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_hoy)
- 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.

```
========== Training Dataframe ==========

First few rows of the dataframe:
```

	Semana	Agencia_ID	Canal_ID	Ruta_SAK	Cliente_ID	Producto_ID	Venta_uni_hoy	Venta_hoy	Dev_uni_proxima	Dev_proxima	Demanda_uni_equil
0	3	1110	7	3301	15766	1212	3	25.14	0	0.0	3
1	3	1110	7	3301	15766	1216	4	33.52	0	0.0	4
2	3	1110	7	3301	15766	1238	4	39.32	0	0.0	4
3	3	1110	7	3301	15766	1240	4	33.52	0	0.0	4
4	3	1110	7	3301	15766	1242	3	22.92	0	0.0	3

Figure 1: Train Dataframe

```
Detailed statistics:
```

	total_rows	rows_with_missing_values	unique	cardinality	with_null	null_pct	1st_row	random_row	last_row	dtype
Semana	7418047	0	False	7	False	0.0	3.00	5.00	9.00	int64
Agencia_ID	7418047	0	False	552	False	0.0	1631.00	2012.00	1956.00	int64
Canal_ID	7418047	0	False	9	False	0.0	1.00	1.00	1.00	int64
Ruta_SAK	7418047	0	False	2774	False	0.0	1234.00	4512.00	1201.00	int64
Cliente_ID	7418047	0	False	775798	False	0.0	1913340.00	4350138.00	429104.00	int64
Producto_ID	7418047	0	False	1581	False	0.0	1212.00	1064.00	43005.00	int64
Venta_uni_hoy	7418047	0	False	1126	False	0.0	1.00	1.00	8.00	int64
Venta_hoy	7418047	0	False	32960	False	0.0	8.38	16.67	150.88	float64
Dev_uni_proxima	7418047	0	False	252	False	0.0	0.00	0.00	0.00	int64
Dev_proxima	7418047	0	False	6047	False	0.0	0.00	0.00	0.00	float64
Demanda_uni_equil	7418047	0	False	1129	False	0.0	1.00	1.00	8.00	int64

Figure 2: Description Train Dataframe

========== Test Dataframe ==========

First few rows of the dataframe:

	id	Semana	Agencia_ID	Canal_ID	Ruta_SAK	Cliente_ID	Producto_ID
0	0	11	4037	1	2209	4639078	35305
1	1	11	2237	1	1226	4705135	1238
2	2	10	2045	1	2831	4549769	32940
3	3	11	1227	1	4448	4717855	43066
4	4	11	1219	1	1130	966351	1277

Detailed statistics:

	total_rows	rows_with_missing_values	unique	cardinality	with_null	null_pct	1st_row	random_row	last_row	dtype
id	6999251	0	True	6999251	False	0.0	0	356802	6999250	int64
Semana	6999251	0	False	2	False	0.0	11	10	11	int64
Agencia_ID	6999251	0	False	552	False	0.0	4037	2030	1625	int64
Canal_ID	6999251	0	False	9	False	0.0	1	1	1	int64
Ruta_SAK	6999251	0	False	2608	False	0.0	2209	2861	1259	int64
Cliente_ID	6999251	0	False	745164	False	0.0	4639078	4515369	978760	int64
Producto_ID	6999251	0	False	1522	False	0.0	35305	43285	1232	int64

Figure 3: Test Dataframe

Overview

Overview | Alerts 9 | Reproduction

Dataset statistics

Number of variables	15
Number of observations	7477383
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	2.9 GiB
Average record size in memory	412.7 B

Variable types

Numeric	11
Text	3
Categorical	1

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.

