Data Intelligence Hackathon: Demand Prediction in the Industry

Marissa E. Luna[1], Ximena A. Cantón[1] and Nubia G. Barajas[1]

[1] *Instituto Tecnológico y de Estudios Superiores de Monterrey, Escuela de Ingeniería y Ciencias, Guadalajara, Jalisco*

Abstract—This paper presents an analysis of historical sales from one of the most important lubricant companies in Mexico, including segmentation by customer, product, and sales unit. Prediction models were trained with XGBoost for kilograms, liters, and pieces, improving performance after removing outliers. Finally, additional internal and external variables are proposed for future improvements in model accuracy.

Keywords—exploratory data analysis, sales prediction, XGBoost, outliers, time series, sales units, performance metrics

I. Introduction

Demand prediction is key to optimizing production, logistics, and procurement in the industry. This work analyzes the sales history of a company specializing in industrial lubricants, with the aim of forecasting future sales quantities by product.

The analysis included an exploratory study to detect patterns and anomalies, followed by the development of predictive models using *XGBoost*, employing variable encoding, outlier removal through IQR, and logarithmic transformations to improve model stability.

II. Methodology

The exploratory data analysis (EDA) was conducted to understand the historical behavior of sales orders recorded by the company, identifying general patterns, irregularities, and potential errors in the data. The main stages of the methodology were as follows:

Dataset Cleaning and Structuring

- The original file contained all fields combined into a single column. Preprocessing was performed to correctly divide it into: *Sales Order*, *Sales Order Creation*, *Customer Code*, *Product*, *Quantity*, and *Sales Unit*.
- Duplicate records and those with a quantity equal to zero were detected and removed.
- Additionally, a systematic error in the data was corrected: when the *Quantity* column contained more than three consecutive zeros at the end (e.g., 270000), these zeros were removed to reflect the actual quantity (270).

Data Type Conversion

- The fields *Quantity* were converted to numeric type and *Sales Order Creation* to date type (`datetime64`).
- To facilitate monthly analysis, a new column *YearMonth* was generated in text format (`YYYY-MM`).

Division by Sales Units

- It was identified that products were sold in three different units: **KG**, **L**, and **PZA**.
- To avoid mixing that could bias the analysis, the dataset was segmented by unit, and each unit was analyzed independently.

Structured Exploratory Analysis

The following dimensions were addressed:

- Customer Analysis
- Product Analysis
- Order Quantity Analysis
- Monthly Temporal Analysis
- Outlier Detection through *boxplots* and robust statistics (IQR)

III. Exploratory Data Analysis

Before beginning the exploratory analysis, a review and cleaning process was carried out on the data. The original dataset contained 31,156 records distributed across 6 key columns: order number, creation date, customer, product, sales unit, and quantity.

No null values were found in any of the columns, indicating that the dataset was complete. However, **4,957 duplicate records** were identified and removed, reducing the total to **26,199 unique entries**. **20 records with a quantity equal to zero** were also detected and discarded as they did not represent real sales.

These records were added to an error dataframe, and it was found that all the articles and customers from these zero

sales had already been part of valid orders, ruling out the possibility of erroneous products or inactive customers.

Subsequently, the column `Sales Order Creation` was converted to the `datetime` data type to allow proper temporal analysis. Finally, the data was segmented into three subsets according to the sales unit: **kilograms (KG)**, **liters (L)**, and **pieces (PZA)**. This allowed for more precise descriptive statistics and visualizations for each product type.

The following basic statistics of quantities sold per unit are presented:

- **Kilograms (KG)** – `12,826 observations`
  - Mean: 596.4 Standard deviation: 1,515
  - Minimum: 1 Median: 135 Maximum: 37,800
- **Liters (L)** – `7,702 observations`
  - Mean: 366.1 Standard deviation: 746.1
  - Minimum: 1 Median: 142 Maximum: 17,280
- **Pieces (PZA)** – `5,651 observations`
  - Mean: 25.9 Standard deviation: 77.2
  - Minimum: 1 Median: 9 Maximum: 4,500

As observed, products sold in kilograms and liters present higher sales volumes and greater dispersion. In contrast, products sold by piece tend to be marketed in smaller quantities, with a strong concentration near the minimum and a median significantly lower than the average, indicating a right-skewed distribution.

