Improving Basket Prediction: Enhancing F1 Score with Advanced Feature Engineering and Modeling Techniques

Improving Basket Prediction: Enhancing F1 Score with Advanced Feature Engineering and Modeling Techniques....1

Problem Statement:....1

Objectives:
After extensive communication and refinement of the initial vague problem statement, we defined the following specific objectives for this project:....1

Constraints:
This project was conducted with the following limitations:....2

Data Overview:....2

Data Preprocessing and Exploration:....2

Modeling:-....6

Final Model Performance and Future Scope:....7

Problem Statement:

I was tasked with improving the performance of Instacart's current basket prediction algorithm. The initial goal provided was general—"enhance the accuracy of predicting users' next baskets".

Objectives:

After extensive communication and refinement of the initial vague problem statement, we defined the following specific objectives for this project:

1. Improve the current algorithm's F1 score, aiming to exceed 0.25 by at least 0.03.
2. Identify and analyze patterns in customer purchasing behavior that can provide insights for more informed decision-making.

Data Overview:

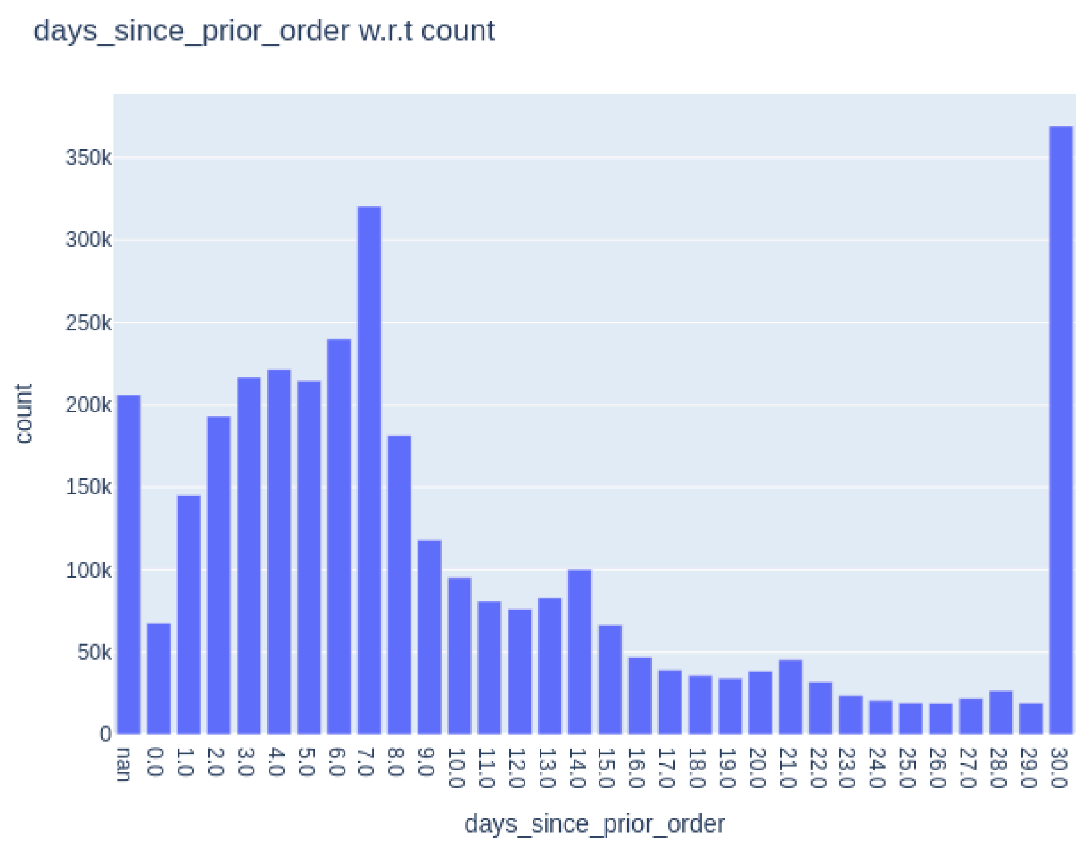
The dataset contains approximately 30 million product order records from over 3 million orders made by around 200,000 unique customers. The data is organized across six CSV files:

1. **aisle.csv**: Information about product aisles.
2. **departments.csv**: Information about product departments.
3. **products.csv**: Detailed product information.
4. **orders.csv**: Contains order-specific details.
5. **prior_product_orders.csv**: Historical records of prior product orders.
6. **train_product_orders.csv**: Product orders used for model training.

