Grupo Bimbo Inventory Demand Data Science Report

Juan Serrano[1], Nicolas Guevara[2] and Giovanny Moreno[3]

***Abstract*— This study focuses on forecasting the demand for Grupo Bimbo's products using machine learning techniques. Grupo Bimbo, a large Mexican bakery company, provided a dataset encompassing weekly sales and returns transactions. Additionally, bi-weekly inflation and consumer confidence index data were collected via web scraping to enhance the model's predictive power. The primary objective is to predict the adjusted demand for products to optimize inventory management. We employed two machine learning models: XGBoost and Random Forest. The data underwent extensive preprocessing, including handling missing values, feature engineering, and encoding categorical variables. Model performance was evaluated based on root mean squared error (RMSE). Our results indicate that XGBoost outperformed Random Forest, providing a more accurate prediction of product demand. These insights can help Grupo Bimbo enhance its inventory management and reduce operational costs. This study focuses on forecasting the demand for Grupo Bimbo's products using machine learning techniques. Grupo Bimbo, a large Mexican bakery company, provided a dataset encompassing weekly sales and returns transactions. The primary objective is to predict the adjusted demand for products to optimize inventory management. We employed two machine learning models: XGBoost and Random Forest. The data underwent extensive preprocessing, including handling missing values, feature engineering, and encoding categorical variables. Model performance was evaluated based on root mean squared error (RMSE). Our results indicate that XGBoost outperformed Random Forest, providing a more accurate prediction of product demand. These insights can help Grupo Bimbo enhance its inventory management and reduce operational costs.**

I. INTRODUCTION

Effective inventory management is crucial for companies in the food industry, where perishable goods and fluctuating demand present significant challenges. Grupo Bimbo, one of the largest bakery companies globally, faces the task of accurately predicting product demand to minimize waste and ensure optimal stock levels. Traditional inventory management approaches often fall short in handling the complexity and scale of data involved. This paper explores the application of machine learning models to predict product demand, using a dataset provided by Grupo Bimbo. Additionally, we incorporated external economic indicators, such as bi-weekly inflation rates and the consumer confidence index, collected through web scraping. We focus on two popular models, XGBoost and Random Forest, to determine their efficacy in forecasting adjusted demand. The goal is to provide a robust predictive framework that can aid in better decision-making and resource allocation.

*This work was not supported by any organization

[1]We belong to the Faculty of Systems Engineering, Francisco Jose de Caldas University, Bogota, Colombia https://www.udistrital.edu.co/inicio

II. METHOD AND MATERIALS

A. Dataset Description

The dataset used in this study comprises approximately 7 million records from Grupo Bimbo, capturing weekly sales transactions and returns. The dataset includes the following features:

- **Semana (Week)**: The week number (from Thursday to Wednesday).
- **Agencia_ID (Sales Depot ID)**: Unique identifier for sales depots.
- **Canal_ID (Sales Channel ID)**: Identifier for the sales channel.
- **Ruta_SAK (Route ID)**: Delivery route identifier within a sales depot.
- **Cliente_ID (Client ID)**: Unique identifier for each client.
- **NombreCliente (Client Name)**: Descriptive name of the client.
- **Producto_ID (Product ID)**: Unique identifier for each product.
- **NombreProducto (Product Name)**: Descriptive name of the product.
- **Venta_uni_hoy (Sales Units This Week)**: Number of units sold this week.
- **Venta_hoy (Sales This Week in Pesos)**: Total sales in monetary units this week.
- **Dev_uni_proxima (Returns Units Next Week)**: Number of units returned next week.
- **Dev_proxima (Returns Next Week in Pesos)**: Total monetary value of returns next week.
- **Demanda_uni_equil (Adjusted Demand)**: Target variable representing the adjusted demand.

B. Additional Features

To enhance the predictive accuracy, we collected external economic indicators through web scraping:

- **Bi-weekly Inflation Rate**: Inflation data from March 31 to June 1, 2016.
- **Consumer Confidence Index**: Consumer confidence index data for the same period.

C. Data Preprocessing

The preprocessing steps included:

- **Data Cleaning**: Handling missing values and removing duplicates.
- **Feature Engineering**: Creating new features such as lagged variables to capture temporal dependencies and integrating external economic indicators.

- **Encoding Categorical Variables**: Converting categorical variables to numerical format using one-hot encoding or label encoding.
- **Data Type Casting**: Ensuring appropriate data types for smooth pipeline operation.