This preprocessing was essential to ensure the quality of the subsequent analysis and to obtain reliable conclusions about the behavior of sales, customers, and products.

a. Customer Analysis

907 unique customers were identified in the dataset.

Most customers make few orders: the median is **4 orders**, but the average is **29**, indicating the presence of customers with highly frequent behavior.

The customer with the highest number of orders registered **1,913 orders**, while several customers have only one.

When observing the quantities sold per customer, relevant differences were identified depending on the unit of measure. For example, in products sold by kilogram, histograms revealed a high concentration of sales in low quantity ranges, with some extreme values influencing the mean.

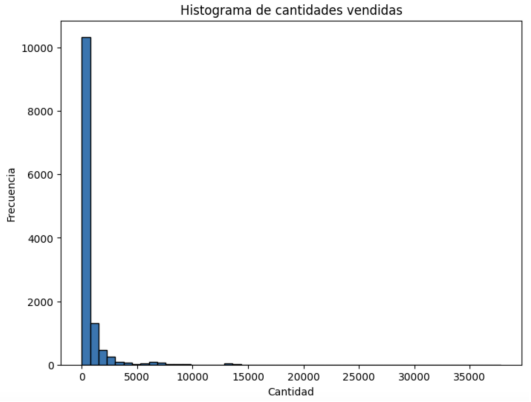

Fig. 1: Histogram of quantities sold in kilograms. A strong concentration of sales in low values is observed, but with a long tail to the right indicating the presence of outliers.

When applying a logarithmic scale to the histograms, the general trend of the data became clearer.

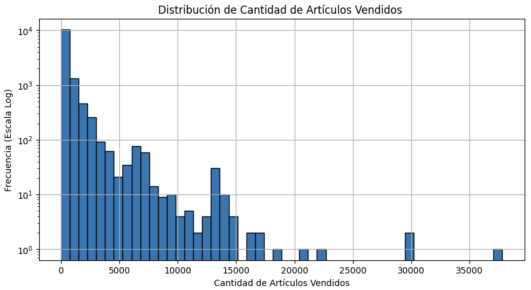

Fig. 2: Logarithmic histogram of quantities sold in kilograms. The transformation allows for a clearer view of the true concentration of the data in low sales ranges.

A similar analysis with products sold in liters showed a similar pattern, reinforcing the trend that most customers purchase small quantities per transaction.

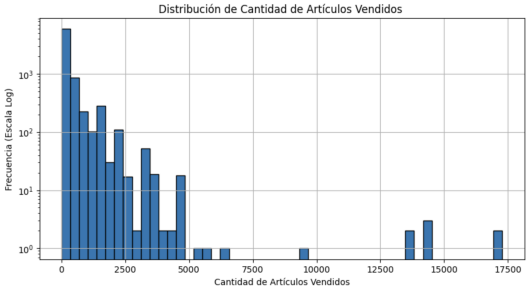

Fig. 3: Histogram of quantities sold in liters. The distribution shows a pattern similar to kilograms: high concentration in low values and presence of extreme values.

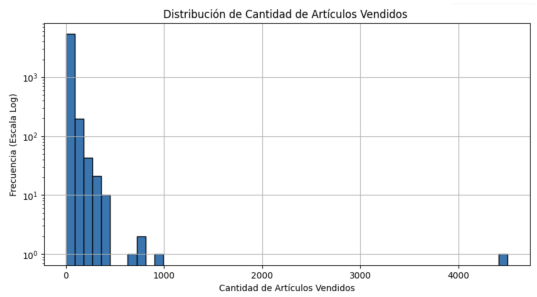

Fig. 4: Histogram of quantities sold in pieces. A high frequency in small values is observed, with some outliers at the extreme right of the distribution.

The distribution of quantities sold in pieces reinforces the trend observed in other types of units of measure. Most purchases are in small quantities, suggesting that customers tend to purchase only what is necessary in each transaction. However, the presence of extreme values indicates that there are some outlier purchases with significantly larger volumes. This could be due to wholesale purchases or customers with specific needs that require a larger than average quantity.

