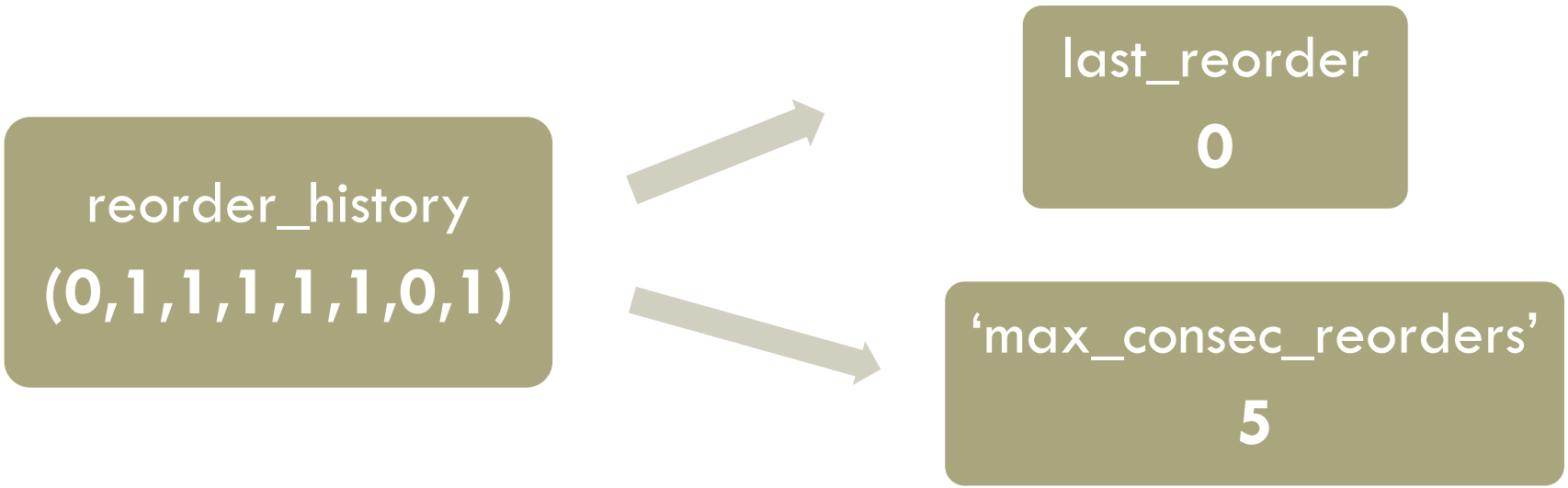
INSTAKART!

