Predicting the Products an Online Grocery Shopper Will Purchase Again

Springboard Capstone Project 2

Yi Li

Introduction

- Online grocery shopping is growing rapidly these years.
- U.S. Online Grocery Survey 2020 showed 52.0% of all respondents had bought groceries online - more than double the shopper numbers from two years ago.
- The coronavirus pandemic is transforming consumers' needs and behaviors, and has encouraged more grocery shoppers to start buying or buying more online.

Potential Client

- Grocery delivery apps in the market:

  Instacart

  Shipt

  Amazon prime now

  Walmart grocery delivery

  ….
- Correctly predicting customers' shopping behavior using machine learning, and incorporate it into the features of the apps will make their consumers' shopping experience more pleasant.

Data

- https: “The Instacart Online Grocery Shopping Dataset 2017”, Accessed from https://www.instacart.com/datasets/grocery-shopping-2017 on <2020/05/>

Basic Structure of the datasets

column	description	dtype
aisle_id	aisle identifier	integer in [1:134]
aisle	the name of the aisle	string

aisle.csv

column	deciption	dtype
department_id	department identifier	integer in [1:21]
department	the name of the department	string

department.csv

column	deciption	dtype
product_id	product identifier	integer in [1:49688]
product_name	name of the product	string
aisle_id	aisle identifier	integer
department_id	department identifier	integer

products.csv

column	deciption	dtype
order_id	order identifier	integer in [1: 3421083]
user_id	customer identifier	integer in [1: 206209]
eval_set	which evaluation set this order belongs in	category(prior/train/test)
order_number	the order sequence number for this user (1 = first, n = nth)	integer in [1:100]
order_dow	the day of the week the order was placed on	integer in [1:7]
order_hour_of_day	the hour of the day the order was placed on	integer in [0:23]
days_since_prior	days since the last order, capped at 30 (with NAs for order_number = 1)	float in [0:30] or NA

orders.csv

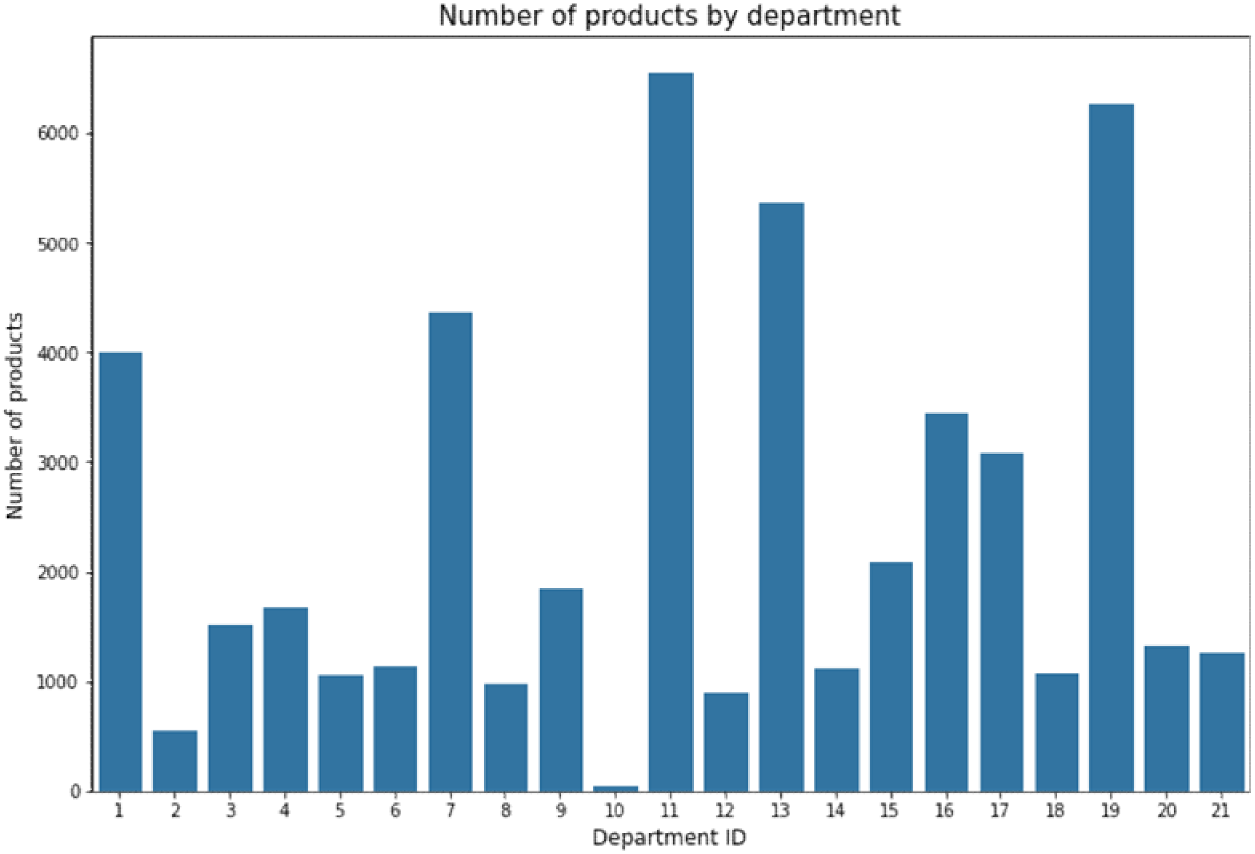
column	deciption	dtype
order_id	order identifier	integer
product_id	customer identifier	integer
add_to_cart_order	order in which each product was added to cart	integer
reordered	1 if this product has been ordered by this user in the past, 0 otherwise	integer(0/1)

