



DATA ANALYSIS & PREDICTIVE MODELLING FOR **INSTACART**

TEAM 9

Dylan Dsouza

Chaitanya Bhargava

Meghna Gupta

Sanjay DV

Ameya Shah



01

ABOUT INSTACART

ABOUT THEM

Instacart is a leading online grocery delivery and pick-up service that connects customers with personal shoppers to deliver fresh groceries and everyday essentials from local stores



WHY DID WE CHOOSE THIS DATASET



DIVERSE DATA

Granular product and order information



MARKET BASKET ANALYSIS

Identify frequently bought combinations



REAL-WORLD RELEVANCE

Simulates real grocery transactions



CUSTOMER BEHAVIOR INSIGHTS

Analyze reorder patterns and habits



TIME-BASED TRENDS

Understand peak shopping times



PREDICTIVE MODELING SCOPE

Build recommendation and inventory models



02

PROJECT OVERVIEW

Objective: To analyze and predict trends in Instacart data.

Datasets: Orders, Products, Departments, Aisles, and other related datasets.

Methods: Data visualization, exploratory analysis, and predictive modeling.

DATA OVERVIEW

Aisles Dataset: Contains aisles and their IDs (e.g., aisle_id, aisle).

Departments Dataset: Contains department names and IDs (e.g., department_id, department).

Orders Dataset: Contains order-related details like order_id, order_hour_of_day, order_dow, etc.

Products Dataset: Contains product details (e.g., product_id, product_name).