M Du

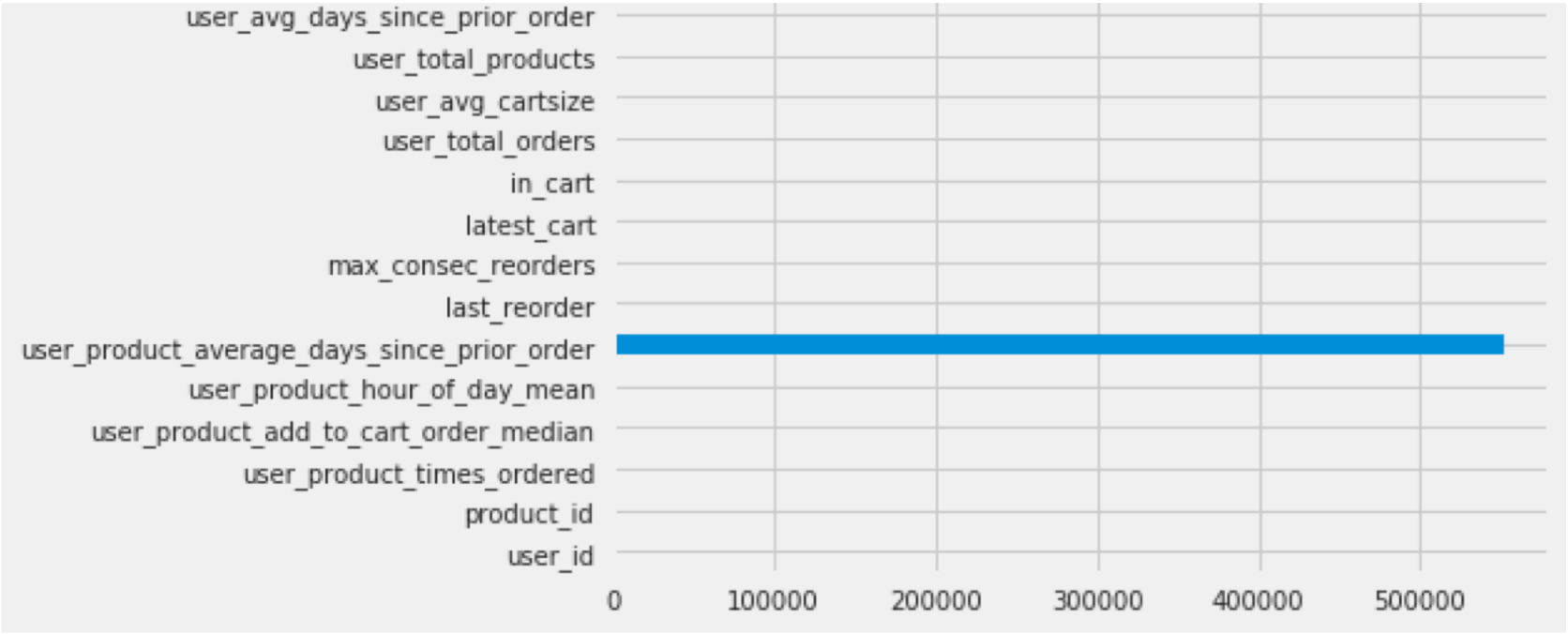
FEATURE ENGINEERING (USER – PRODUCTS)

Frequency of Product Ordered by User	Median Cart Order by User	Mean Hour of day Ordered by User	Mean Days Since Last Order	reorder_history
4	5	14	3.5	(0,1,1,0)
8	1	11	5	(0,1,1,1,1,1,0,1)
2	2	10	31	(0,1)

