MACHINE LEARNING BASED TIME SERIES FORECASTING FOR FOOD DEMAND IN SUPPLY CHAINS

[1]Dr. P. Nagendra Kumar, [2]P. Phani, [3]P.Jagannath, [4]S. Aadhitya Reddy, [5]Sk. Hyder Ali
[1]Professor and HOD, [2,3,4,5]UG Students, Department Of Computer Science & Engineering(AI&ML) Geethanjali Institute Of Science And Technology, Gangavaram (V), Kovur(M), SPSR Nellore (Dt), Andhra Pradesh, India-524137

Abstract
The objective of this project is to develop a time series forecasting and modeling framework for predicting food demand in a supply chain, utilizing regression analysis. Accurate demand forecasting is critical for optimizing food supply chain operations, minimizing waste, and ensuring adequate supply. This study applies advanced time series techniques, such as Random Forest Regressor, XGBoost Regressor, Gradient Boost Machine, Long Short-Term Memory (LSTM) and Bidirectional LSTM (BI-LSTM), to predict future food demand based on historical data. Key features such as seasonal trends, promotional effects, and external factors (e.g., holidays, weather conditions) are incorporated into the model through extensive dataset preparation. The project compares the performance of different forecasting models, and evaluates their accuracy through metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and R Squared Error.. The resulting model provides a data driven approach for food supply chain managers to make informed decisions on inventory management, procurement, and distribution, ultimately improving efficiency and reducing costs in the food industry.
Keywords: Machine learning, time series forecasting, food demand in supply chains, LSTM

Introduction

Due to the consumer's varying needs and increasing levels of competitiveness among companies, most companies in today's market are shifting their focus to demand forecasting for effective demand-supply chain management. Demand forecasts are beyond the scope of any planning decisions, as they directly impact a company's profitability. Inaccurate approximation of demand can either cause too much inventory, which eventually results in a high risk of wastage and high costs to pay or too little inventory, leading to out-of-stocks which ultimately pushes the company's customers to seek services from its competitors. For these very reasons, the use of demand forecasting methods is one of the most fundamental components of the strategic planning and administration of a company's logistics.

Its importance becomes evident as its outcome is used by many subdivisions in the company: the financial department uses it to estimate costs, profit levels, and the required capital; the marketing department uses it to plan its course of action and analyze the impact of diverse marketing strategies on the volume of sales; the purchasing department may devise their plans of short- and long-term investments; and finally, the operations department can manage their plan of purchasing the necessary raw materials, machinery, and labor well in advance. It is, therefore, concordant that forecasts are beneficial, and their high accuracy has the potential to prove lucrative, improve demand-supply chain management, and reduce wastage.

Efficient inventory management is the backbone of a well-functioning food supply chain. It ensures that the right quantity of products is available at the right time, minimizing waste and maximizing profitability. However, many Indian food supply chain companies struggle with inefficient inventory practices, leading to substantial financial losses, supply-demand mismatches, and food wastage. India is one of the world's largest food producers, with agriculture contributing approximately **17-18% of the GDP** and employing nearly **58% of the population**. Despite this, a significant portion of the country's food is wasted due to inefficiencies in inventory management. In India, where an alarming **40% of food is wasted**, the economic toll is staggering, amounting to

Food wastage occurrence in supply chain

Food wastage pervades every stage of the supply chain, contributing significantly to losses in India and globally. In India, a staggering one-third of produced food is lost or spoiled before reaching consumers, as reported by the Food Safety and Standards Authority (FSSAI).

Globally, the United Nations notes that approximately 13% of the world's food production disappears between harvest and retail. Inefficiencies abound throughout the supply chain: from overproduction and poor harvesting techniques, leading to significant losses, to quality control issues and mishandling during processing and packaging. Transportation and storage inefficiencies set the stage for further deterioration and pest infestation.

Even at the retail level, supermarkets contribute to wastage by discarding products nearing expiration dates or with minor imperfections. A holistic approach involving technology, responsible practices, and innovation is essential to create a more efficient, sustainable, and responsible food supply chain.

Case Study: Dalmia Bharat Sugar and Industries

Dalmia Bharat Sugar and Industries Limited, a prominent player in India's sugar industry, has recently faced significant financial challenges attributed to inventory management issues and regulatory constraints. In the first quarter of the 2024 fiscal year, the company reported a 10.8% decline in net profit, falling to ₹547.3 million from ₹613.4 million in the same period the previous year. This downturn occurred despite a 15.2% increase in revenue, which rose to ₹9.6 billion. The primary factors contributing to this profit decline were elevated inventory and finance costs, with total expenses increasing by 19.2% to ₹9.01 billion.

The financial strain continued into the second quarter of 2024, where Dalmia Bharat Sugar experienced a substantial 48.8% drop in profit before tax, decreasing to ₹378.9 million from ₹740.2 million in the corresponding quarter of the previous year. This significant decline was largely due operations grew by 26.6% to ₹9.26 billion, total expenses escalated by 26.8%, leading to a contraction in profit margins from 9% to 7%.

In the third quarter ending December 2024, the company reported a net income of ₹595.1 million, down from ₹649.2 million in the same period the previous year. Despite a notable increase in sales to ₹8.41 billion from ₹5.83 billion, the profit decline underscores the ongoing impact of high inventory costs and regulatory challenges on the company's financial health. The surge in revenue reflects strong demand and expanded market reach, but profitability remains constrained due to rising input expenses and operational inefficiencies. Management has indicated ongoing efforts to streamline supply chain processes and mitigate cost pressures in the coming quarters.

Motivation

The persistent inefficiencies in inventory management within India's food supply chain underscore the urgency of developing robust forecasting and supply chain optimization strategies. Given that nearly 40% of food is wasted annually, leading to financial losses amounting to ₹89,000 crore, it is imperative to adopt innovative solutions that enhance demand forecasting accuracy, minimize wastage, and improve overall operational efficiency. The case of Dalmia Bharat Sugar and Industries exemplifies the devastating financial impact of poor inventory management and regulatory bottlenecks. The company's significant profit decline, despite increased revenues, highlights how unchecked inventory costs and supply-demand imbalances can erode profitability. If a leading industry player is grappling with such issues, the implications for smaller enterprises and the broader food supply network are even more severe.

The food industry operates within narrow margins, and any mismatch between supply and demand can lead to either excess stock, increasing holding costs and wastage, or shortages that push consumers toward competitors. A data-driven, predictive approach to inventory management is crucial to mitigating these risks. Advanced demand forecasting methods leveraging time-series analysis, machine learning, and deep learning can provide precise insights into market fluctuations, seasonality trends, and external factors such as weather conditions and economic policies. demand forecasting. The models adhere to various archetypes but have the same fundamental idea. Traditionally, forecasting models consisted of Linear Regression, Random Forest Regression, etc., suitable for short-term demand situations. But, several boosting algorithms like Gradient Boosting Regressor (GBR), Light Gradient Boosting Machine Regressor (LightGBM), Extreme Gradient Boosting Regressor (XGBoost), and Cat Boost Regressor perform

better than the traditional algorithms when both numerical and categorical features are involved. Also, models like Long-Short Term Memory (LSTMs) and Bidirectional LSTMs have good portability and application scenarios, as they can internally maintain the memory of the input, thus making them well suited for solving problems involving sequential data, such as a time series, and for long-term demand situations.

