Capgemini

User Recommendation Engine & Market Basket Analysis

March 22, 2020

Xuechun Wang

Link: https://github.com/XuechunWang/Cap_Recommendation

User Recommendation Engine & Market Basket Analysis

Capgemini

01 Description of the Problem and Data Sets

02 Insights and Exploratory Analysis

03 Feature Engineering and Clustering Analysis

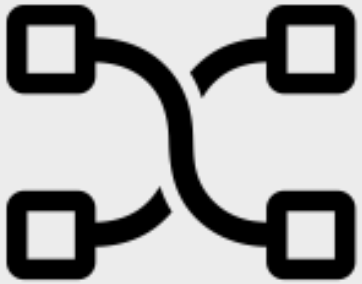
04 Recommendation Engine Modeling and Diagnostics

05 Technology Used and Future Work

Description of the **Problem** and Data Sets

Transaction Data

Data Processing



Data Discovery

Prediction Model
on which product a user will buy again

Description of the Problem and **Data Sets**

Missing Values and Outliers: None

Orders

Order Products
(prior, train, test)
30+m rows

- order_id
- product_id
- add_to_cart_order
- reordered

User Orders
3.4m rows
206K users

- order_id
- user_id
- Eval_set
- order_number
- order_dow
- order_hour_of_day
- days_since_prior

Products
50km rows

- product_id
- product_name
- aisle_id
- department_id

aisles
134 rows

- aisle_id
- aisle

departments
21 rows

- department
- department_id

Insights

and Exploratory Analysis

Insights and Exploratory Analysis - Understanding the Products

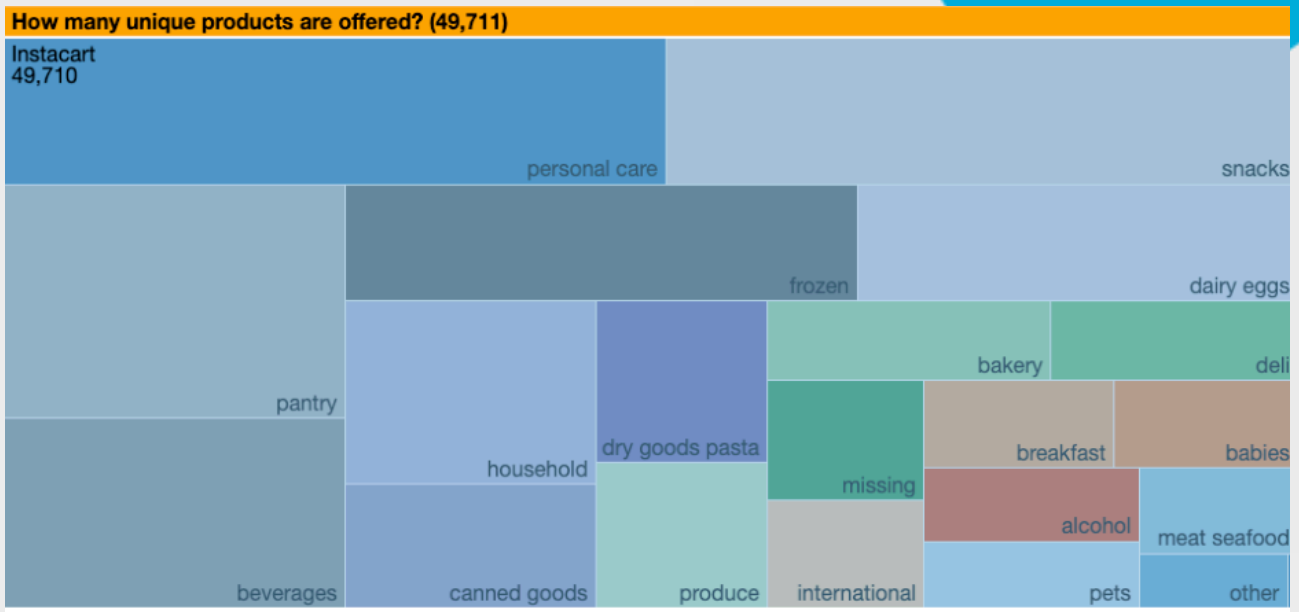
Top 3 Departments

How many unique products:

Personal care

Snacks

Pantry

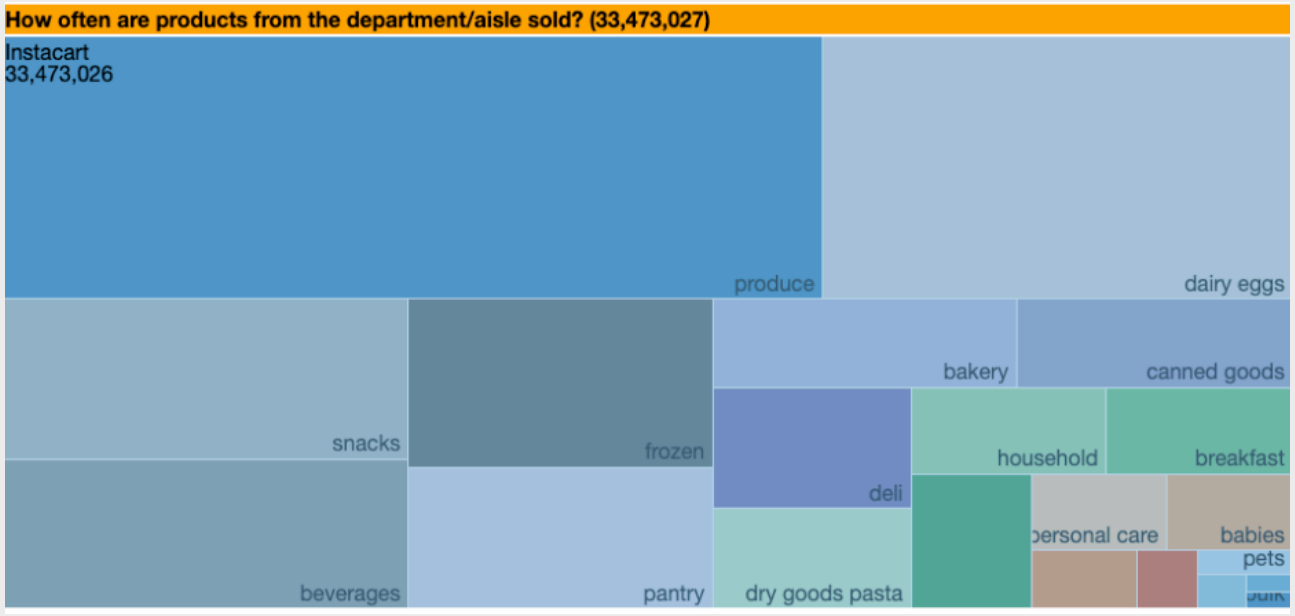


How often purchased:

Produce

Dairy eggs

Snacks



Insights and Exploratory Analysis – Never buy items

Products that are never sold:

	product_id	product_name	aisle_id	department_id
3629	3630	Protein Granola Apple Crisp	57	14
3717	3718	Wasabi Cheddar Spreadable Cheese	21	16
7044	7045	Unpeeled Apricot Halves in Heavy Syrup	88	13
37702	37703	Ultra Sun Blossom Liquid 90 loads Fabric Enhan...	75	17
43724	43725	Sweetart Jelly Beans	100	21
45970	45971	12 Inch Taper Candle White	101	17
46624	46625	Single Barrel Kentucky Straight Bourbon Whiskey	31	7

Capgemini

Insights and Exploratory Analysis - Most frequently buy items

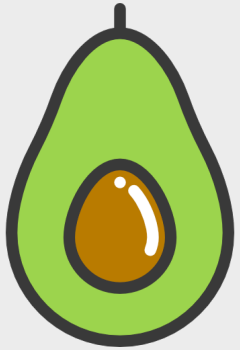
Top List:

1. 
Banana

2. 
Strawberry

3. 
Lemon


spinach

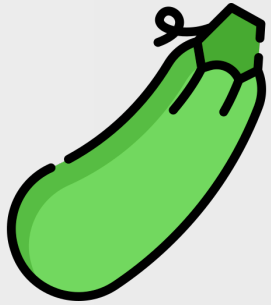

avocado

milk


raspberry


onion

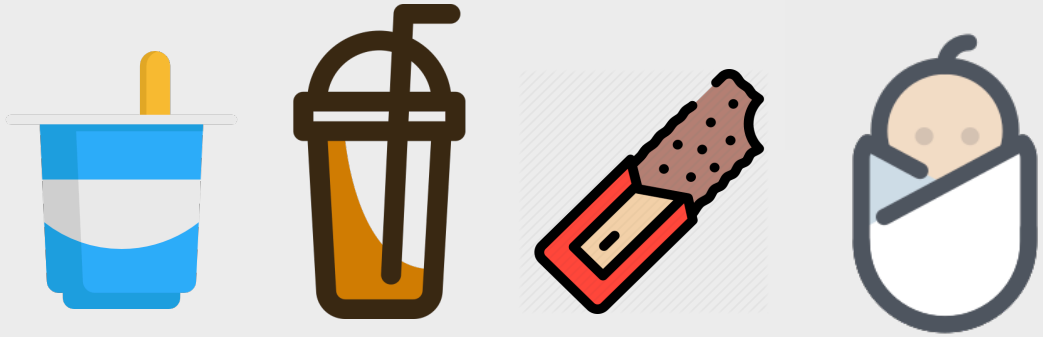

garlic


zucchini

Insights and Exploratory Analysis - Most frequently buy items

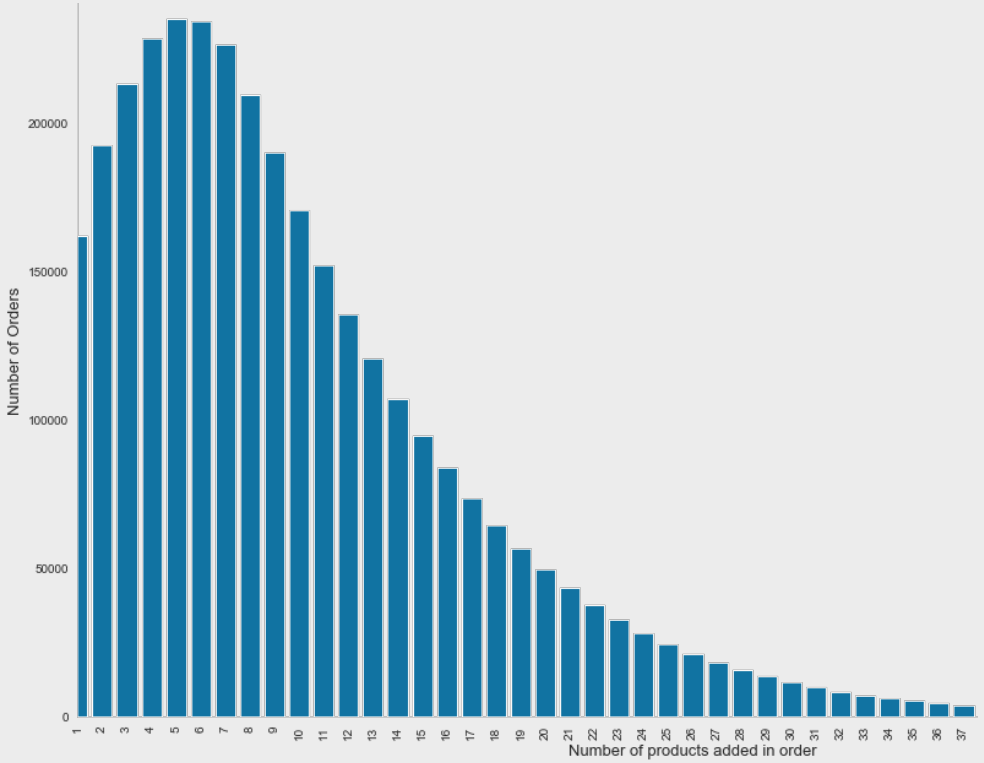
From Association role Analysis - **Apriori Algorithm**

