The First International Conference on
Machine Learning and Knowledge Discovery (MLKD 2024)
Amirkabir University of Technology, December 18-19, 2024

Machine Learning for Food Demand Prediction: A CatBoost Regression Model

Pouria Mirazei* Ali Shahandeh Nokabadi*

Abstract

Accurately forecasting food demand is crucial for optimizing supply chain efficiency and reducing waste. This study presents a machine learning approach utilizing the CatBoost algorithm to predict food demand based on historical data. By leveraging features such as promotion activity, meal types, and customer center information, the model was trained using a dataset aggregated from multiple sources. Data preprocessing involved handling missing values and categorizing features. The model's performance was evaluated using key metrics such as R^2, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). Results demonstrate that the CatBoost Regressor achieves a high prediction accuracy (R^2 = 80.08%) compared with other algorithms such as linear regression (R2 =7.73%), Gradient Boosting Regressor (R2=45.52%), Extreme Gradient Boosting(R2=64.91%), Random Forest Regressor algorithm (R2=46.49%), making it a reliable tool for improving food supply chain decisions.

1 Introduction

In an era characterized by rapid advancements in technology and data analytics, the food industry faces unprecedented challenges and opportunities. With consumer preferences evolving at an accelerated pace, the ability to accurately forecast food demand has become paramount. The necessity for precise demand forecasting is underscored by several critical factors: the perishable nature of food items, the complexities of supply chain logistics, and the pressing need to minimize waste. Inefficient demand predictions can lead to either surplus stock, resulting in food wastage, or shortages, causing lost sales and diminished customer satisfaction. Therefore, developing robust models that accurately predict food demand is not just advantageous but essential for businesses aiming to thrive in today's competitive landscape.

The traditional methods of demand forecasting, such as time series analysis and simple moving averages, often fall short of capturing the intricate patterns in consumer behavior and market dynamics. These methods typically rely on historical sales data and do not account for external factors such as economic fluctuations, seasonal trends, and promotional activities that can significantly impact demand. As a result, businesses may find themselves ill-equipped to respond to sudden shifts in market conditions, ultimately affecting their bottom line. This gap in forecasting accuracy presents a compelling opportunity for the application of machine learning techniques, which offer the potential to revolutionize how food demand is predicted

Machine learning (ML) leverages vast amounts of

* Department of Industrial and System Engineering, Isfahan University of Technology, Isfahan, Iran

historical data and advanced algorithms to identify complex relationships and patterns that traditional methods may overlook. Unlike static models, ML algorithms can adapt in real time, incorporating various variables—including promotions, economic indicators, and regional preferences—into their predictions. This dynamic adaptability not only enhances forecasting accuracy but also empowers food providers to make informed decisions about inventory management, pricing strategies, and customer engagement. By implementing machine learning models for demand forecasting, businesses can optimize their operations, reduce waste, and ultimately improve profitability. Moreover, the integration of machine learning in food demand forecasting aligns with a broader trend towards data-driven decision-making in various sectors. As the volume of available data continues to grow, businesses that harness this information effectively will be better positioned to respond to market changes, meet consumer expectations, and sustain competitive advantage. Therefore, this study seeks to explore the development of a food demand predicting model using machine learning techniques, emphasizing the necessity of adopting innovative approaches to address the challenges of demand forecasting in the food industry. In conclusion, the urgency of this research lies in its potential to enhance the efficiency and effectiveness of food supply chains. As the food industry grapples with the dual challenges of waste reduction and consumer satisfaction, the development of advanced predictive models through machine learning offers a promising solution. By accurately forecasting food demand, businesses can ensure optimal inventory levels, reduce operational costs, and contribute to a more sustainable food system. This study will not only contribute to the academic discourse on machine learning applications but also provide practical insights for food industry stakeholders aiming to leverage technology for improved demand forecasting.[1]

2 Literature Review

- **Machine Learning**

  Machine learning (ML) is a subset of artificial intelligence (AI) that enables systems to learn from data, improve their performance over time, and make predictions or decisions without being explicitly programmed. It encompasses various algorithms and models, including supervised learning, unsupervised learning, and reinforcement learning, each serving different applications across industries. ML algorithms can identify complex patterns and relationships in large datasets, making them particularly valuable for tasks requiring predictive analytics and classification . The growth of big data and advancements in computational power have significantly enhanced the capabilities of ML, allowing businesses to leverage historical data for strategic decision-making . In recent years, ML has gained prominence in diverse fields such as finance, healthcare, and logistics, showcasing its potential to revolutionize traditional practices and optimize operations .[2]-[6]

