Intelligent System for Inventory Demand Forecasting

Authors: Sreeja Pasupuleti, Madhumitha Gannavaram, Sri Sai Lahari Gandrapu

Abstract

This project develops an intelligent inventory management system for Indiana University dining halls, focusing on demand forecasting and waste reduction. Using advanced machine learning models like ARIMA, Holt-Winters, and MLP, it predicts inventory needs accurately. The data pipeline includes preprocessing with StandardScaler, label encoding, and cyclic feature engineering using sine and cosine transformations. Additionally, an LLM enables real-time inventory insights and decision-making. The system significantly improves forecasting accuracy and operational efficiency, demonstrating the potential of AI in inventory management.

Keywords: Inventory Management, Time Series Forecasting, ARIMA, Holt-Winters, Multilayer Perceptron (MLP), Data Preprocessing, StandardScaler, Cyclic Encoding, Feature Engineering, Large Language Model (LLM), Demand Prediction, Operational Efficiency.

1 Introduction

Effective inventory management is essential for optimizing operations and reducing costs, particularly in environments like dining halls where perishable goods play a critical role. Our project addresses challenges in inventory forecasting by leveraging advanced time series machine learning techniques, including ARIMA, Holt-Winters, and Multilayer Perceptron (MLP) models. These algorithms facilitate precise demand forecasting, enabling businesses to minimize overstocking and under-stocking, thus reducing waste and enhancing cost-efficiency.

The project employs a robust data pipeline comprising exploratory data analysis (EDA), feature engineering, and standardization of historical sales data to extract meaningful insights. Additionally, we introduce a novel integration of a Large Language Model (LLM) trained on inventory data to support dynamic decision-making and real-time information retrieval. The proposed system enhances forecasting accuracy and operational efficiency, paving the way for intelligent inventory automation at Indiana University's dining halls.

2 Methodology

A systematic, step-by-step approach was employed to complete the project, beginning with data collection to ensure sufficient data for robust analysis. The data preprocessing stage involved handling missing values, detecting outliers, and standardizing the dataset to prepare it for effective modeling. Several time series forecasting algorithms were implemented to predict order demand.

Exponential Smoothing and its advanced variant, the Holt-Winters method, were utilized to model data trends and seasonal patterns by assigning higher weights to recent observations. The ARIMA (AutoRegressive Integrated Moving Average) model was applied to capture and forecast temporal dependencies by leveraging its components: Autoregression (AR), Differencing (I), and Moving Average (MA). Additionally, a Multilayer Perceptron (MLP), a type of artificial neural network, was implemented to capture non-linear relationships and dependencies in the data. Model performance was evaluated using the Root Mean Squared Error (RMSE) metric to quantify the average difference between predicted and observed values, ensuring the selection of the most accurate model for forecasting future inventory demands.

3 Exploratory Data Analysis (EDA)

This section summarizes the dataset's structure, dimensions, and key statistics, focusing on missing values, distribution of Order_Demand, and unique values in categorical columns like Warehouse and Product_Name. Outlier detection was also performed to ensure data integrity.

3.1 Initial Data Overview and Quality Assessment

The dataset was successfully loaded from a CSV file using Pandas, revealing its structure and the relationships between various product order columns. An assessment of data quality identified some missing values, indicating areas that may require imputation, while data types were confirmed to be appropriate for analysis. Descriptive statistics provided insights into the distribution and variability of Order_Demand, and unique counts highlighted a diverse range of Warehouse and Product_Name entries, indicating a broad selection of products and storage locations.

3.2 Outlier Detection

- Outliers in Order_Demand were identified using the IQR method
- Outlier Flagging: Rows with outliers were marked, revealing a small percentage of data points that could skew analysis, warranting further investigation.
- The Interquartile Range (IQR) method was used to identify outliers in the Order Demand data. The IQR, representing the spread of the middle 50% of the data between the first quartile (Q1) and third quartile (Q3), was calculated. Values outside the range ([Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]) were flagged as outliers. A new column, **Outlier_Flag**, was created to label these points for further analysis.
- The box plot of the log-transformed order demand demonstrates how logarithmic transformation reduces extreme right skewness and minimizes the impact of outliers, providing a clearer representation of the central tendency, interquartile range (IQR), and overall variability. In contrast, the box plot of the untransformed data reveals significant right skewness with numerous extreme outliers extending far beyond the upper whisker,

distorting the distribution and obscuring its central characteristics. The transformation effectively normalizes the scale, enhancing interpretability and revealing a more balanced view of the data.

