THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Instacart Market Basket Analysis

——Data Mining Final Project

Group Member:
Qing Ruan
Zixuan Huang
Ya Liu

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

CONTENTS

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Introduction

01

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Dataset Description

File Name	Column Names
orders.csv	order_id,user_id,eval_set,order_number,order_dow,order_hour_of_day,days_since_prior_order
order_products__*.csv	order_id,product_id,add_to_cart_order,reordered
aisles.csv	aisle_id,aisle
departments.csv	department_id,department
products.csv	product_id,product_name,aisle_id,department_id

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

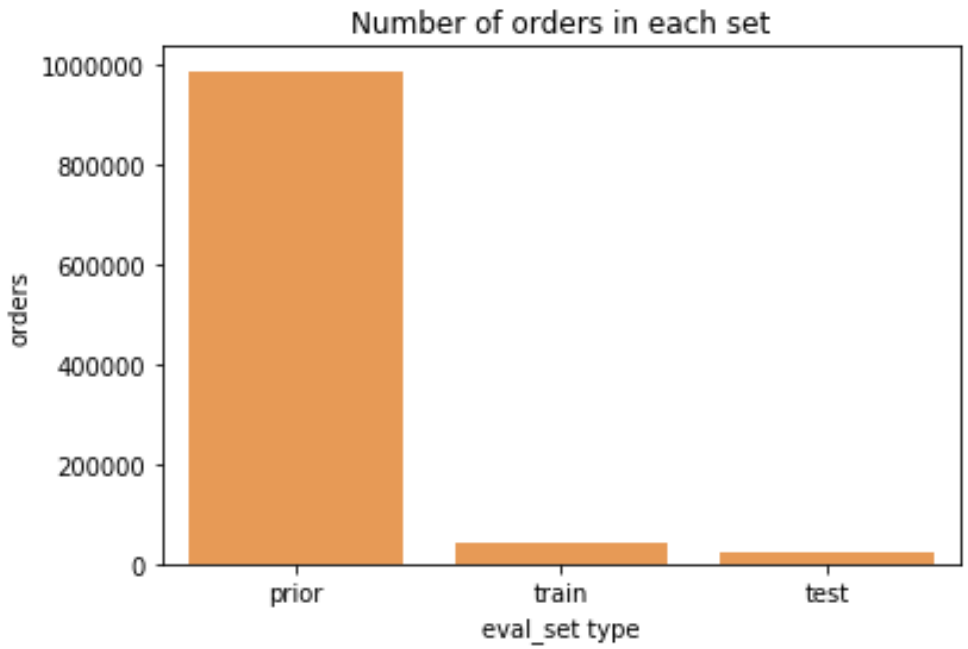
Exploratory Data Analysis

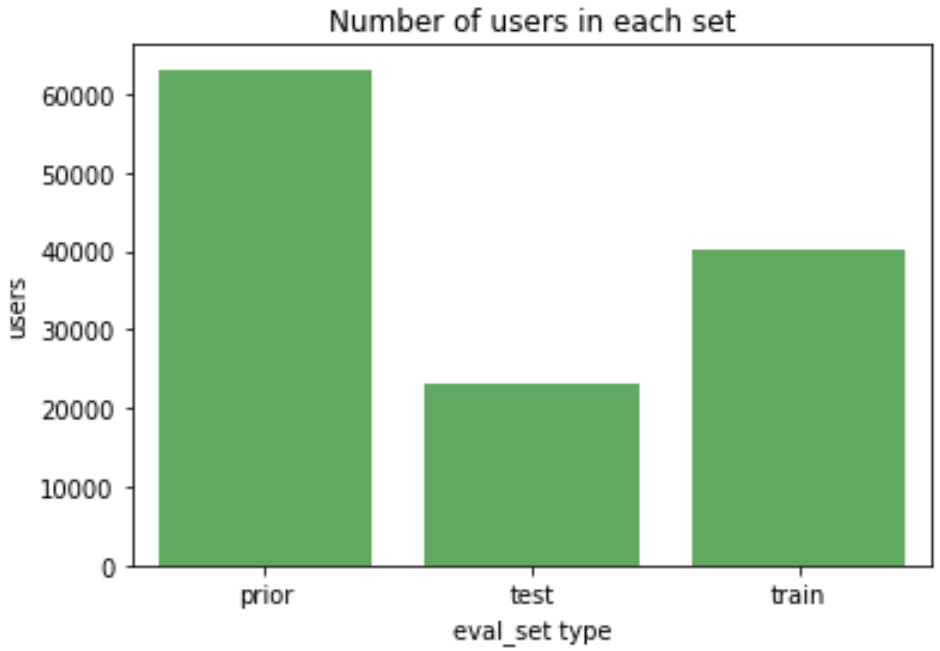
02

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Orders Information

- Three sets.