- **Supervised Learning** Supervised learning is a major category of machine learning in which models are trained on labeled datasets, meaning each input data point is paired with the correct output . This approach aims to learn a mapping from inputs to outputs, enabling the model to make accurate predictions on unseen data . The training phase involves minimizing a loss function that quantifies the difference between predicted and actual outputs, while the testing phase evaluates the model's performance and generalization capability . Common algorithms used in supervised learning include linear regression, support vector machines, and decision trees, each offering unique strengths for various prediction tasks .[7]

- **Demand Prediction**

  Demand prediction involves estimating the future demand for a product or service based on historical data and influencing factors. Accurate demand forecasting is essential for businesses to optimize inventory management, reduce costs, and enhance customer satisfaction . Traditional forecasting methods, such as time series analy-

sis and exponential smoothing, often struggle to adapt to changing market conditions . In contrast, machine learning techniques can analyze large volumes of data and identify complex patterns, leading to more accurate demand predictions . The integration of external factors, such as economic indicators and seasonal trends, further improves the reliability of demand forecasting models .[8]

- **Food Demand Predictio**

  Food demand prediction is a specialized application of demand forecasting that focuses on estimating the future demand for food products in various settings, including restaurants, grocery stores, and supply chains . The ability to predict food demand accurately is crucial for minimizing waste, optimizing inventory, and ensuring product availability . Recent studies have demonstrated the effectiveness of machine learning algorithms in food demand forecasting, outperforming traditional methods in accuracy and adaptability . By incorporating variables such as seasonal trends, promotional campaigns, and consumer preferences, ML models can provide more nuanced insights into food demand dynamics . As the food industry increasingly embraces data-driven decision-making, the relevance of advanced predictive models continues to grow, driving efficiency and sustainability .[9]

3 Conclusion of Literature Review

While previous studies have made significant strides in food demand forecasting using various machine learning algorithms, few have explored the full potential of ensemble methods like CatBoost in this domain. Many approaches overlook the nuances of categorical feature interactions or fail to achieve a balance between prediction accuracy and computational efficiency. This research seeks to address these gaps by employing CatBoost, a robust gradient boosting algorithm, to improve the precision of food demand predictions. By leveraging a unique combination of categorical features and real-world data, this study aims to enhance decisionmaking processes in food supply chain management

4 Methodology

1. **Algorithm Explanation**

   The model utilized in this study is the CatBoost (Categorical Boosting) algorithm, an advanced gradient boosting framework developed by Yandex. CatBoost is designed to handle categorical features natively, which simplifies the preprocessing of data and improves the model's performance on datasets that contain such variables.[10]

   CatBoost builds an ensemble of decision trees through a boosting process, where each tree corrects the errors made by the previous ones. The mathematical formulation for updating the prediction with the k-th tree can be expressed as:[11],[12]

$$f_k(x) = f_{k-1}(x) + \boldsymbol{\gamma} h_k(x)$$

   Here, $f_k(x)$ is the prediction of the model after adding the k-th tree, $= f_{k-1}(x)$ is the prediction before adding the tree, $\boldsymbol{\gamma}$ is the learning rate, and $h_k(x)$ is the output of the new tree.

   CatBoost employs an ordered boosting technique, which mitigates overfitting by permuting the data during training. This allows the model to generalize better when faced with new, unseen data. The loss function optimized by CatBoost varies depending on the task; for regression problems, Mean Squared Error (MSE) is commonly used, defined mathematically as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y}_i)$$

   Where y_i are the true values and $\widehat{y}_i$ are the predicted values. Regularization techniques are also incorporated to prevent overfitting, such as

depth constraints on trees and careful handling of categorical variables.

2. **Dataset Description**

   The dataset used for this study comprises historical food order data, including both numerical and categorical features. It contains information about the orders made over several weeks, such as the center ID, meal ID, week, promotional campaigns, and whether items were featured on the homepage. The dataset is divided into training and testing sets, ensuring that the model can be evaluated on unseen data to assess its predictive capabilities.[13],[14]

   Key Features:

   Numerical Features:

   checkout price: The price of the meal at checkout (e.g., 136.83).

   base price: The base price of the meal before any discounts or promotions (e.g., 152.29).

   op area: The operational area of the center (e.g., 3.7).

