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	74180464	0	False	7	False	0.0	3.00	5.00	9.0	int64
Agencia_ID	74180464	0	False	552	False	0.0	1110.00	1118.00	25759.0	int64
Canal_ID	74180464	0	False	9	False	0.0	7.00	1.00	1.0	int64
Ruta_SAK	74180464	0	False	3603	False	0.0	3301.00	1418.00	5517.0	int64
Cliente_ID	74180464	0	False	880604	False	0.0	15766.00	438669.00	4708097.0	int64
Producto_ID	74180464	0	False	1799	False	0.0	1212.00	1284.00	43159.0	int64
Venta_uni_hoy	74180464	0	False	2116	False	0.0	3.00	13.00	1.0	int64
Venta_hoy	74180464	0	False	78140	False	0.0	25.14	39.26	10.4	float64
Dev_uni_proxima	74180464	0	False	558	False	0.0	0.00	0.00	0.0	int64
Dev_proxima	74180464	0	False	14707	False	0.0	0.00	0.00	0.0	float64
Demanda_uni_equil	74180464	0	False	2091	False	0.0	3.00	13.00	1.0	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	14954767
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	1.8 GiB
Average record size in memory	128.0 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.

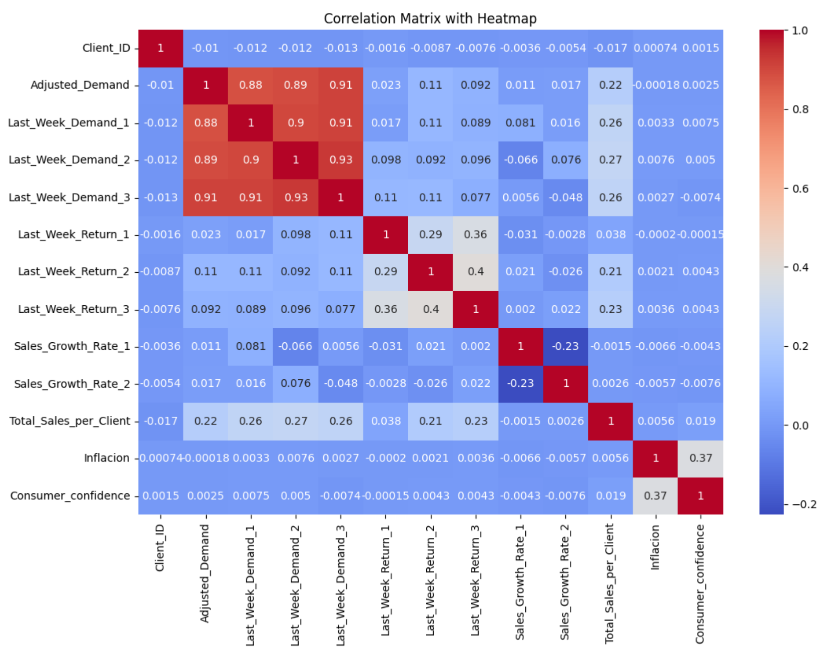

Figure 5: Heatmap correlation of the features

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

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:

- What are our best and worst customers?