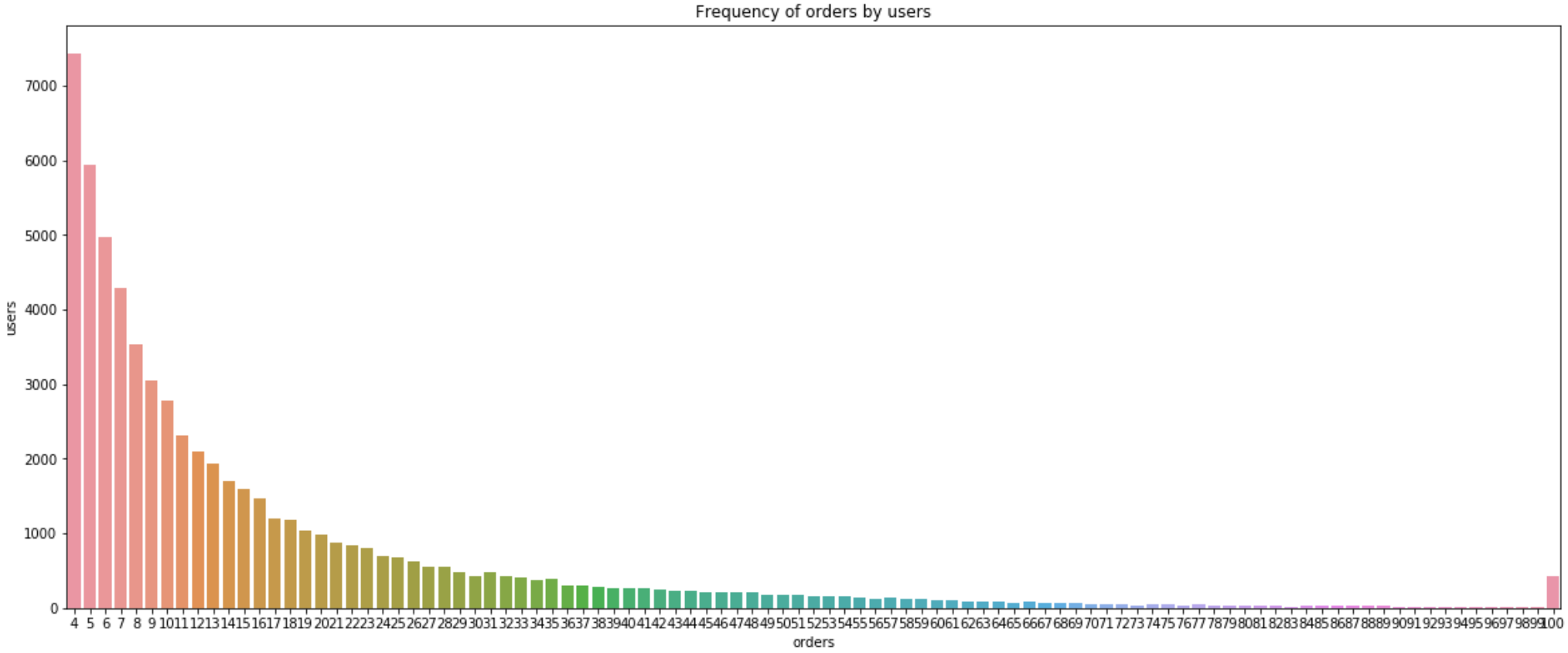




THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Orders Information

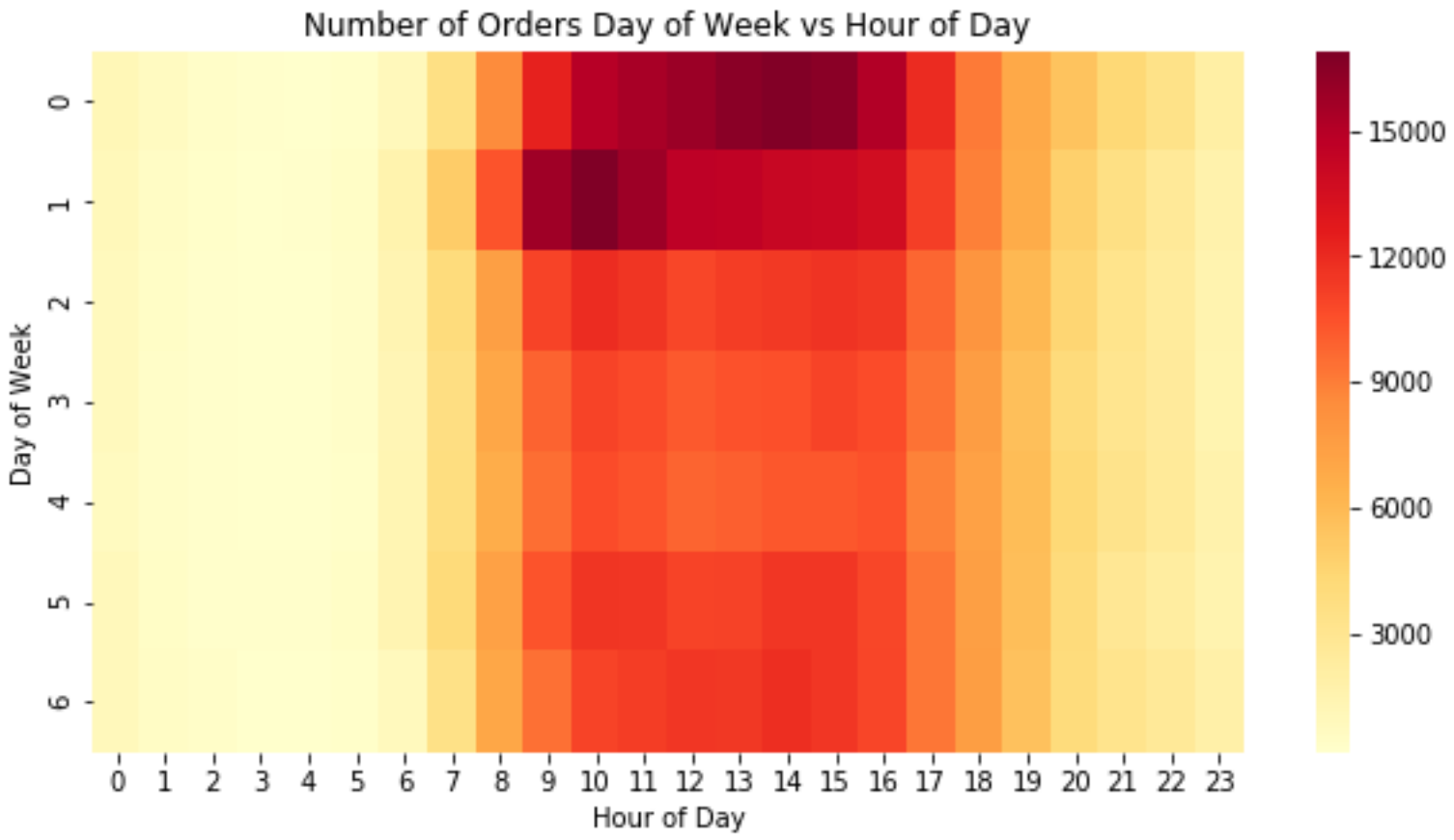
- Frequency of orders



THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Orders Information

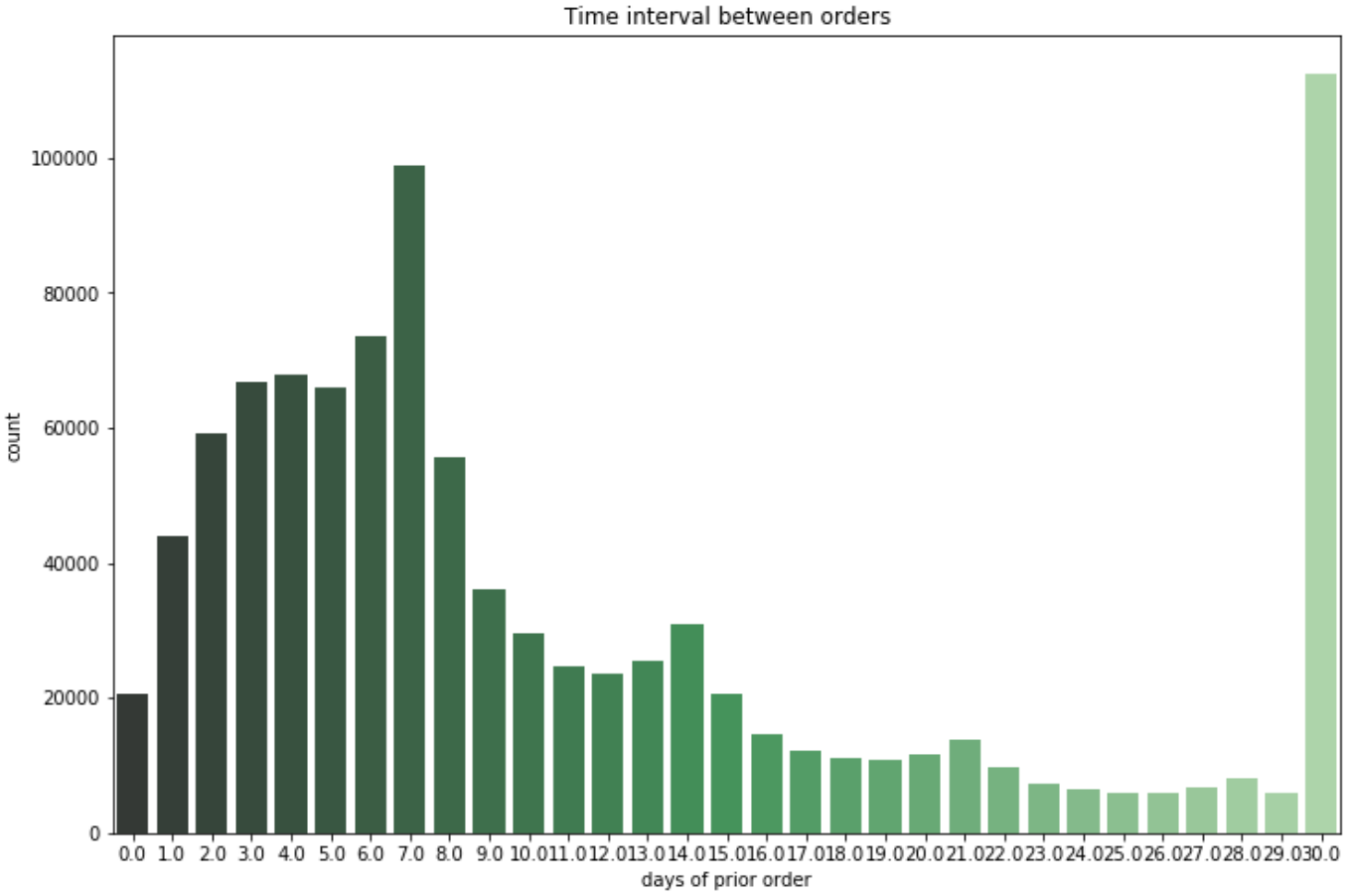
- Times of orders



THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Orders Information

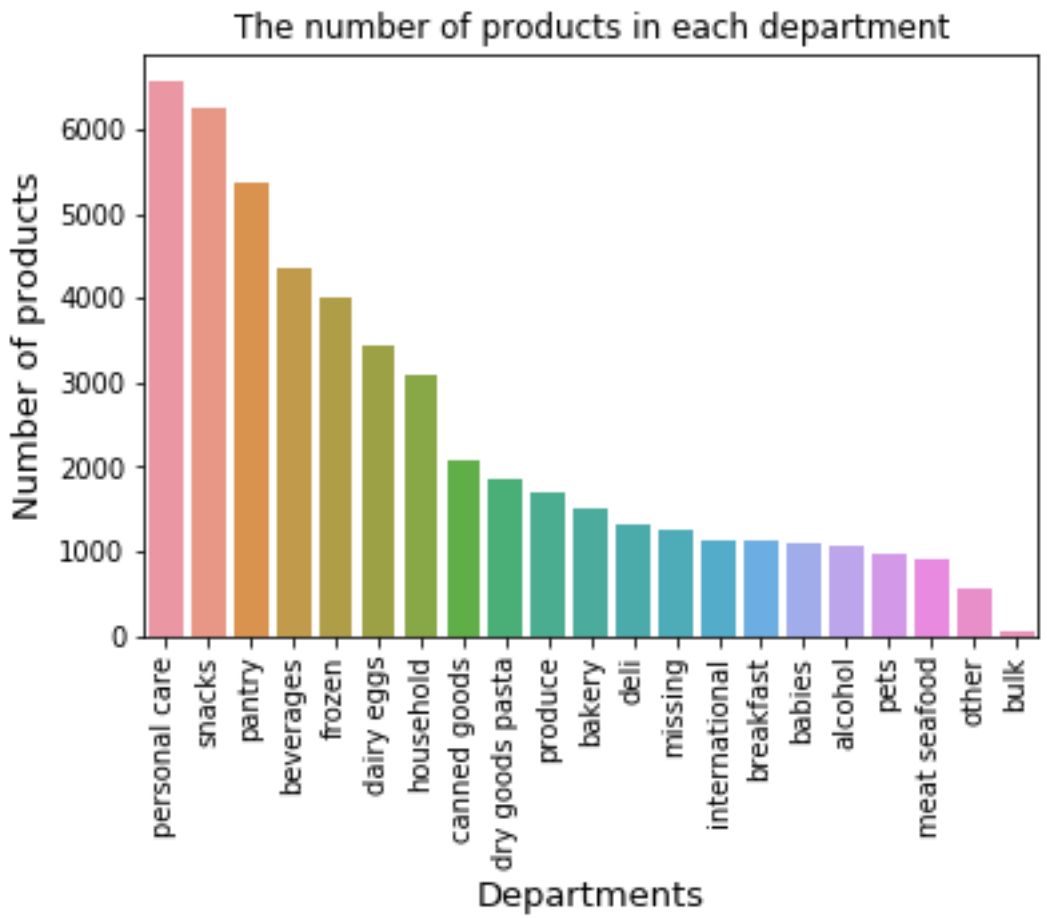
- Times of orders

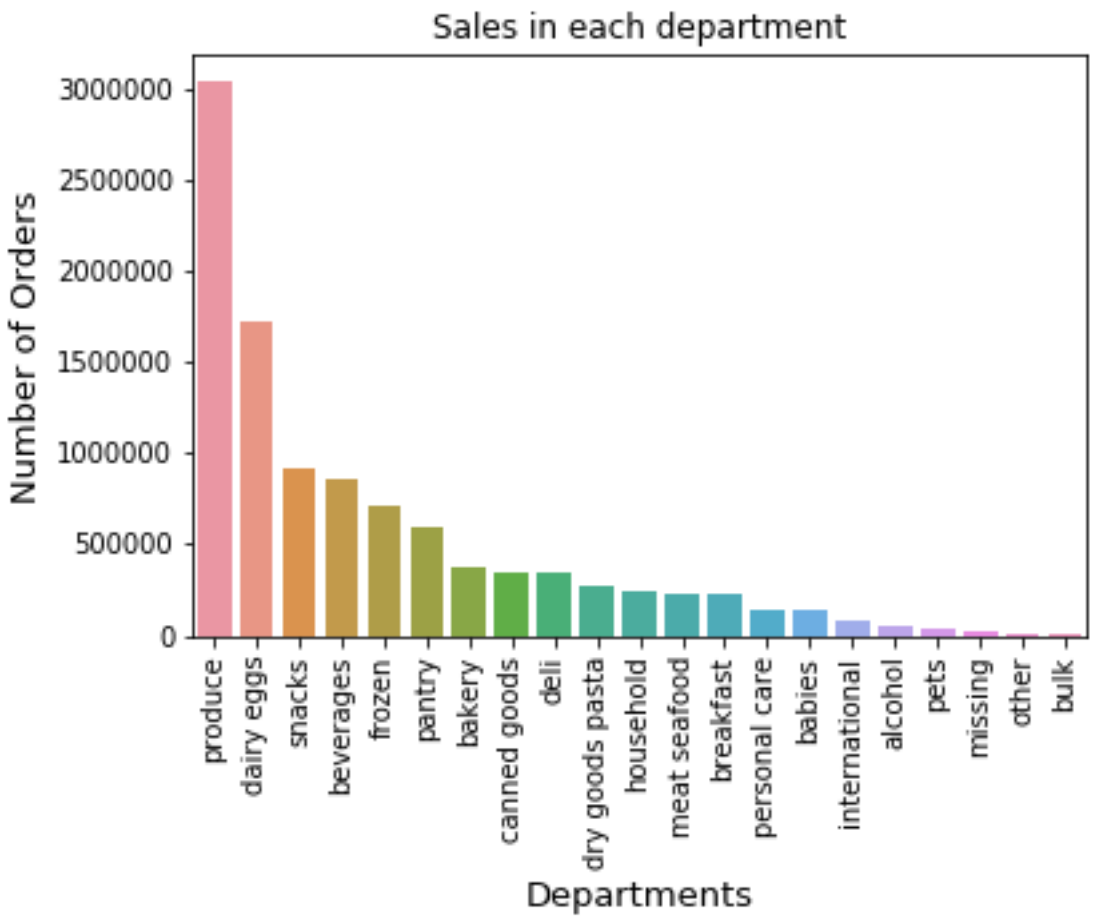


THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Product Information

- Products and Sales

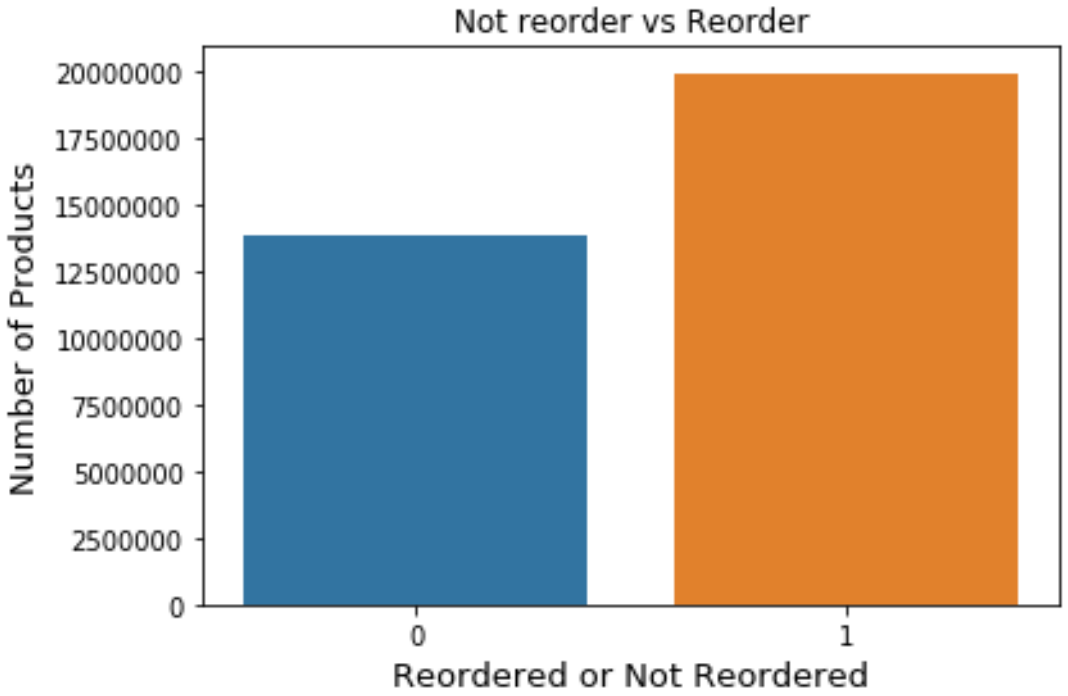


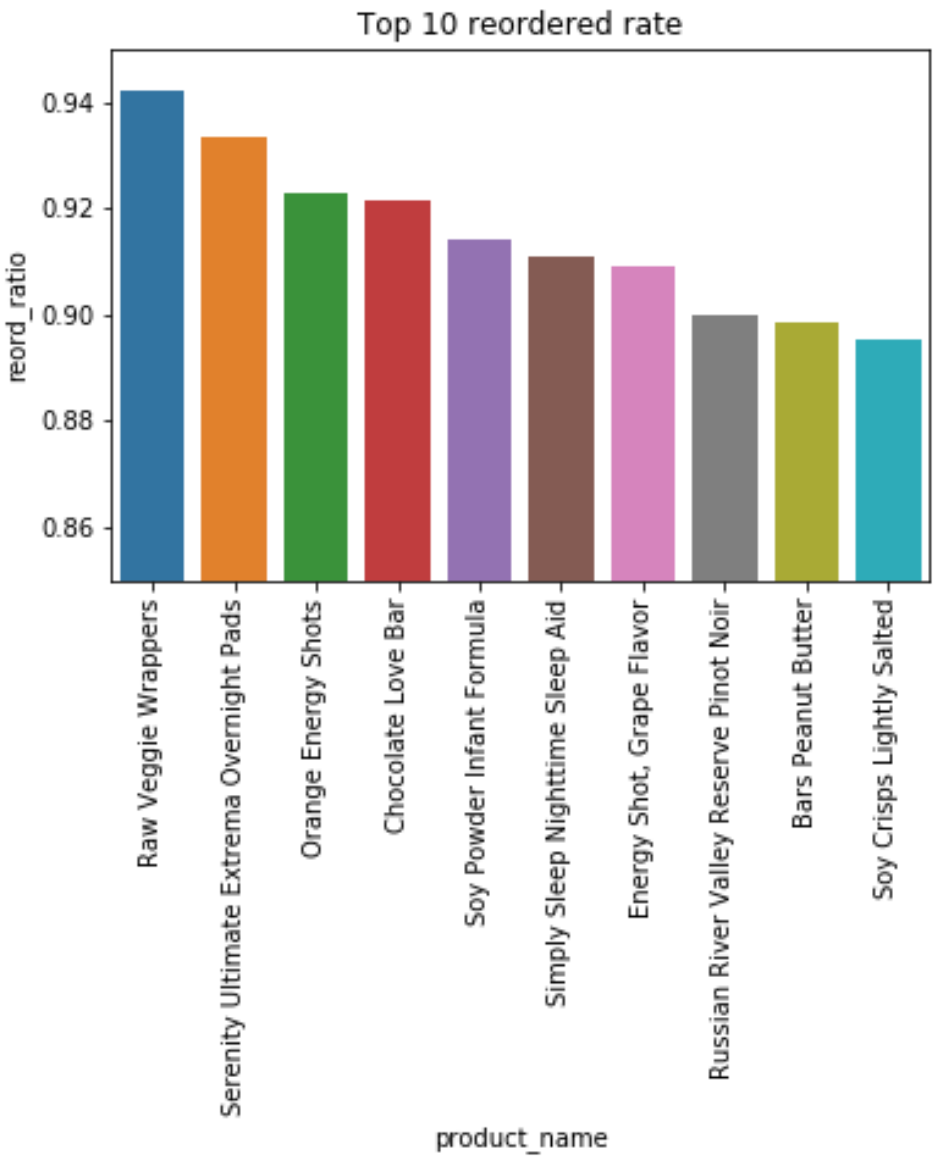


THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Product Information

- Reordered products

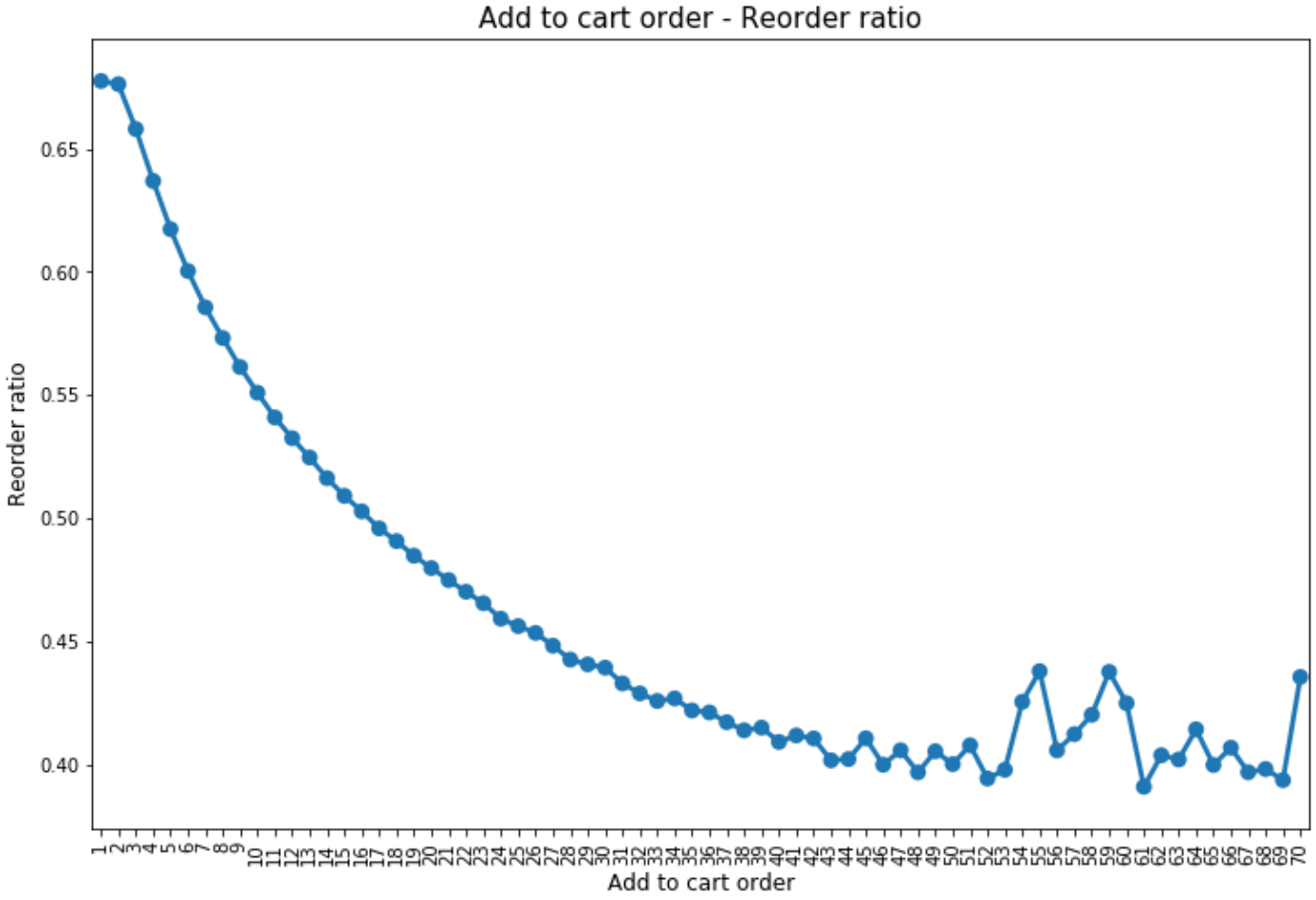




THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Product Information

- Add to cart-reorder ratio



GOAL: PREDICT WHICH PREVIOUSLY PURCHASED PRODUCTS WILL BE IN A USER'S NEXT ORDER