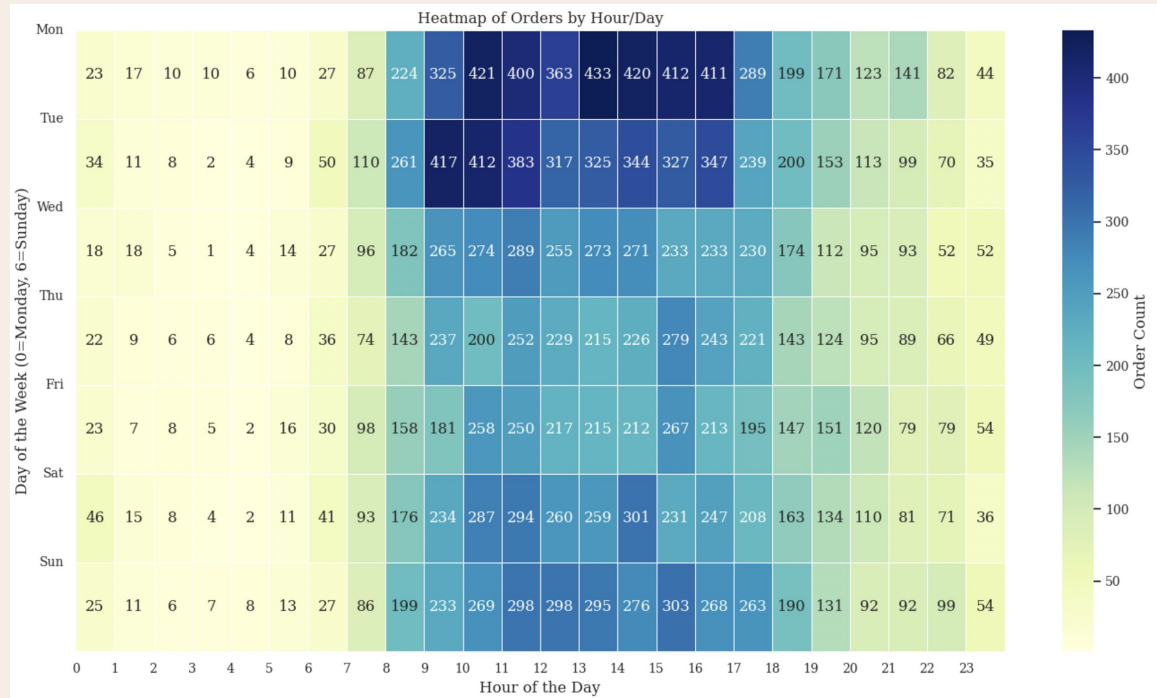


EXPLORATORY DATA ANALYSIS

HEAT MAP

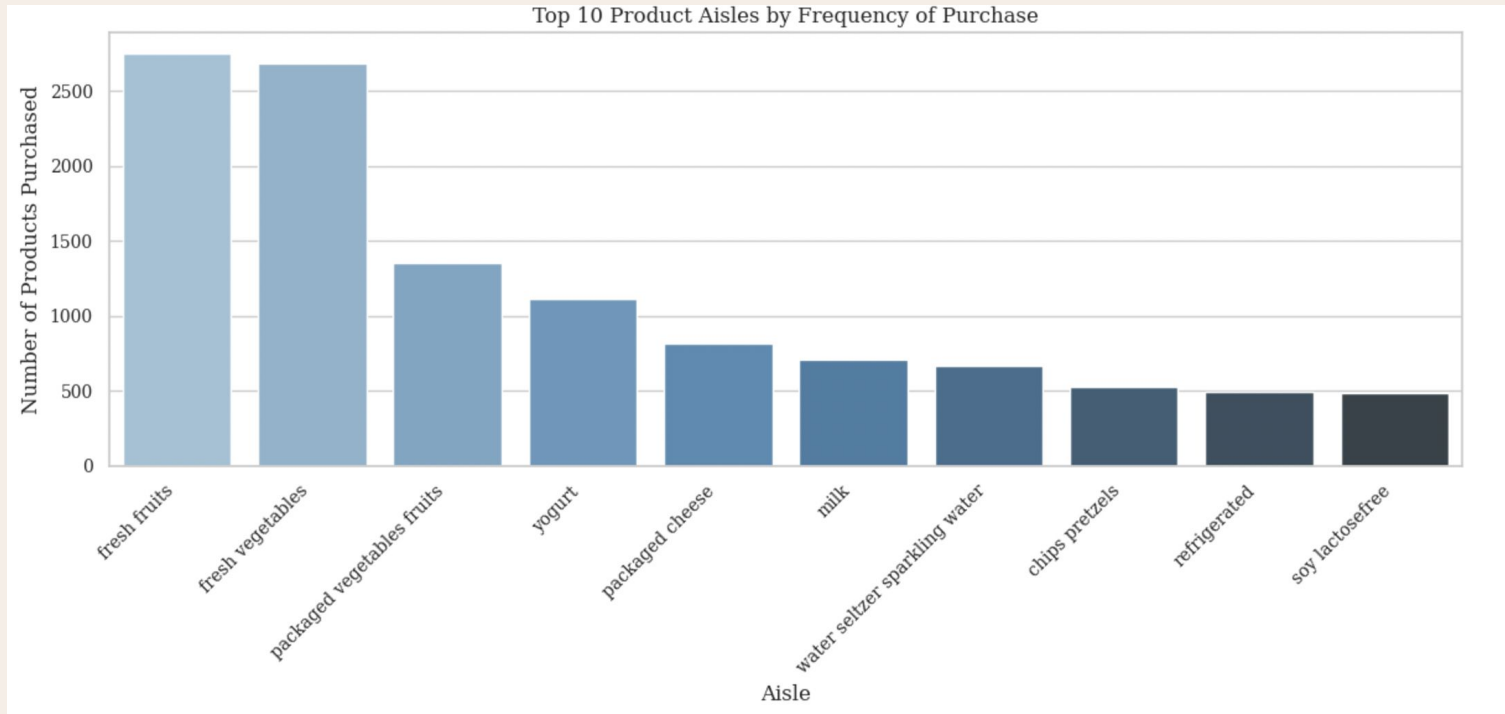
Orders by day and hour.

A heatmap showing order frequency with order_dow (day of the week) on the y-axis and order_hour_of_day on the x-axis.



EXPLORATORY DATA ANALYSIS

BAR CHART Top 10 aisles by frequency of purchase.
A bar chart displaying the top aisles and their order counts.