   Categorical Features:

   center id: Unique identifier for each center (e.g., 679).

   meal id: Unique identifier for each meal (e.g., 1885).

   category: The type of meal (e.g., Beverages).

   cuisine: The cuisine type (e.g., Thai, Indian, Italian).

   emailer for promotion: Binary feature indicating if the meal was part of an email promotion (0/1).

   homepage featured: Binary feature indicating if the meal was featured on the homepage (0/1).

   Target Variable:

   num orders: The number of orders placed for each meal during the corresponding week.

   Dataset Size and Characteristics:

   The dataset contains historical food order data over several weeks, including both categorical and numerical features that influence food demand. The data spans a diverse range of meal categories and cuisines and reflects promotional and seasonal trends. Missing values in numerical features were imputed with column means, and rows with missing target values were removed to ensure data quality.

3. **Purpose of the Model and Study**

   The primary objective of this study is to develop a reliable predictive model for food demand using machine learning techniques. Accurate forecasting of food orders is critical for optimizing inventory management, reducing waste, and enhancing customer satisfaction in the food industry.

   By employing the CatBoost algorithm, the model aims to identify complex patterns and relationships in the historical data that traditional forecasting methods may overlook. The study emphasizes the necessity of leveraging machine learning to adapt to changing market dynamics and consumer preferences, ensuring that food providers can maintain optimal stock levels and make informed decisions regarding their supply chain. Ultimately, this research contributes to improving operational efficiency and profitability in the competitive landscape of the food industry.

4. **model evaluation**

   This study uses several techniques to validate the results, including R^2 , $RMSE$, and MAE[15]

   - R^2 Score (Coefficient of Determination):

     $$R^2 = \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2}$$

     Where y_i are the actual values, $\widehat{y}_i$ are the predicted values, and $\overline{y}_i$ is the mean of the actual values.

   - Mean Absolute Error (MAE):

     $$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y}_i|$$

- Root Mean Squared Error (RMSE):
  As mentioned above, RMSE is calculated as:

$$MAE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}$$

5 Importing Libraries

- **pandas**: A library for data manipulation and analysis, providing data structures like DataFrames to work with tabular data.
- **sklearn.model_selection**:Contains functions for splitting datasets into training and testing sets.
- **CatBoostRegressor**: A regression model from the CatBoost library designed to handle categorical features efficiently.
- **sklearn.metrics**:: Provides functions to evaluate the performance of the model using various metrics.
- **numpy**: A library for numerical operations, often used for handling arrays and mathematical functions.

6 Function Definitions

1. **load_data**
   - This function takes a list of CSV file paths as input.
   - It reads each CSV file into a pandas DataFrame and stores them in a list.
   - The DataFrames are concatenated into a single DataFrame using pd.concat, which combines the data while ignoring the original index (resetting it).
   - Finally,it returns the combined DataFrame.
2. **preprocess_data**
   - This function takes the combined DataFrame as input and performs preprocessing.
   - It strips any whitespace from column names for consistency.
   - If the target variable (num_orders) contains any NaN values, it prints a warning and drops those rows from the DataFrame.
   - It identifies numerical columns and fills any NaN values in these columns with the mean of the respective column.
   - It also identifies categorical columns and replaces NaN values with a placeholder string ('missing').
   - Finally, it separates the features from the target variable and returns them along with the list of categorical columns.
3. **main**
   - This is the main function where the program execution starts.
   - A list of CSV file paths is defined to specify the training data sources.
   - It calls the function to load and combine the CSV files into a single DataFrame.
   - The preprocessing function is called to clean and prepare the data for modeling, returning the features, target variable, and categorical columns.
   - The data is split into training and testing sets, with 10% of the data reserved for testing.
   - An instance of the CatBoostRegressor is created with specified parameters, such as the number of boosting iterations, depth of the tree, learning rate, loss function, and verbosity.
   - The model is fitted on the training data using the specified categorical features.
   - Predictions are made on the test set.

- Finally, evaluation metrics are calculated (R^2 score, Mean Absolute Error, and Root Mean Squared Error) to assess the model's performance, and the results are printed.

7 Model Results Analysis

Based on the output from your model, the following metrics were produced:

1. **Model Accuracy (R^2): 80.08%**

   The R^2 (coefficient of determination) value indicates that approximately 80.08of the variance in the target variable (number of orders) can be explained by the features used in the model. This is a strong indication of the model's predictive capability. In practical terms, this means that the model is quite effective in capturing the underlying patterns in the data, which is crucial for forecasting food demand accurately.

