Analysis of Instacart Orders

Andrew Bingen,

Brief Problem/Data Description - Onyinye

Rationale (possibly including visualization) for choosing specific techniques or modeling approaches

- Associated Rules – Girisha
- Andrew – Random Forests
- Brandon/Victor – Logistic Regression

Association Rules Rationale

- **Goal of Technique:** Identify co-occurring patterns (e.g., time of day + department) in purchasing behavior that influence product reordering.
- **Strategy:**
  - Numeric variables were binned into categories to meet association rule requirements. For example, there were 4 categories for hours of the day (12am to 5am, 6am to 11am, 12pm to 5pm, and 6pm to 11pm)
  - Measured outcomes for when Reordered=Yes and Reordered=No

Association Rules Analysis

Top Predictors of Reordering:

- Products in the Produce and Dairy & Eggs departments
- Reordered frequently within 5–9 days
- Typically bought on weekdays and between 6 AM–5 PM
- Reordering is more likely when users have moderate order history (26to50 orders)

Top Predictors of Not Reordering:

- Pantry, Household, and Personal Care items
- Purchased by users with low order history (0to25 orders)
- Purchased after 25+ days since prior order
- Often purchased during midday on weekends

Top 15 Rules for Reordered = Yes

Rule #	Attribute 1	Attribute 2 (if applicable)	Support	Confidence	Coverage	Lift	Count
1	WeekendWeekday=Weekday	NumberOfUserOrders=26to50	0.064432	0.771052	0.083563	1.288104	89213
2	NumberOfUserOrders=26to50		0.110072	0.768936	0.143148	1.28457	152407
3	department=produce	DaysSinceOrder=5to9	0.057673	0.754673	0.076421	1.260742	79855
4	aisle=fresh fruits		0.080009	0.736225	0.108675	1.229923	110782
5	aisle=fresh fruits	department=produce	0.080009	0.736225	0.108675	1.229923	110782
6	DaysSinceOrder=5to9	HourOfDay=6amto11am	0.058477	0.704597	0.082993	1.177085	80968
7	aisle=fresh fruits	NumberOfUserOrders=0to25	0.057841	0.690798	0.08373	1.154034	80087
8	aisle=fresh fruits	department=produce	0.057841	0.690798	0.08373	1.154034	80087
9	DaysSinceOrder=5to9	WeekendWeekday=Weekend	0.072109	0.6844	0.105361	1.143345	99843
10	DaysSinceOrder=5to9		0.166677	0.682214	0.244318	1.139693	230784
11	department=produce	HourOfDay=6amto11am	0.06342	0.681945	0.092998	1.139244	87812
12	DaysSinceOrder=5to9	WeekendWeekday=Weekday	0.094568	0.680556	0.138958	1.136923	130941
13	department=produce	DayOfWeek=Sunday	0.050819	0.676034	0.075172	1.129369	70365
14	department=produce	DayOfWeek=Sunday	0.050819	0.676034	0.075172	1.129369	70365
15	department=dairy eggs		0.105807	0.674966	0.156759	1.127585	146502

Top 15 Rules for Reordered = No

1	department=pantry	DaysSinceOrder=25to30	0.014941	0.711515	0.020999	1.772558	20688
2	department=pantry	DaysSinceOrder=25to30	0.015404	0.700828	0.02198	1.745935	21329
3	department=pantry	WeekendWeekday=Weekend	0.01259	0.695345	0.018107	1.732276	17433
4	department=pantry	HourOfDay=Noonto5pm	0.016542	0.689827	0.023981	1.718527	22905
5	department=pantry	WeekendWeekday=Weekday	0.010016	0.683539	0.014654	1.702863	13869
6	department=pantry	NumberOfUserOrders=0to25	0.032894	0.683264	0.048142	1.702178	45545
7	department=pantry	WeekendWeekday=Weekday	0.020303	0.675981	0.030035	1.684034	28112
8	department=personal care		0.010327	0.662912	0.015578	1.651475	14299
9	department=pantry	HourOfDay=Noonto5pm	0.018676	0.648095	0.028817	1.614565	25859
10	department=pantry	WeekendWeekday=Weekend	0.01453	0.645946	0.022494	1.609211	20118
11	department=pantry	WeekendWeekday=Weekday	0.011232	0.642485	0.017482	1.600589	15552
12	department=pantry		0.037371	0.636912	0.058675	1.586704	51744
13	department=pantry	WeekendWeekday=Weekday	0.022841	0.631295	0.036181	1.572712	31626
14	department=pantry	HourOfDay=6amto11am	0.011176	0.621112	0.017994	1.547342	15475
15	department=household	NumberOfUserOrders=0to25	0.013327	0.613423	0.021726	1.528188	18453