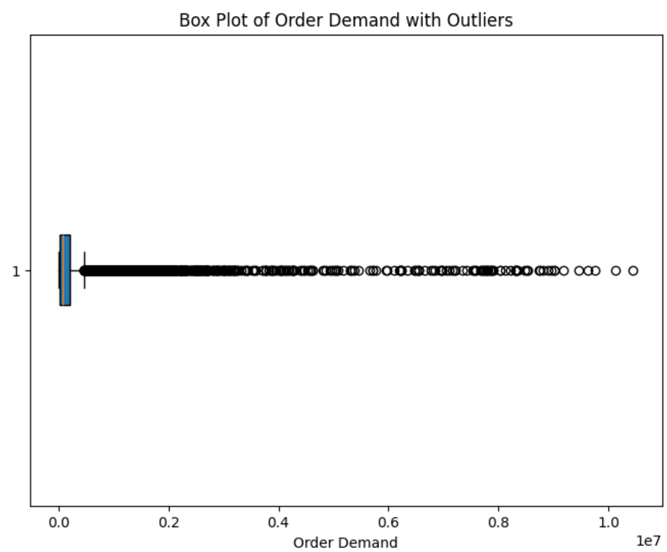

Figure 3.2.1 Box plot of order demand with outliers

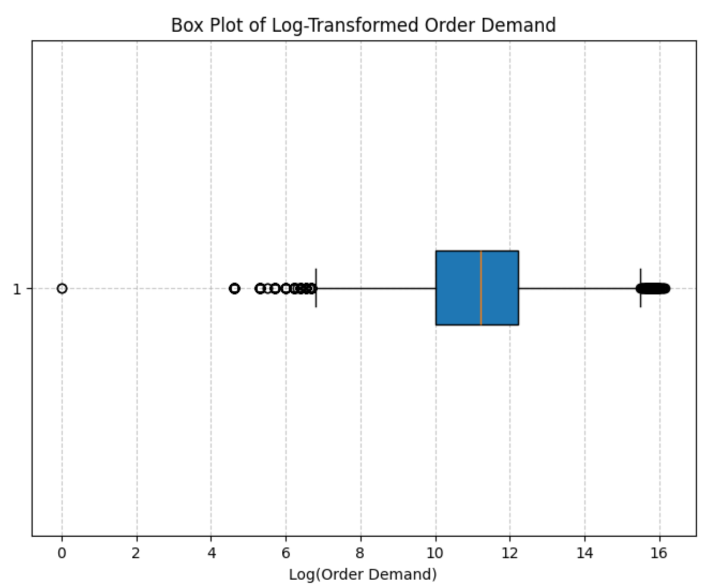

Figure 3.2.2 Box plot of log-transformed order demand

3.3 Visualization of Insights

- Top Products: Aggregated data identified leading products by Order_Demand, highlighting the most popular items.
- Word Cloud Visual representation of the top 30 products illustrated their demand prominence, enhancing interpretability.

Figure 3.3.1 Word Cloud visualization of top 30 popular products

3.4 Correlation Analysis

Explored relationships among product demands:

- Correlation Matrix: Generated for the top 10 products, revealing interesting relationships and dependencies among them.
- Heatmap: Visualized correlation strengths, facilitating easier identification of closely related products.
- The correlation analysis reveals strong relationships between products like **Chocolate Milk and Jalapeno Bagel (0.84)** and **Orange Juice and Double Chocolate Muffin (0.65)**, making them ideal for bundled promotions, especially as breakfast combos. High-demand items such as **2% Milk** also show strong correlations with **Chocolate Milk (0.79)** and **Uproot Soy Milk (0.70)**, emphasizing the need for optimized stock levels to prevent shortages. Products with weak correlations, like **Low Fat Strawberry Yogurt**, cater to niche segments and can benefit from targeted marketing. Bundling, targeted promotions during peak hours, and cross-selling opportunities—such as pairing **Whipped Butter with Orange Juice**—can help maximize sales and improve overall product visibility.