2. **Mean Absolute Error (MAE): 93.47**

   The MAE represents the average absolute difference between the predicted values and the actual values of the target variable. An MAE of approximately 93.47 suggests that, on average, the model's predictions deviate from the actual number of orders by around 93 orders. While this value might seem large, its impact should be contextualized based on the overall scale of orders in your dataset. If the average order count is significantly higher, this level of error could be acceptable for making operational decisions.

3. **Root Mean Squared Error (RMSE): 167.46**

   The RMSE provides a measure of the model's prediction error, where larger errors are penalized more than smaller ones due to the squaring effect. An RMSE of about 167.46 indicates that the model's predictions are, on average, about 167 orders away from the actual number of orders. Similar to the MAE, the significance of this value should be assessed in relation to the scale of your data.

In our analysis, several machine learning models were evaluated, including Linear Regression, Gradient Boosting Regressor, Extreme Gradient Boosting, Random Forest Regressor, and CatBoost. The models were compared based on their R^2, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) scores.

1. **Linear Regression:**
   - $R^2 = 7.73\%$

Linear regression performed poorly due to its linear nature, which failed to capture the complex, nonlinear relationships in the data. This model struggled particularly with categorical features and interactions between them. The low R^2 value reflects this inadequacy, as linear regression could not adequately predict the demand based on multiple influencing factors.

2. **Gradient Boosting Regressor:**
   - $R^2 = 45.52\%$

Gradient Boosting showed improved performance, but it still faced difficulties in capturing the interactions between categorical features like meal id and center id. This led to a suboptimal fit, explaining only 45.52% of the variance in the target variable.

3. **Extreme Gradient Boosting (XGBoost):**
   - $R^2 = 64.91\%$

XGBoost performed better, but still fell short of capturing the full complexity of the dataset, as it does not handle categorical features natively. Feature engineering was required for this model, which may have resulted in loss of predictive power.

4. **Random Forest Regressor:**
   - $R^2 = 46.49\%$

Random Forest also exhibited a moderate R^2 score, but its performance was impacted by overfitting, especially with high-dimensional categorical data. Its inability to model categorical interactions as effectively as CatBoost contributed to the lower R^2.

5. **CatBoost:**
   - $R^2 = 80.08\%$

CatBoost significantly outperformed all other models, achieving an R^2 of 80.08%, indicating a strong ability to capture the interactions between categorical features. Unlike other models, CatBoost handles categorical features natively, which improves its performance on datasets like this one that include

features such as meal id and center id. Its ability to reduce overfitting through ordered boosting also contributed to the higher accuracy.

8 Benefits of the Model

- **High Accuracy:**The R^2 value indicates that your CatBoost model can explain a substantial portion of the variance in food demand. This high level of accuracy is beneficial for making informed decisions about ordering ingredients like meat and other perishable items.
- **Robust to Overfitting:**CatBoost, as a gradient boosting algorithm, is designed to be robust against overfitting, especially with categorical features. This is important in the food industry, where patterns in demand can vary greatly over time and among different food items.
- **Ease of Use with Categorical Features:**The ability of CatBoost to handle categorical features directly without extensive preprocessing (like one-hot encoding) simplifies the modeling process. This efficiency is crucial in a real-world application where datasets can be large and complex.
- **Prediction Reliability:** With a relatively low MAE and RMSE, your model is capable of making reliable predictions. This reliability can help in optimizing inventory management, reducing waste, and improving overall efficiency in food supply chains.

9 Conclusion

This study highlights the integration of machine learning into food demand forecasting, demonstrating how advanced algorithms like CatBoost can address the challenges of managing supply chains in the food industry. By leveraging a unique combination of numerical and categorical features, the proposed approach simplifies preprocessing while delivering high predictive accuracy. The CatBoost model achieved an R^2 score of 80.08%, significantly outperforming traditional models such as Linear Regression, Gradient Boosting, and Random Forest. This performance underscores its ability to handle complex feature interactions and deliver reliable forecasts for large-scale operations.

Beyond predictive accuracy, the analysis of feature importance provided actionable insights into key factors influencing food demand, such as promotional activities, meal categories, and operational characteristics. These findings offer practical guidance for optimizing inventory management and targeting promotional efforts, ultimately contributing to waste reduction and cost savings. The resilience of the model to overfitting and its ease of implementation further enhance its suitability for real-world applications in supply chain management.