Association Rules

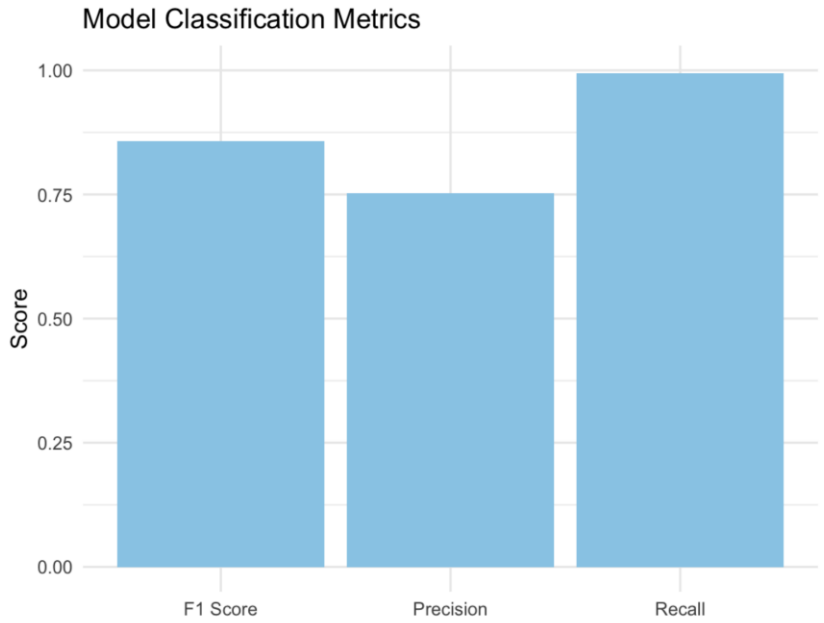
- Discussion of Lift, Confidence, and Support

Random Forests Rationale

- Handles categorical data and nonlinear relationships
- Captures interactions automatically
- Provides feature importance for interpretability
- Well-suited for classification tasks like predicting reorders

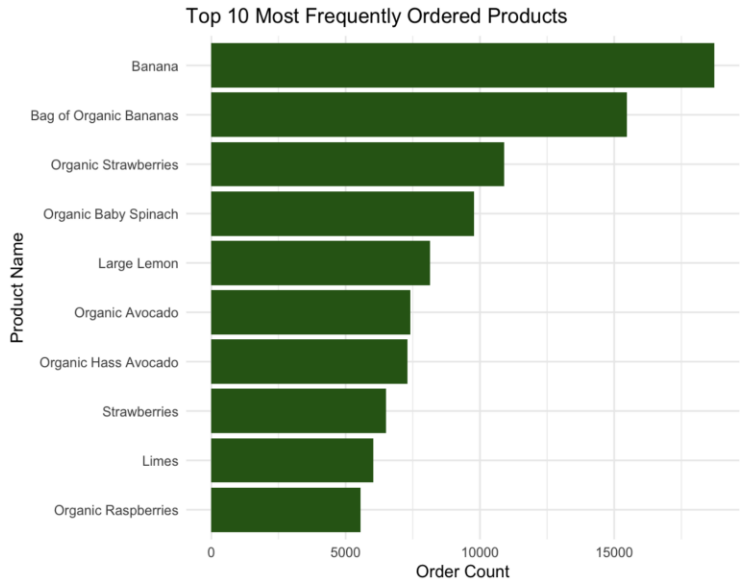
Random Forests Analysis and Results Summary

