Kaggle Challenge Instacart

Anwar Habeeb

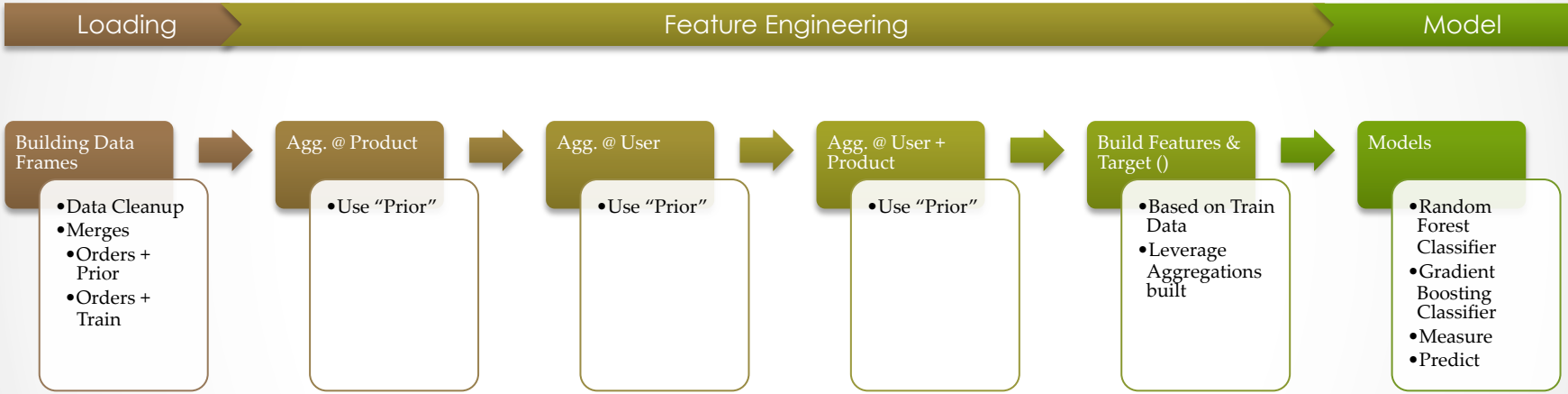
Problem

- Instacart, a grocery ordering and delivery app, aims to make it easy to fill your refrigerator and pantry with your personal favorites and staples when you need them. After selecting products through the Instacart app, personal shoppers review your order and do the in-store shopping and delivery for you.

- Use transactional data to develop models that predict which products a user will buy again, try for the first time, or add to their cart next during a session.
  - **3 Million** Instacart Orders, Open Sourced.

- Using this anonymized data on customer orders **to predict which previously purchased products will be in a user's next order.**