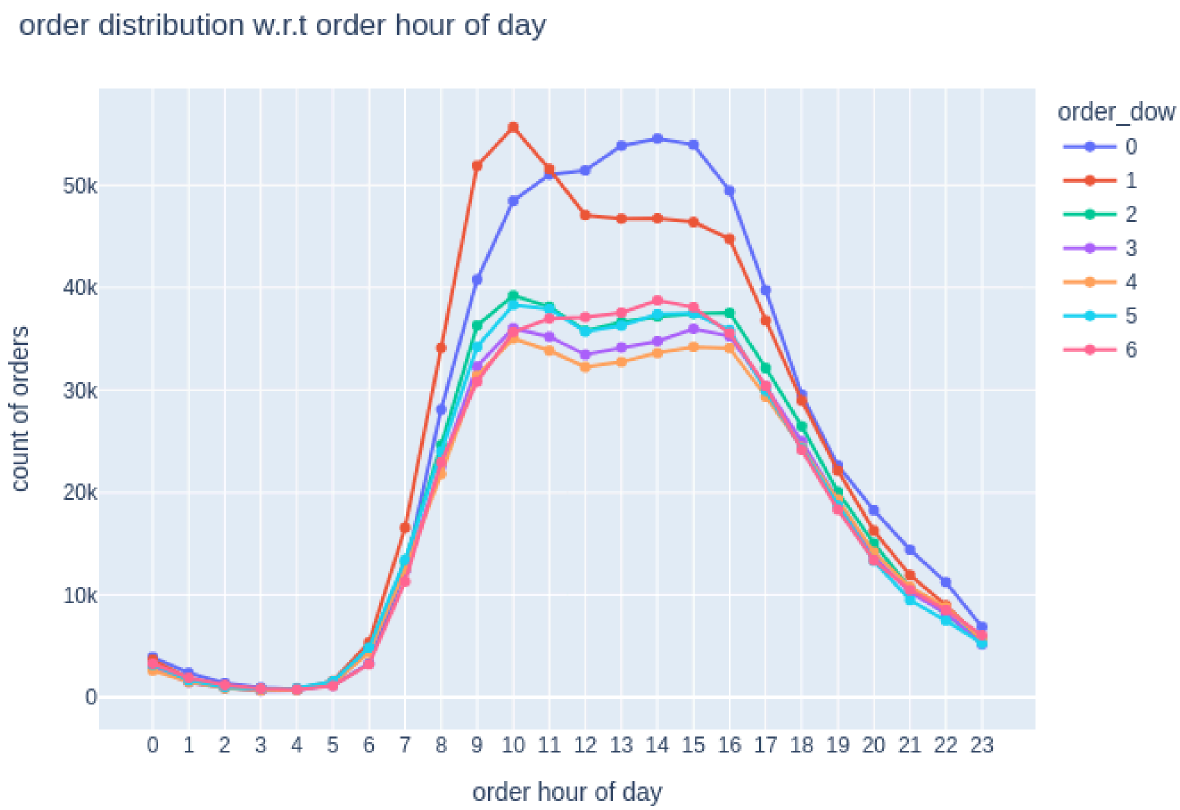
Data Preprocessing and Exploration:

The dataset provided was preprocessed and balanced, with no missing values, as it had been cleaned by a previous team during the creation of the original algorithm. Key findings from data exploration include:

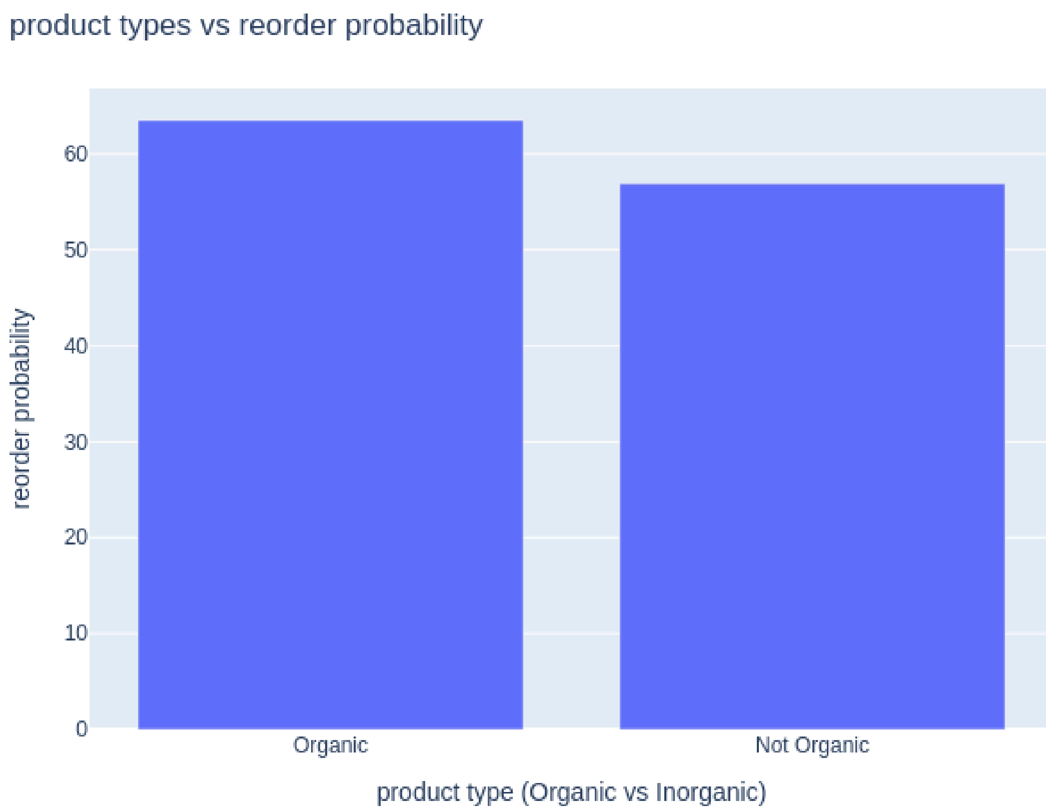
1. **Reordering Patterns**: Users tend to reorder on the same day, the 7th day, or the 30th day after a previous order.