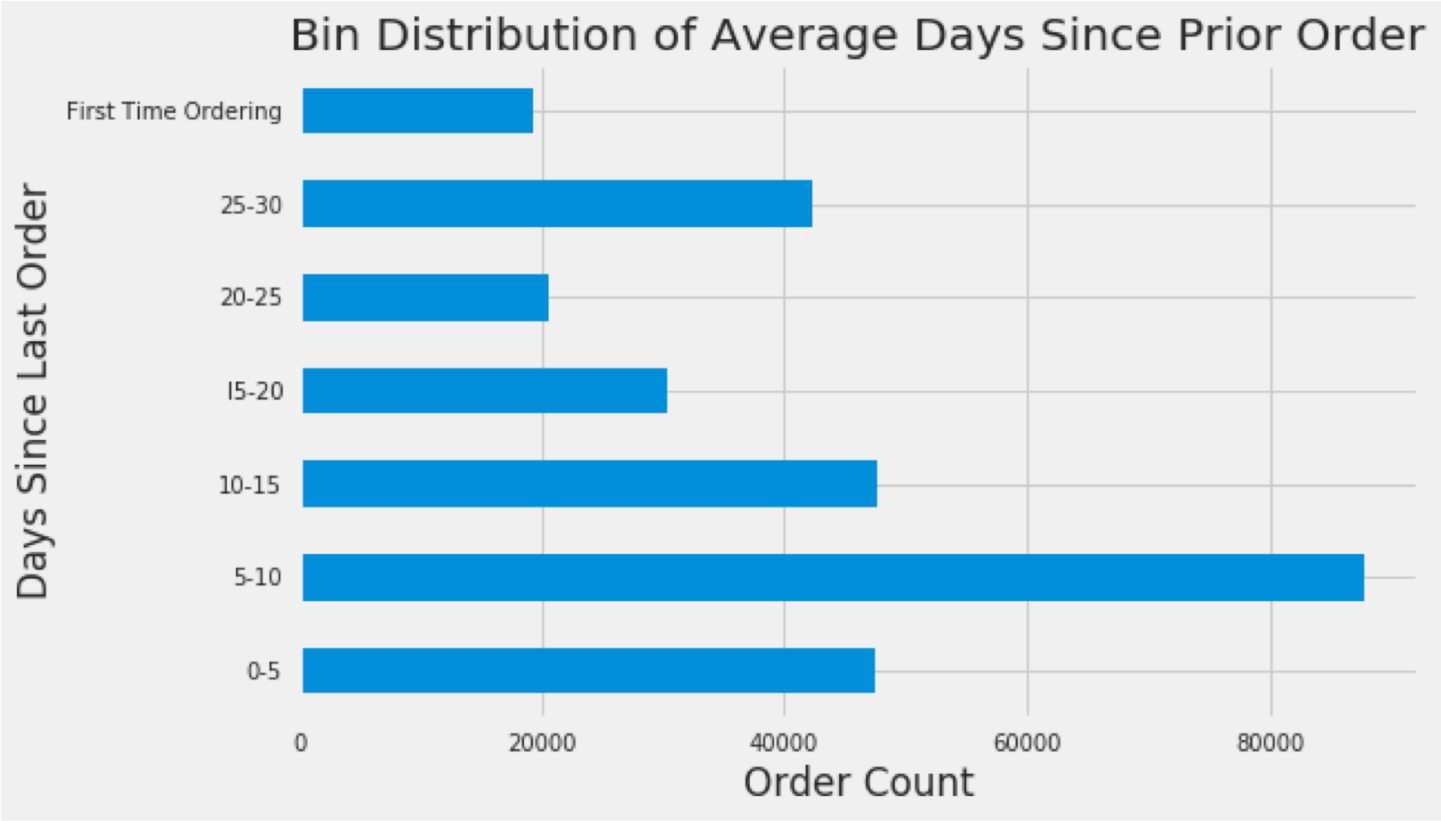
FEATURE ENGINEERING



NAN VALUES


user_avg_days_since_prior_order
user_total_products
user_avg_cartsize
user_total_orders
in_cart
latest_cart
max_consec_reorders
last_reorder
user_product_average_days_since_prior_order
user_product_hour_of_day_mean
user_product_add_to_cart_order_median
user_product_times_ordered
product_id
user_id
0
100000
200000
300000
400000
500000

NAN VALUES


Bin Distribution of Average Days Since Prior Order
First Time Ordering
25-30
20-25
I5-20
10-15
5-10
0-5
Days Since Last Order
0
20000
40000
60000
80000
Order Count

FEATURE ENGINEERING (USER SPECIFIC)

Total Orders

Average Cartsize

Total Unique Products Ordered

Average Days Between Orders

FEATURE ENGINEERING (AISLE AND DEPARTMENT)

Departments: 'alcohol', 'babies', 'bakery', 'beverages', …

Aisles: 'air fresheners candles', 'asian foods', 'baby accessories', …

Indicator Variable

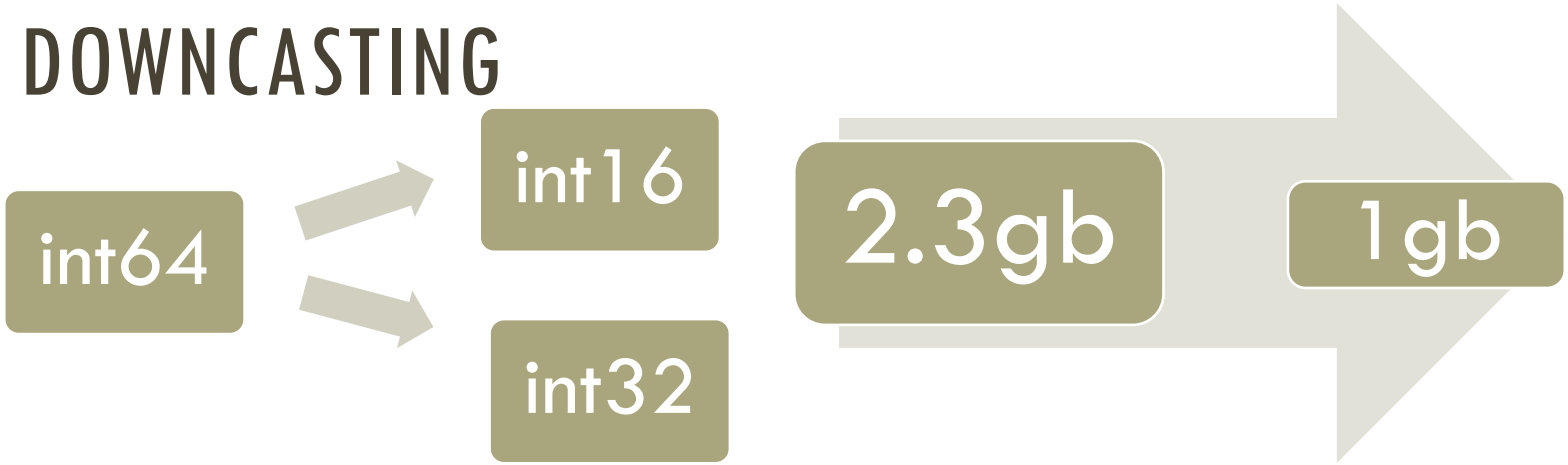
DOWNSIZING (SUBSETTING)