- One flavor of an item being purchased with another flavor from the same family being added to the order
  - Especially on **Yogurt, Beverage, and Food Bar, Baby Food**

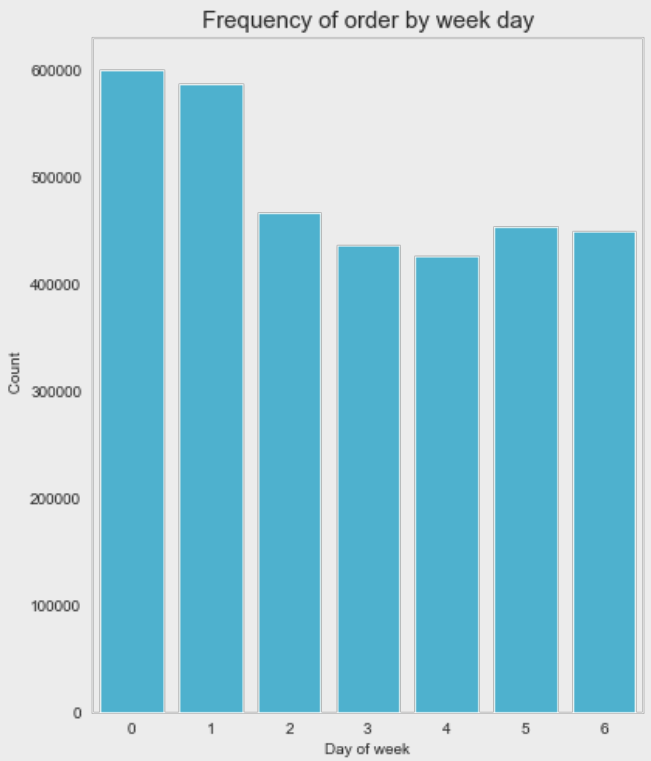


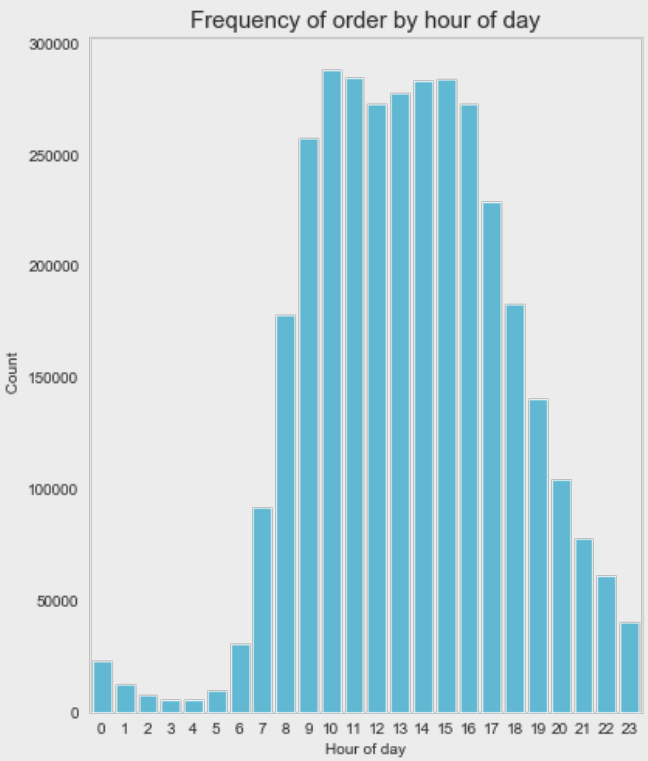
- Recognized some common patterns:
  - Ex. **Moisture Conditioner - Hair Shampoos**