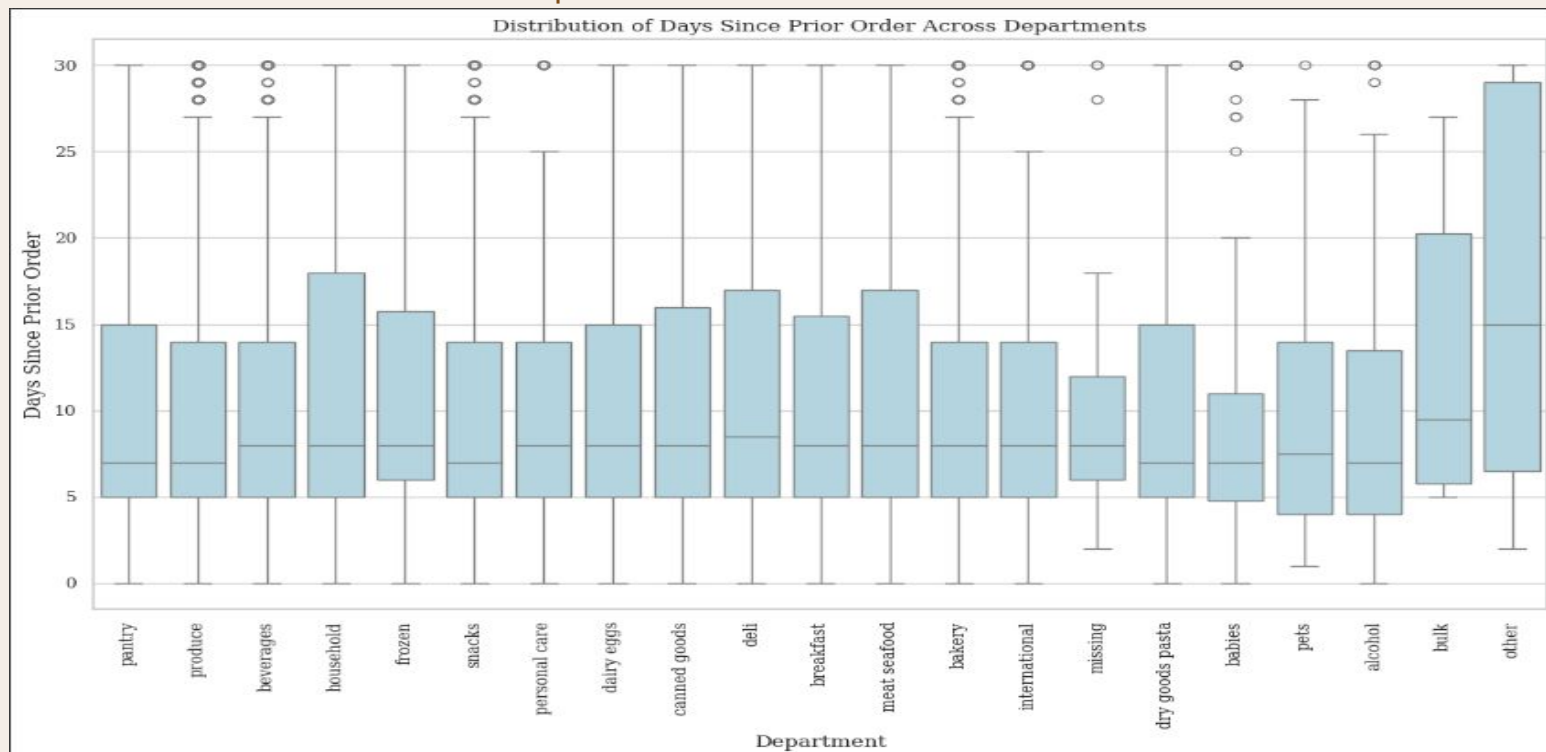


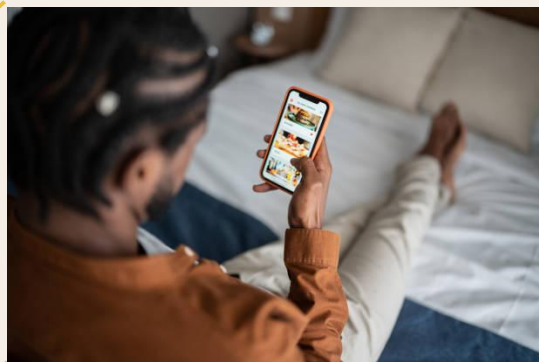
EXPLORATORY DATA ANALYSIS

BOX PLOT

Days since prior order by department.

A box plot comparing the distribution of days_since_prior_order across departments.





03

KEY

INSIGHTS

Ordering Patterns: Peak ordering hours observed during the late morning and early afternoon and most orders occur on weekends.

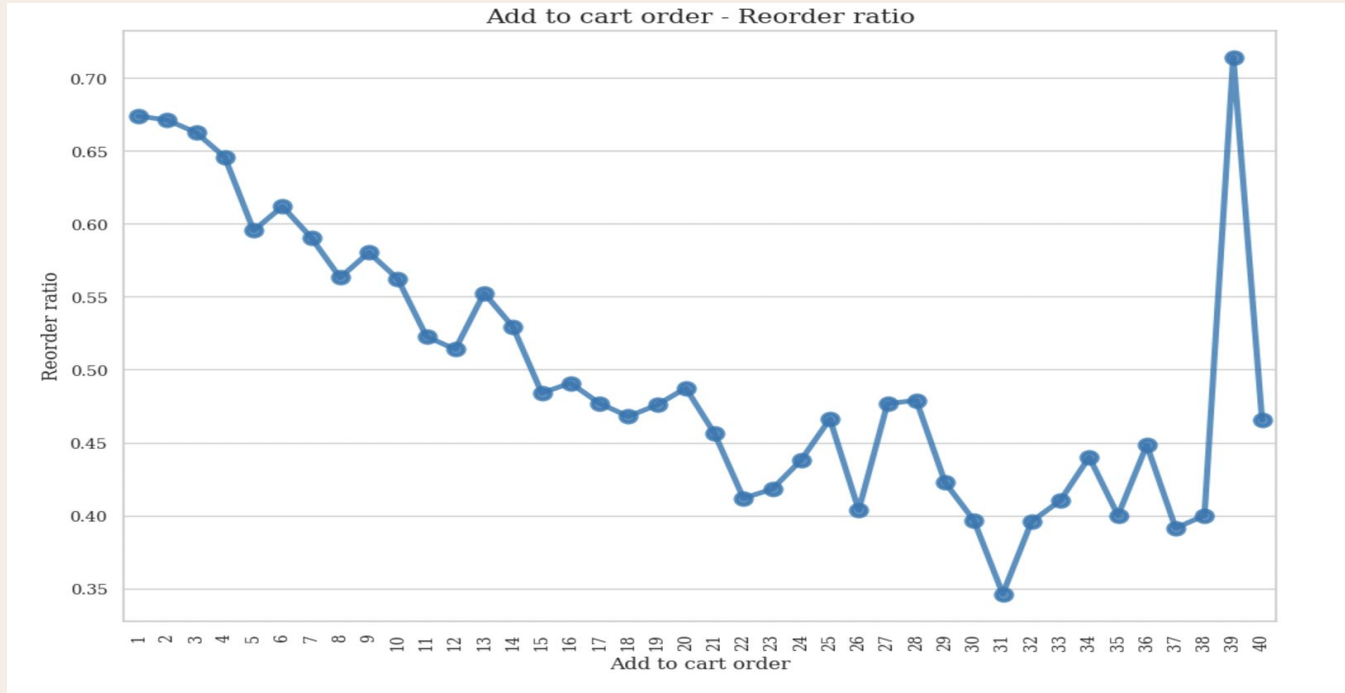
Popular Categories: Certain aisles, such as beverages and fresh produce, dominate the order frequency.

