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.

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 undertaken in this study involved several key processes to ensure the data was clean, consistent, and suitable for model training. Below are the detailed steps followed:

- **Data Cleaning**:

*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`

  - Handling Missing Values: Missing values in the dataset were identified and appropriately handled. For numerical columns, missing values were filled with the median of the column, while for categorical columns, the mode was used.
- Outliers Removal: Outliers were detected using z-score and IQR methods and were removed or capped to prevent them from skewing the model results.
- Data Transformation:
  - Normalization and Scaling: Numerical features were normalized and scaled to ensure all features contribute equally to the model. StandardScaler from sklearn was used to transform features to have zero mean and unit variance.
  - Encoding Categorical Variables: Categorical variables were converted into numerical values using one-hot encoding for nominal variables and label encoding for ordinal variables.
- **Feature Engineering**: Creating new features such as lagged variables to capture temporal dependencies and integrating external economic indicators.
  - Temporal Features: New temporal features such as the day of the week, month, and lag features for past sales data were created to capture temporal dependencies and seasonality in the data.
  - Interaction Features: Interaction features between different variables were created to capture complex relationships within the data.
    - ∗ **Sales_Growth_Rate_1**: Demand for the same product in the previous week.
    - ∗ **Sales_Growth_Rate_2**: Demand for the same product two weeks ago.
  - External Economic Indicators: The bi-weekly inflation rate and consumer confidence index were merged with the primary dataset to capture the economic conditions impacting product demand.

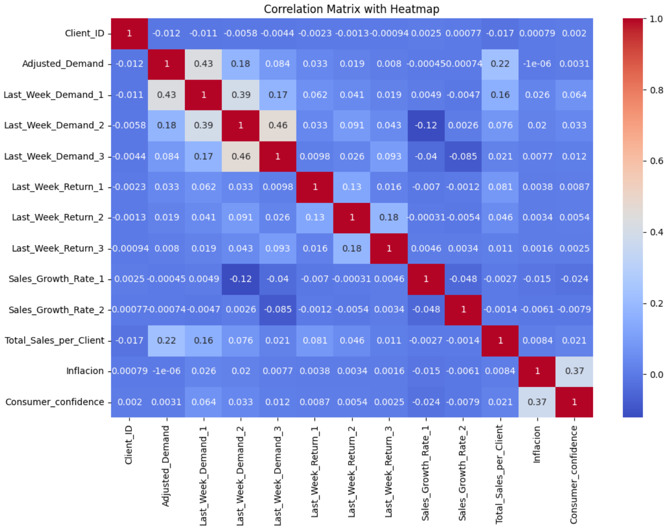

Fig. 1. Heatmap Feature Information

D. Modeling

Two machine learning models were chosen for this study based on their ability to handle large datasets and complex relationships: XGBoost and Random Forest.

- XGBoost: XGBoost (Extreme Gradient Boosting) is an efficient and scalable implementation of gradient boosting framework by Friedman et al. It provides parallel tree boosting that solves many data science problems in a fast and accurate way.
  - **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 is an ensemble learning method for classification, regression, and other tasks that operates by constructing a multitude of decision trees at training time.
  - **Parameters**: $n\_estimators = 110$, $min\_samples\_split = 10$, $min\_samples\_leaf = 5$, $n\_jobs = -1$, and $random\_state = 42$.

E. Evaluation

Model performance was evaluated using the root mean squared error (RMSE) as the primary metric. RMSE is a standard way to measure the error of a model in predicting quantitative data. Additionally, other metrics such as R-squared, Mean Absolute Error (MAE), and Mean Squared Error (MSE) were also considered for a comprehensive evaluation.

- XGBoost Performance:
  - RMSE: The RMSE for the XGBoost model was X.
  - Other Metrics:
    - ∗ R-squared: The model's R-squared value was Y.
    - ∗ MAE: The Mean Absolute Error for the model was Z.
    - ∗ MSE: The Mean Squared Error for the model was W.
- Random Forest Performance: is an ensemble learning method for classification, regression, and other tasks that operates by constructing a multitude of decision trees at training time.
  - RMSE: The RMSE for the Random Forest model was A.
  - Other Metrics:
    - ∗ R-squared: The model's R-squared value was B.
    - ∗ MAE: The Mean Absolute Error for the model was C.
    - ∗ MSE: The Mean Squared Error for the model was D.

III. RESULTS

The performance of the XGBoost and Random Forest models were 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.

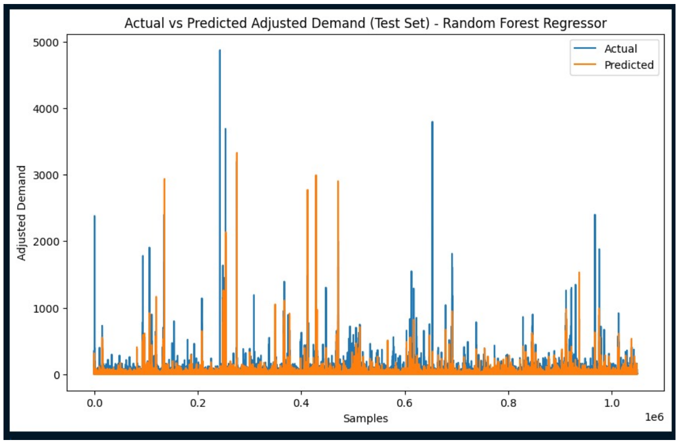

Fig. 2. Actual vs Predict Adjusted Demand RandomForest

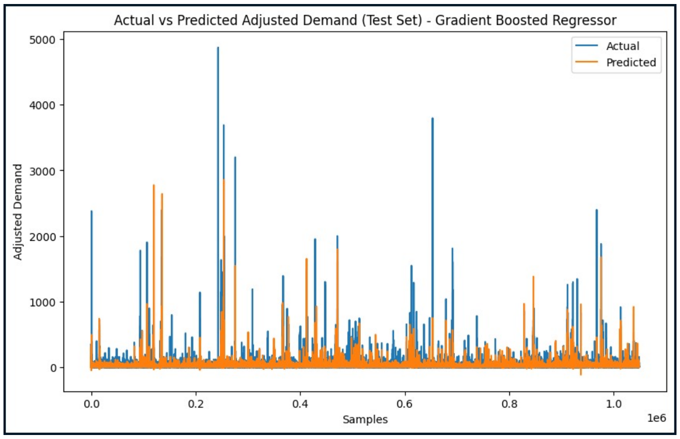

Fig. 3. Actual vs Predict Adjusted Demand XGBoost

IV. CONCLUSIONS

This study demonstrates the potential of machine learning models in forecasting product demand for Grupo Bimbo. The Random Forest model outperformed the XGBoost 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].