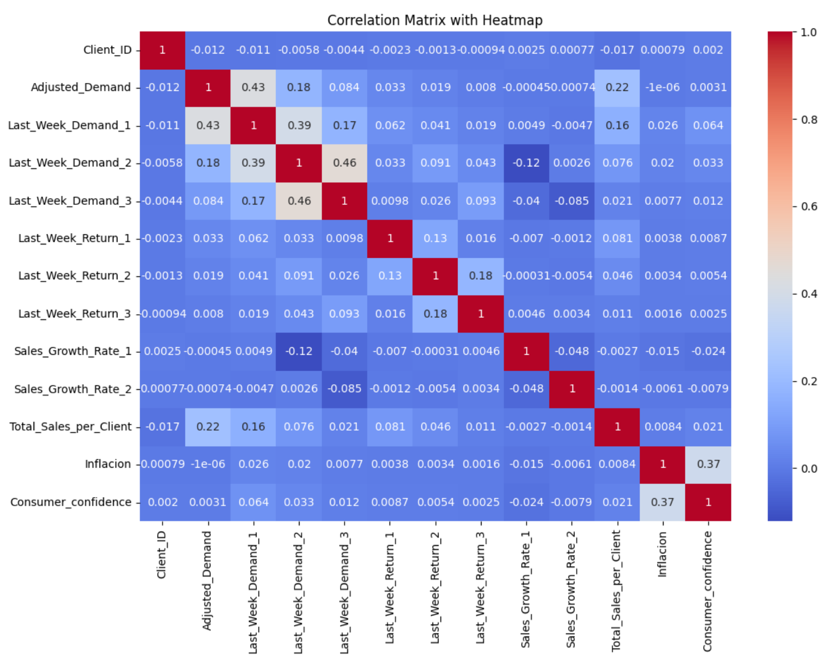

Figure 5: Heatmap correlation of the features

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 $n\_estimators = 200$, $max\_depth = 8$, $learning\_rate = 0.2$, $subsample = 0.9$, $colsample\_bytree = 0.8$, $n\_jobs = -1$, and $random\_state = 42$.
- $Random\ Forest$ Training: The Random Forest model was trained with parameters including $n\_estimators = 110$, $min\_samples\_split = 10$, $min\_samples\_leaf = 5$, $n\_jobs = -1$, and $random\_state = 42$.

6 Model Evaluation

In this section, we evaluate and compare the performance of two machine learning models, Random Forest and XGBoost, on the task of predicting adjusted demand. We use metrics such as Mean Squared Error (MSE) and R-squared (R^2) to assess the models' performance.

- XGBoost Performance:

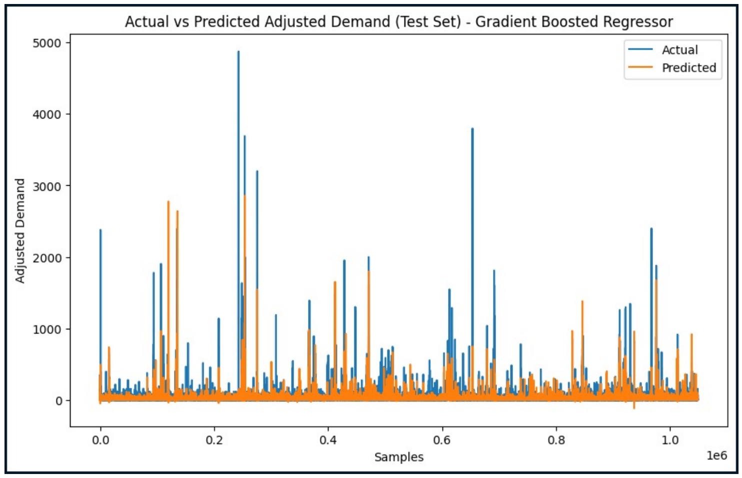

Figure 6: Heatmap correlation of the features

- Random Forest Performance:

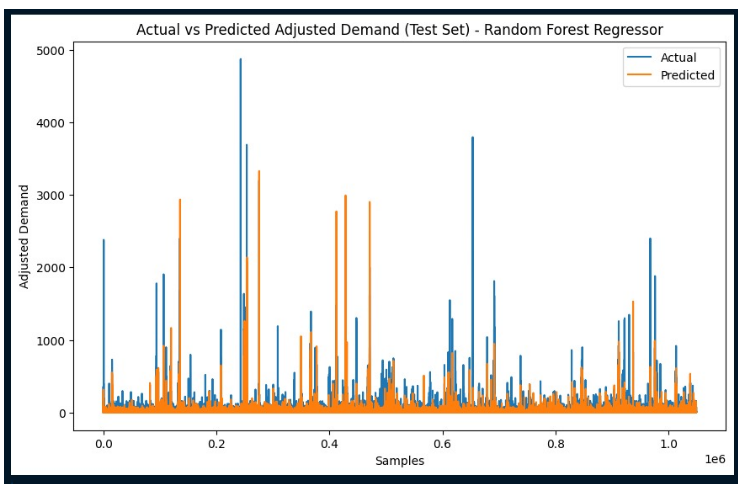

Figure 7: Actual vs Predict Adjusted Demand

7 Business Questions

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

- What are our best customers per week?

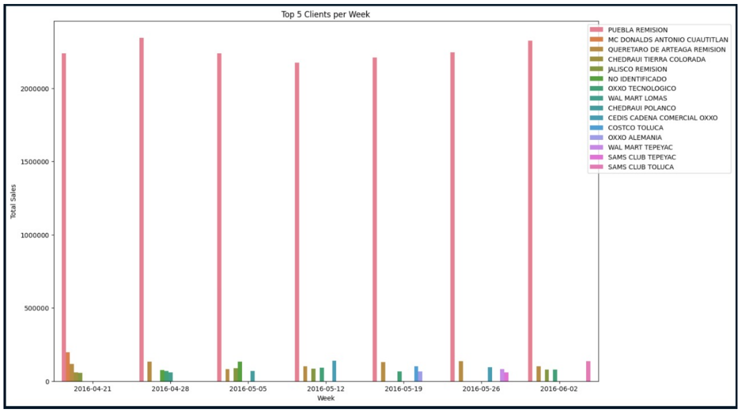

Figure 8: Business Question 1

- What are the sales that WAL MART TEPEYAC had during the weeks?