Original File: 32,434,489 rows

Grouped Dataframe: 8,474,661 rows

Extracted ~5% of data: 423772 rows

DOWNCASTING



FEATURE TESTING (BASELINE)

70/30 train/test split

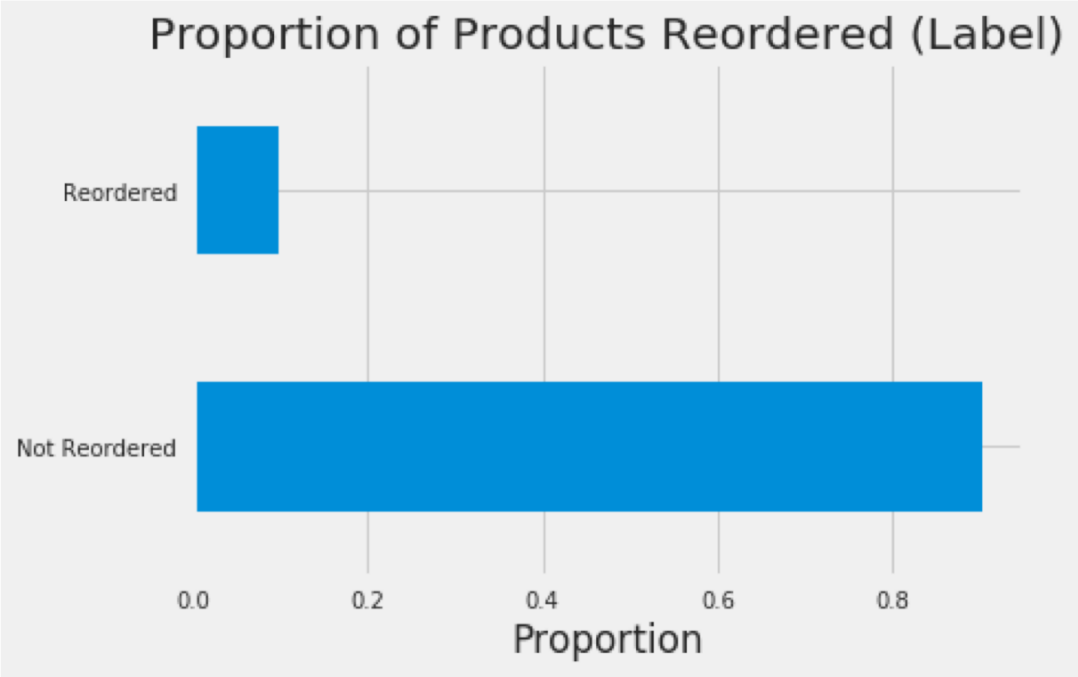
Logistic Regression with single feature: ‘user_total_orders’

ALL ZEROES!

```
/home/ubuntu/anaconda3/lib/python3.6/site-packages/sklearn/m
etrics/classification.py:1137: UndefinedMetricWarning: F-sco
re is ill-defined and being set to 0.0 due to no true sample
s.
  'recall', 'true', average, warn_for)
```