Order_products__prior.csv

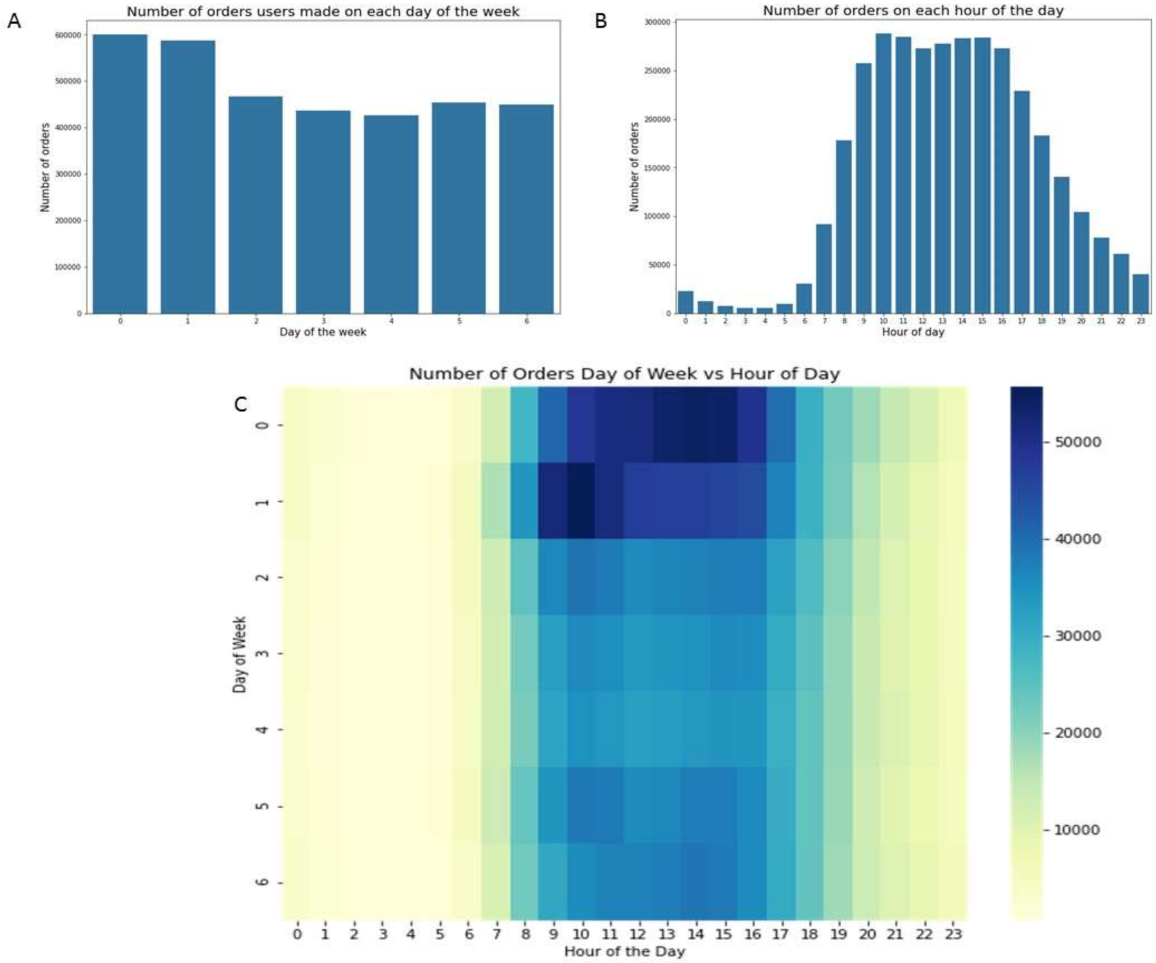
Order_products__train.csv

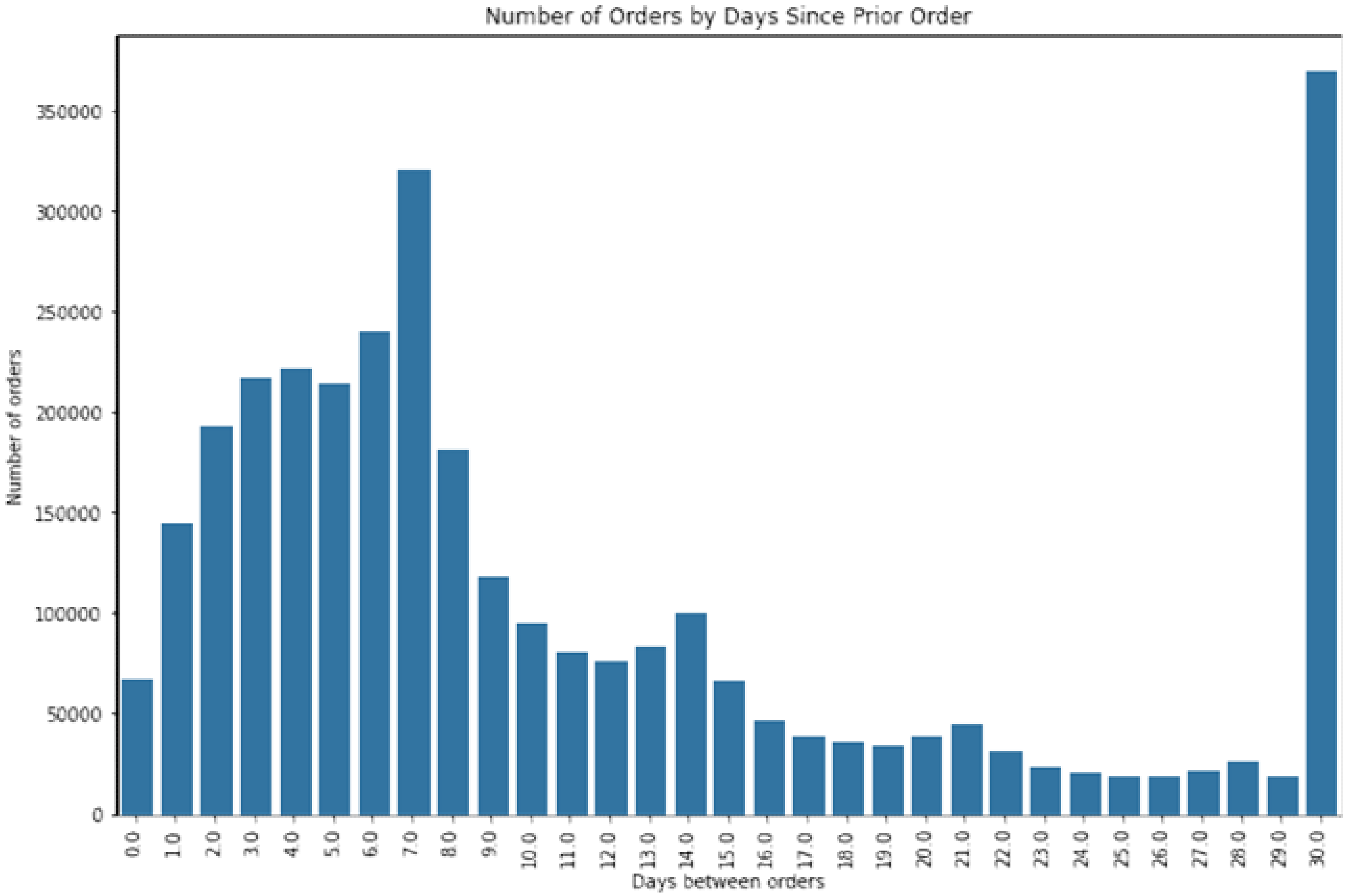
Exploratory Data Analysis and Statistical Inference

- Number of products by department
- Number of orders by time
- Number of orders by days since prior orders
- Count of orders by number of products in the order
- Count of orders by number of reordered products in the order
- Number of reordered products in an order by day of week or hour of day