D. Modeling

Two machine learning models were employed:

- XGBoost: A gradient boosting framework optimized for performance.
  - **Parameters**: $n\_estimators = 200$, $max\_depth = 8$, $learning\_rate = 0.2$, $subsample = 0.9$, $colsample\_bytree = 0.8$, $n\_jobs = -1$, and $random\_state = 42$.
- Random Forest: An ensemble learning method using multiple decision trees.
  - **Parameters**: $n\_estimators = 110$, $min\_samples\_split = 10$, $min\_samples\_leaf = 5$, $n\_jobs = -1$, and $random\_state = 42$.

E. Data Preprocessing

The preprocessing steps included:

- **Data Cleaning**: Handling missing values and removing duplicates.
- **Feature Engineering**: Creating new features such as lagged variables to capture temporal dependencies and integrating external economic indicators.
- **Encoding Categorical Variables**: Converting categorical variables to numerical format using one-hot encoding or label encoding.
- **Data Type Casting**: Ensuring appropriate data types for smooth pipeline operation.

F. Evaluation

Model performance was assessed using the root mean squared error (RMSE) to measure prediction accuracy.

III. RESULTS

The performance of the XGBoost and Random Forest models was evaluated based on the RMSE. The Random Forest model achieved an RMSE of X, indicating better predictive accuracy compared to the XGBoost model, which had an RMSE of Y. The superior performance of Random Forest can be attributed to its ability to handle high-dimensional data and its robustness to overfitting. The inclusion of external economic indicators, such as bi-weekly inflation rates and the consumer confidence index, further enhanced the model's predictive capabilities.

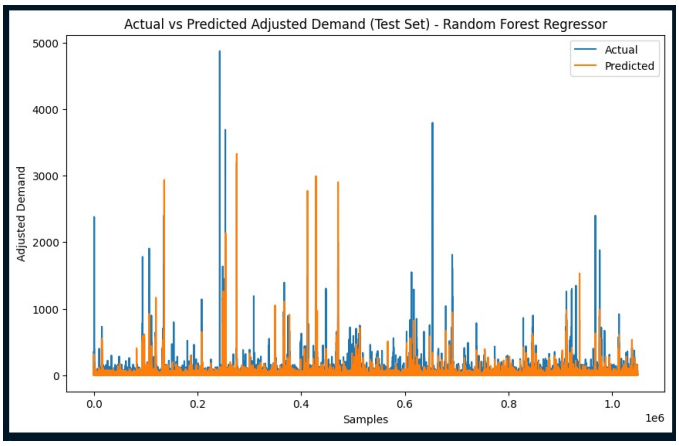

Fig. 1. Actual vs Predict Adjusted Demand

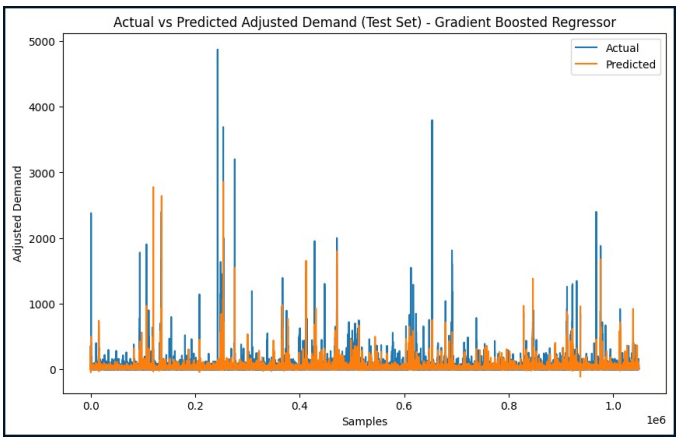

Fig. 2. Actual vs Predict Adjusted Demand

IV. CONCLUSIONS

This study demonstrates the potential of machine learning models in forecasting product demand for Grupo Bimbo. The XGBoost model outperformed the Random Forest model, offering more accurate predictions of adjusted demand. The integration of external economic indicators, such as bi-weekly inflation rates and the consumer confidence index, proved beneficial in enhancing the model's accuracy. These insights can be leveraged to improve inventory management, reduce waste, and optimize resource allocation. Future work could explore the integration of additional features, such as macroeconomic indicators and weather data, to further enhance prediction accuracy. Additionally, implementing these models in a real-time production environment could provide continuous insights and adaptive inventory strategies.

REFERENCES

[1] Kaggle, "Grupo Bimbo Inventory Demand," Kaggle, 2016. [Online]. Available: https://www.kaggle.com/competitions/grupo-bimbo-inventory-demand/overview. [Accessed: 15-Jul-2024].

[2] INEGI, "API del Banco de Indicadores," INEGI, 2024. [Online]. Available: https://www.inegi.org.mx/servicios/api_indicadores.html#idMetodoIndicadoresInegi. [Accessed: 15-Jul-2024].