Problem Statement

India's food supply chain is plagued by inefficiencies in inventory management, leading to excessive food wastage, financial losses, and supply-demand mismatches. Despite being one of the world's largest food producers, the country loses nearly 40% of its total food production annually, amounting to economic losses of approximately ₹89,000 crore, which equates to 1% of the nation's GDP. This wastage occurs at multiple levels, including production, processing, storage, transportation, and retail, exacerbating food insecurity and economic instability.

The issue is particularly evident in the case of Dalmia Bharat Sugar and Industries, where poor inventory management, coupled with regulatory challenges such as export bans, led to soaring stockpiles and financial strain. The company witnessed a staggering 48.8% decline in profit before tax in Q2 2024 due to a nearly 37% rise in inventory costs. Despite an increase in revenue, the high cost of maintaining unsold stock eroded profitability. This case highlights how businesses operating in the food sector are vulnerable to supply chain inefficiencies, government policies, and demand fluctuations.

Traditional inventory management systems often fail to account for dynamic market conditions such as seasonal demand variations, external economic indicators, and unexpected regulatory changes. Without accurate forecasting models, companies struggle to optimize intelligence, and time-series forecasting models. By leveraging historical demand patterns, seasonality trends, and external factors such as weather conditions and economic policies, businesses can optimize inventory levels, reduce wastage, and enhance profitability. This study aims to explore and implement data-driven demand forecasting techniques to enable companies to make informed decisions, ensuring a more resilient and sustainable food supply chain in India.

Fig-1: Supply Chain Workflow

Time Series Analysis

Time series analysis is a statistical technique used to analyze time-ordered data points to identify trends, patterns, and dependencies over time. It is widely applied in various fields, including finance, healthcare, economics, weather forecasting, and machine learning. Time series data consists of observations recorded at consistent time intervals, such as daily stock prices, monthly sales figures, or hourly temperature readings. Unlike cross-sectional data, where observations are independent, time series data exhibits temporal dependencies, making its analysis unique and complex.

A fundamental aspect of time series analysis is identifying key components such as trend, seasonality, and cyclic patterns. The trend represents the overall long-term movement in the data, either upward or downward. Seasonality refers to periodic fluctuations occurring at regular intervals, such as increased retail sales during holidays. Cyclical patterns are variations that happen Several techniques are employed in time series analysis, ranging from traditional statistical methods to advanced machine learning models. Classical methods are widely used for forecasting by

capturing linear dependencies and smoothing fluctuations. More advanced approaches, such as Long Short-Term Memory (LSTM) networks, leverage deep learning to model complex, nonlinear relationships in time series data. Additionally, feature engineering, such as differencing to remove trends or using lag variables, is essential for improving model accuracy. As the availability of high-frequency data grows, advancements in time series modeling continue to enhance forecasting accuracy, making it a powerful tool in various domains.

Time Series

A time series is a sequence of data points collected, recorded, or measured at successive, evenly-spaced time intervals.

Each data point represents observations or measurements taken over time, such as stock prices, temperature readings, or sales figures. Time series data is commonly represented graphically with time on the horizontal axis and the variable of interest on the vertical axis, allowing analysts to identify trends, patterns, and changes over time. This type of data is crucial in various fields, including finance, meteorology, and economics, where understanding past patterns can help predict future outcomes. Analysts often apply statistical techniques such as autoregressive models to extract meaningful insights from time series data. Additionally, advancements in machine learning and deep learning have enabled more sophisticated forecasting models, improving accuracy in predicting future trends.

Time series data poses unique challenges like seasonality, noise, and missing values. Analysts often apply preprocessing techniques such as smoothing, differencing, and normalization to better capture underlying patterns and trends over time. Proper handling of these challenges is essential to ensure accurate analysis and reliable forecasting. Additionally, maintaining the temporal order of data is crucial, as even small shifts can lead to misleading insights or incorrect predictions.

Components of Time Series

Trend: Trend represents the long-term movement or directionality of the data over time. It captures the overall tendency of the series to increase, decrease, or remain stable. Trends can be linear, indicating a consistent increase or decrease, or nonlinear, showing more complex patterns.

Seasonality: Seasonality refers to periodic fluctuations or patterns that occur at regular intervals within the time series. These cycles often repeat annually, quarterly, monthly, or weekly and are typically influenced by factors such as seasons, holidays, or business cycles.

Cyclic variations: Cyclical variations are longer-term fluctuations in the time series that do not have a fixed period like seasonality. These fluctuations represent economic or business cycles, which can extend over multiple years and are often associated with expansions and contractions in economic activity.

Irregularity (or Noise): Irregularity, also known as noise or randomness, refers to the unpredictable or random fluctuations in the data that cannot be attributed to the trend, seasonality, or cyclical variations. These fluctuations may result from random events, measurement errors, or other unforeseen factors. Irregularity makes it challenging to identify and model the underlying patterns in the time series data.

Time series data is influenced by components like trend, seasonality, cyclic variations, and irregularity. Understanding these elements helps in identifying patterns and improving the accuracy of analysis and forecasting.

Line Plots: Line plots are a common way to visualize time series data, connecting data points over time to reveal trends, cycles, and fluctuations. They help analysts quickly identify patterns and changes in the data.

Seasonal Plots: Seasonal plots break down time series data into seasonal components, helping to visualize patterns within specific time periods. By grouping data by season, month, or week, these plots make it easier to detect recurring trends and compare seasonal behavior across different years. This visualization is especially useful for identifying consistent peaks or dips tied to specific times of the year, aiding in more accurate forecasting and planning.

Example:

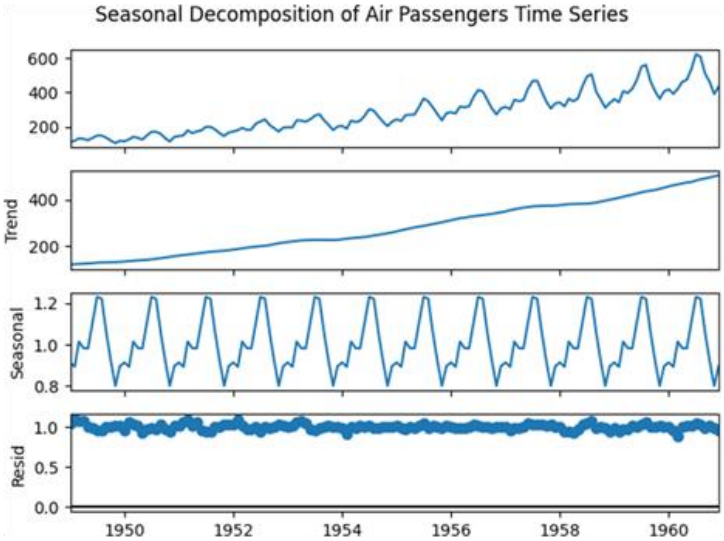


Fig.2. Seasonal Plots of Air Passengers

Example:

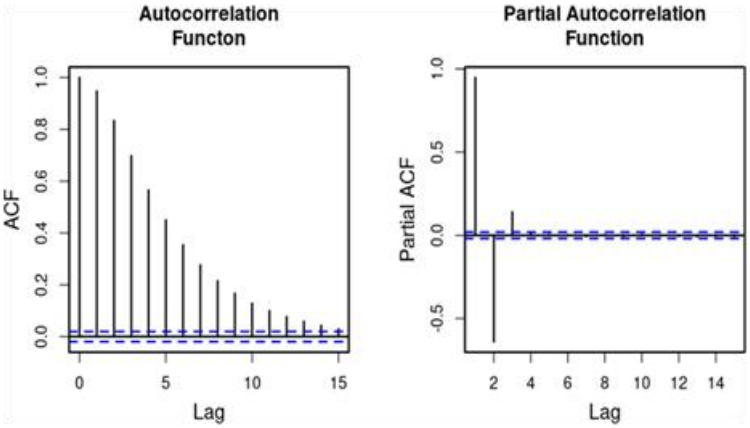


Fig.3. Auto-correlation and Partial Autocorrelation Plot

Spectral Analysis: Spectral analysis techniques, such as periodograms and spectrograms, visualize frequency components within time series data, useful for identifying periodicity and cyclical patterns

Time Series Forecasting

Time Series Forecasting is a statistical technique used to predict future values of a time series based on past observations. In simpler terms, it's like looking into the future of data points plotted over time. By analyzing patterns and trends in historical data, Time Series Forecasting helps make informed predictions about what may happen next, assisting in decision-making and planning for the future

Different Time Series Forecasting Algorithms

Autoregressive (AR) Model: Autoregressive (AR) model is a type of time series model that predicts future values based on linear combinations of past values of the same time series. In an AR(p) model, the current value of the time series is modeled as a linear function of its previous p values, plus a random error term. The order of the autoregressive model (p) determines how many past values are used in the prediction.

Vector Autoregression (VAR) Models: VAR models extend autoregression to multivariate time series data by modeling each variable as a linear combination of its past values and the past values of other variables. They are suitable for analyzing and forecasting interdependencies among multiple time series.

Theta Method: A simple and intuitive forecasting technique based on extrapolation and trend fitting.

Gaussian Processes Regression: Gaussian Processes Regression is a Bayesian non-parametric approach that models the distribution of functions over time. It provides uncertainty estimates along with point forecasts, making it useful for capturing uncertainty in time series forecasting.

Generalized Additive Models (GAM): A flexible modeling approach that combines additive components, allowing for nonlinear relationships and interactions.

Random Forests: Random Forests is a machine learning ensemble method that constructs multiple decision trees during training and outputs the average prediction of the individual trees. It can handle complex relationships and interactions in the data, making it effective for time series forecasting.

Gradient Boosting Machines (GBM): GBM is another ensemble learning technique that builds multiple decision trees sequentially, where each tree corrects the errors of the previous one. It excels in capturing nonlinear relationships and is robust against overfitting.

State Space Models: State space models represent a time series as a combination of unobserved (hidden) states and observed measurements. These models capture both the deterministic and stochastic components of the time series, making them suitable for forecasting and anomaly detection.

Dynamic Linear Models (DLMs): DLMs are Bayesian state-space models that represent time series data as a combination of latent state variables and observations. They are flexible models capable of incorporating various trends, seasonality, and other dynamic patterns in the data.

Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) Networks: RNNs and LSTMs are deep learning architectures designed to handle sequential data. They can capture complex temporal dependencies in time series data, making them powerful tools for forecasting tasks, especially when dealing with large-scale and high-dimensional data.

Performance Metrics:

Performance metrics are quantitative measures used to evaluate the accuracy and effectiveness of time series forecasts. These metrics provide insights into how well a forecasting model performs in predicting future values based on historical data. Common performance metrics which can be used for time series include:

Mean Absolute Error(MAE): Measures the average magnitude of errors between predicted and actual values.

Mean Squared Error(MSE): Computes the average squared differences between predicted and actual values.

Root Mean Squared Error(RMSE): The square root of MSE, providing a measure of the typical magnitude of errors.

R^2 Coefficient: R^2, or the coefficient of determination, measures how well a model explains the variance in the target variable, with values closer to 1 indicating a better fit.

Literature Survey

[1] S. Yadav, T.-M.Choi, S.Luthra, A.Kumar, andD.Garg, ‘‘Using Internet of Things (IoT) in agri-food supply chains: A research framework for social good with network clustering analysis’’

The integration of the Internet of Things (IoT) in Agri-Food Supply Chains (AFSCs) has gained significant attention due to its potential to enhance efficiency, traceability, and sustainability. Numerous studies have explored IoT applications in AFSCs, focusing on aspects such as food safety, performance measurement, supply chain resilience, transparency, and coordination. A systematic review of 346 research articles from the Web of Science (WoS) database was conducted, employing network analysis using VOS viewer software to categorize key themes. The analysis identified seven major clusters, addressing critical areas such as agri-food safety, sustainability, performance evaluation, resilience against disruptions, integration, transparency, and the barriers to IoT adoption. Based on these insights, a research framework has been proposed to guide future studies and help policymakers, researchers, and industry professionals optimize AFSC operations for improved social welfare.

[2]. Zheng, L. Wang, L. Wang, S. Wang, J.-F. Chen, and X. Wang, ‘‘Solving stochastic online food delivery problem via iterated greedy algorithm with decomposition-based strategy,’’

Online Food Delivery (OFD) services have grown rapidly, offering convenience to customers and market expansion for restaurants. However, challenges like high demand, uncertainty in food preparation time, and strict delivery constraints persist. Researchers have proposed an iterated greedy algorithm with a decomposition-based strategy to optimize delivery efficiency. This includes rider filtration, risk-aware optimization, and machine learning-based evaluation techniques. Scenario-sampling and adaptive selection strategies enhance computational efficiency. An upper bound for stochastic time cost improves prediction accuracy. Experiments on real-world datasets validate

the effectiveness of these approaches in optimization OFD logistics.

[3]. V. K. Shrivastava, A. Shrivastava, N. Sharma, S. N. Mohanty, and C. R. Pattanaik, ''Deep learning model for temperature prediction: A case study in New Delhi ''

Temperature prediction is a critical area of research, with various machine learning models being employed for improved accuracy. Multivariate Polynomial Regression (MPR) has been used for capturing non-linear temperature trends, but its performance declines as the number of input features increases. Deep Neural Networks (DNNs) have shown better accuracy in handling complex meteorological data. In this study, three DNN models (DNNM-1, DNNM-2, and DNNM-3) with varying input parameters are implemented. The results indicate that while MPR models perform well with fewer features, DNNM-3 outperforms all models by leveraging a high-dimensional dataset. This aligns with existing literature that highlights deep learning's superiority in climate forecasting.