Reordering Trends: Departments like dairy and snacks show higher reorder ratios compared to others.

Customer Behavior: Customers tend to reorder frequently within a week for daily essentials.

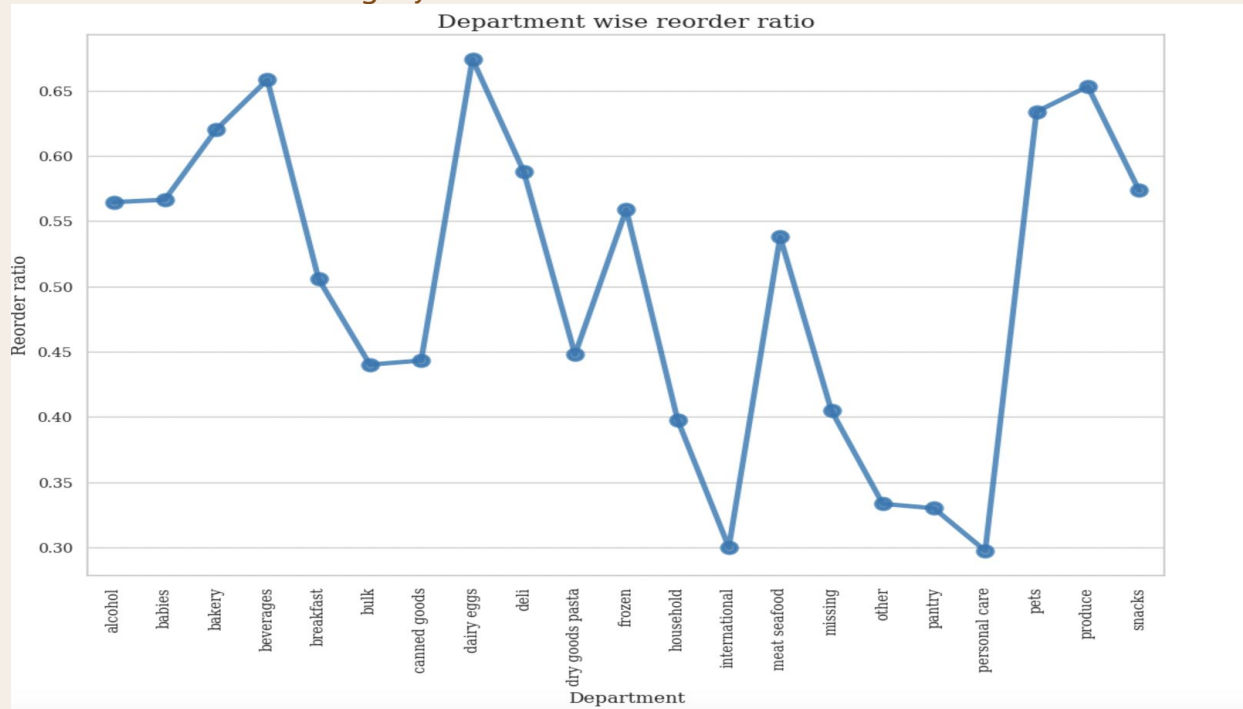
REORDERING ANALYSIS

POINT PLOT Analyzes how the position of a product in the cart affects its reorder likelihood. Identifies customer priorities during the shopping process



REORDERING ANALYSIS

POINT PLOT Highlights the reorder ratios across different departments. Shows the likelihood of products being reordered based on their category





LOGISTIC REGRESSION ANALYSIS



Logistic regression was used to predict the likelihood of a product being reordered.

Model Performance Metrics:



- Accuracy: 70.27% - The model correctly predicts reorders 70% of the time.
- Precision: 72.62% - Of all predicted reorders, 72.62% were actual reorders.
- Recall: 79.98% - Captures nearly 80% of all reorders (important for sensitivity).
- ROC-AUC: 75.36%

Actionable Insights:

Focus on high-reorder items like snacks, milk, and bakery goods for inventory optimization.

Tailor marketing strategies to promote low-reorder items through discounts or bundles.

Use predictive features to enhance personalized recommendations.