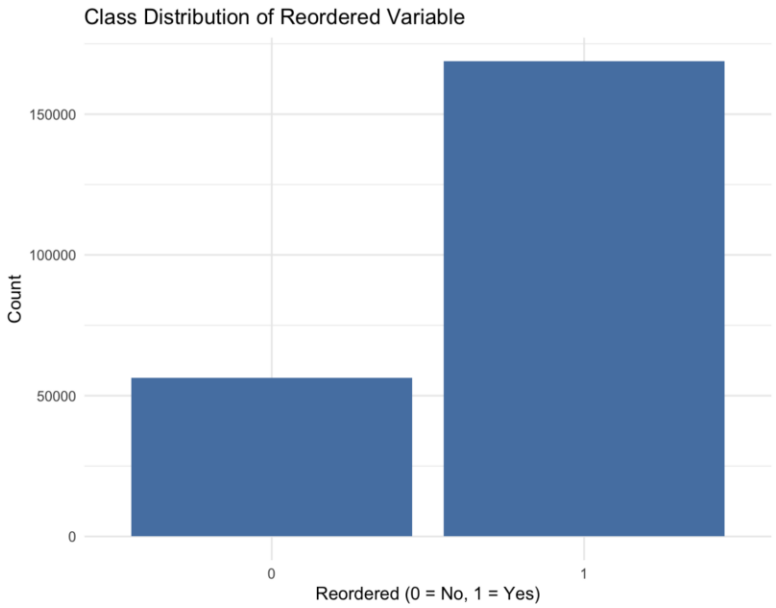
- **Model Setup**
  - Filtered for top 50 most frequently ordered products
  - Tuned Random Forest with 200 trees, mtry = 3
- **Performance Metrics**
  - Accuracy (Test Set): **74.96%**
  - AUC Score: **0.5529**
  - Threshold-adjusted Accuracy (0.4): **75.02%**
  - **Precision:** 75.23%
  - **Recall:** 99.49%
  - **F1 Score:** 85.67%
- **Key Insights**
  - Top predictive features: product_name, add_to_cart_order
  - High recall suggests strong sensitivity to reordered items
  - Balanced F1 score reflects solid classification performance despite imbalance



Random Forests Visual Insights

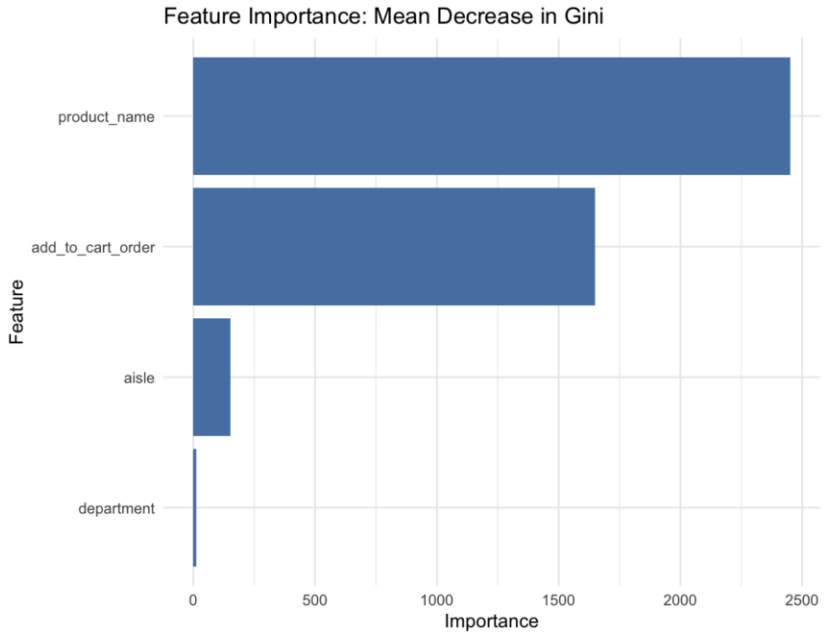
- Reordered class was significantly more frequent, creating class imbalance that influenced model sensitivity and led us to adjust the classification threshold
- Top ordered items — such as bananas and organic produce — reflect high-frequency, perishable products that justified narrowing our focus to the top 50
- Helped justify product-level filtering and feature choices





Random Forests Feature Importance

- product_name is the most influential predictor
- Cart position (add_to_cart_order) is also highly important
- Aisle and department were less impactful
- Suggests product-specific behavior drives reordering



Model Comparison

2. Exploration of different techniques

While I've tried to expose you to a wide range of analytical techniques, we obviously can't cover everything. You might choose one or a few techniques (Naïve Bayes classification, neural networks, etc.) and provide a brief lesson on the basics of those approaches. The presentation might include: identify the technique(s) you're covering, a brief description of each technique and how it works, an illustration of each technique applied to a data set (preferably the same data set across techniques, if that's appropriate), and comments on the strengths/weaknesses and similarities/differences among the techniques.

Summary of your analysis and results

- Associated Rules – Girisha
- Andrew – Random Forests
- Brandon/Victor – Logistic Regression

Any additional comments (additional data you wish you had, questions you
had about the data but didn’t have answers to, etc.)