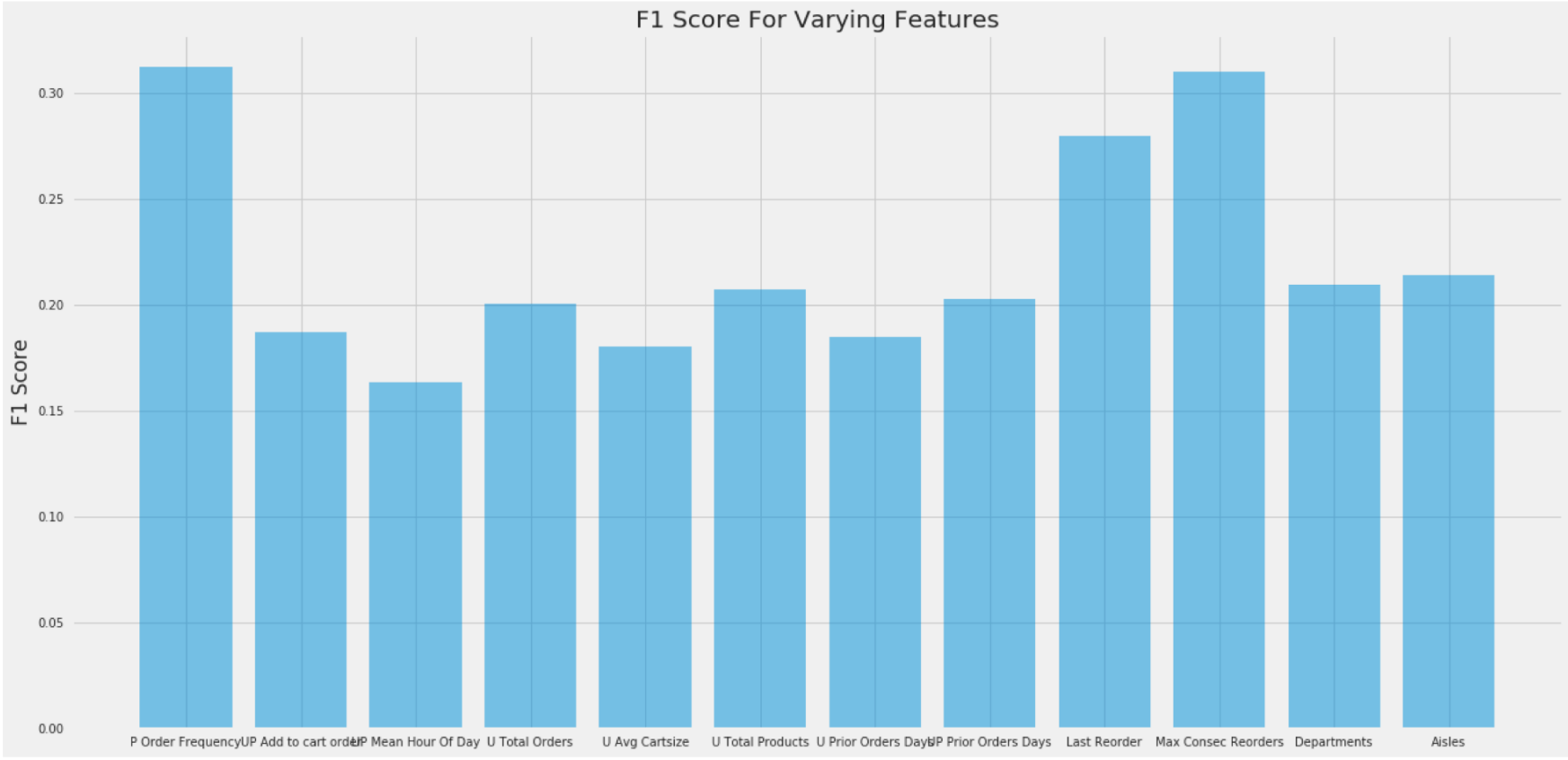
```
Simple Logistic Regression; Test F1: 0.000, Test AUC: 0.598
```

LABEL IMBALANCE



F1 Score After Resampling: **0.201**

TEST ALL FEATURES (INDIVIDUALLY)