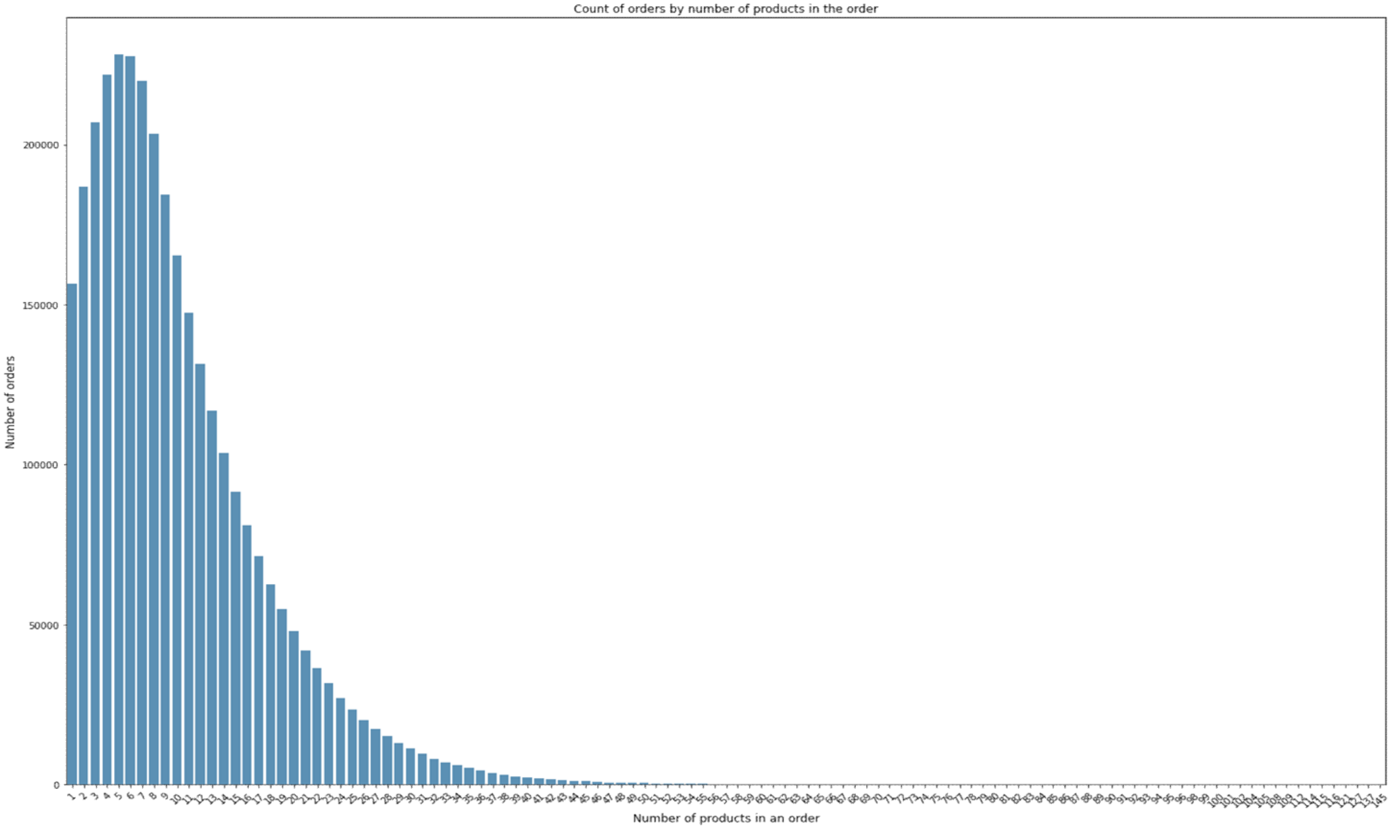

Number of products by department
Number of products
0
1000
2000
3000
4000
5000
6000
Department ID
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21