LOGISTIC REGRESSION ANALYSIS

TOP FEATURES INFLUENCING REORDERS

Positive Coefficients (Increase Reorder Likelihood):

Aisle_Trail Mix Snack Mix (+1.09)- Frequently reordered snacks.

Aisle_Breakfast Bakery (+0.86)- Reflects perishability and high reorder demand.

Aisle_Milk (+0.83)- Staple item with frequent reorders.

Negative Coefficients (Decrease Reorder Likelihood):

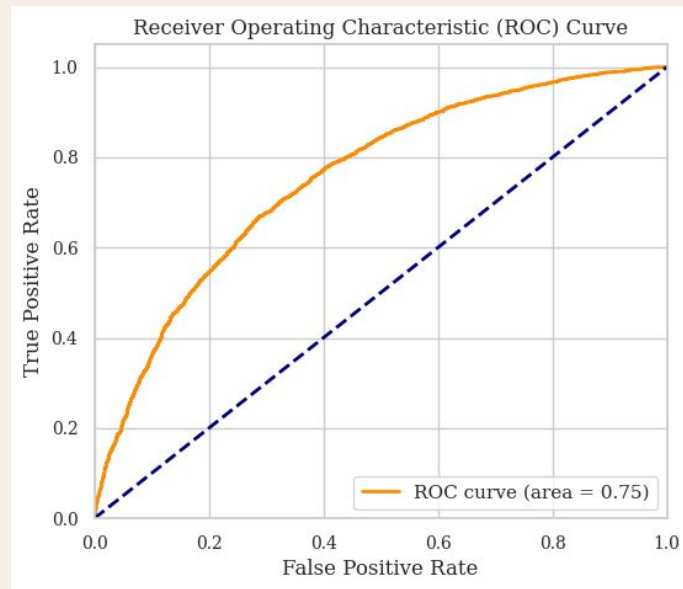
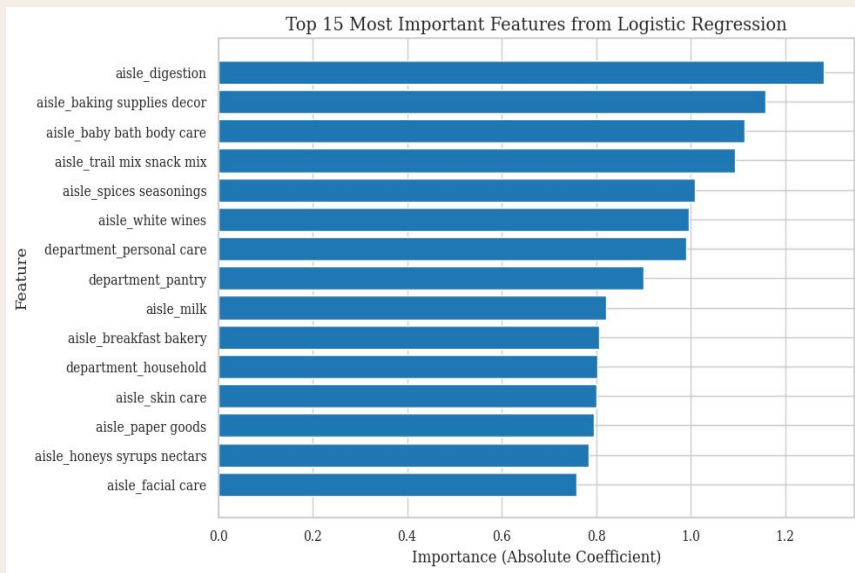
Aisle_Digestion Aids (-1.28)- Niche products with less reorder demand.

Aisle_Baking Supplies (-1.15)- Long shelf-life leads to fewer frequent reorders.

Top 15 Most Important Features:

	Feature	Coefficient	Importance
56	aisle_digestion	-1.282348	1.282348
75	aisle_baking supplies decor	-1.158857	1.158857
78	aisle_baby bath body care	-1.113326	1.113326
94	aisle_trail mix snack mix	1.094117	1.094117
80	aisle_spices seasonings	-1.009148	1.009148
48	aisle_white wines	0.995468	0.995468
103	department_personal care	-0.990009	0.990009
104	department_pantry	-0.900828	0.900828
64	aisle_milk	0.821108	0.821108
72	aisle_breakfast bakery	0.807311	0.807311
108	department_household	-0.802008	0.802008
84	aisle_skin care	0.801748	0.801748
42	aisle_paper goods	0.794713	0.794713
23	aisle_honeys syrups nectars	0.783707	0.783707
59	aisle_facial care	0.758341	0.758341

IMPORTANT FEATURES & ROC CURVE





04

PREDICTIVE ANALYSIS

Decision Tree Classifier: Simple and interpretable.