TEST ALL FEATURES (INDIVIDUALLY)

P Order Frequency	Max Consec Reorders	Last Reorder	Aisles
0.3119	0.3101	0.2798	0.2136

LOGISTIC REGRESSION WITH ALL FEATURES

Feature	Coefficient
Max Consecutive Reorders	0.413161
Average Cart Size	0.31279
Last Reorder	0.248632
Product Order Frequency	0.215626
Fresh Fruit Aisle	0.081017
Produce Department	0.072444
Water Seltzer Aisle	0.061491
20-25 days since last order	0.05745
Milk Aisle	0.051247
15-20 days since last orer	0.050933

F1 Score: 0.32624

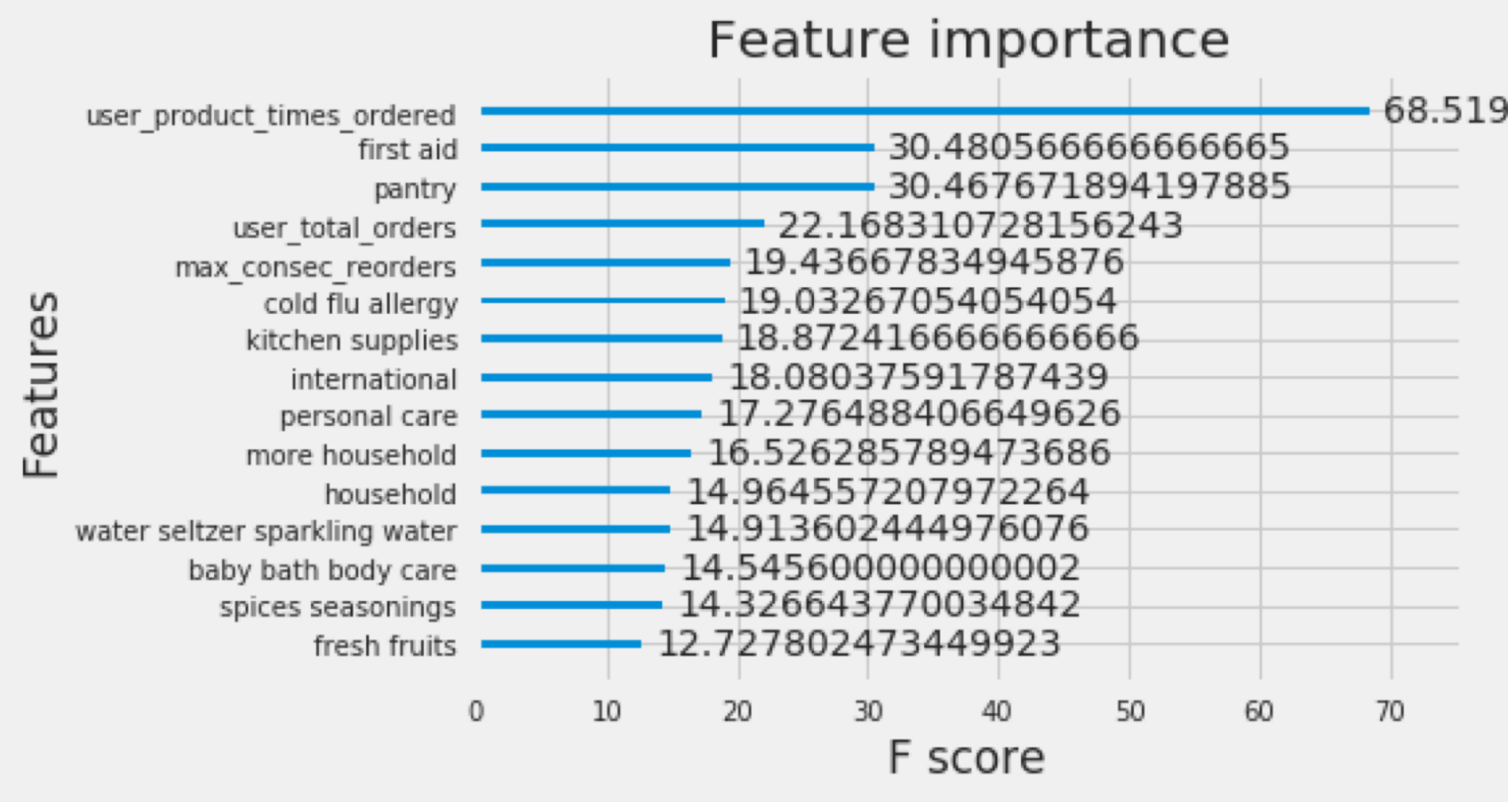
RANDOM FOREST

F1 Score: 0.37117

XGBOOST

F1 Score: 0.3396

XGBOOST


Feature importance
user_product_times_ordered
first aid
pantry
user_total_orders
max_consec_reorders
cold flu allergy
kitchen supplies
international
personal care
more household
household
water seltzer sparkling water
baby bath body care
spices seasonings
fresh fruits
68.519
30.480566666666665