Number of orders by time

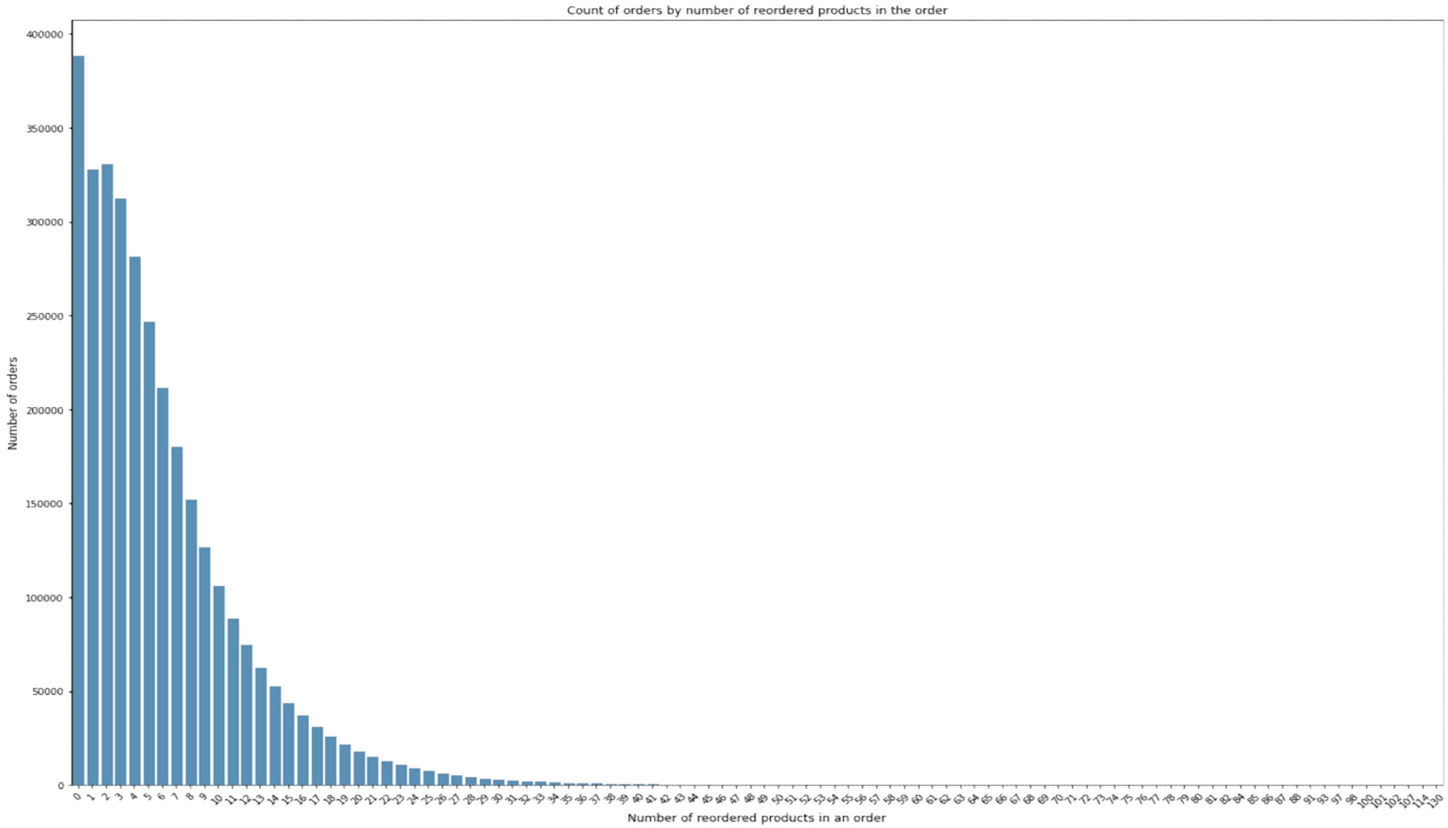



Number of Orders by Days Since Prior Order
Number of orders
0
50000
100000
150000
200000
250000
300000
350000
Days between orders
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
13.0
14.0
15.0
16.0
17.0
18.0
19.0
20.0
21.0
22.0
23.0
24.0
25.0
26.0
27.0
28.0
29.0
30.0

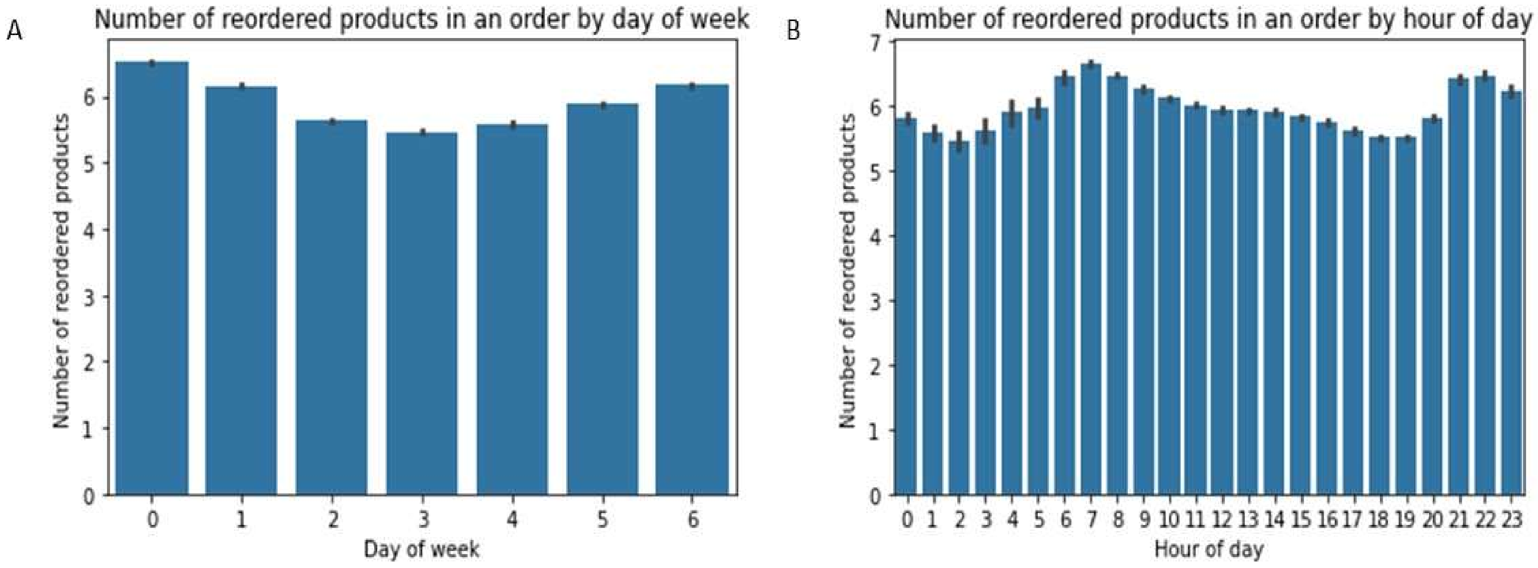
Count of Orders by Number of Products in the Order



Count of Orders by Number of Reordered Products in the Order



Number of Reordered Products in an Order by Day of Week or Hour of Day



Feature Engineering

- User features
- Product features
- User product interaction features
- Last order features