Distribution of Orders by Customer

The following chart shows the distribution of the number of orders placed by each customer:

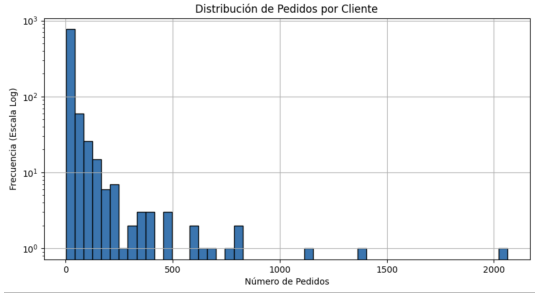

Fig. 5: Histogram of orders by customer.

Key observations:

- Right-skewed distribution: The majority of customers made a small number of orders. As the number of orders increases, the number of customers decreases rapidly.
- Occasional vs. frequent customers: There is a broad base of customers who made few orders, suggesting sporadic or trial purchasing behavior. In contrast, there is a small group of customers who made between 500 and over 2000 orders, indicating high purchase frequency and strategic value for the business.
- Valuable outliers: Some atypical customers made more than 1000 orders. Although they represent less than 1% of the total, their contribution to revenue may be disproportionately high.

It is recommended to segment the customer base based on their order frequency to design differentiated strategies:

- Loyalty and retention for frequent customers.
- Reactivation or promotion campaigns for occasional customers.

Top 10 Customers with Most Orders

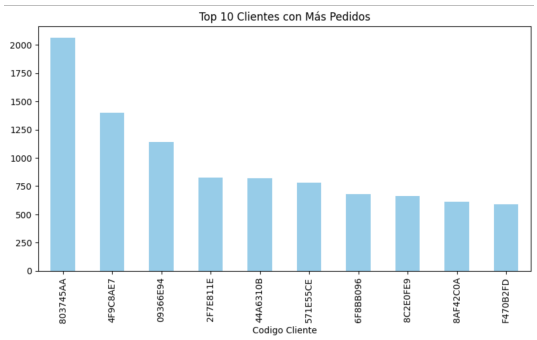

Fig. 6: Top 10 customers with the most orders. Each bar represents a customer identified by their code, sorted from highest to lowest number of orders.

Key observations:

- High concentration of orders in a few customers: The customer with the code `803745AA` stands out with more than 2000 orders, followed by others with volumes ranging from about 600 to 1400 orders.
- Gradual decrease: There is a stepped decrease in the number of orders as we move down the ranking, reinforcing the importance of the top spots.
- Key customers for the business: These 10 customers likely represent a significant portion of the total order volume. Their sustained purchasing behavior suggests established business relationships or a high level of loyalty.

These customers are strategic for the business. It is recommended to provide them with close follow-up, either through loyalty programs, personalized attention, or preferential commercial terms. It is also worth analyzing whether these customers belong to the same segment, region, or industry type, to identify growth opportunities in similar profiles.

b. Article Analysis

A total of **889 unique articles** were identified within the dataset. For easier analysis, the articles were grouped by unit of sale: kilograms (kg), liters (L), and pieces (PZA). This allowed comparison not only of the frequency of sale of each article but also the quantity sold by their unit of measurement.

Descriptive Statistics by Unit

The following shows the main statistics of the quantity sold per article by unit:

- **Kilograms (KG):** mean of 17,075, median of 1,080, and maximum of 897,102. There is a strong right-skew.
- **Liters (L):** mean of 8,623, median of 1,670, and maximum of 207,724. The distribution follows a similar pattern to KG.
- **Pieces (PZA):** mean of 808, median of 18, and maximum of 33,548. High concentration in low values.

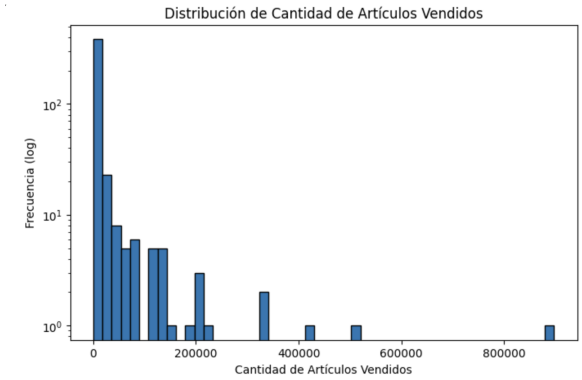

Fig. 7: Distribution of quantity sold per article (KG). A strong right-skew is observed: most articles are sold in small quantities, but some exceed 100,000 units.