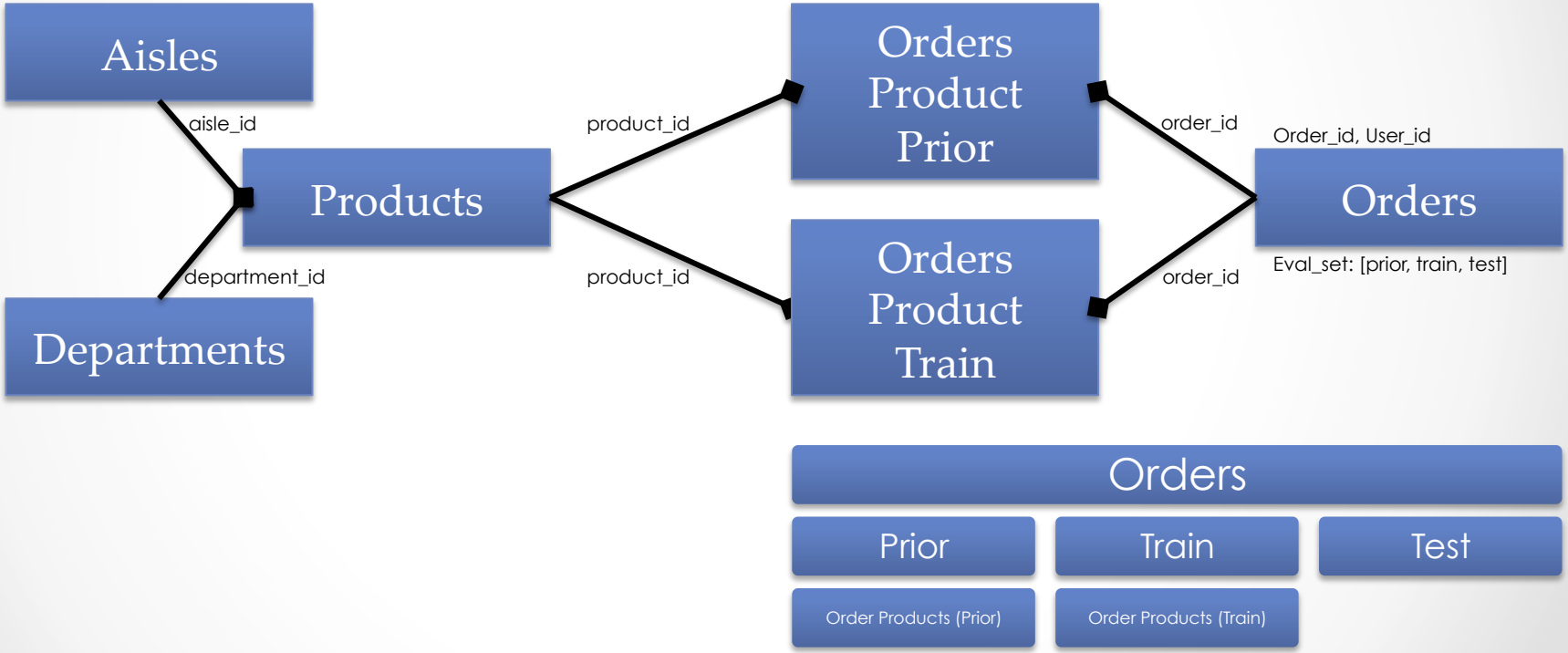
Approach


Loading
Feature Engineering
Model
Building Data Frames
•Data Cleanup
•Merges
•Orders + Prior
•Orders + Train
Agg. @ Product
•Use "Prior"
Agg. @ User
•Use "Prior"
Agg. @ User + Product
•Use "Prior"
Build Features & Target ()
•Based on Train Data
•Leverage Aggregations built
Models
•Random Forest Classifier
•Gradient Boosting Classifier
•Measure
•Predict