[4]. Ismail Shah, Faheem Jan, and Sajid Ali "Functional Data Approach for Short-Term Electricity Demand Forecasting"

Electricity demand forecasting is crucial for managing power systems, especially in liberalized markets, but it remains challenging due to factors like extreme values, spikes, periodic trends, and holiday effects. Since demand is determined a day in advance, accurate one-day-ahead forecasts are essential. Recent studies have explored functional data analysis (FDA) as a promising yet underutilized approach in energy forecasting. FDA-based models divide demand data into deterministic and stochastic components, with the former modeled using generalized additive models and the latter using functional autoregressive (FAR), FAR with exogenous variables (FARX), and classical autoregressive (AR) models. Using Nord Pool electricity market data, researchers found that the FARX model achieved the best accuracy, with a Mean Absolute Percentage Error (MAPE) of 2.74, compared to 6.27 for FAR and 9.73 for AR models. These results highlight the effectiveness of functional modeling in electricity forecasting and suggest that integrating exogenous variables, such as weather conditions and market factors, could further improve predictive performance.

[5]. Jingyi Ding, Ziqing Chen, Li Xiaolong and Baoxin Lai "Sales Forecasting Based on CatBoost"

Sales forecasting plays a crucial role in the retail industry, enabling businesses to make informed decisions using advanced machine learning and deep learning techniques. Recent studies have explored various models for improving sales prediction accuracy, with boosting algorithms gaining popularity due to their efficiency. One such approach is CatBoost, a gradient boosting algorithm designed for handling categorical data effectively. In a study using the Walmart sales dataset, one of the largest in this field, researchers implemented a CatBoost-based sales forecasting system with extensive feature engineering to enhance prediction accuracy and speed. Experimental results demonstrated that CatBoost outperformed traditional machine learning models like Linear Regression and SVM, achieving a Root Mean Squared Error (RMSE) of 0.605. Additionally, the model requires less fine-tuning, improving its generalization ability across different datasets. These findings suggest that boosting-based methods, particularly CatBoost, offer a promising solution for scalable and efficient sales forecasting in retail.

[6] Y. Zhang, L. Wang, X. Chen, Y. Liu, S. Wang, and L. Wang, ''Prediction of winter wheat yield at county level in China using ensemble learning,''

Wheat is one of the three staple foods in China, and plays an important role in maintaining food supply and security. Timely and accurate wheat yield information on the national scale is of great significance for agricultural decision making and sustainable developments. The grain yield is affected by multiple factors, such as soil quality, weather conditions, field management practices, agricultural subsidy policy, and the market prices of grain. For example, high grain prices usually encourage farmers to invest more resources to achieve higher yields . In addition, many of these factors have a non-linear relationship with yield and some factors such as weather, soil, and applied nutrients are interrelated in a complex agroecosystem. Thus, wheat yield forecasting over large spatial regions still remains a challenge.

prices: A comparison of alternative approaches,'' J. Math., vol. 2022, Jul. 2022, Art. no. 3581037.

Electricity demand and price forecasting are key components for the market participants and system operators as precise forecasts are necessary to manage power systems effectively. However, forecasting electricity demand and prices are challenging due to their specific features, such as high frequency, volatility, long trend, nonconstant mean

and variance, mean reversion, multiple seasonalities, calendar effects, and spikes/jumps. Thus, the main aim of this study is to propose models that can efficiently forecast electricity demand and prices. To this end, the time series (demand/price) is divided into two components. The first component is considered a deterministic component that includes a trend, yearly, seasonal, and weekly periodicities, calendar effects, and lagged exogenous information and is modeled by parametric and nonparametric approaches. The second component is known as a stochastic (residual) component that is estimated using univariate autoregressive (AR) and multivariate vector autoregressive (VAR) models. Besides descriptive statistics, a statistical significance test is also used to evaluate the models' forecasting accuracy.

[8] I. Shah, S. Akbar, T. Saba, S. Ali, and A. Rehman, ''Short-term forecasting for the electricity spot prices with extreme values treatment,'' IEEE Access, vol. 9, pp. 105451–105462, 2021.

Nowadays, modeling and forecasting electricity spot prices are challenging due to their specific features, including multiple seasonalities, calendar effects, and extreme values (also known as jumps, spikes, or outliers). This study aims to provide a comprehensive analysis of electricity price forecasting by comparing several outlier filtering techniques followed by various modeling frameworks. To this end, extreme values are first treated with five different filtering techniques and are then replaced by four different outlier replacement approaches. Next, the spikes-free series is divided into deterministic and stochastic components. On the other hand, the stochastic component accounts for the short-run dynamics of the price time series and is modeled using different univariate and multivariate models.

series forecasting: Current status and future directions,'' Int. J. Forecasting, vol. 37, no. 1, pp. 388–427, 2021.

Recurrent Neural Networks (RNN) have become competitive forecasting methods, as most notably shown in the winning method of the recent M4 competition. However, established statistical models such as ETS and ARIMA gain their popularity not only from their high accuracy, but they are also suitable for non-expert users as they are robust, efficient, and automatic. In these areas, RNNs have still a long way to go. We present an extensive empirical study and an open-source software framework of existing RNN architectures for forecasting, that allow us to develop guidelines and best practices for their use. For example, we conclude that RNNs are capable of modelling seasonality directly if the series in the dataset possess homogeneous seasonal patterns, otherwise we recommend a deseasonalization step. Comparisons against ETS and ARIMA demonstrate that the implemented (semi-)automatic RNN models are no silver bullets, but they are competitive alternatives in many situations.

[10] N. Bibi, I. Shah, A. Alsubie, S. Ali, and S. A. Lone, ''Electricity spot prices forecasting based on ensemble learning,'' IEEE Access, vol. 9, pp. 150984–150992, 2021.

Efficient modeling and forecasting of electricity prices are essential in today's competitive electricity markets. However, price forecasting is not easy due to the specific features of the electricity price series. This study examines the performance of an ensemble-based technique for forecasting short-term electricity spot prices in the Italian electricity market (IPEX). To this end, the price time series is divided into deterministic and stochastic components. The deterministic component that includes long-term trends, annual and weekly seasonality, and bank holidays, is estimated using semi-parametric techniques. On the other hand, the stochastic component considers the short-term dynamics of the price series and is estimated by time series and various machine learning algorithms. Based on three standard accuracy measures, the results indicate that the ensemble-based model outperforms the others, while the random forest and ARMA are highly competitive.

XGBoost

XGBoost (Extreme Gradient Boosting) is an advanced machine learning algorithm designed for speed, efficiency, and high predictive accuracy. It is widely used in regression, classification, and ranking tasks, including time-series forecasting such as food demand prediction. XGBoost is based on gradient boosting, which builds an ensemble of weak learners (decision trees) in a sequential manner, optimizing each step to minimize prediction errors.