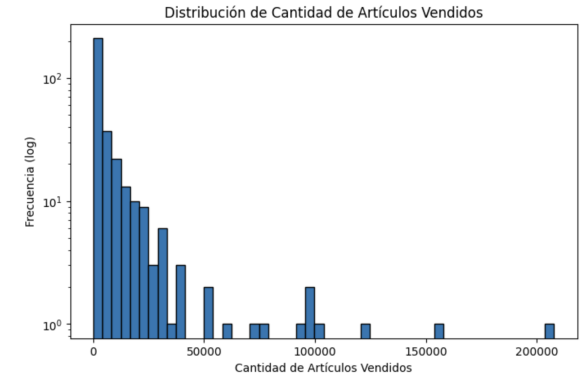

Fig. 8: Distribution of quantity sold per article (L). The concentration in low values is also evident, although with less dispersion than in KG.

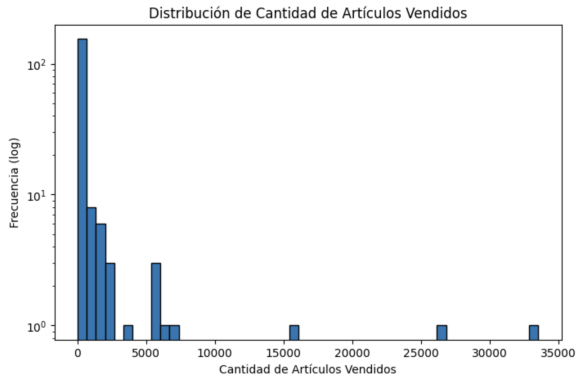

Fig. 9: Distribution of quantity sold per article (PZA). A very high concentration in small sales is shown, with a long tail of articles exceeding 10,000 units.

Top Sold Articles by Quantity

The following shows the **10 most sold articles** by total quantity, broken down by unit of measurement.

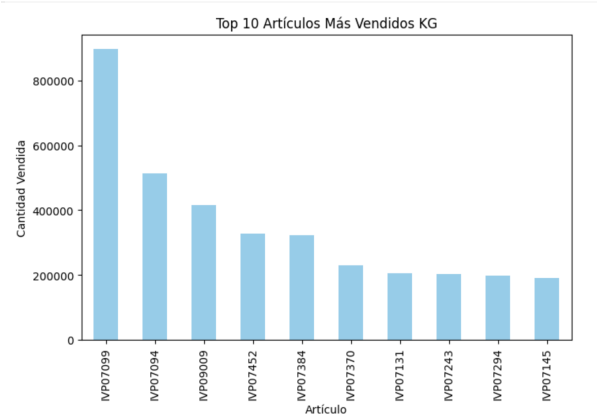

Fig. 10: Top 10 most sold articles in KG. The article `IVP07099` clearly dominates with more than 800,000 kg sold.

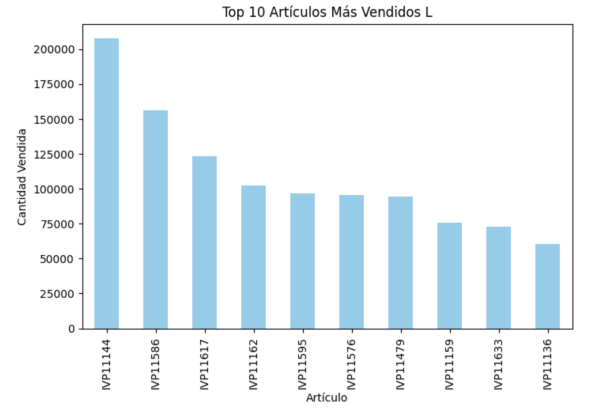

Fig. 11: Top 10 most sold articles in L. The article `IVP1144` leads the list with more than 200,000 liters sold.

