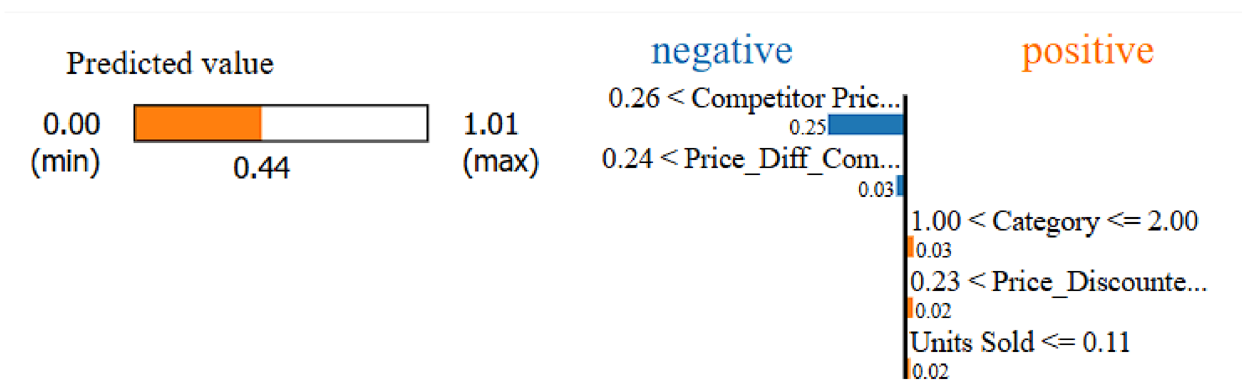

Fig 11. LIME Summary Plot (Dynamic Pricing)

The LIME plot in Fig.11 explains the prediction of a machine learning model for a specific instance. The predicted value is 0.44, as shown on the left. Features are categorized into negative (blue), which reduce the prediction, and positive (orange), which increase it. Competitor Pricing and Price Diff Competitor negatively impact the prediction, while Category, Price Discounted, and Units Sold contribute positively. The right side lists feature values, showing Competitor Pricing at 0.44 and Price Discounted at 0.38, influencing the final prediction. Then, Table 8 gives an overall summary of the predicted values which influences the final prediction.

TABLE IX

LIME SUMMARY TABLE

Feature	Values
Competitor Pricing	0.44
Price Diff Competitor	0.47
Category	2.00
Units Ordered	0.38
DayOfWeek	0.03

VI. DISCUSSIONS

In this paper, we are now discussing the hypothesis and every challenge that we analyzed before step by step.

— Hypothesis Test

- In the initial phase of model training, statistical hypothesis testing was employed to validate key assumptions about the data. The Pearson correlation test examined the relationship between Inventory Level and Units Sold, where the null hypothesis (H_0) proposed no correlation. The resulting correlation coefficient of 0.590 and a p-value of 0 indicated a moderate yet statistically significant positive correlation. It leads to the rejection of the null hypothesis and confirms that inventory levels meaningfully influence units sold. This type of relationship is crucial for building predictive models, as it shows the dependency between stock availability and sales performance.
- To evaluate regional variation in demand, a One-Way ANOVA test was conducted with Demand Forecast as the dependent variable and Region as the independent factor. The results in Table II, revealed a p-value of 0.614, which is well above the conventional 0.05 significance threshold. So, the demand patterns are consistent across different geographic regions. the result explains that regional segmentation may not be necessary for demand forecasting, allowing for more generalized models without sacrificing accuracy.

- A Two-Way ANOVA test was further applied to analyze the joint effect of Competitor Pricing and Region on Units Sold. As shown in Table III, the p-values for both main effects and their interaction (ranging from 0.493 to 0.993) far exceeded the 0.05 threshold. The results show that Competitor Pricing does not significantly influence sales differently across regions. This explains that internal pricing and inventory factors have a more direct impact on sales volume than external competitive dynamics which shows the feature selection process for model development.
- Additional hypothesis tests evaluated the temporal consistency and spatial variance of sales related metrics. The comparison of Demand Forecast accuracy over two time periods using Mean Absolute Error (MAE) showed negligible difference (8.29 vs. 8.32), leading to the acceptance of the null hypothesis that forecast accuracy remained stable over time. Similarly, Levene's test for homogeneity of variance across stores yielded a p-value of 0.163, suggesting uniform sales variance among stores. As summarized in Table IV, only one hypothesis (H_{01}) was rejected, reinforcing the overall stability of sales and demand patterns across time and space, and affirming that the model can be trained effectively with minimal adjustments for regional or temporal variability.