How XGBoost Works

Step 1: Base Model Initialization

Step 2: Gradient Boosting – Learning from Errors

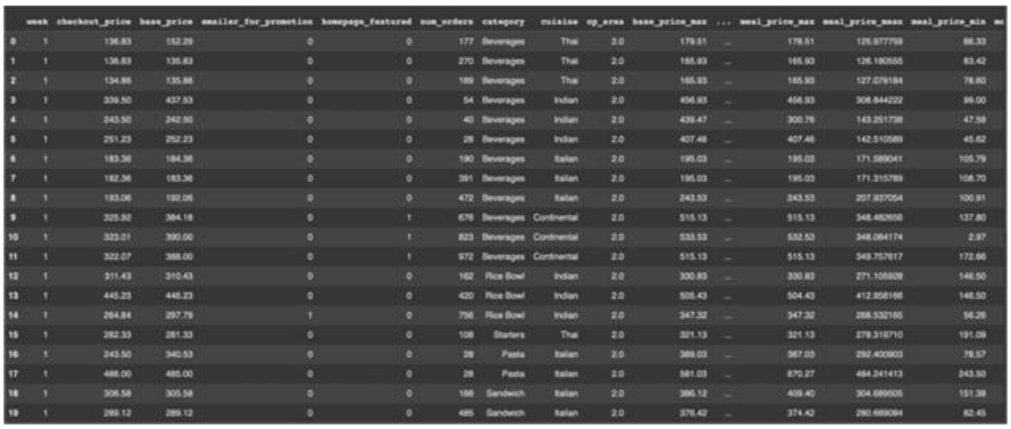
Step 3: Weighted Tree Addition

Step 4: Regularization for Overfitting Prevention **Step 5:**
Optimal Tree Splitting Using Gain Calculation **Step 6:** Pruning
Trees for Efficiency

Gradient Boosting Machine (GBM) is a powerful and widely used machine learning algorithm that falls under the category of ensemble learning methods. It is particularly effective for both regression and classification tasks. GBM works by building a sequence of decision trees, where each new tree attempts to correct the errors made by the previous ones. The idea is to combine multiple weak learners, typically shallow decision trees, in such a way that they together form a strong predictive model. The process starts with an initial prediction, and then the algorithm calculates the residuals — the differences between the actual and predicted values.

Dataset

In this section, we describe our proposed methods by first introducing the existing and calculated features and then the preprocessing techniques. **The ‘‘Food Demand Forecasting’’ dataset** released by Genpact, an American professional services firm, comprises 145 weeks’ worth of weekly orders for 50 distinct meals.



	week	checkout_price	base_price	emailer_for_promotion	homepage_featured	num_orders	category	cuisine	op_area	base_price_max	...	meal_price_max	meal_price_mean	meal_price_min	m
0	1	136.83	152.29	0	0	177	Beverages	Thai	2.0	179.51	...	179.51	125.977759	86.33	
1	1	136.83	135.83	0	0	270	Beverages	Thai	2.0	165.93	...	165.93	126.180555	83.42	
2	1	134.86	135.86	0	0	189	Beverages	Thai	2.0	165.93	...	165.93	127.079184	78.80	
3	1	339.50	437.53	0	0	54	Beverages	Indian	2.0	456.93	...	456.93	308.844222	99.00	
4	1	243.50	242.50	0	0	40	Beverages	Indian	2.0	439.47	...	300.76	143.251738	47.59	
5	1	251.23	252.23	0	0	28	Beverages	Indian	2.0	407.46	...	407.46	142.510589	45.62	
6	1	183.36	184.36	0	0	190	Beverages	Italian	2.0	195.03	...	195.03	171.589041	105.79	
7	1	182.36	183.36	0	0	391	Beverages	Italian	2.0	195.03	...	195.03	171.315789	108.70	
8	1	193.06	192.06	0	0	472	Beverages	Italian	2.0	243.53	...	243.53	207.937054	100.91	
9	1	325.92	384.18	0	1	676	Beverages	Continental	2.0	515.13	...	515.13	348.482668	137.80	
10	1	323.01	390.90	0	1	823	Beverages	Continental	2.0	533.53	...	532.53	348.064174	2.97	
11	1	322.07	388.00	0	1	972	Beverages	Continental	2.0	515.13	...	515.13	349.757617	172.66	
12	1	311.43	310.43	0	0	162	Rice Bowl	Indian	2.0	330.83	...	330.83	271.105809	146.50	
13	1	445.23	446.23	0	0	420	Rice Bowl	Indian	2.0	505.43	...	504.43	412.958166	146.50	
14	1	264.84	257.79	1	0	756	Rice Bowl	Indian	2.0	347.32	...	347.32	268.532166	56.26	
15	1	282.33	281.33	0	0	108	Starters	Thai	2.0	321.13	...	321.13	278.318710	191.09	
16	1	243.50	340.53	0	0	28	Pasta	Italian	2.0	389.03	...	367.03	292.400903	78.57	
17	1	486.00	485.00	0	0	28	Pasta	Italian	2.0	581.03	...	670.27	484.241413	243.50	
18	1	306.58	305.58	0	0	186	Sandwich	Italian	2.0	386.12	...	409.40	304.689505	151.38	
19	1	289.12	289.12	0	0	485	Sandwich	Italian	2.0	376.42	...	374.42	280.689084	82.45	

Fig.4. Dataset

Random Forest Regressor

Random Forest Regressor is a powerful machine learning algorithm used for regression tasks, including time-series forecasting such as food demand prediction. It is an ensemble learning method that builds multiple decision trees and combines their outputs to make more accurate predictions. By aggregating the results from multiple trees, Random Forest reduces overfitting and improves generalization.

Steps of Random Forest Regressor

Ensemble of Decision Trees: Random Forest is based on the concept of decision trees, where each tree is trained on a subset of the data. The final prediction is obtained by averaging the predictions of all trees. Unlike a single decision tree, which can be prone to overfitting, Random

Bootstrapping & Bagging: Random Forest uses a technique called bootstrap aggregation (bagging) to create diverse trees. Instead of training all trees on the same dataset, it randomly selects subsets of data (with replacement) to train each tree. This helps in reducing bias and improving model robustness.

Feature Randomness: While training each tree, Random Forest does not use all available features. Instead, it randomly selects a subset of features for each tree. This ensures that trees do not rely too heavily on a particular feature and that the overall model remains stable and accurate.

Prediction Aggregation: For regression problems, the final prediction is obtained by taking the average of predictions made by individual decision trees. This ensemble approach makes Random Forest more resistant to noise and less prone to overfitting than a single decision tree.

Key Advantages of Random Forest Regressor(Proposed System)

Handles Non-Linear Relationships: Random Forest does not assume any predefined relationship between features and the target variable. It can capture complex, non-linear patterns in data, making it suitable for forecasting demand fluctuations.

Robust to Outliers and Missing Values: Since Random Forest aggregates multiple trees, it is less sensitive to outliers. Missing values can be handled efficiently by averaging predictions from different trees.

Reduces Overfitting: Unlike individual decision trees, which tend to overfit on training data, Random Forest maintains high accuracy on test data by averaging multiple predictions. This prevents the model from memorizing noise in the data.

Feature Importance Ranking: Random Forest provides an importance score for each feature, helping identify the most influential factors in demand forecasting. Features such as past demand, pricing, and promotional discounts can be analyzed to understand their impact on future orders**.**

Works Well on Large Datasets: Since Random Forest distributes computation across multiple trees, it performs efficiently on large datasets without a significant loss in speed.