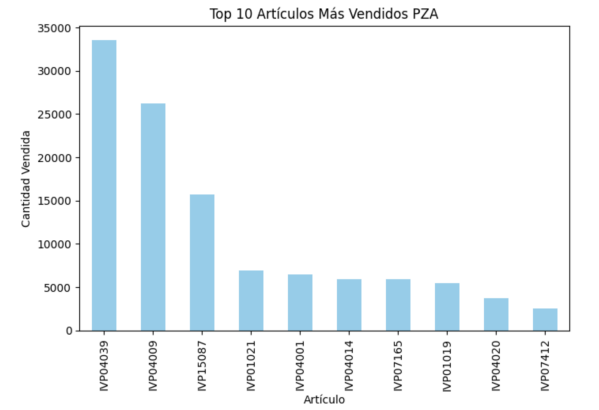

Fig. 12: Top 10 most sold articles in PZA. The article `IVP04039` exceeds 30,000 pieces sold, far ahead of the others.

Top-Selling Items by Number of Orders

In addition to total volume, items that appear most frequently in orders were identified. This helps to distinguish high-recurrence products, even if they are sold in small volumes.

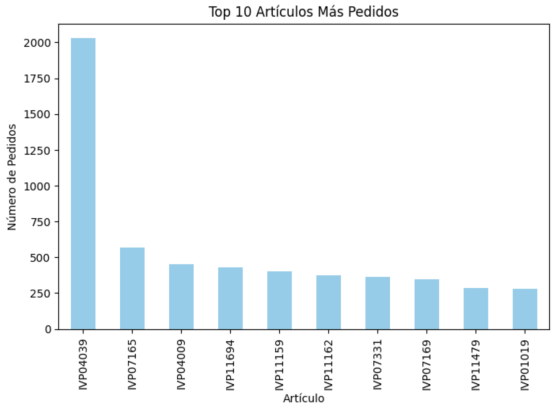

Fig. 13: Top 10 items with the most orders. `IVP04039` leads with more than 2,000 orders registered, positioning it as a high-frequency purchase item.

Conclusions from Item Analysis

- The distributions of quantity sold per item are strongly skewed to the right, especially in pieces, indicating that a few products account for the majority of the volume.
- Some items are leaders both in volume and frequency of orders, but there are also products that, although not sold in large quantities, have high turnover.
- These differences are important for inventory management, promotions, and forecasting future demand.

c. Temporal Analysis

An analysis of sales evolution over time was conducted using the `Date` variable and aggregating total quantity sold by day. This approach helps detect behavioral patterns, potential seasonality, or unusual events affecting demand.

Trend by Unit of Measure

Sales were disaggregated by unit of measure: kilograms, liters, and pieces. This allowed us to observe differences in the temporal behavior patterns of each product type.

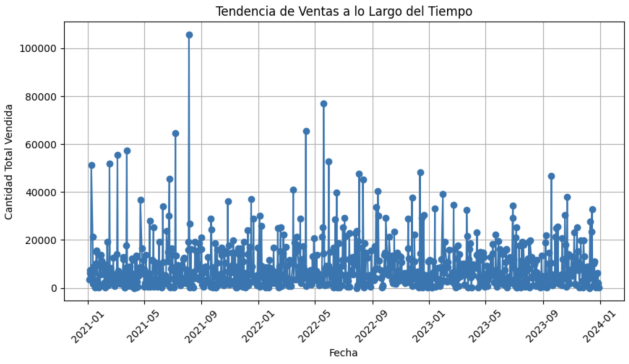

Fig. 14: Sales trend for products sold in kilograms. Multiple sales peaks are observed, especially in the first quarters of 2021 and 2022, which could suggest seasonality patterns or specific campaigns.

The kilograms series shows strong daily variability, with values ranging from 0 to more than 100,000 units sold in a single day. This dispersion suggests that sales are not evenly distributed over time.

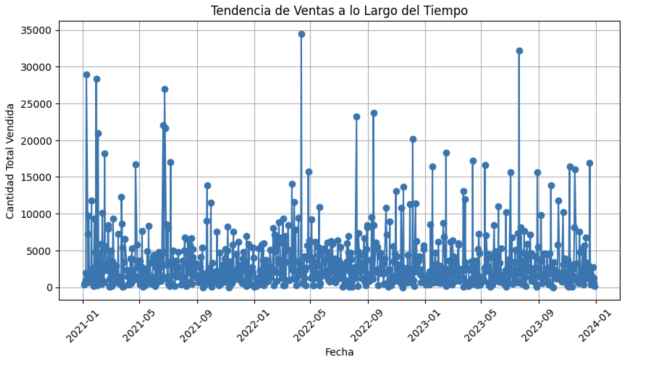

Fig. 15: Sales trend for products sold in liters. A more stable distribution is seen, with some peaks that could be related to campaigns or specific customers.