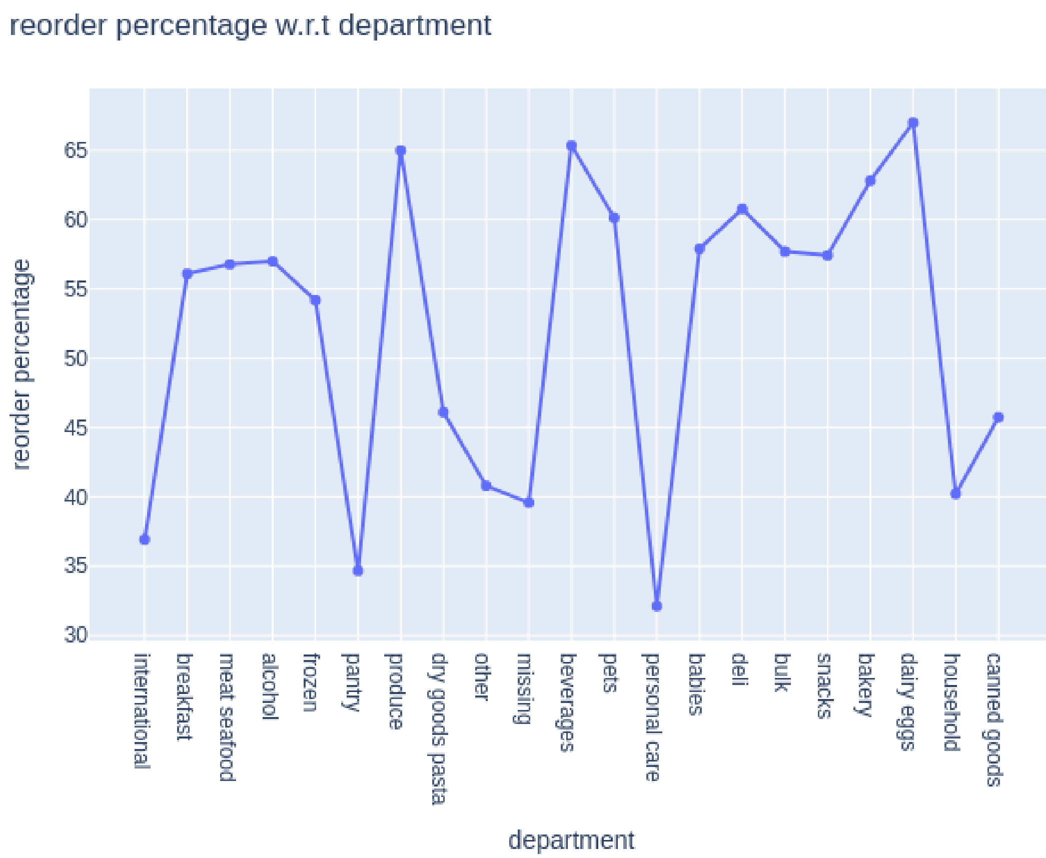
2. **Order Timing**: A significant volume of orders is placed between 8:00 AM and 4:00 PM, indicating peak purchasing hours.



3. **Product Type Preference**: Organic products are reordered 8% more frequently than non-organic items.



4. **Department Reorder Rates**: Categories like Dairy & Eggs, Produce, Beverages, and Bakery have high reorder rates, exceeding 65%. Conversely, Personal Care and Pantry items have lower reorder rates, below 35%.



Modeling:-

- **Data Processing**: Started with **Polars** for faster data manipulation; switched to **PySpark** for distributed computing as dataset size grew.
- **Feature Engineering**: Focused on **user**, **product**, **user-product relations**, and **time-based features**; most features performed well.
- **Validation**: Created a **time-based validation set** using the latest orders to account for the problem's temporal nature.
- **Modeling**: Used **XGBoost**, **H2O**, and **LightGBM** with **distributed computing** and **GPU acceleration** to handle large data and speed up training.

Final Model Performance and Future Scope:

- **Final Score**: The model achieved an F1 score of **0.30**, surpassing the project's success threshold of **0.27**.
- **Future Improvements**:
  - Further **feature engineering** could enhance model performance.
  - Exploring advanced architectures like **LSTMs**, **GRUs**, and **Transformers** may provide additional improvements, especially in capturing sequential patterns.