Ensemble Models: Random Forest and Gradient Boosting for better performance

DECISION TREE RESULTS

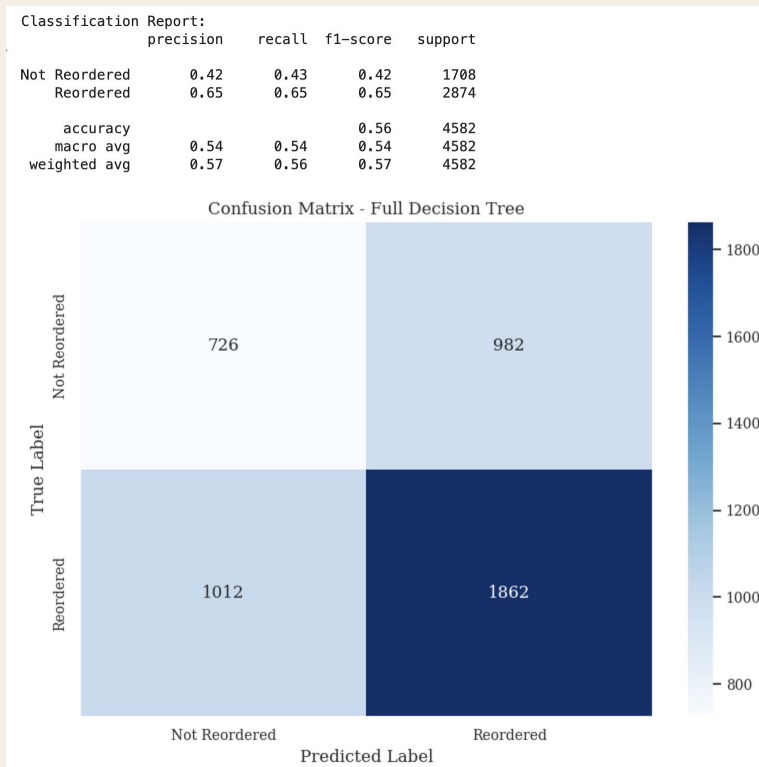
CONFUSION MATRIX(FULL)

Performance Metrics:

1. Accuracy- 56.48%
2. Precision- 65% (for reordered products).
3. Recall- 65% (for reordered products).
4. F1 Score- 0.65 (balanced metric).

Confusion Matrix:

1. True Positives (Reordered)- 1862.
2. False Positives (Not reordered but predicted reordered)- 982.
3. False Negatives- 1012.



DECISION TREE RESULTS

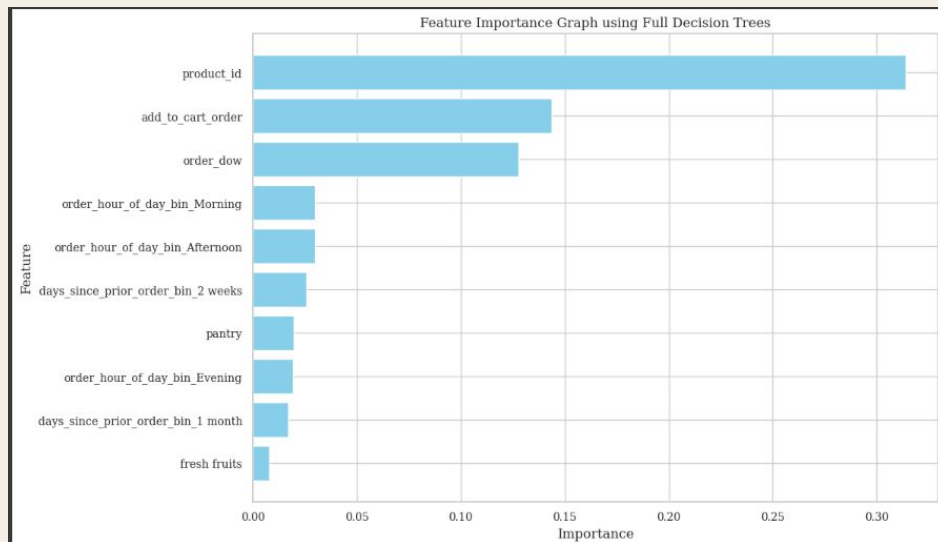
FEATURE IMPORTANCE(FULL)

Top Features:

1. Product_ID - Key identifier influencing reorders.
2. Add_to_Cart_Order - Higher priority items added early are reordered more.
3. Order_DOW - Day of the week affects purchasing trends.
4. Order_Hour_of_Day_Bin - Morning and evening orders are critical.