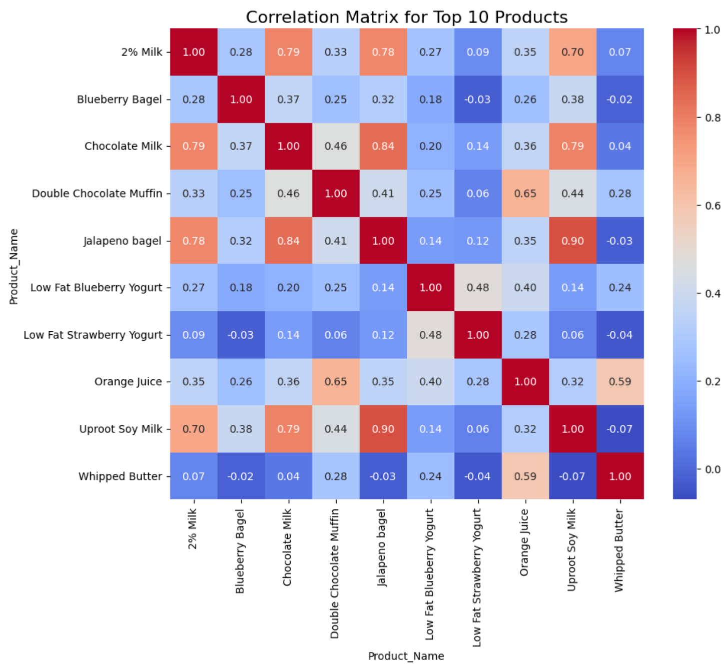

Figure 3.4.1 Correlation Matrix for Top 10 products

3.5 Temporal Analysis of Demand

Analyzed demand trends over time:

- Date Mapping: Converted month numbers to specific dates, enhancing the temporal analysis by aligning data with calendar years.
- Stacked Bar Chart: Illustrated Order_Demand trends for the top 10 products, effectively showcasing how demand fluctuated over time.
- The order demand for the top 10 products remained stable from **2019 to 2023**, indicating consistent customer preferences and a mature market. Minor growth in **2021 and 2022** suggests effective promotions or increased consumer engagement during those years. Products like **Chocolate Milk** and **2% Milk** consistently dominated demand, highlighting their importance in the product portfolio and the need for optimized inventory management. The early decline in **2024** may be due to incomplete data or shifting customer behavior, emphasizing the need for continuous monitoring and adaptability. The stable trends also provide a reliable basis for **sales forecasting** and supply chain optimization, while customer loyalty toward top-performing products can be leveraged through strategic promotions and bundling.

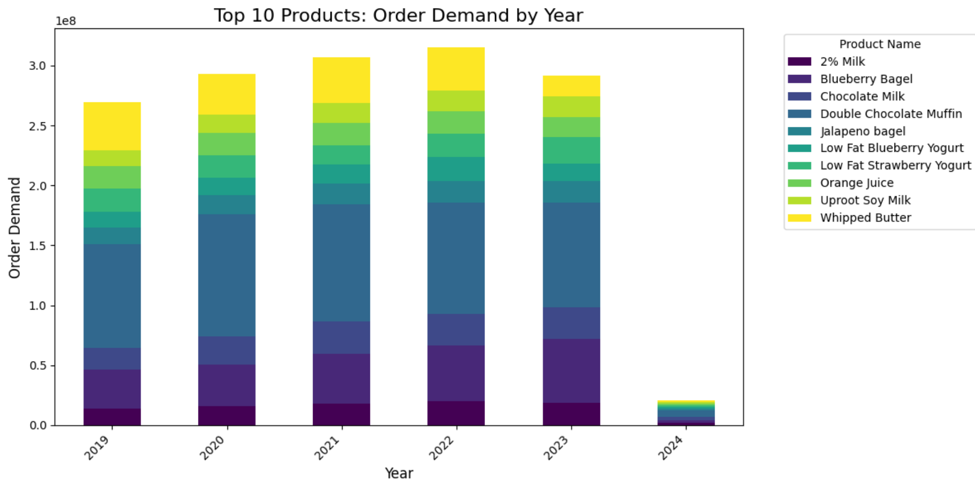

Figure 3.5.1 Order demand for Top 10 products

4 Data Preprocessing

4.1 Data Cleaning

- Preparing the Dataset: Processed and refined the raw dataset, making it suitable for subsequent preprocessing and modeling stages.
- Handling Missing Data: Identified and addressed missing or null values to ensure the dataset was complete and reliable for analysis.
- Maintaining Data Consistency: Checked and corrected inconsistencies in data types and value ranges to ensure uniformity across the dataset.

4.2 Handling Categorical Data

- Columns Processed: Product_Code and Warehouse.
- Method: Label Encoding
  - It is often used for algorithms that require numerical input while preserving category identity
  - Converts categorical variables into numerical values (e.g., Product_Code A → 0, B → 1)
  - The encoded columns replace their original categorical versions, and encoders are stored in a dictionary for future reference.

4.3 Standardizing/Normalizing Numeric Columns

- Column Processed: Order_Demand.
- Method: StandardScaler
    - Ensures that the Order_Demand column is on a comparable scale with other features, which is critical for distance-based algorithms or those sensitive to scale.
    - Transforms the Order_Demand column to have a mean of 0 and a standard deviation of 1. The formula is:

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}$$

    - A new column, Order_Demand_Scaled, is added to the dataset.