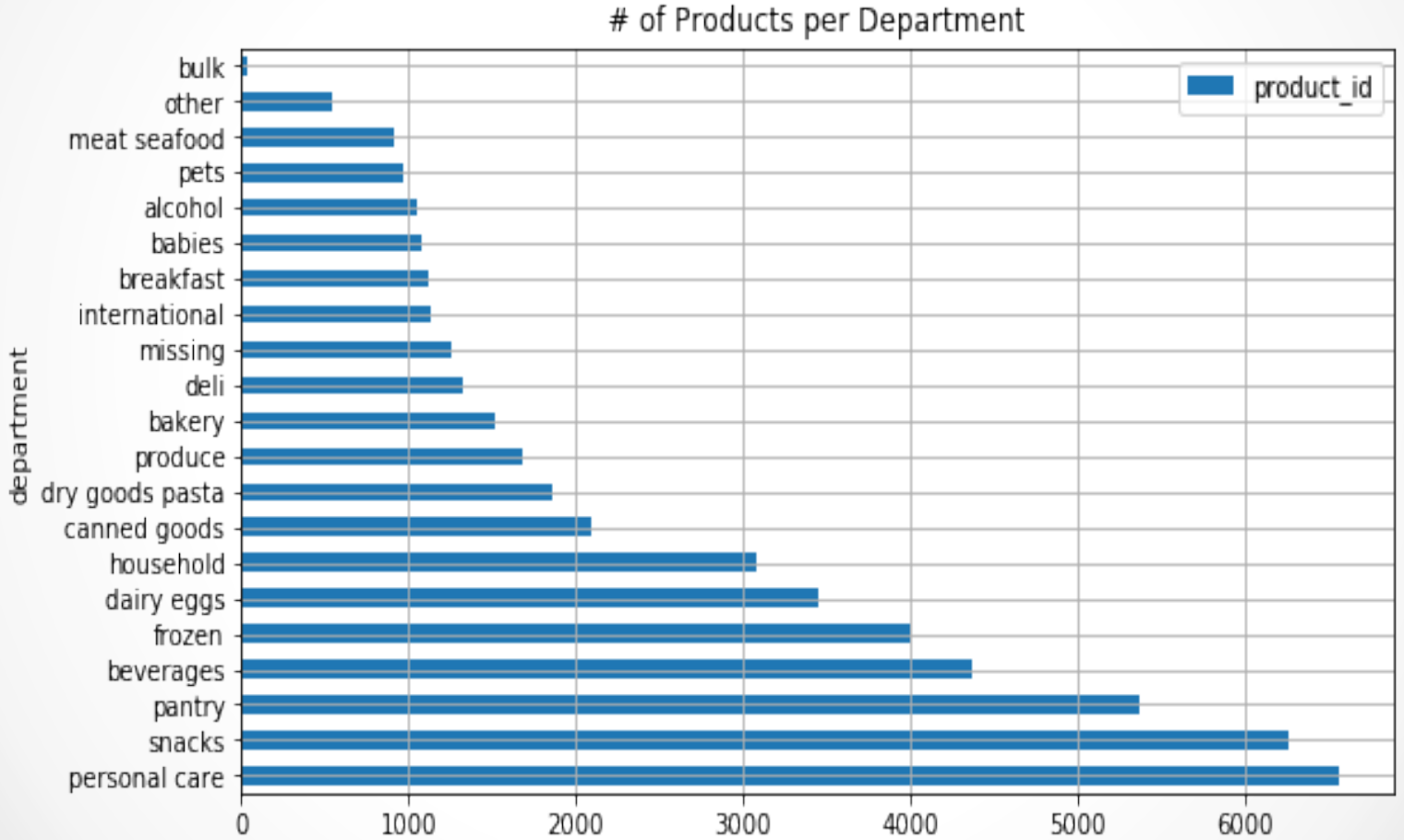
Data Model

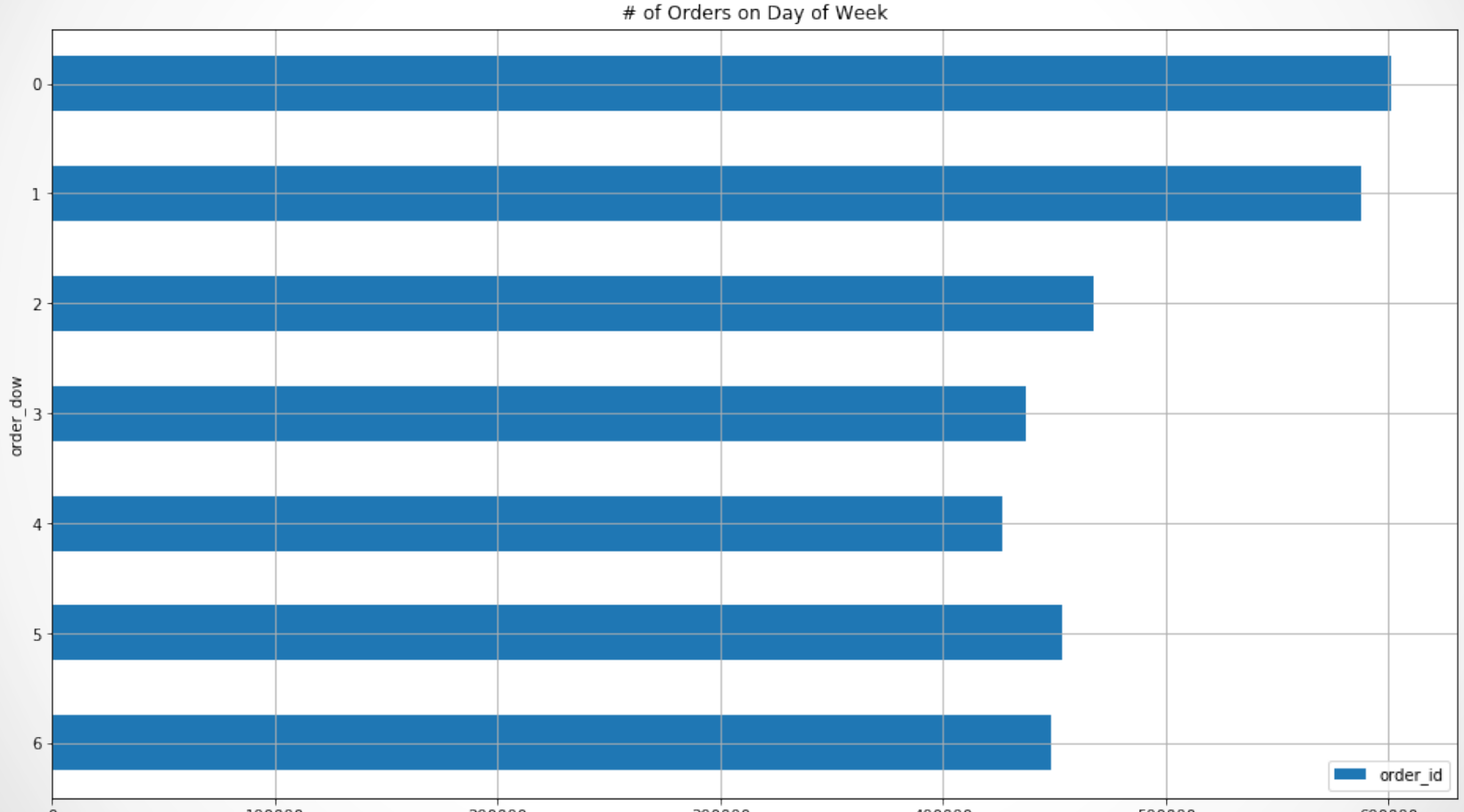
Data Model

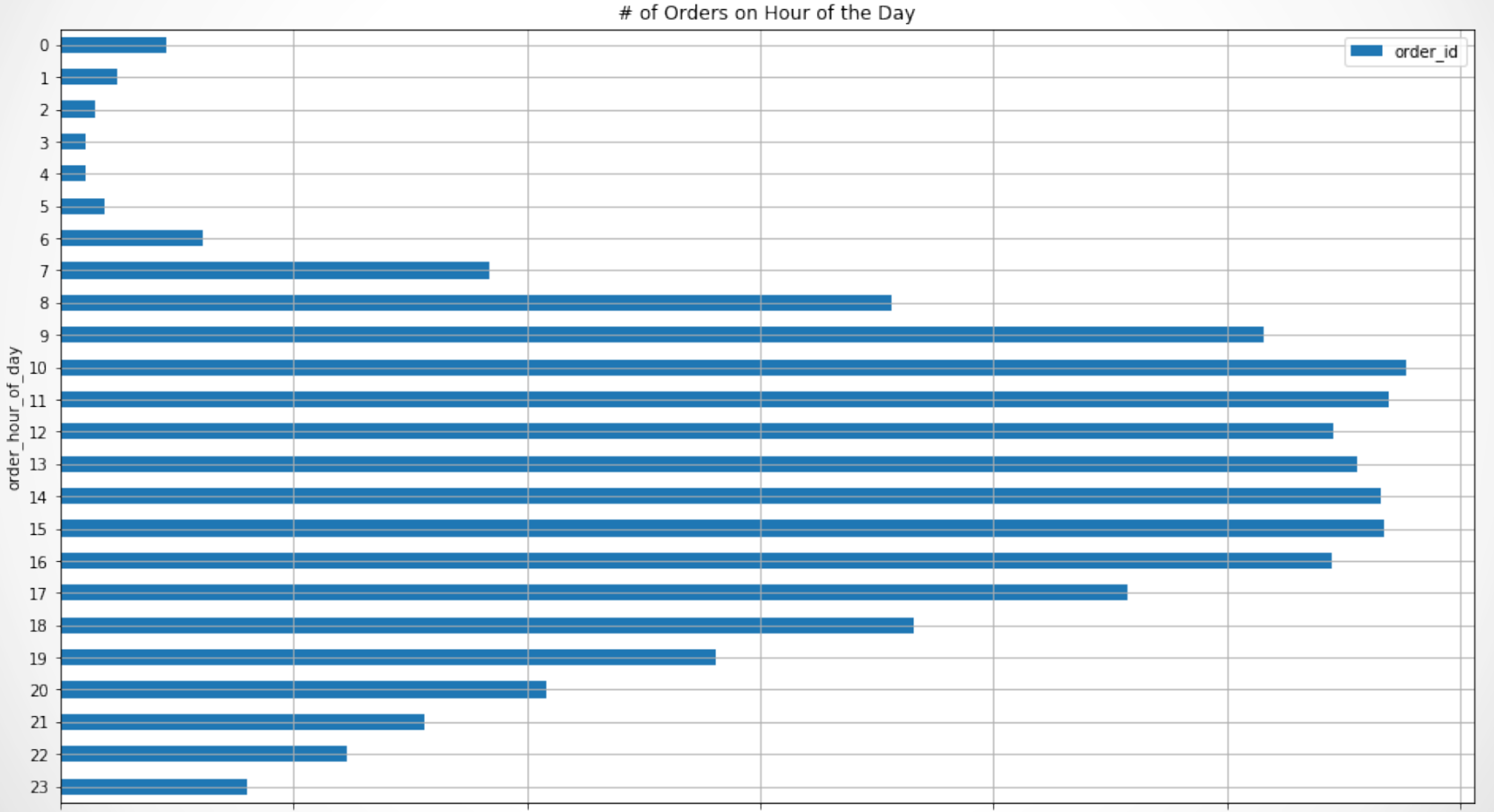
1:M