0 MEANS WILL NOT BUY, 1 MEANS WILL BUY

14 FEATURES WE CREATED

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Data Preprocessing

03

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Data Preprocessing

- Almost no missing value
- except for `days_since_prior_order` in orders.csv
  - 1048575 missing values
  - means these orders are first ordered by the user.
- replace NaN with -1 to indicate that it is a different level.

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Feature Engineering

04

FEATURE SELECTION

User Features

Product Features

User and Product Features

USER FEATURES

n_orders_users

The number of orders for each user

n_productss_users

The number of products for each user

avg_products_users

Average number of products per user ordered

dow_most_user

The day on which each user ordered most frequently

times_h

The time of a day on which each user ordered most frequently

reorder_ratio_user

Reordered ratio per user

shopping_freq

Shopping frequency for each user

USER FEATURES

	user_id	n_orders_users	n_products_users	avg_products_users	dow_most_user	hod_most_user	reorder_ratio_user	shopping_freq
0	1	10	59	5.900000	4	7	0.694915	17.600000
1	2	14	195	13.928571	2	9	0.476923	14.142857
2	3	12	88	7.333333	0	16	0.625000	11.083333
3	4	5	18	3.600000	4	15	0.055556	11.000000
4	5	4	37	9.250000	3	18	0.378378	10.000000

We combine all the user features above into a new data frame. ‘user_id’ is the key variable in this data frame.

PRODUCT FEATURES

times_bought_prod

Ordering frequency for each product.

reorder_ratio_prod

Reordered ratio for each product.

position_cart_prod

Average sequence in the cart for each product.

reorder_ratio_prod

Reordered ratio for each department

PRODUCT FEATURES

	product_id	times_bought_prod	reorder_ratio_prod	position_cart_prod	department_id	reorder_ratio_dept
0	1	1852	0.613391	5.801836	19	0.438319
1	2	90	0.133333	9.888889	13	0.242846
2	3	277	0.732852	6.415162	7	0.471714
3	4	329	0.446809	9.507599	1	0.418642
4	5	15	0.600000	6.466667	13	0.242846