As consumer preferences and market dynamics continue to evolve, this research demonstrates the transformative potential of machine learning in enabling data-driven decision-making. Future work will extend this methodology to other industries, such as retail and healthcare, to validate its adaptability and explore additional opportunities for improving operational efficiency.

Future Research Directions

The model developed for food demand forecasting has the potential to be applied in various other sectors beyond food. In retail sales forecasting, it can predict demand for diverse products, enabling retailers to optimize inventory levels and enhance customer satisfaction by reducing stockouts. In supply chain management, the model can anticipate fluctuations in demand, allowing companies to plan procurement and logistics more effectively. Additionally, it can be used for production planning in manufacturing, helping companies align their output with expected demand, thus minimizing overproduction and waste. In the healthcare sector, the model can forecast the need for medical supplies and equipment based on patient admission patterns, ensuring adequate resources are available. Overall, the model's versatility enables it to contribute significantly to improving operational efficiency and decision-making across various industries.

References

[1] Qin Y, Tang J, Li T, Qi X, Zhang D, Wang S, et al. Cultivated Land Demand and Pressure in Southeast Asia from 1961 to 2019: A Comprehensive Study on Food Consumption. Foods. 2023;12(19).

[2] Weller DL, Love TMT, Wiedmann M. Interpretability Versus Accuracy: A Comparison of Machine Learning Models Built Using Different Algorithms, Performance Measures, and Features to Predict E. coli Levels in Agricultural Water. Front Artif Intell. 2021;4:628441.

[3] Beam KS, Zupancic JAF. Machine learning: remember the fundamentals. Pediatr Res.2023;93(2):291-2.

[4] Maleki F, Ovens K, Najafian K, Forghani B, Reinhold C, Forghani R. Overview of Machine Learning Part 1: Fundamentals and Classic Approaches. Neuroimaging Clin N Am. 2020;30(4):e17-e32.

[5] Makridakis S, Spiliotis E, Assimakopoulos V. Statistical and Machine Learning forecasting methods: Concerns and ways forward. PLoS One. 2018;13(3):e0194889.

[6] Galata Z, Kloukina I, Kostavasili I, Varela A, Davos CH, Makridakis M, et al. Amelioration of desmin network defects by alphaB-crystallin overexpression confers cardioprotection in a mouse model of dilated cardiomyopathy caused by LMNA gene mutation. J Mol Cell Cardiol.2018;125:73-86.

[7] Schroer HW, Just CL. Feature Engineering and Supervised Machine Learning to Forecast Biogas Production during Municipal Anaerobic Co-Digestion. ACS ES T Eng. 2024;4(3):660-72.

[8] Li N, Arnold DM, Down DG, Barty R, Blake J, Chiang F, et al. From demand forecasting to inventory ordering decisions for red blood cells through integrating machine learning, statistical modeling, and inventory optimization. Transfusion. 2022;62(1):87-99.

[9] Goli A, Khademi-Zare H, Tavakkoli-Moghaddam R, Sadeghieh A, Sasanian M, Malekalipour Kordestanizadeh R. An integrated approach based on artificial intelligence and novel meta-heuristic algorithms to predict demand for dairy products: a case study. Network. 2021;32(1):1-35.

[10] Montesinos-Lopez OA, Gonzalez HN, Montesinos-Lopez A, Daza-Torres M, Lillemo M, Montesinos-Lopez JC, et al. Comparing gradient boosting machine and Bayesian threshold BLUP for genome-based prediction of categorical traits in wheat breeding. Plant Genome. 2022;15(3):e20214.

[11] Ahn JM, Kim J, Kim K. Ensemble Machine Learning of Gradient Boosting (XGBoost, LightGBM, CatBoost) and Attention-Based CNN-LSTM for Harmful Algal Blooms Forecasting. Toxins (Basel). 2023;15(10).

[12] Xu S, Liu S, Wang H, Chen W, Zhang F, Xiao Z. A Hyperspectral Image Classification Approach Based on Feature Fusion and Multi-Layered Gradient Boosting Decision Trees. Entropy (Basel). 2020;23(1).

[13] Odion D, Shoji K, Evangelista R, Gajardo J, Motmans T, Defraeye T, et al. A GIS-based interactive map enabling data-driven decision-making in Nigeria's food supply chain. MethodsX. 2023;10:102047.

[14] Winer BJ. Statistics and data analysis: trading bias for reduced mean squared error. Annu Rev Psychol. 1978;29:647-81.

[15] Chicco D, Warrens MJ, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput Sci. 2021;7:e623.