— Time Series Forecasting

- To forecast daily product demand across stores, we employed a hybrid ARIMA-LSTM model that effectively combines ARIMA's ability to model linear trends with LSTM's strength in capturing nonlinear temporal dependencies. This approach was chosen to handle the complexities in retail sales data which includes seasonality, sudden demand shifts, and long-term patterns. By training on daily data from January 2022 to January 2024, the model was able to learn both historical behavior and project future demand with high accuracy.
- The model performance was evaluated using standard regression metrics. It achieved a Mean Absolute Error (MAE) of 0.0218, Mean Squared Error (MSE) of 0.0007, and Root Mean Squared Error (RMSE) of 0.0267 with an R^2 score of 0.9884. These values reflect a minimal error margin and indicate that the hybrid model explains over 98% of the variation in daily units sold. So, it confirms its effectiveness for demand forecasting tasks.
- As shown in Fig.6, the Units Sold Over Time visualization presents the model's predictions from early 2022 through early 2024. The X-axis spans daily dates, while the Y-axis captures units sold, ranging from around 100 to just under 200. The central forecast line is surrounded by a blue-shaded confidence interval which highlights the model's ability to quantify uncertainty and adapt to fluctuations in the data over time.
- Despite strong results, integrating ARIMA with LSTM required careful preprocessing to align the model assumptions—stationarizing data for ARIMA and scaling for LSTM. Additionally, handling daily sales data which means addressing anomalies such as out-of-stock days and promotional spikes. Overall, this hybrid approach not only enhances forecasting accuracy but also supports more informed inventory management and operational planning across retail stores.

— Inventory Optimization

- To optimize inventory levels and mitigate issues of stockouts and overstocking, we explored five machine learning models—Linear Regression, Decision Tree, Random Forest, Gradient Boosting, and XGBoost. According to Table VI, Gradient Boosting achieved the best performance with the lowest MSE (8.477) and a high R^2 of 0.994, indicating strong predictive accuracy. While most models achieved similar R^2 values, the Decision Tree underperformed slightly with a higher MSE of 141.085 and R^2 of 0.988, suggesting it was less effective in capturing complex inventory patterns.

- Despite Gradient Boosting top performance, Random Forest was chosen for its superior interpretability using Explainable AI techniques. With an MSE of 73.408 and an R^2 of 0.994, it offered a strong balance between accuracy and transparency. Its tree-based structure made it well-suited for providing clear, actionable insights, an essential requirement for real-world inventory decisions where understanding why a prediction is made is just as important as the prediction itself.

- Explainable AI analysis in Fig.7 highlighted that Demand Forecast was the most impactful feature driving inventory decisions, followed closely by Inventory Level. These features align with practical inventory management priorities, emphasizing internal demand signals over external influences. Other features such as Sales RollingMean 7 and Sales Lag 1 had limited effect, while Competitor Pricing, Price, and Discount were found to be largely insignificant in shaping inventory predictions.

- Further interpretability was achieved through LIME in Fig.8, which analyzed feature influence at an individual prediction level. It revealed that Demand Forecast had the strongest negative impact, especially when falling between 53.76 and 113.02 units, reducing the predicted inventory need. Inventory Level and Sales RollingMean 7 also contributed negatively, while Discount, Price, and Sales Lag 7 added modest positive influence. These local insights help tailor inventory actions for specific dates and locations, supporting more nuanced and effective inventory control.

— Dynamic Pricing

- In tackling the dynamic pricing challenge, we employed K-Fold cross-validation (5 folds) to rigorously assess model stability and Randomized Search CV for hyperparameter optimization across various models. This combination ensured not only accuracy but also robustness in model performance. The evaluated models included Linear Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, and LightGBM. According to the results summarized in Table 7, all models except Linear Regression achieved near-perfect fits with R^2 scores of 1.00, confirming their suitability for capturing price variation patterns across products and stores.