Next, we merged these product features together. ‘product_id’ is the key variable in this data frame.

USER & PRODUCT FEATURES

times_bought_up

Times of each product bought by each user.

reorder_ratio_up

The ratio at which each product is reordered by each user.

ratio_last4_orders_up

The ratio of each product bought in each user's last four orders

USER & PRODUCT FEATURES

	user_id	product_id	times_bought_up	reorder_ratio_up	ratio_last4_orders_up
0	1	196	10	1.000000	1.0
1	1	10258	9	1.000000	1.0
2	1	10326	1	0.166667	NaN
3	1	12427	10	1.000000	1.0
4	1	13032	3	0.333333	0.5

Next, we merged these new features we just created

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Get Features and Target

Features

Target

	user_id	product_id	times_bought_up	reorder_ratio_up	ratio_last4_orders_up	n_orders_users	n_products_users	avg_products_users	dow_most_user	hod_most_user	reorder_ratio_user	shopping_freq	times_bought_prod	reorder_ratio_prod	position_cart_prod	reorder_ratio_dept	reordered
0	1	196	10	1.000000	1.00	10	59	5.900000	4	7	0.694915	20.259259	35791	0.776480	3.721774	0.471714	1.0
1	1	10258	9	1.000000	1.00	10	59	5.900000	4	7	0.694915	20.259259	1946	0.713772	4.277492	0.438319	1.0
2	1	10326	1	0.166667	0.00	10	59	5.900000	4	7	0.694915	20.259259	5526	0.652009	4.191097	0.412660	0.0
3	1	12427	10	1.000000	1.00	10	59	5.900000	4	7	0.694915	20.259259	6476	0.740735	4.760037	0.438319	0.0
4	1	13032	3	0.333333	0.50	10	59	5.900000	4	7	0.694915	20.259259	3751	0.657158	5.622767	0.466878	1.0
5	1	13176	2	0.222222	0.00	10	59	5.900000	4	7	0.694915	20.259259	379450	0.832555	5.095947	0.412660	0.0
6	1	14084	1	0.100000	0.00	10	59	5.900000	4	7	0.694915	20.259259	15935	0.810982	5.792595	0.505622	0.0
7	1	17122	1	0.166667	0.00	10	59	5.900000	4	7	0.694915	20.259259	13880	0.675576	6.257421	0.412660	0.0
8	1	25133	8	1.000000	1.00	10	59	5.900000	4	7	0.694915	20.259259	6196	0.740155	7.001614	0.505622	1.0
9	1	26088	2	0.200000	0.00	10	59	5.900000	4	7	0.694915	20.259259	2523	0.539041	6.495838	0.438319	1.0
10	1	26405	2	0.200000	0.00	10	59	5.900000	4	7	0.694915	20.259259	1214	0.441516	3.116969	0.250641	1.0