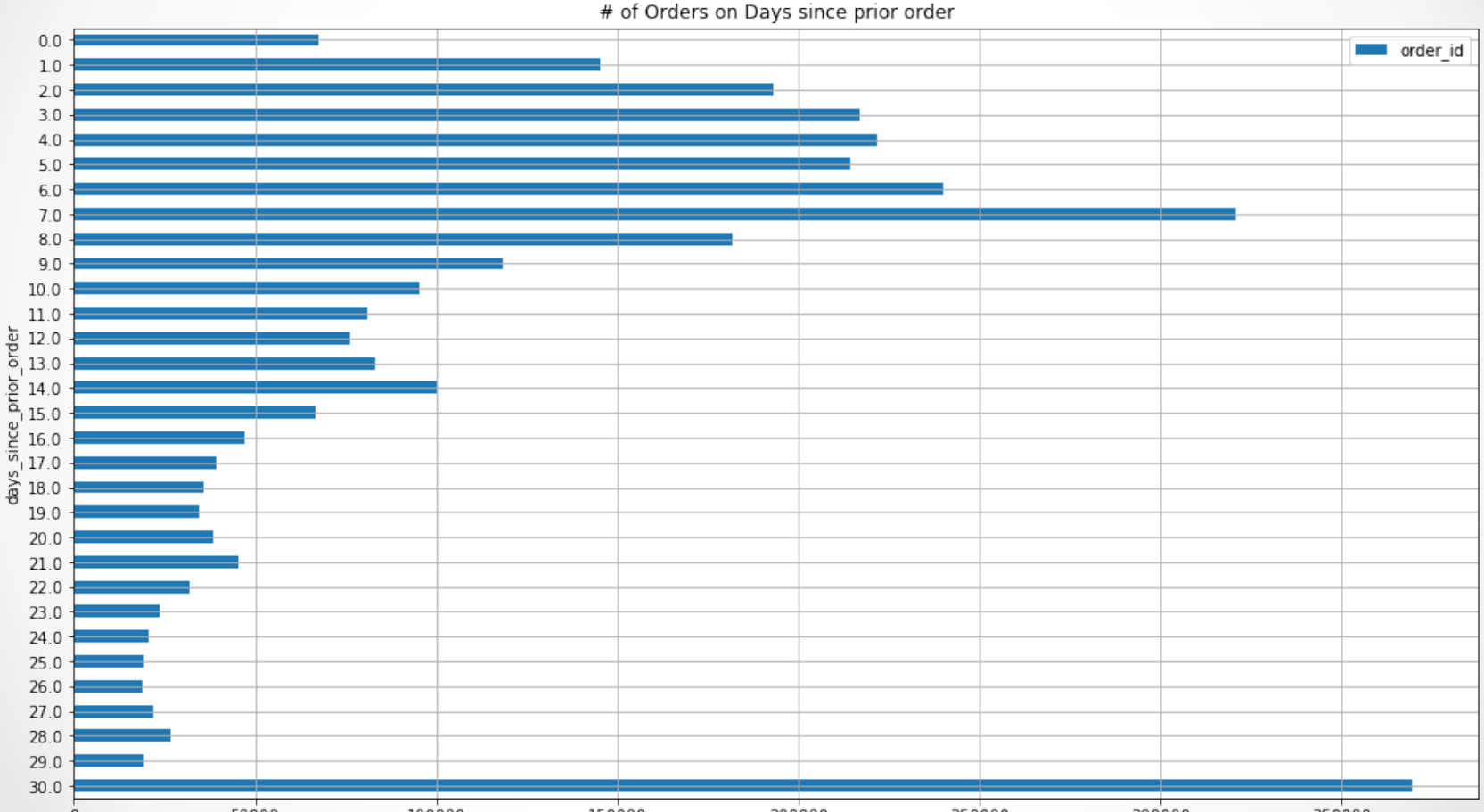


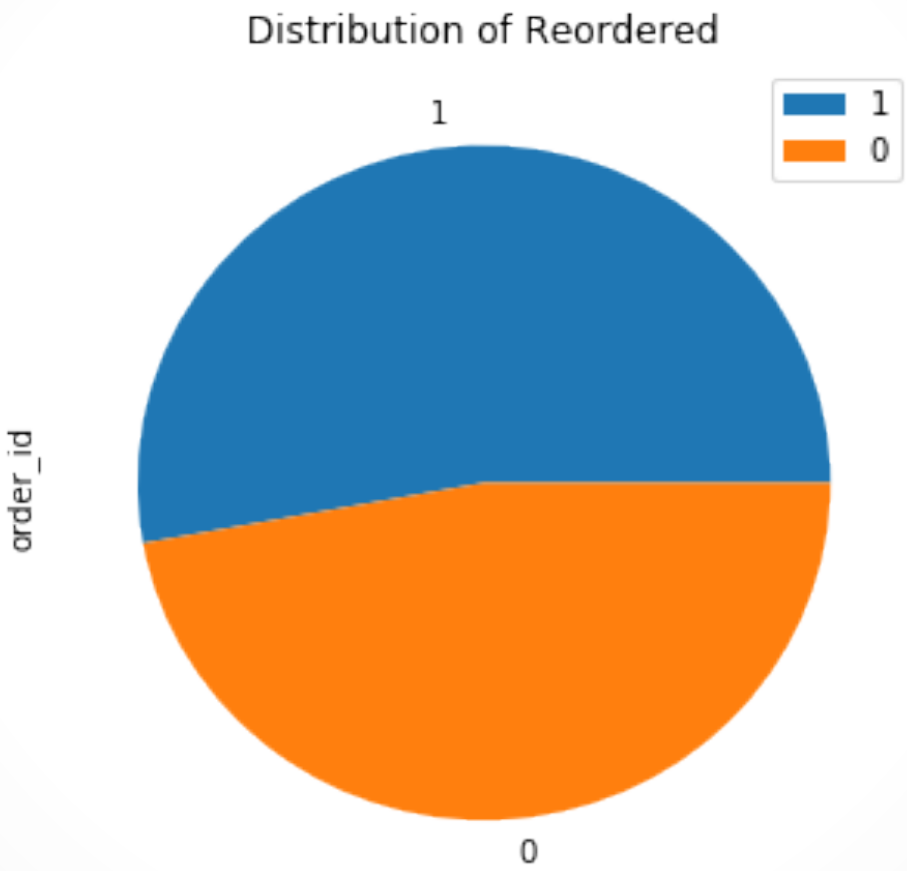
Understanding The Dataset

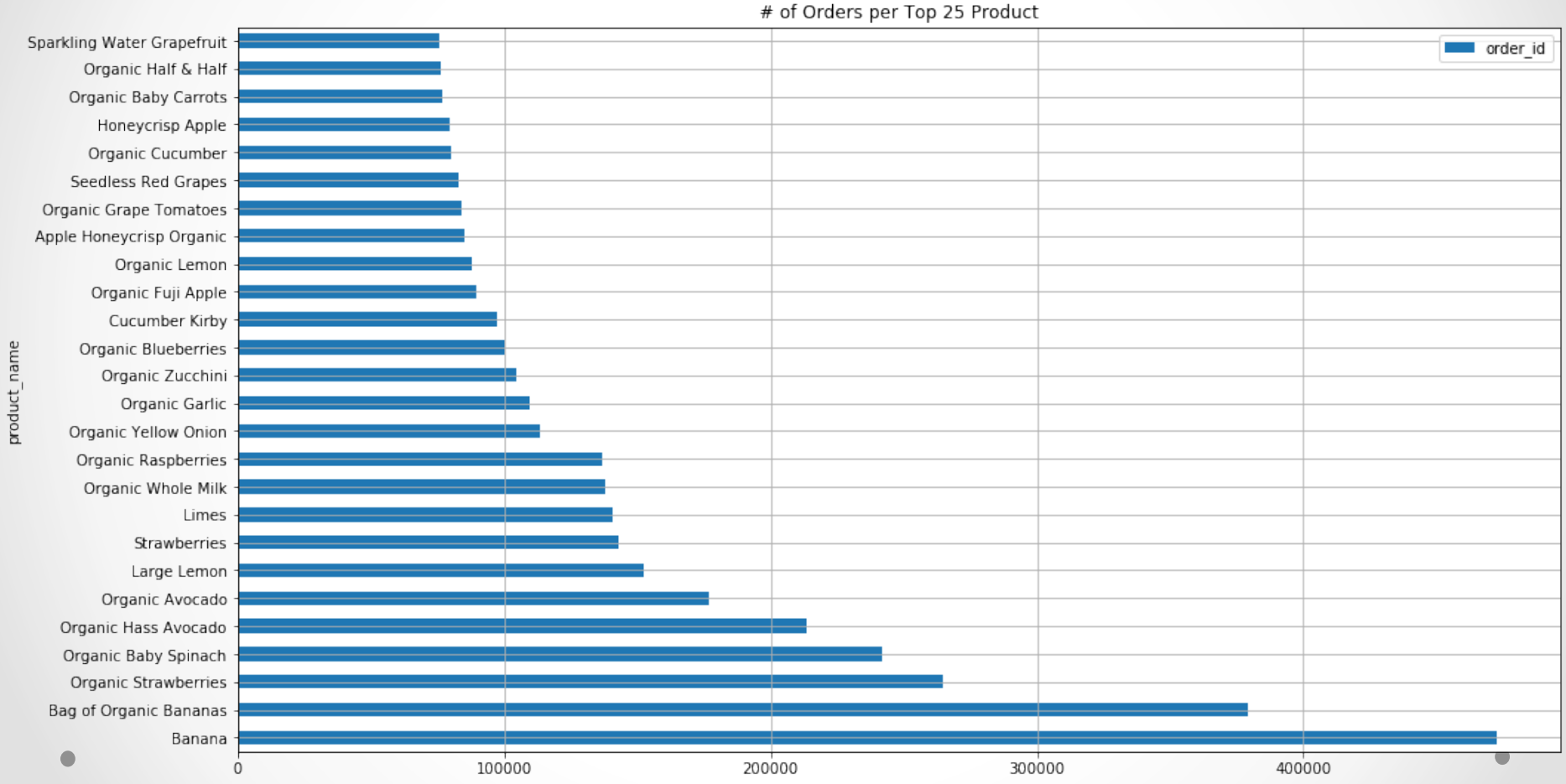

# of Products per Department
product_id
department
bulk
other
meat seafood
pets
alcohol
babies
breakfast
international
missing
deli
bakery
produce
dry goods pasta
canned goods
household
dairy eggs
frozen
beverages
pantry
snacks
personal care
0
1000
2000
3000
4000
5000
6000


# of Orders on Day of Week
order_dow
0
1
2
3
4
5
6
0
100000
200000
300000
400000
500000
600000
order_id


# of Orders on Hour of the Day
order_id
order_hour_of_day
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
0
50000
100000
150000
200000
250000
300000