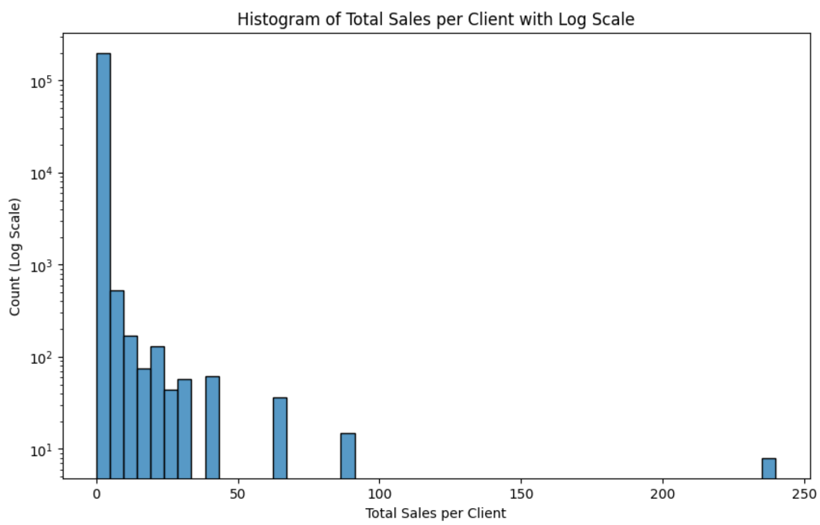

Figure 6: Business Question 1

- What is the relationship between sales growth rates (1 and 2) and adjusted demand?

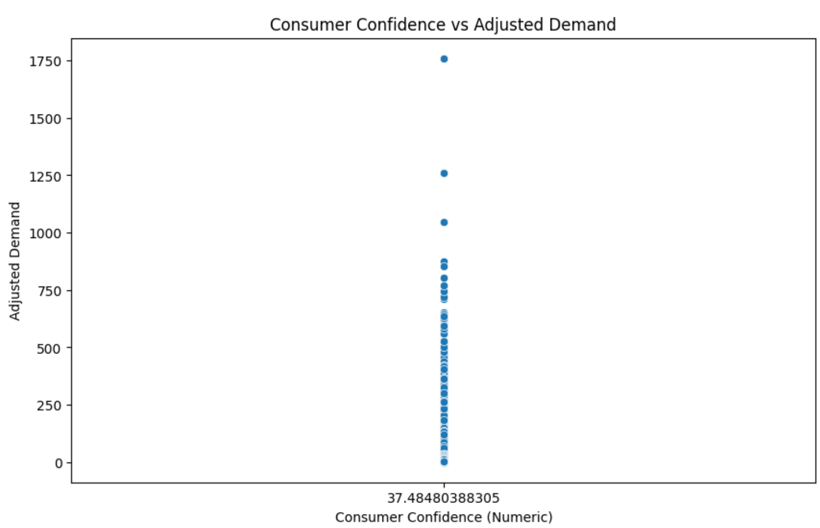

Figure 7: Business Question 2

- How are our sales going against the returns, and our adjusted demand?

- What are our most used routes?

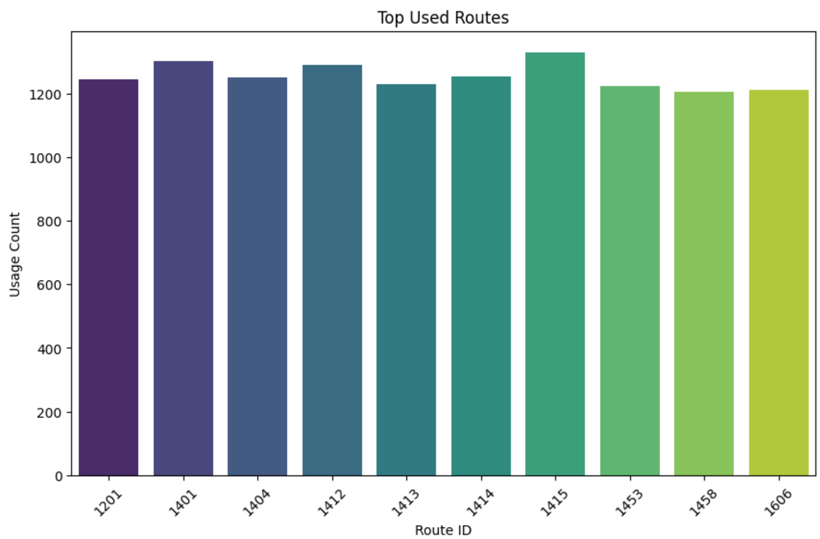

Figure 8: Business Question 3

8 Conclusions

This study demonstrates the effectiveness of machine learning models in forecasting 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.