- Among these models, Random Forest demonstrated the lowest RMSE (0.0009), suggesting high predictive accuracy. However, XGBoost was selected as the best model for further analysis due to its faster execution time and strong performance, making it ideal for real-time pricing strategies. XGBoost was also integrated with Explainable AI tools to interpret predictions clearly and quickly, a key requirement in dynamic pricing where rapid and explainable decisions are vital. This dual advantage of speed and interpretability gave XGBoost an edge over slightly more accurate models.

- As shown in Fig.9, the learning curve for the XGBoost model plots training and validation loss (MSE) against the number of training examples. Initially, there is a noticeable gap between training and validation loss—indicative of overfitting—where the training loss starts very low (~0.0000032) and validation loss relatively higher (~0.0000095). However, as the training size increases, this gap diminishes, with both losses converging and decreasing over time, reflecting improved generalization. This learning pattern confirms that XGBoost is not only fitting well but also learning effectively and avoiding overfitting with increased data.

- To understand individual predictions, we used LIME visualization, as shown in Fig.11, which breaks down the contributing factors to a specific prediction outcome (0.44). Features like Competitor Pricing and Price Diff Competitor had a negative impact, lowering the predicted value (shown in blue), while Category, Price Discounted, and Units Sold positively influenced the prediction (shown in orange). These insights are consistent with pricing logic, where competitive pricing pressure tends to reduce the value, while discounting and high demand increase it. Table 8 complements this by summarizing the aggregated influence of features across predictions, providing a holistic view of the decision dynamics in the pricing model.

VII. CONCLUSION AND FUTURE WORK

Unbalanced demand forecasting not only hampers retail business but also produces unoptimized inventory, causing long-term financial losses. On the other hand, dynamic pricing helps businesses compete with others, resulting in an extended version of sales outcomes. Additionally, optimizing inventory is the best solution to prevent dead stock. Moreover, maintaining a low inventory turnover ratio requires excellent demand forecasting and inventory optimization. To address these challenges, we first created five hypotheses. Only the first hypothesis showed a significant correlation between inventory level and units sold, while the others failed to reject the null hypothesis. Our vision was clear: determine effective demand forecasting and optimize inventory as adequately as possible, alongside implementing dynamic pricing. To improve our analysis, we enhanced the data with additional features to find better correlations. After level encoding and scaling, we split the data for training. Finally, we applied unsupervised learning techniques to discover hidden patterns and identify groupings through clustering. Thus, we used the Elbow method with K-Means algorithms and PCA visualization. For time series demand forecasting and inventory optimization, the targeted attribute is unit sold, and we have applied Hybrid ARIMA LSTM in demand forecasting. In inventory optimization, we have applied various models such as Linear Regression, Decision Tree, Random Forest, Gradient Boosting, and XGBoost. Among them, the best-performing and most accurate fit model is Random Forest. For dynamic pricing, XGBoost's performance was unbeatable. After selecting the best performative model, we fine-tuned the prediction and applied SHAP and lime for better-predicted entity visualization. Though we have gained some achievements, we also have limitations that might be improved in the future. Our paper is bound by the available historical data, which can not capture all external factors. So, we can apply those methods to enhance data with additional external factors. Again, the model did not show any improvements over time as it may not be enough to capture evolving trends. We also failed to build a recommendation system and real-time app that can be done in the future. Furthermore, the hybrid model can also be performed through XAI, which we failed to perform.

REFERENCES

[1]Sihotang, J. (2023). Optimization of Inventory Ordering Decision in Retail Business using Exponential Smoothing Approach and Decision Support System. *International Journal of Mechanical Computational and Manufacturing Research*, *12*(2), 46-52.

[2] Li, M., & Mizuno, S. (2022). Dynamic pricing and inventory management of a dual-channel supply chain under different power structures. *European Journal of Operational Research*, *303*(1), 273-285.

[3] Punia, S., & Shankar, S. (2022). Predictive analytics for demand forecasting: A deep learning-based decision support system. *Knowledge-Based Systems*, *258*, 109956.