In food demand forecasting, Random Forest is useful for predicting the number of meal orders based on historical data. It considers multiple factors such as past demand trends, pricing, promotions, and regional variations. Since demand fluctuations are influenced by various external factors, Random Forest helps by capturing complex relationships between these factors. However, since time-series forecasting relies heavily on sequential data, Random Forest may not perform as well as deep learning models like LSTM and Bi-LSTM, which are designed to learn long-term dependencies. In cases where time-dependent trends are crucial, Random Forest is often used as a baseline model, while LSTM and Bi-LSTM may not provide more accurate results.

availability of robust machine learning tools and frameworks such as Python, R, Facebook Prophet, and Scikit-learn, which are well-suited for time series forecasting tasks. Most organizations already possess large volumes of relevant data through POS systems or supply chain management software. The computational demands of training these models can be efficiently met using scalable cloud platforms like Google Cloud, AWS, or Azure. Furthermore, the system can be seamlessly integrated into existing supply chain infrastructures, providing real-time insights for operational teams. Given the maturity of current technologies and data availability, the technical implementation of the model is entirely feasible.

Feasibility Study

Operational Feasibility:

Operationally, the system is designed to enhance existing workflows rather than replace them, ensuring smooth adoption within the organizationThe modern food supply chain faces significant challenges in maintaining a balance between supply and demand, particularly due to fluctuating consumer behavior, seasonal trends, and perishability of products. Traditional forecasting methods often fall short in adapting to these dynamic conditions. This project proposes the implementation of a Machine Learning (ML)-based Time Series Forecasting model to predict food demand accurately. By leveraging historical sales data, weather patterns, holidays, and external factors, the model can help stakeholders make informed decisions on inventory, procurement, and distribution.

Although the implementation of a machine learning system involves initial investments in data processing, skilled personnel, and computational resources, the long-term benefits significantly outweigh the costs. Accurate demand forecasting can drastically reduce overstocking and understocking situations, minimizing losses due to spoilage and missed sales. Additionally, optimized inventory planning translates into reduced holding costs and more efficient procurement strategies, such as bulk buying at discounted rates. With the availability of cloud-based infrastructure and open-source ML libraries, the cost of implementation can be further minimized. Overall, the economic feasibility of the project is strong, promising considerable cost savings and improved profitability over time.

Technical Feasibility

From a technical perspective, the proposed project is highly viable. There is widespread . Stakeholders such as inventory managers and supply planners can be trained to use the system with minimal effort, especially when presented with user-friendly dashboards and interpretable forecasts. The model can be retrained periodically to adapt to new trends and data patterns, ensuring continued reliability and accuracy. Moreover, the solution is scalable across various food categories and geographical regions, making it suitable for both small-scale operations and large supply networks. Therefore, from an operational standpoint, the project is feasible with high potential for impact and minimal resistance to adoption.

SYSTEM DESIGN

WORKFLOW OF THE SYSTEM

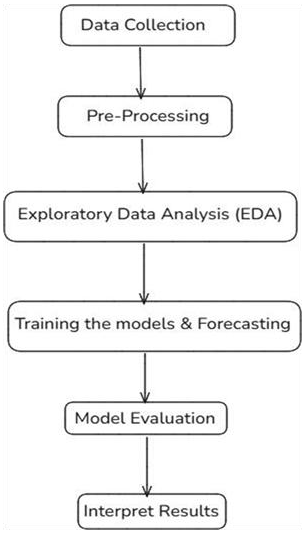


Fig.5. Workflow Of the System

The system design workflow for this project, **"Machine Learning-Based Time Series Forecasting for Food Demand in Supply Chains,"** outlines a structured approach for building an effective forecasting framework. The objective is to accurately predict future food demand using historical data and advanced regression models, ultimately supporting better decision-making in inventory management, procurement, and distribution.

The process begins with **Data Collection**, where historical food demand data is gathered along with relevant external variables such as holidays, promotional events, and weather conditions. These features are essential in capturing real-world factors that influence demand patterns.

Following this, the **Pre-Processing** stage involves preparing the data for analysis. This includes cleaning the dataset by handling missing values, removing inconsistencies, creating time-based features like lag values, and encoding categorical variables. Proper preprocessing ensures the dataset is structured, informative, and ready for modeling.

Next is **Exploratory Data Analysis (EDA)**, which provides valuable insights into trends, seasonality, and correlations within the data. Through visualizations and statistical summaries, this step helps in understanding how different factors impact food demand and guides the selection of model features.

In the **Training the Models & Forecasting** phase, a range of machine learning and deep learning algorithms are applied, including **Random Forest Regressor**, **XGBoost Regressor**, **Gradient Boost Machine**, **Long Short-Term Memory (LSTM)**, and **Bidirectional LSTM (Bi-LSTM)**. These models are trained using the processed dataset to learn complex patterns and generate accurate forecasts of future demand.

Once trained, the models are assessed during the **Model Evaluation** stage using performance metrics such as **Root Mean Square Error (RMSE)**, **Mean Absolute Error (MAE)**, and **R Squared Error**. These metrics help compare the models' accuracy and reliability, allowing the selection of the most effective forecasting approach.

Finally, the **Interpret Results** step involves analyzing the forecasting outputs and translating them into actionable insights. These predictions enable food supply chain managers to make data-driven decisions, reduce operational costs, minimize waste, and ensure a balanced and efficient supply

ARCHITECTURAL DESIGN

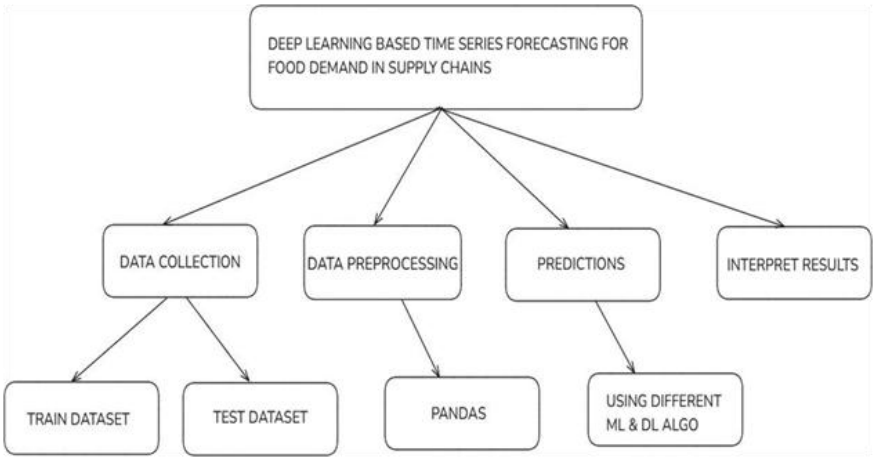


Fig.6.Architectural Design

Step 1 : Data Collection

This module is responsible for gathering the raw data needed to train and evaluate forecasting models.

Train Dataset: This part of the data is used to teach the ML/DL models. It includes historical food demand, dates, holidays, promotions, and weather-related variables.

Test Dataset: This data is separated from the training set and is used to test how well the model performs on unseen information. It simulates real-world future data the model will eventually predict.

Step 2 : Data Preprocessing

Data Preprocessing Contains various methods to clean the data and give it to the model.

Below are a few cases where Data preprocessing is required.

Missing values occur due to incomplete records or system errors.

Solution:

Forward Fill – Fill missing values with the previous valid entry.

Mean Imputation – Replace missing values with the average of that feature.

Handling Outliers

Outliers are extreme values that can **distort model predictions**.

Solution:

Use **box plots and Z-score analysis** to detect and remove extreme values.