4.4 Handling Temporal Data

- Column Processed: Month.
- Method: Sine and Cosine Transformations
    - Encodes the cyclic nature of months (e.g., January is close to December). The transformations preserve periodicity:

$$\text{Sin} = \sin\left(2\pi \frac{\text{Month}}{12}\right), \quad \text{Cos} = \cos\left(2\pi \frac{\text{Month}}{12}\right)$$

    - Allows machine learning algorithms to understand that December (12) is followed by January (1), a relationship lost with simple numerical encoding.
    - Two new columns, Month_Sin and Month_Cos, are added to represent the cyclic month information.

4.5 Feature Engineering

- Derive new features based on existing ones to improve model performance or provide additional insights.
- Provides aggregated demand statistics for each product, which can serve as new predictive features.
- Groups the dataset by Product_Name and calculates the mean and median Order_Demand for each product.
- A DataFrame aggregated_stats is created with columns Product_Name, Demand_Mean, and Demand_Median.

5 Model Building

5.1 Holt-Winters

- Holt-Winters Exponential Smoothing is a time series forecasting method that extends simple exponential smoothing by adding components for trend and seasonality. It is effective for capturing patterns in data with both a consistent seasonal structure and a changing trend over time.
- The monthly scaled order demand data is aggregated and smoothed using Holt-Winters Exponential Smoothing, employing an additive trend and seasonality to accurately capture long-term trends and seasonal patterns while minimizing random fluctuations.
- The model projects demand for the subsequent 12 months, leveraging the identified seasonal and trend components to produce reliable future demand estimates.

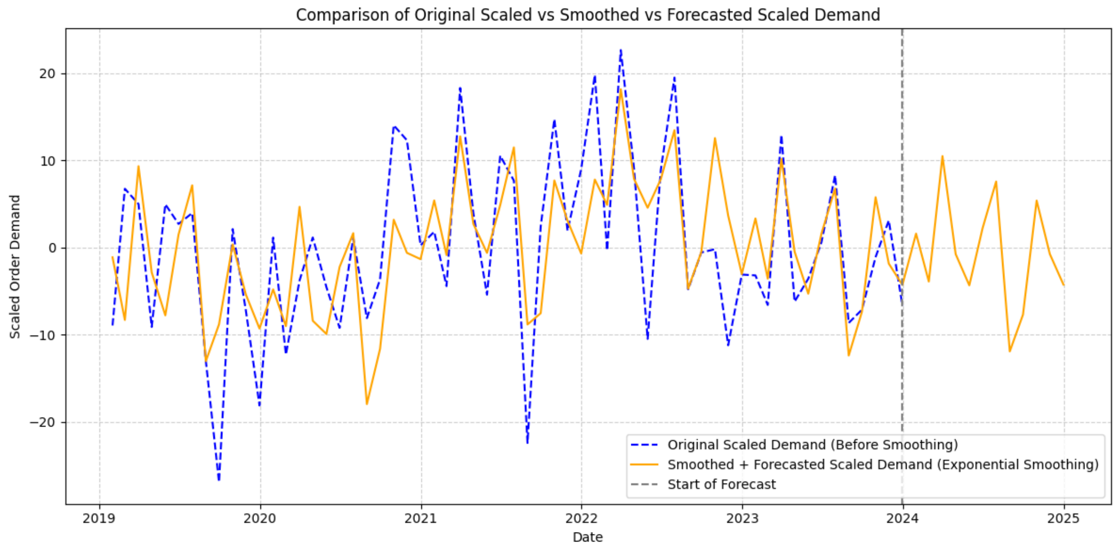

Figure 5.1 Comparison between the original order demand and forecasted demand

5.2 ARIMA (Autoregressive Integrated Moving Average)

- ARIMA is a popular time series forecasting model that combines autoregression, differencing for stationarity, and moving averages to capture patterns in the data. SARIMA (Seasonal ARIMA) extends ARIMA by incorporating seasonal components to model data with seasonal fluctuations over fixed periods.
- The process involves applying the Augmented Dickey-Fuller test to assess the stationarity of the smoothed demand data. ACF and PACF plots are used for initial parameter identification, followed by the **auto_arima** function to automatically select optimal ARIMA/SARIMA parameters, including seasonal adjustments for monthly data.