[4] Hasan, M. R., Kabir, M. A., Shuvro, R. A., & Das, P. (2022). A comparative study on forecasting of retail sales. *arXiv preprint arXiv: 2203.06848*.

[5] Haque, M. S., Amin, M. S., & Miah, J. (2023). Retail demand forecasting: a comparative study for multivariate time series. *arXiv preprint arXiv: 2308.11939*.

[6] Singhal, K., Singh, V., & Kaul, A. (2023). Smart retail: Utilizing machine learning for demand prediction, price strategy, and inventory management. *IEEE Transactions on Engineering Management, 70*(4), 1771-1783.

[7] Falatouri, T., Darbanian, F., Brandtner, P., & Minner, S. (2023). Predictive analytics for demand forecasting–a comparison of SARIMA and LSTM in retail SCM. *Procedia Computer Science, 217*, 105-114.

[8] Taparia, V., Mishra, P., Gupta, N., & Singh, A. (2022). Improved demand forecasting of a retail store using a hybrid machine learning model. *Journal of Green Engineering, 12*(10), 11556-11575.

[9] Bi, X., Adomavicius, G., Li, W., & Qu, A. (2022). Improving sales forecasting accuracy: A tensor factorization approach with demand awareness. *INFORMS Journal on Computing, 34*(6), 2617-2633.

[10] Kumar, P., Choubey, D., & Amosu, O. R. (2022). *AI-enhanced inventory and demand forecasting: Using AI to optimize inventory management and predict customer demand*. ResearchGate.

[11] Baloian, N., Frez, J., & Pino, J. A. (2022). Retail indicators forecasting and planning. In *Proceedings of the International Conference on Information Systems and Technologies (ICIST)* (pp. 1-6).

[12] Saha, P., Gudheniya, N., Mitra, R., Das, D., & Narayana, S. (2023). Demand forecasting of a multinational retail company using deep learning frameworks. *Procedia Computer Science, 217*, 115-124.

[13] Hübner, A., Hense, J., & Dethlefs, C. (2023). The revival of retail stores via omnichannel operations: A literature review and research framework. *European Journal of Operational Research, 305*(3), 965-984.

[14] Wang, C. H., & Gu, Y. W. (2022). Sales forecasting, market analysis, and performance assessment for US retail firms: A business analytics perspective. *Applied Sciences, 12*(17), 8480.

[15] Martins, E., & Galegale, N. V. (2022). Retail sales forecasting information systems: Comparison between traditional methods and machine learning algorithms. In *IADIS International Conference e-Business and e-Government 2022* (pp. 1-8).

[16] Best, J., Glock, C. H., Grosse, E. H., & Rekik, Y. (2023). On the causes of positive inventory discrepancies in retail stores. *The International Journal of Retail & Distribution Management, 51*(1), 1-18.

[17] Esrar, H., Zolfaghariania, H., & Yu, H. (2023). Inventory management practices at a big-box retailer: A case study. *The International Journal of Retail & Distribution Management, 51*(2), 175-194.

[18] Wellens, A. P., Boute, R. N., & Udenio, M. (2024). Simplifying tree-based methods for retail sales forecasting with explanatory variables. *European Journal of Operational Research, 311*(3), 1147-1162.

[19] 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.

[20] Ganguly, P., & Mukherjee, I. (2024). Enhancing retail sales forecasting with optimized machine learning models. *arXiv preprint arXiv:2410.13773*.

[21] Hossam, A., Ramadan, A., Magdy, M., Abdelwahab, R., Ashraf, S., & Mohamed, Z. (2024). Revolutionizing retail analytics: Advancing inventory and customer insight with AI. *arXiv preprint arXiv: 2405.00023*.

[22] Jain, H., Dattpalsinh, V., Ray, S. K., & Vishal, D. (2023). Sales prediction using machine learning. *SSRN Electronic Journal*.

[23] Chauhan A. S. "Retail Store Inventory Forecasting Dataset" 2024 https://www.kaggle.com/datasets/anirudhchauhan/retail-store-inventory-forecasting-dataset