30.467671894197885
22.168310728156243
19.43667834945876
19.03267054054054
18.872416666666666
18.080375917874 39
17.276488406649626
16.526285789473686
14.964557207972264
14.913602444976076
14.545600000000002
14.326643770034842
12.727802473449923
Features
0
10
20
30
40
50
60
70
F score

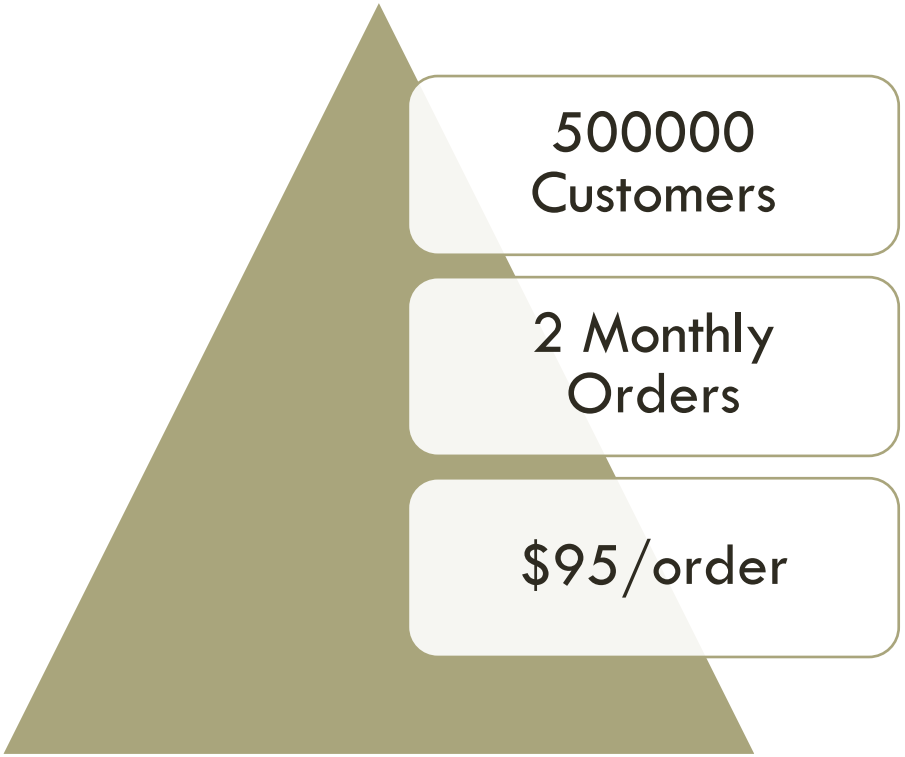
XGBOOST GAIN SCORES

user_product_times_ordered	68.5191
first aid	30.4805
pantry	30.4676
user_total_orders	22.1683
max_consec_reorders	19.4366

COMPARE AND CONQUER

	Logistic Regression	RandomForest	XGBoost
F1	0.3262	**0.3734**	0.3440
Recall	0.2142	**0.3145**	0.2769

CONQUER WHAT?



$95 Million Revenue/Month

~10% reorders

31% Recall

Targetable: $3 Million

CONQUER WHAT?

Increase Frequency

- Infer other reorder products based on correlations with similar buyer profiles

Increase Consistency

- Providing subscription models (with free shipping)
- Email Reminders