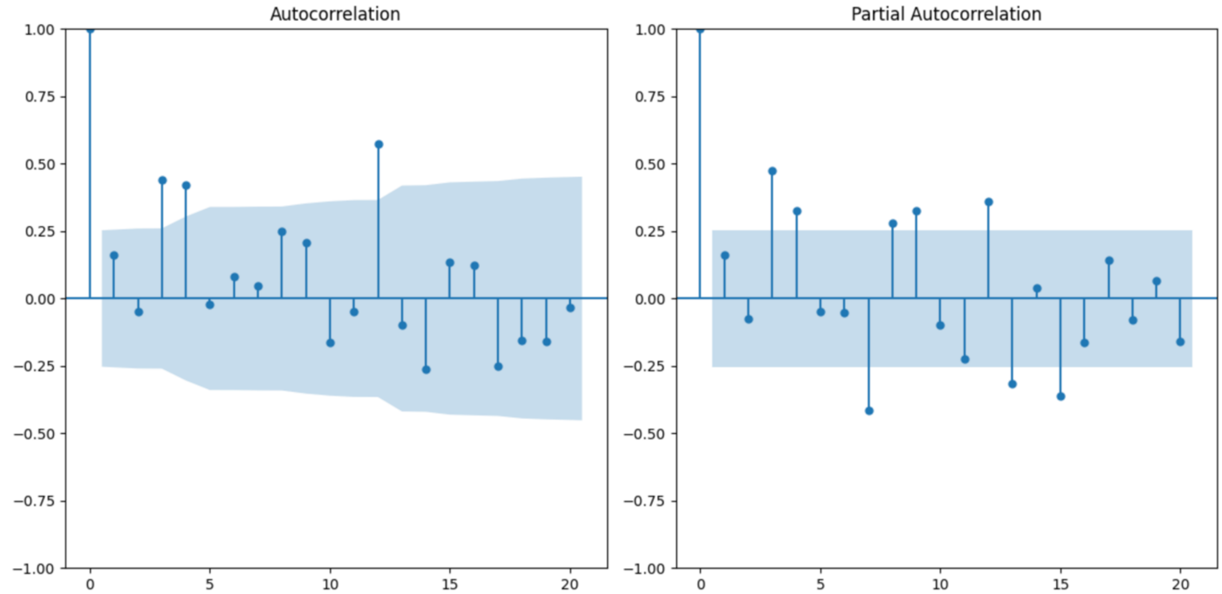

Figure 5.2.1 Autocorrelation (ACF) and Partial Autocorrelation (PACF) plots for finding initial parameters

- The optimized ARIMA/SARIMA model is fitted to the data, generating a 12-month demand forecast

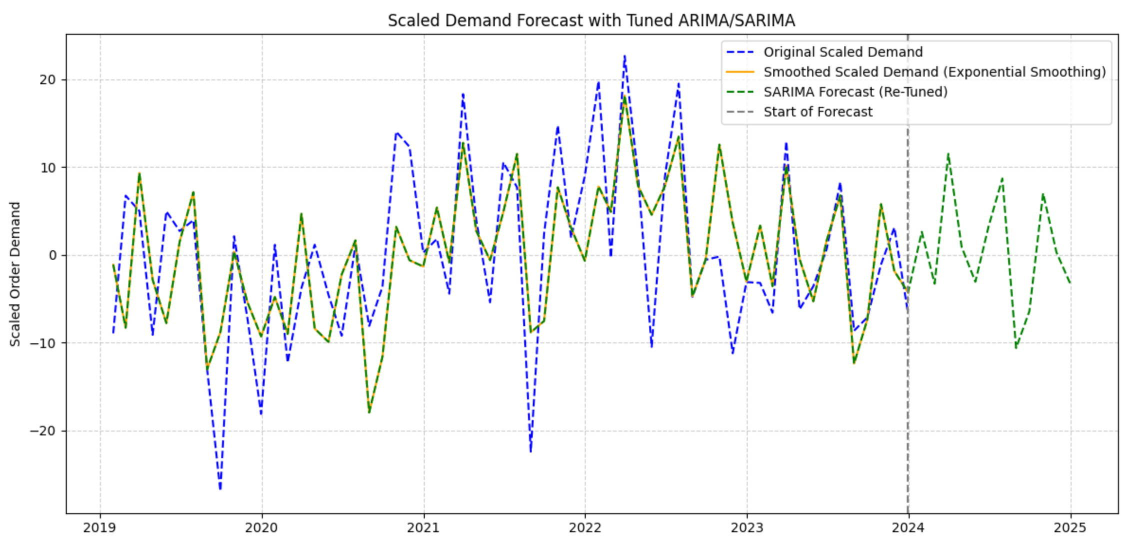

Figure 5.2.2 Comparision between original and forecasted order demand using SARIMA

5.3 Multilayer Perceptron (MLP)

- MLP (Multilayer Perceptron) is a type of neural network model used for regression and classification tasks, consisting of multiple layers of nodes (neurons) that can learn complex patterns in data.
- The data is preprocessed using MinMaxScaler to normalize the smoothed demand values, and lagged features are created to incorporate past values for predicting future demand. The MLPRegressor model is trained using a 5-fold TimeSeriesSplit for cross-validation, with a hidden layer architecture of 100 nodes and ReLU activation for non-linear modeling.

