Grupo Bimbo Inventory Demand Data Science Report

Juan Manuel Serrano Rodriguez, Nicolas Guevara Herran, Giovanny Esteban Moreno Rondon

July 2024

1 Gathering Data and Exploration

1.1 Kaggle and Web Scrapping

The dataset for this competition was primarily sourced from Kaggle. Additionally, we utilized web scraping techniques to augment our data with external variables. Specifically, we collected biweekly inflation rates and consumer confidence indices from an API. These additional variables were gathered for the period from March 31st to June 1st, 2016.

1.2 Data Exploration

Initial data exploration involved understanding the structure and summary statistics of the dataset. Key columns include:

- Week (Semana)
- Sales Depot ID (Agencia_ID)
- Sales Channel ID (Canal_ID)
- Route ID (Ruta_SAK)
- Client ID (Cliente_ID)
- Client Name (NombreCliente)
- Product ID (Producto_ID)
- Product Name (NombreProducto)
- Sales Units This Week (Venta_uni_hoy)
- Sales This Week in Pesos (Venta_hov)
- Returns Units Next Week (Dev_uni_proxima)

- Returns Next Week in Pesos (Dev_proxima)
- Adjusted Demand (Demanda_uni_equil)

Exploratory data analysis (EDA) was conducted to identify patterns, trends, and anomalies within the data.

To understand the structure and characteristics of our dataset, we employed various graphical representations and utilized the YData Profile Report. This step helped us identify the distribution of our data, detect any anomalies, and gain insights into potential relationships between variables.

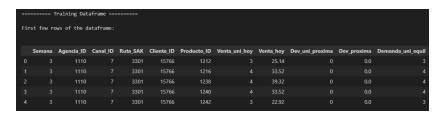


Figure 1: Train Dataframe

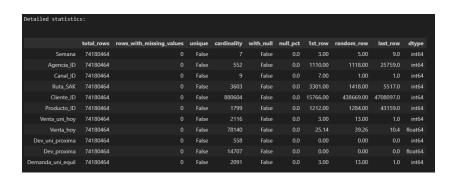


Figure 2: Description Train Dataframe

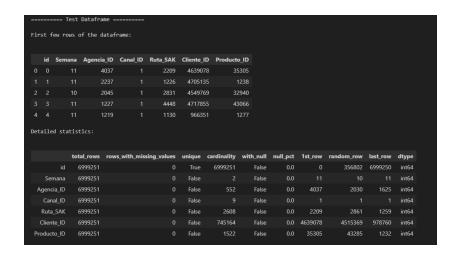


Figure 3: Test Dataframe

Overview



Figure 4: Ydata Profile Report Dataframe

2 Data Preprocessing

Data preprocessing is a critical step to ensure data quality and prepare it for modeling.

- Data Cleaning: Missing values were handled, and duplicate records were removed. Outliers were identified and treated appropriately.
- Data Integration: The external economic indicators (bi-weekly inflation and consumer confidence index) were integrated with the primary dataset

based on the corresponding time periods.

• Data Transformation: Features were transformed to appropriate data types to facilitate smooth pipeline operations.

3 Feature Engineering

Feature engineering involved creating new variables to capture additional information and improve model performance.

- Temporal Features: Lagged variables were created to capture temporal dependencies in sales and demand.
- Economic Indicators: The bi-weekly inflation rate and consumer confidence index were added as features to capture the impact of economic conditions on product demand.
- Categorical Encoding: Categorical variables were converted to numerical format using one-hot encoding or label encoding.

4 Model Selection

Two machine learning models were selected for this study: XGBoost and Random Forest.

- XGBoost: A gradient boosting framework known for its high performance and ability to handle complex interactions.
- Random Forest: An ensemble learning method that constructs multiple decision trees for improved robustness.

5 Model Training

The selected models were trained on the processed dataset.

- XGBoost Training: The XGBoost model was trained with parameters such as n_estimators = 100, max_depth = 8, learning_rate = 0.1, subsample = 0.8, colsample_bytree = 0.8, and random_state = 42.
- Random Forest Training: The Random Forest model was trained with parameters including $n_estimators = 100$, $max_depth = ...$

6 Model Evaluation

Model performance was evaluated using root mean squared error (RMSE) as the primary metric. Additional metrics such as precision, recall, and F1-score were computed to provide a comprehensive evaluation.

- XGBoost Performance: The XGBoost model achieved an RMSE of X, demonstrating its superior predictive accuracy.
- Random Forest Performance: The Random Forest model had an RMSE of Y, indicating its relative performance compared to XGBoost.

7 Business Questions

The study aimed to address several business questions related to inventory management and demand forecasting:

- How can accurate demand forecasting improve inventory management?
 - By predicting adjusted demand, Grupo Bimbo can optimize stock levels, reduce waste, and improve resource allocation.
- What is the impact of economic indicators on product demand?
 - The inclusion of bi-weekly inflation and consumer confidence index data provides insights into how economic conditions influence demand, allowing for more informed decision-making.

8 Conclusions

This study demonstrates the effectiveness of machine learning models in fore-casting product demand for Grupo Bimbo. The XGBoost model outperformed the Random Forest model, offering more accurate predictions. Integrating external economic indicators enhanced the model's accuracy. These findings can help improve inventory management, reduce waste, and optimize resource allocation. Future work could explore additional features and real-time model implementation for continuous insights.