Machine Learning

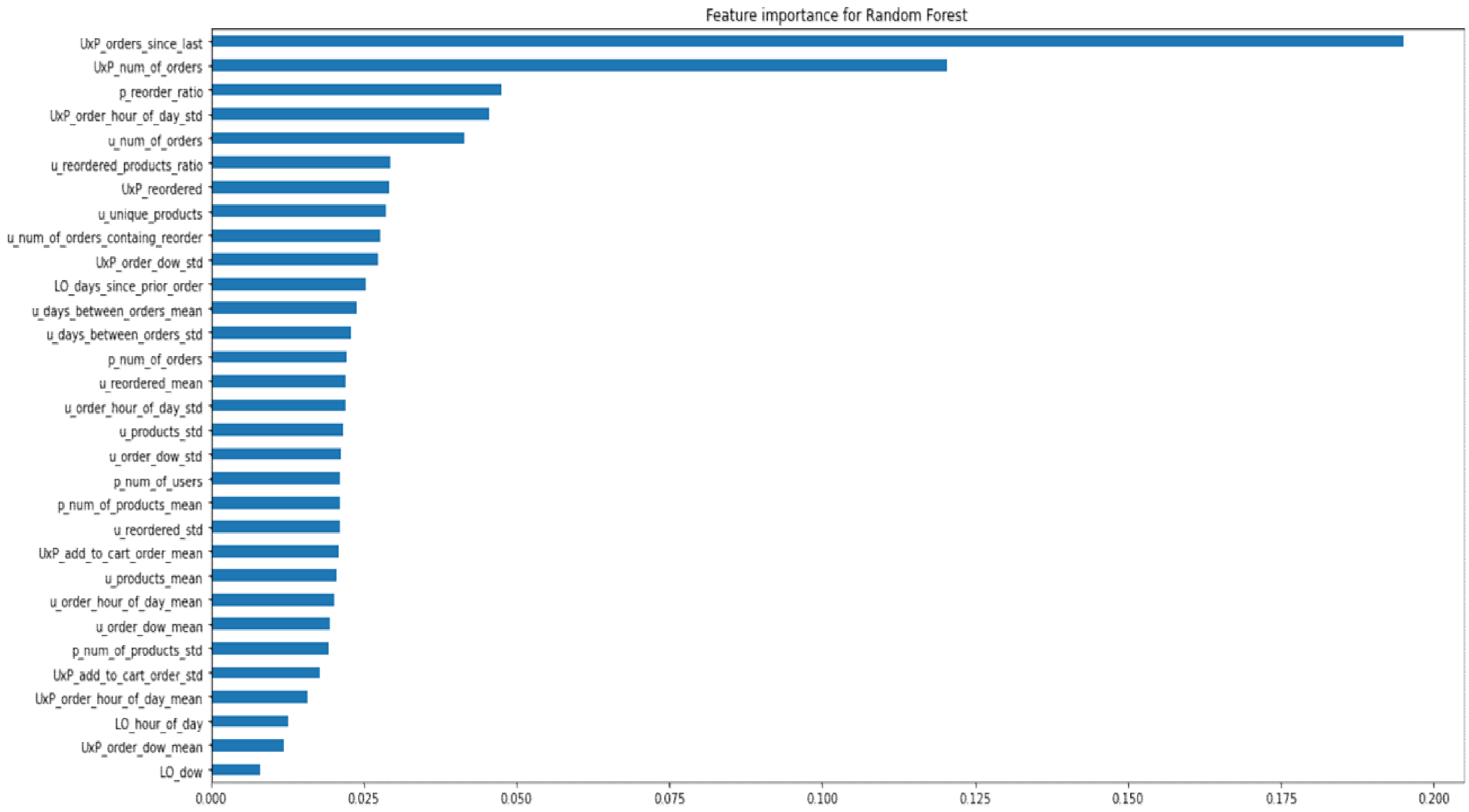
- Classification metrics

Term	Formula
Accuracy	(TP + TN)/(P+N)
Recall	TP/(TP+FN)
Precision	TP/(TP+FP)
F-measure	(2 x recall x precision) / (recall+precision)