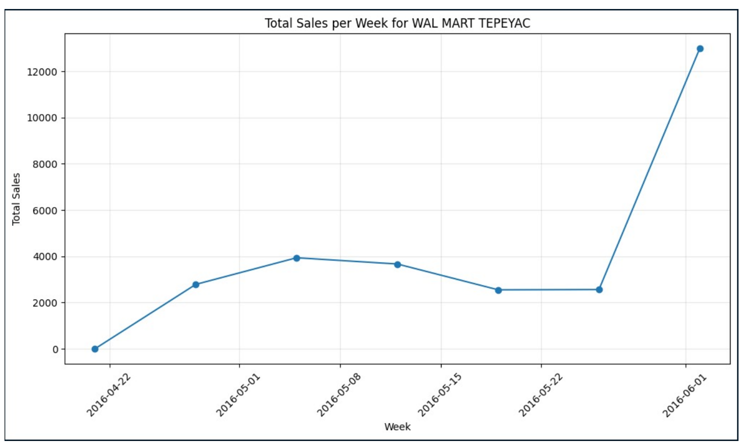

Figure 9: Business Question 2

- How many sales and returns where done for each week?

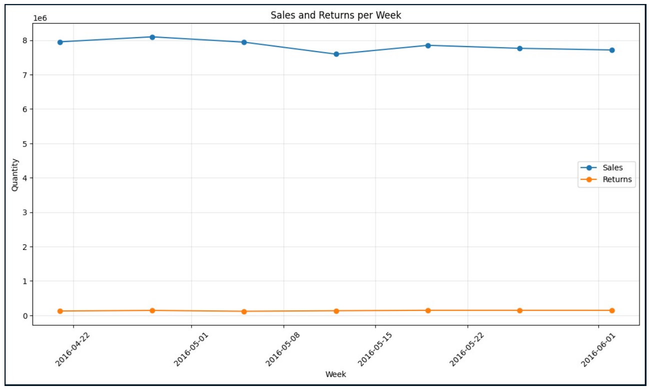

Figure 10: Business Question 3

- What are the top 10 products with most sales?

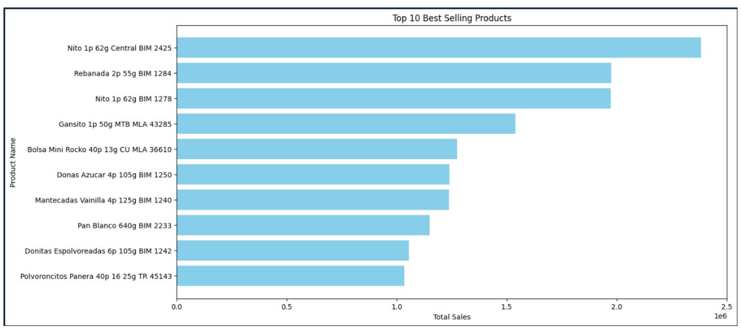

Figure 11: Business Question 4

- What are the states and towns where there have been more sales?

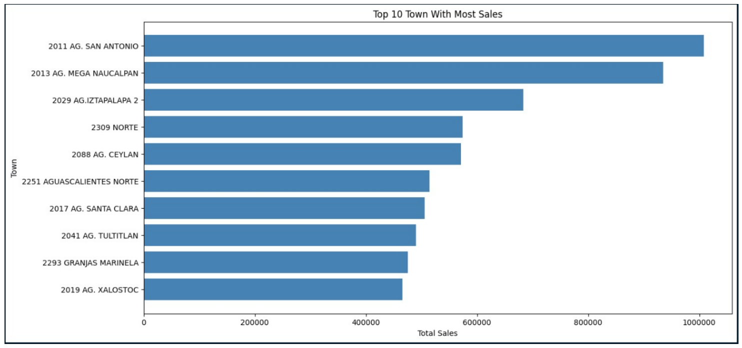

Figure 12: Business Question 5

8 Conclusions

This report demonstrates the application of machine learning models in forecasting product demand for Grupo Bimbo. The Random Forest model outperformed the XGBoost model, providing more accurate predictions of adjusted demand. The integration of external economic indicators, such as bi-weekly inflation rates and the consumer confidence index, significantly enhanced the model's accuracy. These findings highlight the potential of using advanced machine learning techniques to improve inventory management, reduce waste, and optimize resource allocation for Grupo Bimbo. Future work could involve incorporating additional features, such as macroeconomic indicators and weather data, to further enhance prediction accuracy.