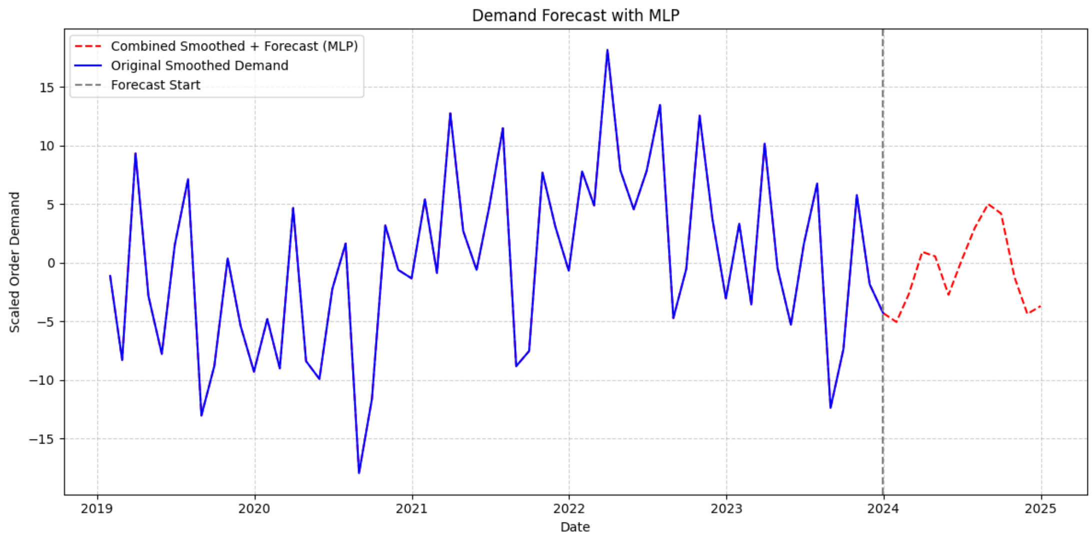

Figure 5.3 Comparison of original order demand and forecasted demand using MLP

5.4 Large Language Model (LLM)

- LLMs are advanced AI models trained on vast datasets to generate human-like text, enabling tasks such as translation, summarization, and conversational AI.
- LLaMA(Large Language Model by Meta AI) which is an efficient LLM architecture designed for research, offering competitive performance on small scale hardware.
- Ollama, which is a framework or platform that helps in running LLMs on local machines. It is used for running LLama model for our use case.
- The forecasted results and historic data is trained on LLM model and information about inventory can be retrieved using user prompts.

```
df.chat("what is the total Order_Demand for Product_1281")
```

```
{'type': 'number', 'value': 22952000}
22952000
```

```
df.chat("what is the total Order_Demand for Product_1281 in month 1")
```

```
{'type': 'number', 'value': 262000}
262000
```

Figure 5.4 User prompt vs output by LLM trained on this inventory data.

6 Results and Discussion

6.1 Comparison of Model Performance:

Among the tested models, the Multi-Layer Perceptron (MLP) Regressor achieved the lowest error metrics (RMSE: 6.52, MAE: 5.54) compared to SARIMA (RMSE: 9.91, MAE: 7.70) and Holt-Winters (RMSE: 9.91, MAE: 7.70). This indicates that MLP effectively handles the underlying non-linear patterns and complex dependencies in the dataset. Conversely, SARIMA and Holt-Winters models, while suitable for seasonal and linear trends, struggle to capture non-linearities, resulting in higher error rates.

6.2 Impact of Preprocessing and Smoothing:

The application of exponential smoothing, as observed in the third chart, reduced the noise in the original scaled demand and enhanced alignment between historical and forecasted data. This preprocessing step was particularly beneficial for machine learning models like MLP, which rely on structured input data. By minimizing variability, the smoothing process helped MLP to better learn the data patterns, thereby improving its forecasting accuracy.

6.3 Why MLP Outperforms Traditional Models:

The MLP's architecture allows it to capture intricate, non-linear relationships in the dataset, which traditional statistical models like SARIMA and Holt-Winters cannot address. The dataset likely contains irregular variations, interactions between features, and non-linear trends that are better suited for neural network-based models. Additionally, MLP's adaptability to diverse input features and its ability to generalize from training data contribute to its superior accuracy for this particular dataset.