Insights:

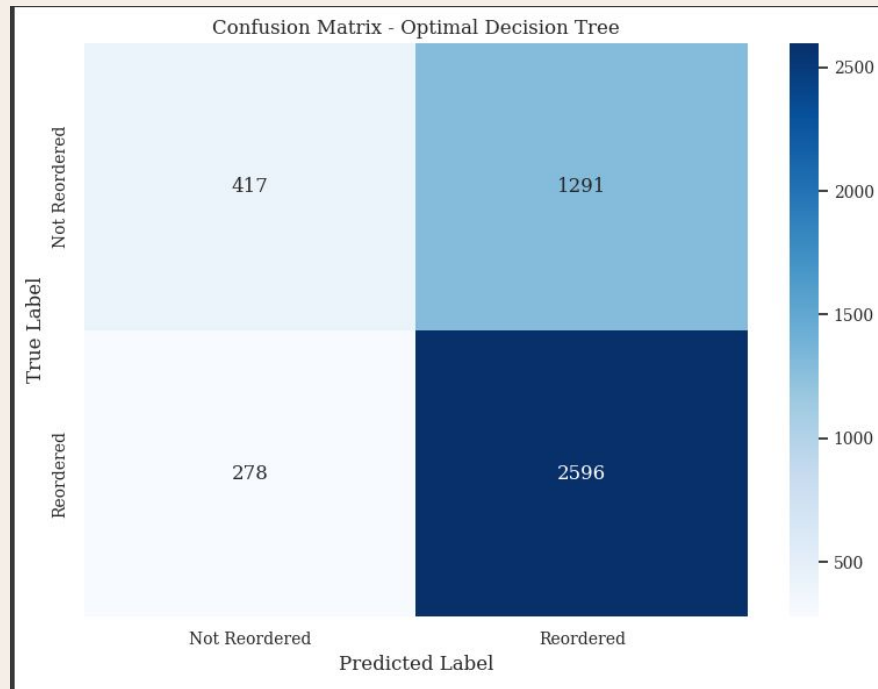
1. Customer behavior patterns are reflected in reorder habits.
2. Time and order position are key influencers.



DECISION TREE RESULTS

CONFUSION MATRIX(OPTIMAL)

1. Performance Metrics:
 - a. Accuracy: 65.76%
 - b. Precision (Reordered Products): 0.67
 - c. Recall (Reordered Products): 0.90
 - d. F1 Score: 0.77
2. Confusion Matrix:
 - a. True Positives (Reordered): 2,874
 - b. False Positives (Incorrectly Predicted Reordered): 1,708
 - c. False Negatives (Incorrectly Predicted Not Reordered): 1,708



DECISION TREE RESULTS

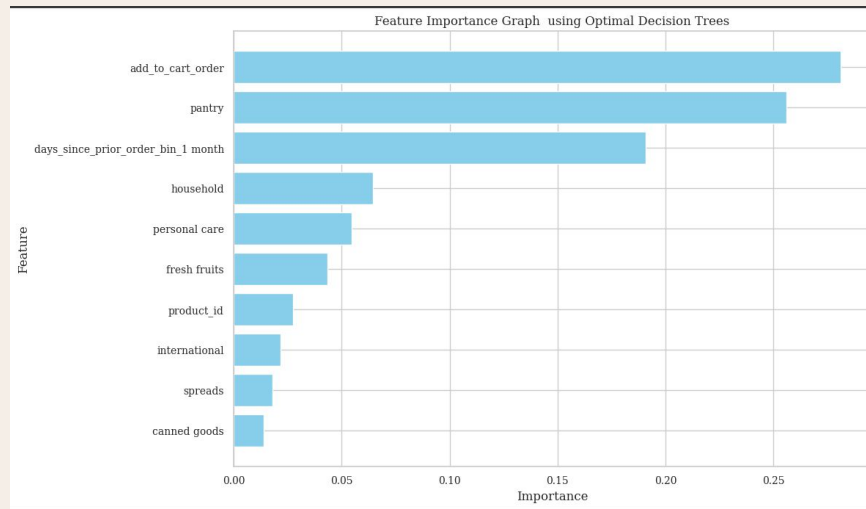
FEATURE IMPORTANCE(OPTIMAL)

Key Features:

1. Add_to_Cart_Order- Retains its high influence as in the first chart.
2. Pantry- Suggests department-specific behavior impacting reorder rates.
3. Days_Since_Prior_Order_Bin- Gaps in ordering behavior influence reorder likelihood.
4. Personal Care and Fresh Fruits- Essential categories for prediction.
5. The optimized tree places greater emphasis on department-level insights (e.g., Pantry, Personal Care).

Insights:

1. Features like Add_to_Cart_Order and Order_Hour_Bin reflect customer urgency and routine behaviors.
2. Department and category-level factors play a critical role in reorder patterns for essential products.



ENSEMBLE MODEL COMPARISONS

RANDOM FOREST & GRADIENT BOOSTING MODELS

Bagging: combines multiple versions of a predicted model to reduce variance and improve the accuracy of classification

Random Forest: Uses multiple decision trees for improved stability.
Handles overfitting better than a single tree.

Boosting: Sequentially builds trees to minimize error.
Provides the best predictive performance among models.

Performance Comparison: Boosting has the lowest RMSE: 0.4543.

Regression statistics

Mean Error (ME) : -0.0009
Root Mean Squared Error (RMSE) : 0.4767
Mean Absolute Error (MAE) : 0.4065

Regression statistics

Mean Error (ME) : -0.0008
Root Mean Squared Error (RMSE) : 0.4772
Mean Absolute Error (MAE) : 0.4069

Regression statistics

Mean Error (ME) : 0.0014
Root Mean Squared Error (RMSE) : 0.4543
Mean Absolute Error (MAE) : 0.4211

GRADIENT BOOSTING MODEL

FEATURE IMPORTANCE & KEY DRIVERS

Top Features (SHAP Analysis):

- Add_to_Cart_Order - Most influential in predicting reorders.
- Days_Since_Prior_Order_Bin_1_Month - Long gaps reduce reorder likelihood.
- Department_Pantry and Department_Dairy Eggs - Critical for customer essentials.