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Model Development and Evaluation

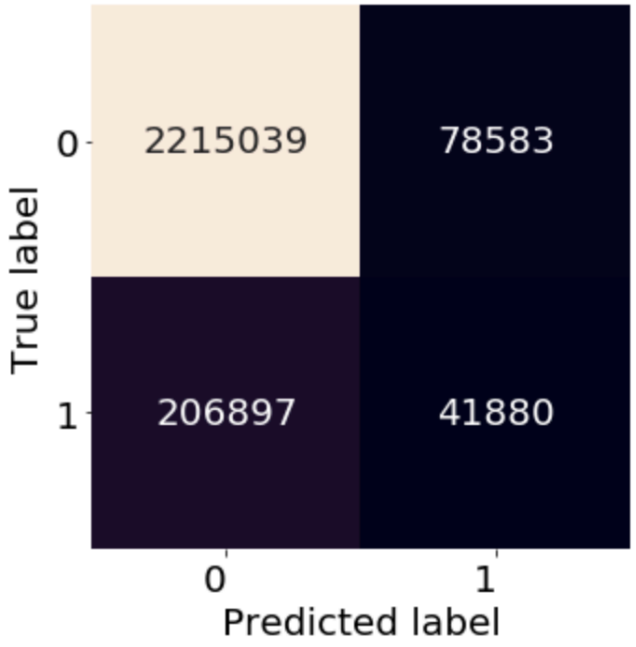
05

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

A Problem Happens

- When developing logistic regression, we found that the model accuracy is high, but the confusion matrix shows it is not a good model.

```
Classification Report:
              precision    recall  f1-score   support

         0.0       0.91      0.99      0.95   1529168
         1.0       0.62      0.11      0.19    165765

    accuracy                           0.91   1694933
   macro avg       0.77      0.55      0.57   1694933
weighted avg       0.88      0.91      0.88   1694933


Accuracy :  90.6327860747298


ROC_AUC :  79.50233767175897
```



THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Oversampling

- Target ‘reordered’ is a binary variable. Level 0 accounts for 90%, while level 1 only accounts for 10%.
- Even without modeling, we can have 90% accuracy.

target	frequency
0	2,334,883
1	253,930

target	frequency
0	2,334,883
1	2,334,883

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Train and Test Split

- Data shape (4669766, 15)

- Train dataset 70%, test dataset 30%

- Train dataset shape (3268836, 15)
- Test dataset shape (1400930, 15)

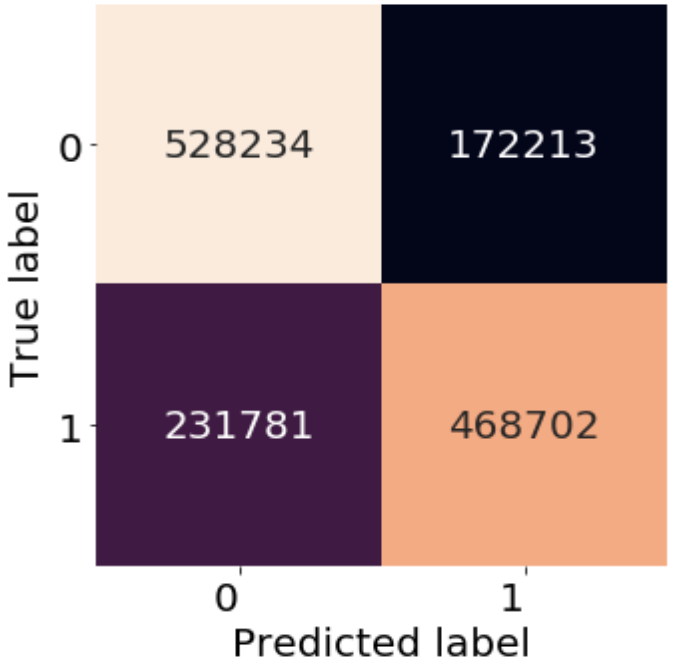
- We use train dataset to train the model, and test dataset to predict.

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Logistic Regression

- Misclassification rate = $\frac{False\ Positives + False\ Negatives}{Total\ instances} = \frac{172213+231781}{1400930}$ = 28.84%.

```
Classification Report:
              precision    recall  f1-score   support

         0.0       0.70      0.75      0.72    700447
         1.0       0.73      0.67      0.70    700483

    accuracy                           0.71   1400930
   macro avg       0.71      0.71      0.71   1400930
weighted avg       0.71      0.71      0.71   1400930


Accuracy :  71.16244209203886


ROC_AUC :  77.71164554475179
```

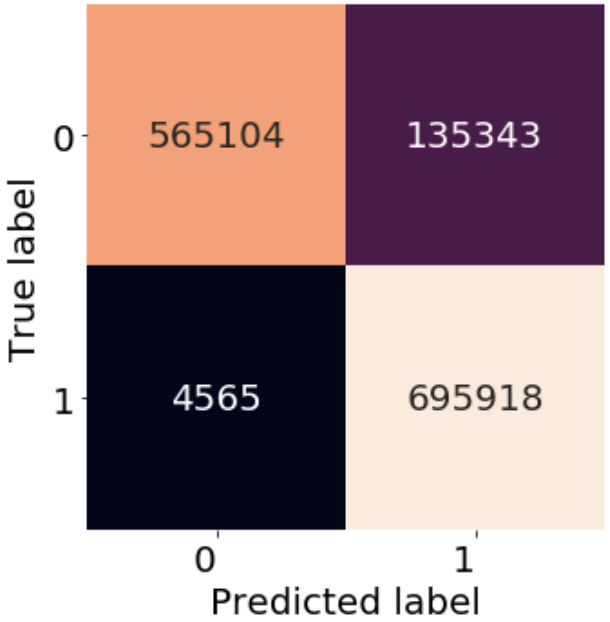


THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

K-Nearest-Neighbor

- K = 3
- Misclassification rate = 9.99%.

```
Classification Report:
              precision    recall  f1-score   support

         0.0       0.99      0.81      0.89    700447
         1.0       0.84      0.99      0.91    700483

    accuracy                           0.90   1400930
   macro avg       0.91      0.90      0.90   1400930
weighted avg       0.91      0.90      0.90   1400930


Accuracy :  90.01320551348033


ROC_AUC :  99.97228979749896
```