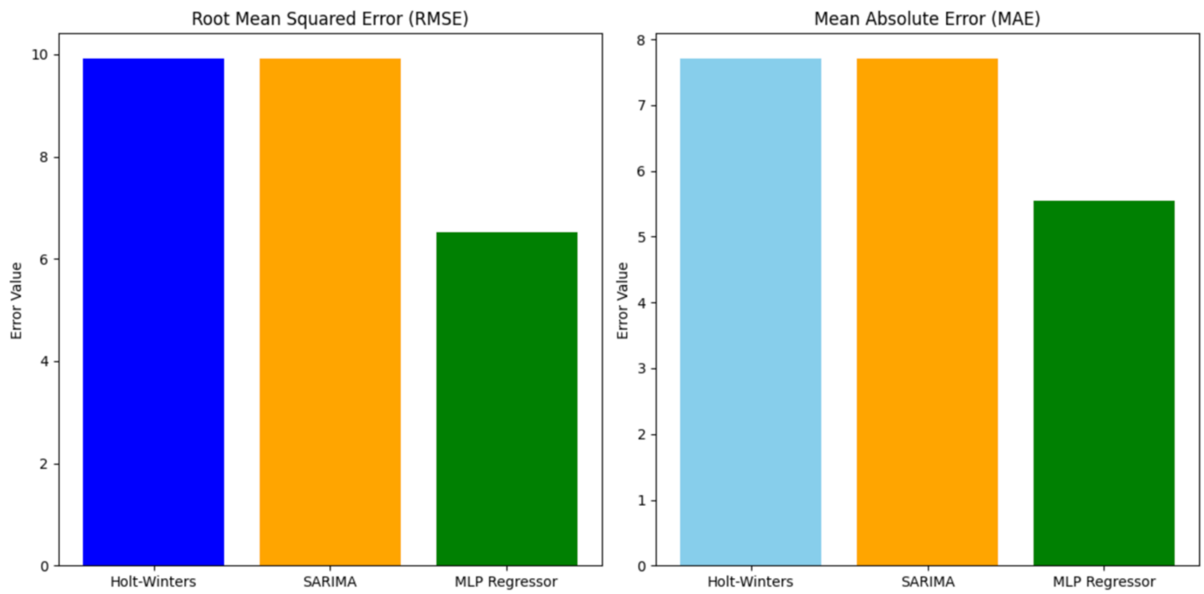

Figure 6.3.1 Comparing the RMSE and MAE of the machine learning algorithms

6.4 Results of predictions for the next 12 months using MLP

Since MLP demonstrated the lowest error rate among all the implemented machine learning algorithms, we have chosen it to forecast the order demand for the top 10 products over the next 12 months.