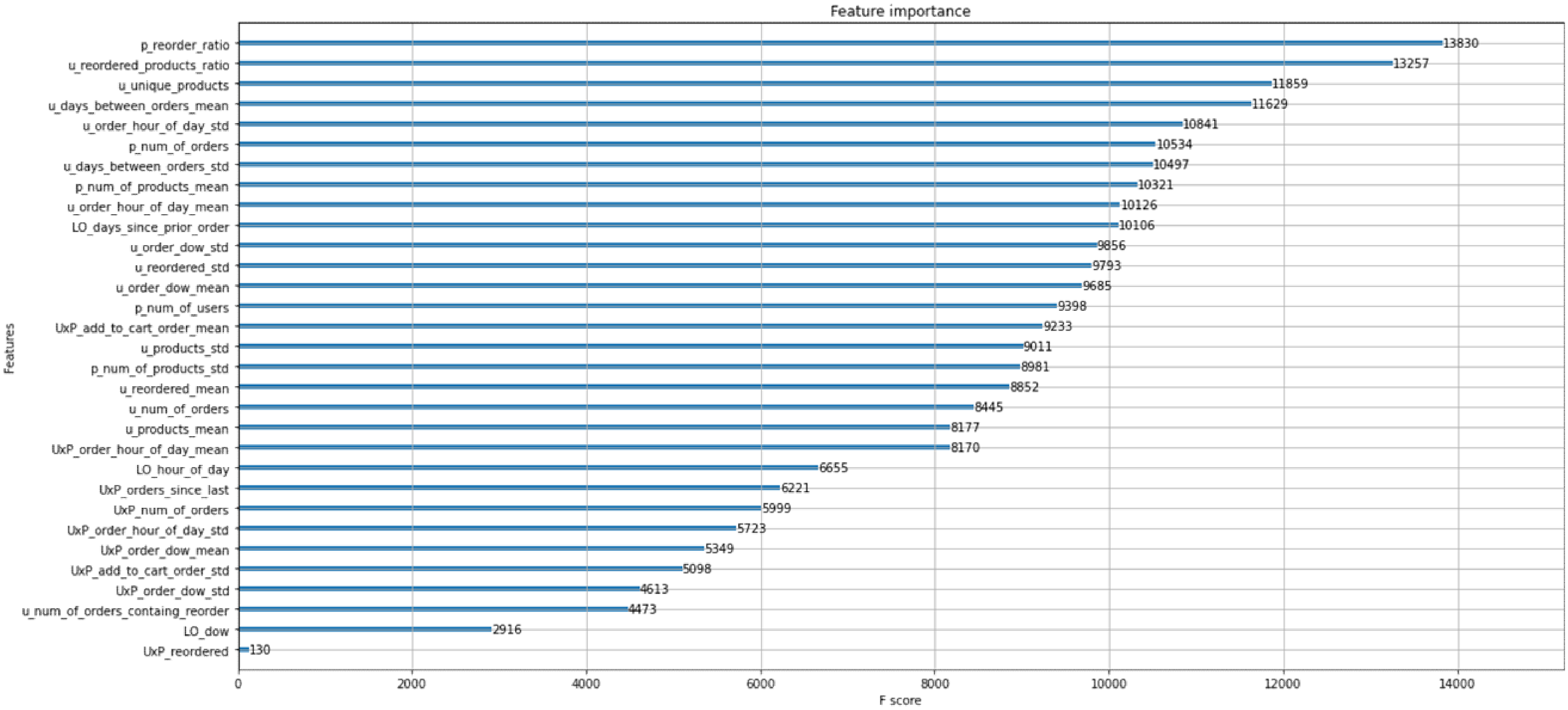
Machine Learning Models Comparison

Classifier /Performance	Random Forest	XGboosting
Accuracy	0.88	0.89
Recall	0.46	0.44
Precision	0.41	0.44
F1	0.44	0.44