Apply **log transformation** if necessary.

Feature Engineering

To improve model accuracy, new features are created:

1. **Lag Features** – Previous weeks' demand is used to predict future demand.
2. **Moving Averages** – Smooth fluctuations to identify demand trends.
3. **Price Difference** – (base_price - checkout_price) to assess discount impact.
4. **Region & City Features** – Adds location-based demand patterns.
5. **Encoding Categorical Variables**

a. Machine learning models require **numerical data**, so categorical features like meal_id, city_code, and region_code must be **converted**.

b. **Label Encoding:** Converts categorical values into numbers (e.g., city_A → 1, city_B → 2).

c. **Embedding Layers (for Deep Learning):** LSTMs and Bi-LSTMs can use **embedding layers** to process categorical features efficiently.

2. Scaling Numerical Features

Machine learning and deep learning models perform better when features are **scaled properly**.

Standardization (Z-score Normalization): Ensures numerical features have **zero mean and unit variance**.
Min-Max Scaling: Normalizes values to a **fixed range (0 to 1)**.

Pandas

A Python library used for manipulating datasets. It helps in tasks like

a. Handling missing values
b. Creating lag features (to include historical demand)
c. Encoding categorical data (e.g., holidays, weekdays)
d. Normalizing numerical features for neural networks

Step 3 : Predictions

This is the core of the forecasting process, where we use Random Forest Regressor to predict the output

Random Forest Regressor – A tree-based ensemble model that handles non-linear relationships and interactions well.

Step 4 : Interpret Results

After predictions are made, the results need to be analyzed and interpreted to provide value.

Performance Evaluation Metrics:

RMSE (Root Mean Square Error): Measures how far predicted values are from actual values.
MAE (Mean Absolute Error): Measures average magnitude of prediction errors.
R^2 Score: Indicates how well the model explains the variability of the target variable.

RESULTS & ANALYSIS

Evaluation metrics are essential tools used to assess the performance and accuracy of machine learning models and algorithms. These metrics provide quantitative measures that enable researchers and practitioners to evaluate the effectiveness of their methods and make informed decisions about model selection and optimization. Moreover, the choice of evaluation metrics depends on the nature of the problem being addressed and the desired outcome. By utilizing a combination of evaluation metrics, practitioners can gain comprehensive insights into the overall performance of their models and make informed decisions regarding their deployment and optimization strategies. These Evaluation metrics play a crucial role in not only validating the performance of machine learning models but also in comparing different models and algorithms. They help in identifying the strengths and weaknesses of a model, guiding the refinement process for better outcomes. Common evaluation metrics include Mean Absolute Error (MAE), Mean Squared Error (MSE), accuracy, and execution time. Each metric serves a specific purpose in evaluating different aspects of model performance, such as prediction accuracy, error magnitude, and computational efficiency.

EVALUATION METRICS

Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) is a metric used to measure the average absolute difference between predicted values and actual values. It is calculated by taking the average of the absolute differences between each predicted value and its corresponding actual value. The formula for

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |x_i - x| \qquad (2)$$

The Mean Squared Error (MSE) is a metric used to measure the average squared difference between predicted values and actual values. It is calculated by taking the average of the squared differences between each predicted value and its corresponding actual value. The formula for MSE is:

$$MSE = \frac{1}{N}\sum_{j=0}^{N-1} (X_j - Y_j)^2$$

X_jShows the upload image
Y_jShows the download image

Root Mean Squared Error (RMSE)

The Root Mean Squared Error (RMSE) is a commonly used metric to measure the average magnitude of the prediction errors. It is the **square root of the Mean Squared Error (MSE)**. RMSE gives an idea of how spread out the residuals (prediction errors) are. A lower RMSE value indicates better model performance.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i - P_i)^2} \qquad (3)$$

Where

Σ - It represents the "sum".

d_i- It represents the predicted value for the i^{th}

p_i- It represents the predicted value for the i^{th}

n - It represents the sample size.

R Squared (R^2 Score / Coefficient of Determination)

The R Squared (R^2) score is a statistical measure that represents the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from **0 to 1**, where 1 indicates perfect predictions and 0 indicates that the model does no better than the mean of the target values.

The formula for R^2 is:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \qquad (4)$$

$SS_{res} = \sum (y_i - \widehat{y_i})^2 : Residual\ sum\ of\ squares$

$SS_{tot} = \sum (y_i - \widehat{y})^2 : Total\ sum\ of\ squares$

y_i: $Actual\ values$

$\widehat{y_i}$: $Predicted\ values$

$\bar{y}$: $Mean\ of\ actual\ values$

An R^2 close to 1 indicates a strong correlation between the actual and predicted values, meaning the model explains most of the variability in the target variable.

Comparison

A comprehensive comparison of the results obtained using different techniques was conducted to evaluate the performance of our system. By contrasting the accuracy, MAE, MSE, RMSE, R^2 and execution time metrics across various forecasting approaches, we gained valuable insights into the strengths and limitations of each method. The comparison highlighted the efficacy of the Forecasting in accurately capturing the demand expressed, as evidenced by its superior performance in terms of accuracy and computational efficiency.

Models	Random Forest Regressor	XG Boost regressor	Gradient Boosting Machine	LSTM	Bi-LSTM
Mean Absolute Error	26.98	90.16	112.42	113.93	112.48
Mean Squared Error	3288.73	27493.5	47988.15	41391.48	59658.67
Root Mean Squared Error	57.34	165.81	219.06	203.44	244.25
R^2 Score	0.9791	0.8257	0.695	0.7377	0.621