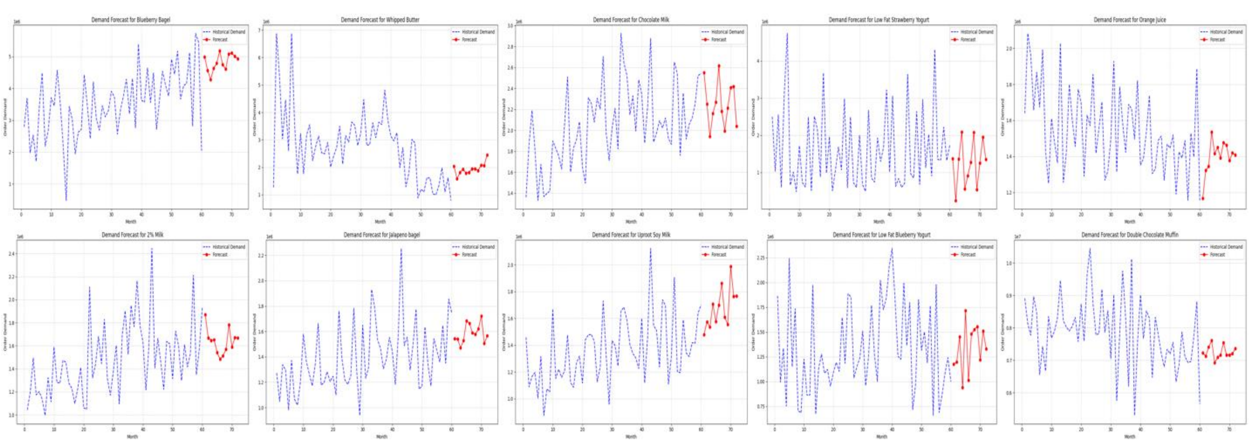

Figure 6.4.1 Predictions of top 10 products for the next 12 months

Product_Name	Forecast	RMSE	MAE
Double Chocolate Muffin	[7226188, 7131176, 7408677, 7605385, 6926521, 7087912, 7164631, 7547996, 7160989, 7164831, 7203054, 7368537]	0.03539098712441644	0.14784867708688135
Blueberry Bagel	[4992449, 4565193, 4276812, 4630482, 4786614, 5184159, 4742638, 4606208, 5083111, 5111938, 5008471, 4932584]	0.06964670290636665	0.2019835815123347
Whipped Butter	[2039927, 1578640, 1813142, 1931407, 1786758, 1815482, 1946900, 1941520, 1879317, 2079951, 2059560, 2450803]	0.011385207143102778	0.0883124548508522
Chocolate Milk	[2550388, 2250488, 1941391, 2161621, 2268893, 2617463, 2181733, 1996026, 2214918, 2410073, 2419025, 2040392]	0.05323864698423983	0.17859458537761586
Low Fat Strawberry Yogurt	[1372542, 245122, 1358825, 2090835, 556455, 908077, 1273799, 2079070, 538036, 1256516, 1945751, 1346583]	0.06385195441099686	0.16206151519133066
Orange Juice	[1167061, 1322920, 1343858, 1535879, 1415642, 1449558, 1392153, 1476951, 1461365, 1376543, 1418802, 1407144]	0.04177883629715706	0.12331837012073633
2% Milk	[1869840, 1667770, 1645179, 1651016, 1540283, 1483361, 1513515, 1569261, 1782186, 1589483, 1670626, 1668591]	0.03163166793938394	0.13887915632657788
Jalapeno bagel	[1544327, 1538780, 1471674, 1528321, 1683391, 1663813, 1588507, 1577116, 1619179, 1722384, 1507023, 1564559]	0.04090052175495335	0.17573210127802386
Uproot Soy Milk	[1479728, 1574753, 1536158, 1706213, 1576060, 1700208, 1863045, 1608559, 1555239, 1986994, 1764886, 1767274]	0.051010972017451195	0.1797303078425936
Low Fat Blueberry Yogurt	[1177312, 1196289, 1452748, 941225, 1714918, 1015659, 1479982, 1526034, 1555335, 1218392, 1511993, 1331935]	0.308882584568009	0.4376049774842068

Figure 6.4.2 Order demand calculation along with the RMSE and MAE values

1. Best Performance: Whipped Butter demonstrates superior forecasting accuracy with the lowest error rates (RMSE: 0.011, MAE: 0.088), while several products including Double Chocolate Muffin (RMSE: 0.035) and 2% Milk (RMSE: 0.032) also show strong predictive performance.

2. Areas for Improvement: Low Fat Blueberry Yogurt exhibits significantly higher error rates (RMSE: 0.309, MAE: 0.438) compared to other products, indicating a clear outlier in forecast accuracy that requires attention. This substantial deviation suggests the need for specialized forecasting approaches for this particular product.

7 Limitations and Future Work

The system's predictive accuracy is constrained by its reliance on historical data and the lack of integration with external variables such as sales trends and seasonal patterns. Future enhancements will include developing a dynamic and adaptive user interface, fine-tuning the LLM for inventory-specific datasets, and incorporating sales data and external features to refine the forecasting model's robustness and scalability.

8 Conclusion

This study effectively applied advanced time series forecasting methodologies to address inventory demand prediction challenges, focusing on optimizing stock levels and aligning demand with operational requirements. The implementation of models such as Holt-Winters, ARIMA, and MLP demonstrated robust predictive performance, with MLP achieving superior accuracy. Rigorous data preprocessing, including outlier detection, standardization, and feature engineering, ensured high-quality input for modeling. Additionally, the integration of a locally hosted LLM facilitated intelligent insights and streamlined data retrieval, enhancing usability for inventory management. However, the system's dependency on historical data underscores the importance of incorporating external variables and sales data to improve scalability and predictive robustness. Future enhancements in adaptive modeling and an interactive user interface will further elevate its applicability in complex, dynamic inventory systems.

9 References

[1] Hussein, Burhan & Kasem, Asem & Omar, Saiful & Siau, nor zainah. (2019). A Data Mining Approach for Inventory Forecasting: A Case Study of a Medical Store: Proceedings of the Computational Intelligence in Information Systems Conference (CIIS 2018). 10.1007/978-3-030-03302-6_16.

[2] Matsumoto, Mitsutaka & Ikeda, Akira. (2015). Examination of demand forecasting by time series analysis for auto parts remanufacturing. Journal of Remanufacturing. 5. 10.1186/s13243-015-0010-y.

[3] Shah, Vaishvi & Gupta, Vivek & Gandhi, Tushya & Gupta, Rajeev. (2023). Efficient Medical Supply Chain Forecasting Using Time Series Analysis. 1-6. 10.1109/CVMI59935.2023.10465213.

[4] Fattah, Jamal & Ezzine, Latifa & Aman, Zineb & Moussami, Haj & Lachhab, Abdeslam. (2018). Forecasting of demand using ARIMA model. International Journal of Engineering Business Management. 10. 184797901880867. 10.1177/1847979018808673.

[5] HAli, M., Zied, M., Boylan, J., & Syntetos, A. (2017). Supply chain forecasting when information is not shared. European Journal of Operational Research.

[6] Cheikhrouhou, N., Marmier, F., & Ayadi, O. W. (2011). A collaborative demand forecasting process with event-based fuzzy judgements.