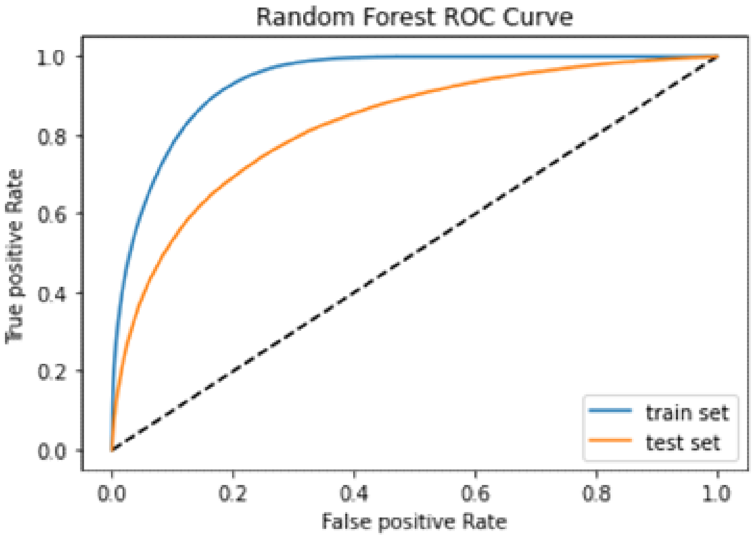
Feature Importance for Random Forest

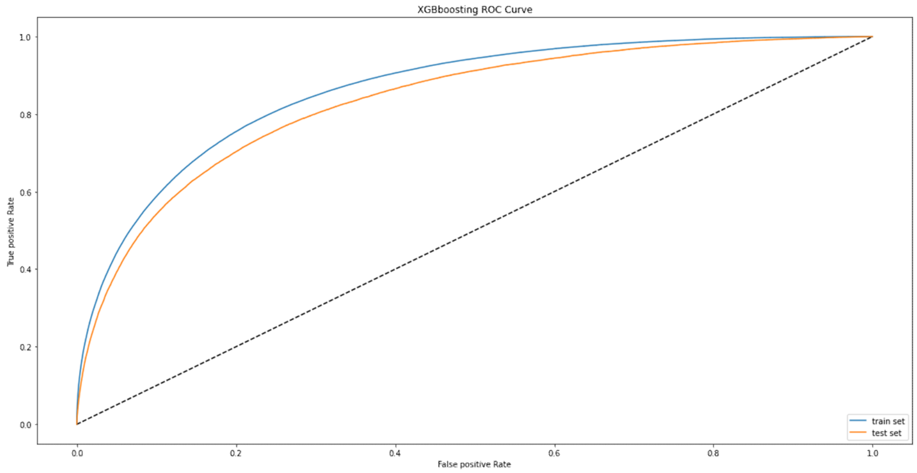


Feature Importance for XGboosting

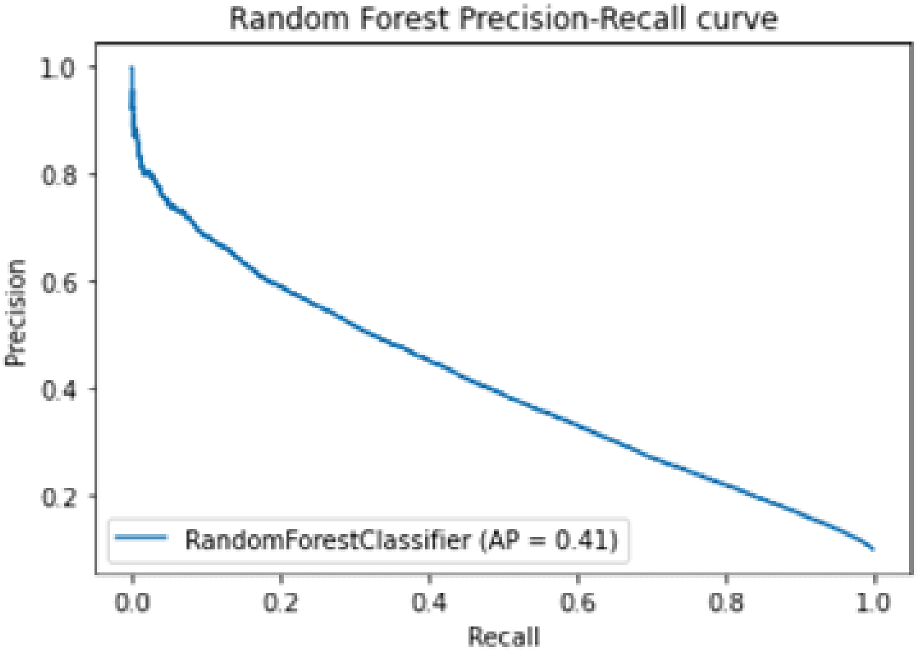


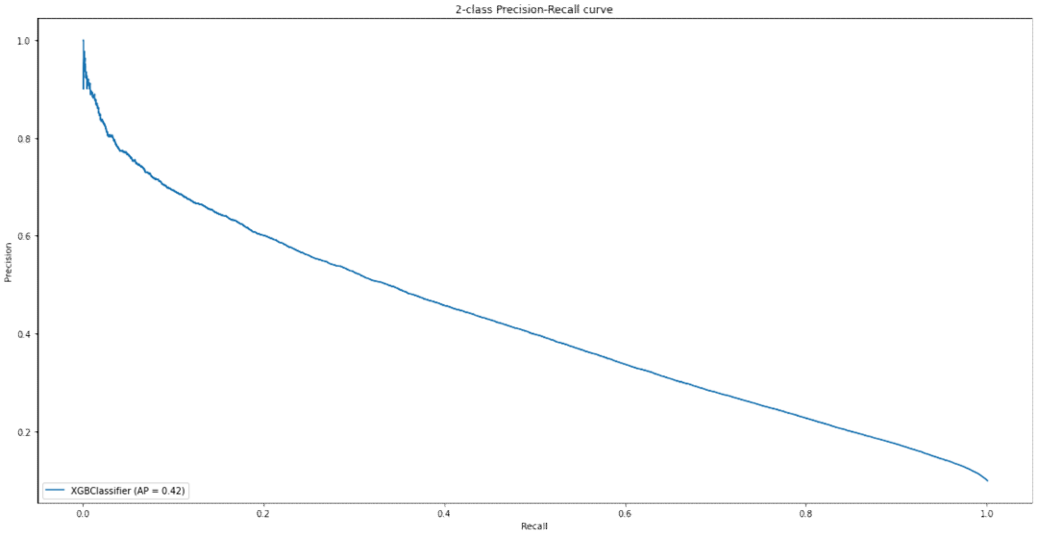
ROC curve


Random Forest ROC Curve
1.0
0.8
0.6
0.4
0.2
0.0
True positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive Rate
train set
test set


XGBboosting ROC Curve
1.0
0.8
0.6
0.4
0.2
0.0
True positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive Rate
train set
test set

Precision-recall Curve





Summary and Ongoing Works

- Modify features, UxP_orders_since_last → UxP_days_since_last
- More UxP interaction features
- Modeling after feature selection
- Random boosting with scale_pos_weight = 1, manually set the prediction threshold

Acknowledgements

- My springboard mentors
- Springboard staff and community