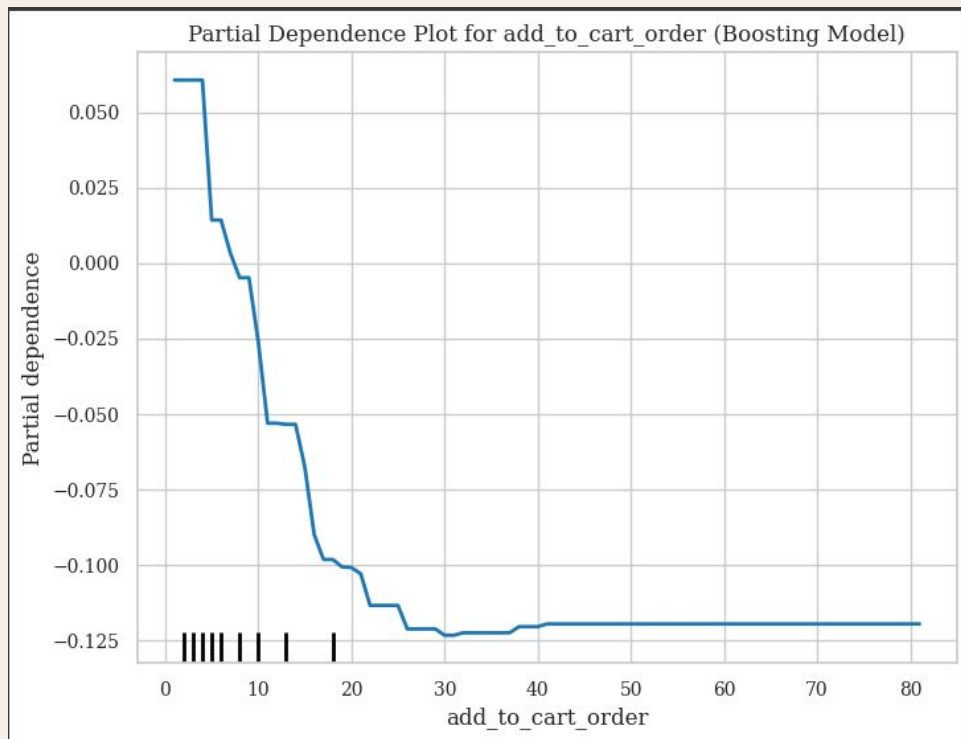
Insights:

SHAP analysis highlights actionable insights into customer preferences.



GRADIENT BOOSTING MODEL

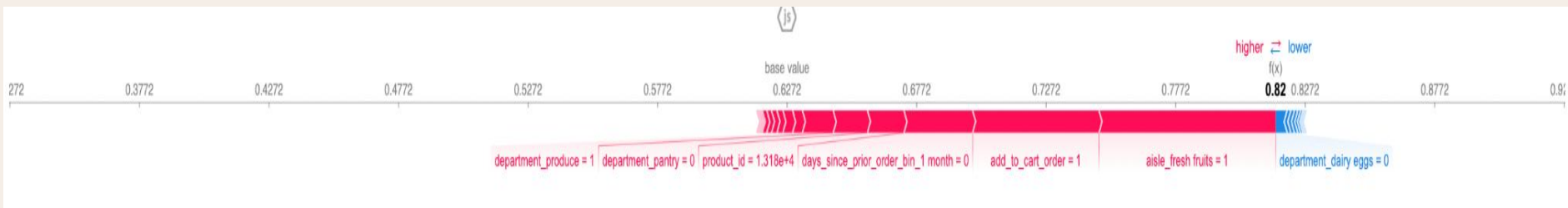
PARTIAL DEPENDENCY PLOT



GRADIENT BOOSTING MODEL

ACTIONABLE INSIGHTS



- **Inventory Management:** Focus on high-reorder products like milk, snacks, and fresh produce.
- **Marketing Strategies:** Personalized recommendations for less-reordered products. Bundle offers for frequently purchased categories.
- **Customer Retention:** Use insights from reordering trends to drive loyalty programs.





RECOMMENDATIONS



- Instacart could let grocery stores know about frequently ordered items and the grocery store could rearrange those items
 - Products that show higher reorder rates can be targeted for promotions, discounts to increase customer loyalty.
 - If certain items show high reorder rates, Instacart could consider expanding partnerships with the brands to offer exclusive products or promotions.
- 
- 

THANKS!

TEAM 9

Dylan Dsouza

Chaitanya Bhargava

Meghna Gupta

Sanjay DV

Ameya Shah

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics and images by **Freepik**