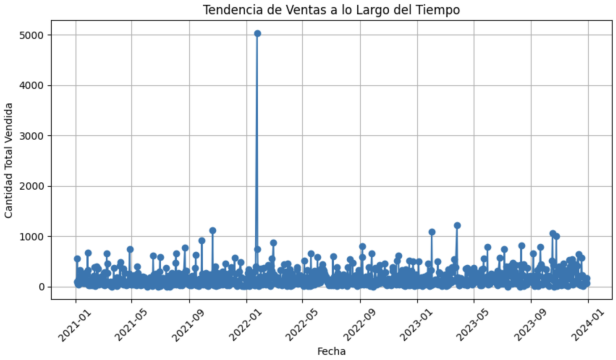

Fig. 16: Sales trend for products sold in pieces. Activity is generally low, with some significant occasional spikes.

Conclusions from Temporal Analysis

- Sales in kilograms show high variability, with certain periods of higher intensity that could be related to seasonality, promotions, or specific customer needs.
- The behavior varies significantly by unit of measure. Sales in liters are more stable, while in pieces, volumes are low and more sporadic.
- This type of analysis is essential as a basis for demand forecasting models or for logistical and commercial planning.

IV. Predictive Models

a. XGBoost

Model Correction and Adjustment

A predictive model was developed using the **XGBoost** algorithm with the aim of estimating future sales quantity per item. To improve its performance, a correction was applied by removing **outliers** using the interquartile range (IQR), which significantly reduced the data dispersion.

Each model was trained separately for the units of sale: kilograms (KG), liters (L), and pieces (PZA). Categorical variables were encoded using `LabelEncoder`, and a 70% training and 30% testing split was used.

Model Evaluation

The following table shows the metrics obtained for the training and testing sets:

TABLE 1: XGBOOST MODEL PERFORMANCE METRICS BY UNIT OF SALE

Unit	Set	MAE	RMSE	R^2
KG	Training	64.28	129.86	0.797
	Testing	106.85	194.34	0.528
L	Training	38.76	72.56	0.850
	Testing	89.24	143.09	0.405
PZA	Training	2.18	4.25	0.911
	Testing	5.87	9.83	0.506

In all cases, the performance on the training set is better than on the testing set, indicating some degree of **overfitting**. However, the results are reasonable considering the inherent variability of sales series and the simplicity of the model.

Future Sales Prediction

With the trained models, predictions were made for the next 30 days for the best-selling items in each unit. The results are shown in the following graphs:

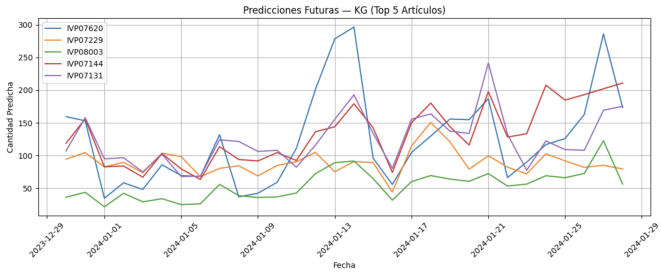

Fig. 17: Future sales prediction for the best-selling items in KG

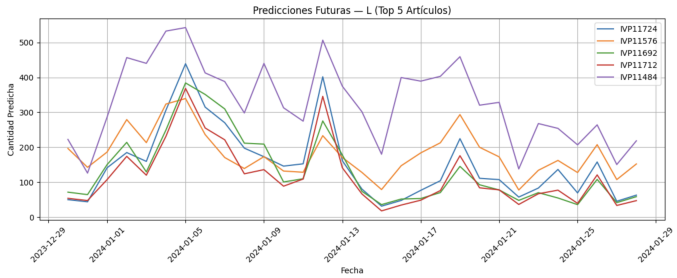

Fig. 18: Future sales prediction for the best-selling items in L

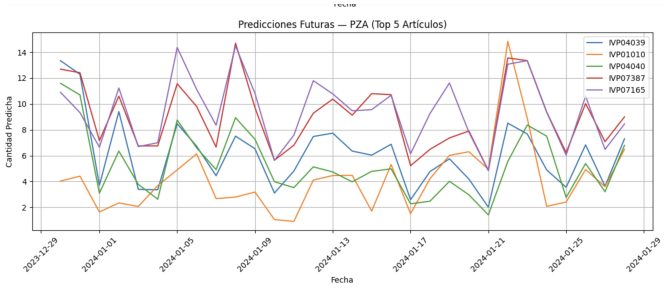

Fig. 19: Future sales prediction for the best-selling items in PZA

Proposal for Additional Variables to Improve the Model

To increase the model's accuracy, it is proposed to incorporate additional variables that better capture the context in which the sales occur. These include:

- **Historical trends:** Moving averages, sales from the same period last year, or seasonality indicators.
- **External factors:**
  - Presence of *promotions* or special offers
  - *Holidays* or key dates in the sector
  - *Supplier availability* or changes in raw materials
- **Economic conditions:** Inflation, exchange rates, or logistics costs that affect purchasing behavior.
- **Additional internal data:**
  - Level of *inventory available* on each date
  - *Credit terms* or commercial conditions by client

These variables would enrich the dataset, reduce the model's error, and improve its ability to predict future scenarios in a more robust manner.

Cross-Validation (K-Fold)

To evaluate the model's stability and reduce the risk of a fortunate or unfortunate data split, 5-fold cross-validation was applied. This process is not intended to train a new model, but to verify if the current hyperparameter configuration is reliable and generalizes well.

The results of cross-validation are shown in the following table:

TABLE 2: CROSS-VALIDATION RESULTS (5 FOLDS) BY UNIT OF SALE

Unit	Average RMSE	RMSE per Fold
KG	1.00	[0.98, 0.97, 1.01, 1.00, 1.01]
L	1.02	[0.99, 1.00, 1.05, 1.04, 1.01]
PZA	0.73	[0.72, 0.72, 0.72, 0.76, 0.71]

Cross-validation allowed us to verify that:

- There was no significant overfitting.
- The results were consistent across different combinations of training and test data.
- It was not necessary to adjust the current hyperparameters, as stable and satisfactory performance was achieved.

This approach provided objective evidence to maintain the current model configuration without further changes, which increases confidence in its ability to generalize.

Conclusion: The hyperparameters were evaluated using 5-fold cross-validation. A stable average RMSE was observed, with no significant variance between folds, leading to the decision to maintain this configuration for the final model.

V. Conclusions

This analysis led to the development of prediction models for each unit of sale (KG, L, and PZA) with satisfactory and stable results. Through the exploratory data analysis (EDA), systematic errors in the recorded quantities were identified, highlighting the importance of preprocessing and cleaning as a prior step to modeling.

After applying outlier removal, categorical variable encoding, and a logarithmic transformation of the target variable, models were trained using the *XGBoost* algorithm, achieving good levels of accuracy. The cross-validation (K-Fold) showed low variance between the folds, confirming that the models generalize adequately without overfitting.

The trained model by unit allowed future demand predictions for the next 30 days, highlighting items with high sales volume and identifying relevant consumption patterns. These predictions have the potential to support strategic decisions related to inventory, production, and customer service.

In summary, the developed model represents a solid first step toward a scalable demand prediction solution at Interlub. Future improvements may include incorporating external variables (seasonal, macroeconomic, or marketing campaigns), as well as deeper analysis of customer or product family behavior.

A. Appendix

Documentation of libraries used in the code development: [1] [2] [3] [4]

References

[1] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Duchesnay, *Scikit-learn: Machine Learning in Python*, 2011. [Online]. Available: https://scikit-learn.org/

[2] T. Chen and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, 2016. [Online]. Available: https://xgboost.readthedocs.io/

[3] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, *Array programming with NumPy*, 2020. [Online]. Available: https://numpy.org/

[4] J. D. Hunter, *Matplotlib: A 2D Graphics Environment*, 2007. [Online]. Available: https://matplotlib.org/