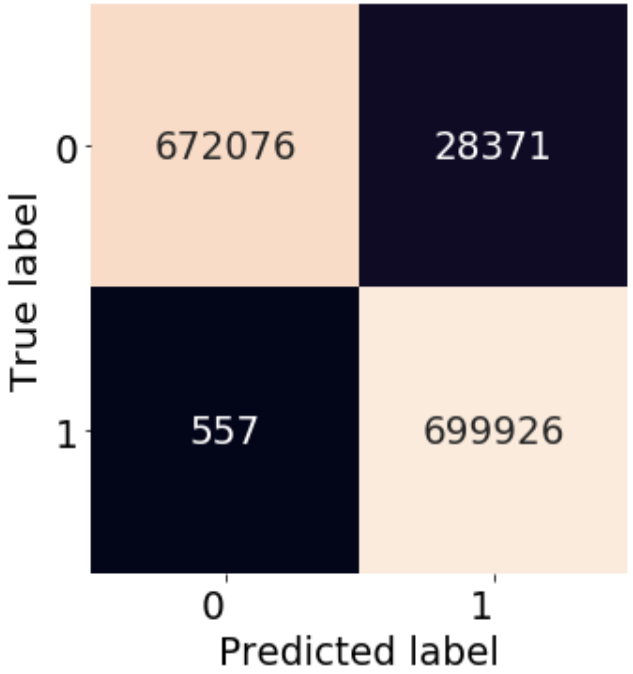


THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Random Forest

- n_estimators=100
- Misclassification rate = 2.06%.

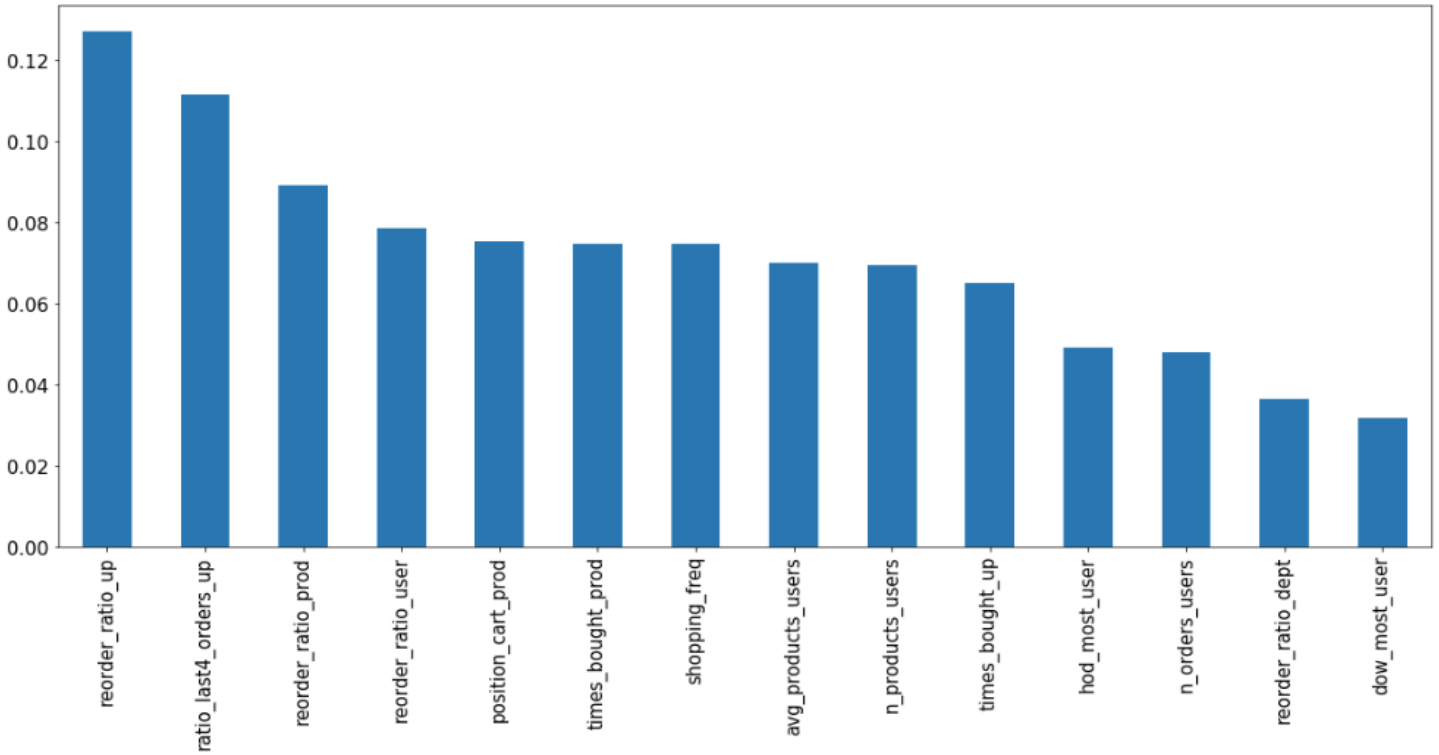
```
Classification Report:
              precision    recall  f1-score   support

         0.0       1.00      0.96      0.98    700447
         1.0       0.96      1.00      0.98    700483

    accuracy                           0.98   1400930
   macro avg       0.98      0.98      0.98   1400930
weighted avg       0.98      0.98      0.98   1400930


Accuracy :  97.93508597859993


ROC_AUC :  99.97318652685728
```



THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Random Forest

- Feature importance plot



THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

06

Summary

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Summary

Goal

- Predict whether a user will buy a product in the next order.

Process

- We create 14 features and merge them all into one dataset.
- Then, we use oversampling to deal with the imbalance of the dataset.
- After splitting train and test dataset, we train logistic regression, KNN, random forest and other classifiers.

Improvements

- Hyperparameter tuning
- Gradient Boosting

Thanks for listening