Table.1. Comparison of different models results

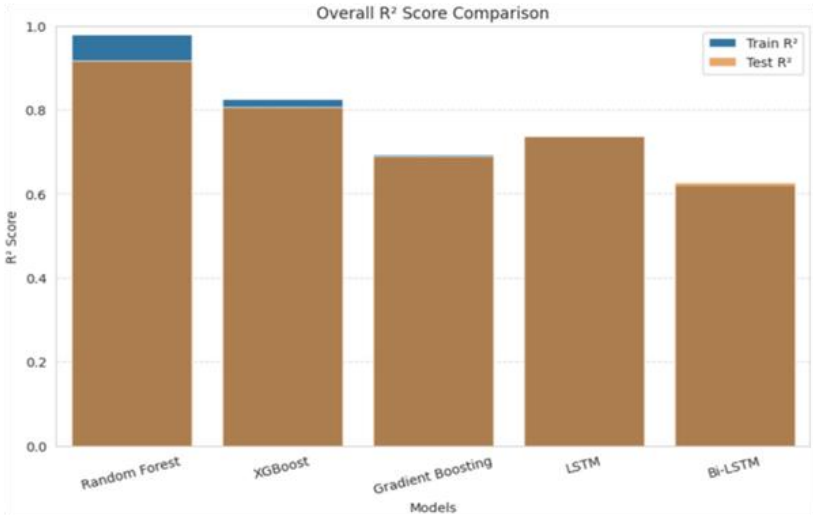


Fig 13: Graphical view of model comparison

Screenshots

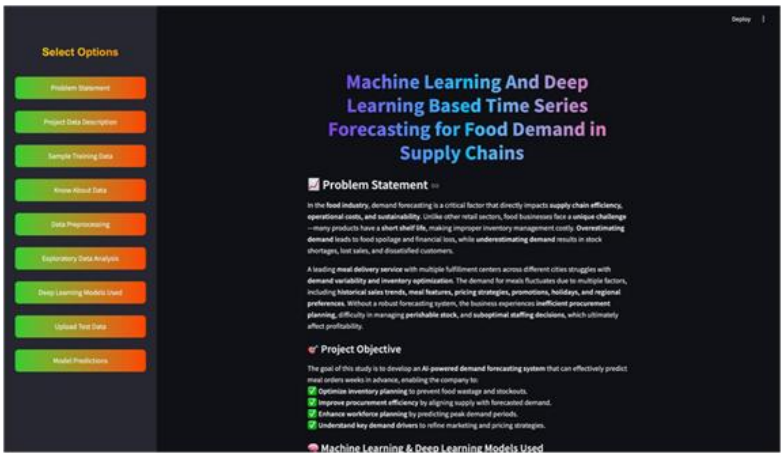


Fig 14: Landing Page of our Project

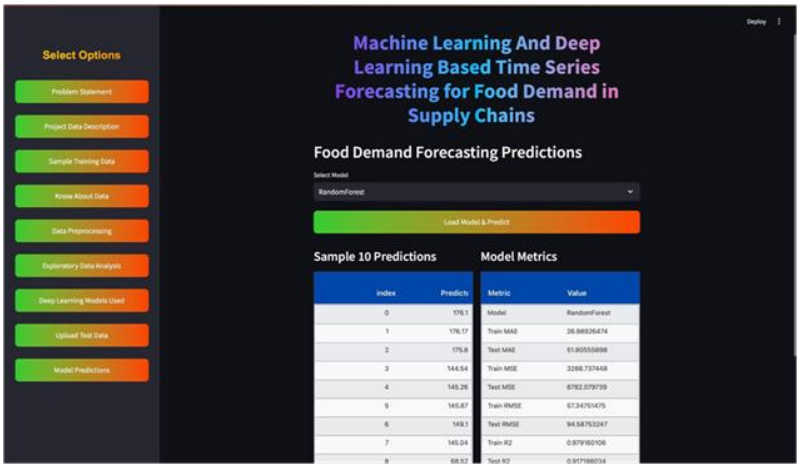


Fig 16: Model Prediction and Metrics

Conclusion

This research project aimed to address one of the most critical challenges in modern supply chain management—accurate forecasting of food demand. By applying machine learning techniques, specifically the **Random Forest**

Regressor, we successfully demonstrated that data-driven models can effectively capture complex patterns and temporal dependencies within historical data, leading to more reliable demand predictions. The implemented model was able to handle non-linear relationships and was robust against outliers and noise, making it suitable for real-world datasets where food demand is influenced by various unpredictable factors. Through this approach, we were able to show that machine learning not only enhances the **accuracy of forecasts** but also contributes to **reducing food waste**, **optimizing inventory levels**, and **improving overall operational efficiency** in the supply chain. In conclusion, this study demonstrates the practical value of machine learning in solving real-world supply chain problems and sets the foundation for more intelligent, scalable, and adaptive forecasting systems in the food industry. The outcomes highlight the potential of predictive analytics to transform supply chain planning into a more efficient and sustainable process.

FUTURE SCOPE

The current implementation of food demand forecasting using Random Forest Regressor has shown promise in capturing non-linear patterns and handling complex feature interactions. However, there is significant scope to enhance the model's performance and applicability in future research. One key direction is the exploration of deep learning models such as LSTM (Long Short-Term Memory) and Bi-LSTM, which are well-suited for time series data and can capture long-term dependencies more effectively than tree-based models. These models can improve forecasting accuracy, especially in highly dynamic environments where seasonality, trends, and irregular patterns exist. Future work can also include the development of hybrid models that combine the strengths of both machine learning and deep learning approaches, such as integrating feature importance from Random Forest with sequence learning in LSTM networks. Additionally, incorporating external variables like weather, holidays, promotions, and location-based data could enrich the model and improve prediction accuracy. Another promising area is the implementation of real-time forecasting systems using streaming data, which can be integrated with IoT devices in warehouses and retail outlets. This would allow for dynamic, on-the-fly adjustments to inventory and supply planning. Moreover, the project can be extended to include automated model selection and hyperparameter tuning using AutoML tools, making the system more adaptive and scalable. Finally, incorporating explainable AI (XAI) techniques would make the forecasting model more transparent and interpretable for supply chain managers and stakeholders, facilitating trust and informed decision-making.

References

1. S.Yadav,T.-M.Choi,S.Luthra,A.Kumar,andD.Garg,‘‘Using Internet of Things (IoT) in agri-food supply chains: A research framework for social good with network clustering analysis,’’ IEEE Trans. Eng. Manag., vol. 70, no. 3, pp. 1215–1224, Mar. 2023
2. J. Zheng, L. Wang, L. Wang, S. Wang, J.-F. Chen, and X. Wang, ‘‘Solving stochastic online food delivery problems via iterated greedy algorithm with decomposition-based strategy,’’ IEEE Trans. Syst., Man, Cybern., Syst., vol. 53, no. 2, pp. 957–969, Feb. 2023
3. [9] V. K. Shrivastava, A. Shrivastava, N. Sharma, S. N. Mohanty, and C. R. Pattanaik, ‘‘Deep learning model for temperature prediction: A case study in New Delhi,’’ J. Forecasting, vol. 43, no. 1, Feb. 2023
4. I. Shah, F. Jan, and S. Ali, ‘‘Functional data approach for short-term electricity demand forecasting,’’ Math. Problems Eng., vol. 2022, Jun. 2022
5. J. Ding, Z. Chen, L. Xiaolong, and B. Lai, ‘‘Sales forecasting based on CatBoost,’’ in Proc. 2nd Int. Conf. Inf. Technol. Comput. Appl. (ITCA), Dec. 2024
6. Y. Zhang, L. Wang, X. Chen, Y. Liu, S. Wang, and L. Wang, ‘‘Prediction of winter wheat yield at county level in China using ensemble learning,’’ Prog. Phys. Geogr., Earth Environ., vol. 46, no. 5,
7. pp. 676–696, Oct. 2022.
8. I. Shah, H. Iftikhar, and S. Ali, ‘‘Modeling and forecasting electricity demand and prices: A comparison of alternative approaches,’’ J. Math., vol. 2022, Jul. 2022, Art. no. 3581037.
9. I. Shah, S. Akbar, T. Saba, S. Ali, and A. Rehman, ‘‘Short-term forecasting for the electricity spot prices with

extreme values treatment,'' IEEE Access, vol. 9, pp. 105451–105462, 2021.

10. H.Hewamalage, C.Bergmeir, andK.Bandara,''Recurrentneuralnetworks for time series forecasting: Current status and future directions,'' Int. J. Forecasting, vol. 37, no. 1, pp. 388–427, 2021.
11. N. Bibi, I. Shah, A. Alsubie, S. Ali, and S. A. Lone, ''Electricity spot prices forecasting based on ensemble learning,'' IEEE Access, vol. 9, pp. 150984–150992, 2021.