# of Orders on Days since prior order
order_id
days_since_prior_order
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
13.0
14.0
15.0
16.0
17.0
18.0
19.0
20.0
21.0
22.0
23.0
24.0
25.0
26.0
27.0
28.0
29.0
30.0
0
50000
100000
150000
200000
250000
300000
350000


Distribution of Reordered
1
0
order_id
1
0


# of Orders per Top 25 Product
order_id
product_name
Sparkling Water Grapefruit
Organic Half & Half
Organic Baby Carrots
Honeycrisp Apple
Organic Cucumber
Seedless Red Grapes
Organic Grape Tomatoes
Apple Honeycrisp Organic
Organic Lemon
Organic Fuji Apple
Cucumber Kirby
Organic Blueberries
Organic Zucchini
Organic Garlic
Organic Yellow Onion
Organic Raspberries
Organic Whole Milk
Limes
Strawberries
Large Lemon
Organic Avocado
Organic Hass Avocado
Organic Baby Spinach
Organic Strawberries
Bag of Organic Bananas
Banana
0
100000
200000
300000
400000

Model

Choices

Random Forest Classifier

Gradient Boosting Classifier

Metrics

Cross Val Score Comparison

Random Forest

Random Forest :

Score: 0.852013953353

ROC AUC: 0.78906167567

Accuracy: 0.849758501971

Recall: 0.118334225001

Feature importance:	features	importance
12	up_total_orders	0.336196
13	up_max_add_to_cart_order	0.097800
2	p_reordered_rate	0.091143
14	order_number	0.088994
5	u_total_orders	0.087582
1	p_sum_reordered	0.069906
0	p_total_orders	0.048231
11	up_max_order_number	0.044468
10	up_max_order_id	0.038993
7	u_total_products	0.034003
6	u_avg_days_since_prior_order	0.018562
17	days_since_prior_order	0.014353
9	u_avg_cart	0.013312
3	aisle_id	0.008460
8	u_total_distinct_items	0.002738
16	order_hour_of_day	0.001784
4	department_id	0.001742
15	order_dow	0.001734

Gradient Boosting

Grandiant Boosting :

Score: 0.997289572245

ROC AUC: 0.754068356894

Accuracy: 0.840725429577

Recall: 0.22290055707

Feature importance:	features	importance
10	up_max_order_id	0.123060
13	up_max_add_to_cart_order	0.108933
6	u_avg_days_since_prior_order	0.096258
9	u_avg_cart	0.078599
2	p_reordered_rate	0.076827
8	u_total_distinct_items	0.070864
7	u_total_products	0.061259
12	up_total_orders	0.050924
0	p_total_orders	0.049493
1	p_sum_reordered	0.047270
11	up_max_order_number	0.046568
17	days_since_prior_order	0.042338
16	order_hour_of_day	0.039318
3	aisle_id	0.029388
15	order_dow	0.028233
5	u_total_orders	0.022189
14	order_number	0.021393
4	department_id	0.007087

Confusion Matrix

Random Forest

n = 275776	Predicted NO	Predicted YES	
Actual NO	TN =	FP =	
Actual YES	FN =	TP =	

Accuracy (TP+TN) / total:	
Error Rate 1 - Accuracy:	
Recall or True Positive Rate TP / Actual YES:	
False Positive Rate FP / Actual NO:	
Specificity 1 – False Positive Rate:	
Precision TP / Predicted YES:	
Prevalence Actual YES / total:	

Gradient Boosting

n = 275776	Predicted NO	Predicted YES	
Actual NO	TN =	FP =	
Actual YES	FN =	TP =	

Accuracy (TP+TN) / total:	
Error Rate 1 - Accuracy:	
Recall or True Positive Rate TP / Actual YES:	
False Positive Rate FP / Actual NO:	
Specificity 1 – False Positive Rate:	
Precision TP / Predicted YES:	
Prevalence Actual YES / total:	

Next Steps

- Huge Potentials to enrich the features to improve model's performance.
- Building label could be improved.
- For validation of the model's in lieu of the test results to compare with, Prior data could be sampled for test set.

Thank You!