Insights and Exploratory Analysis - Understanding the orders



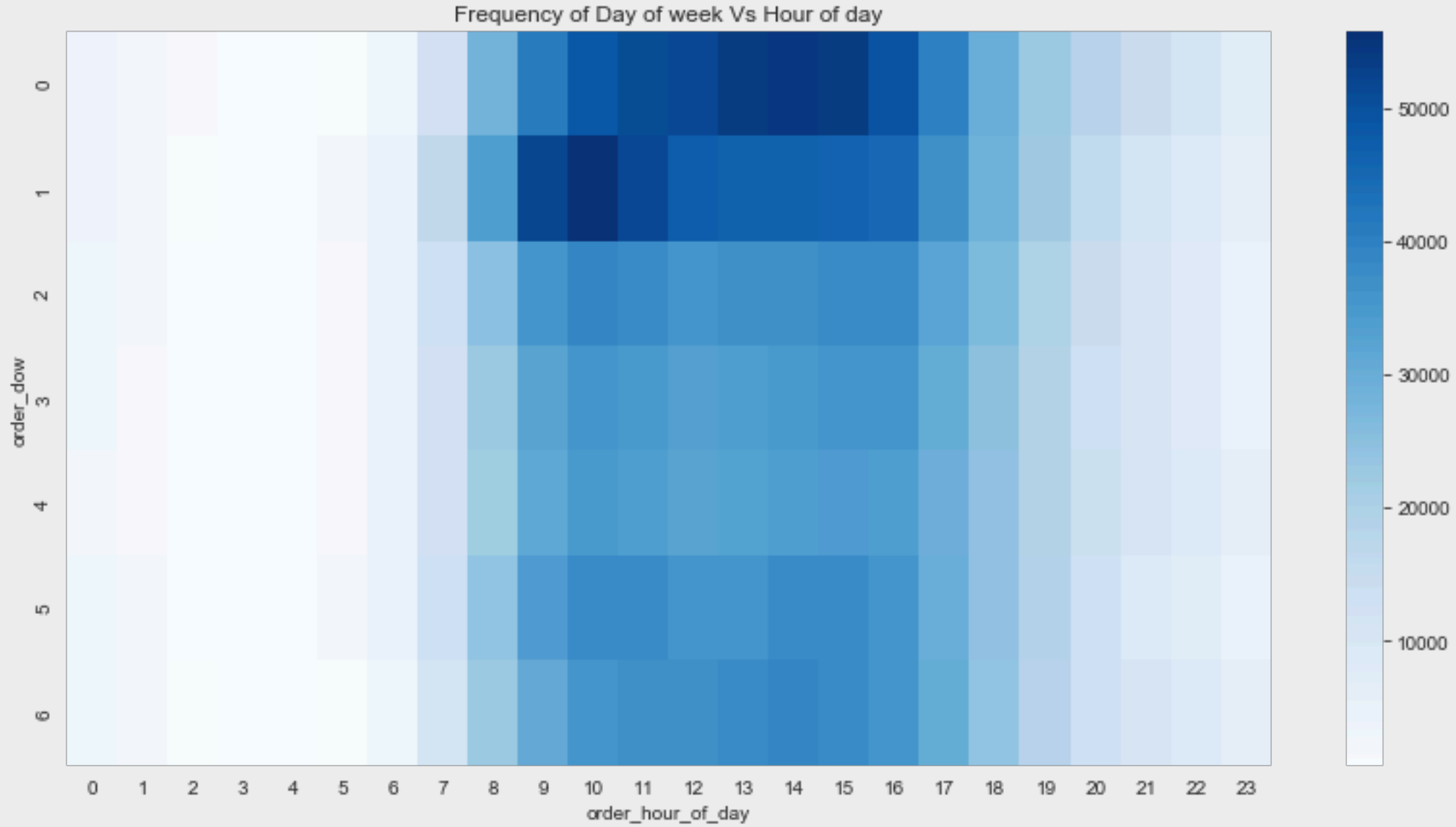
Distribution of number of products per order
Similar to Poisson distribution



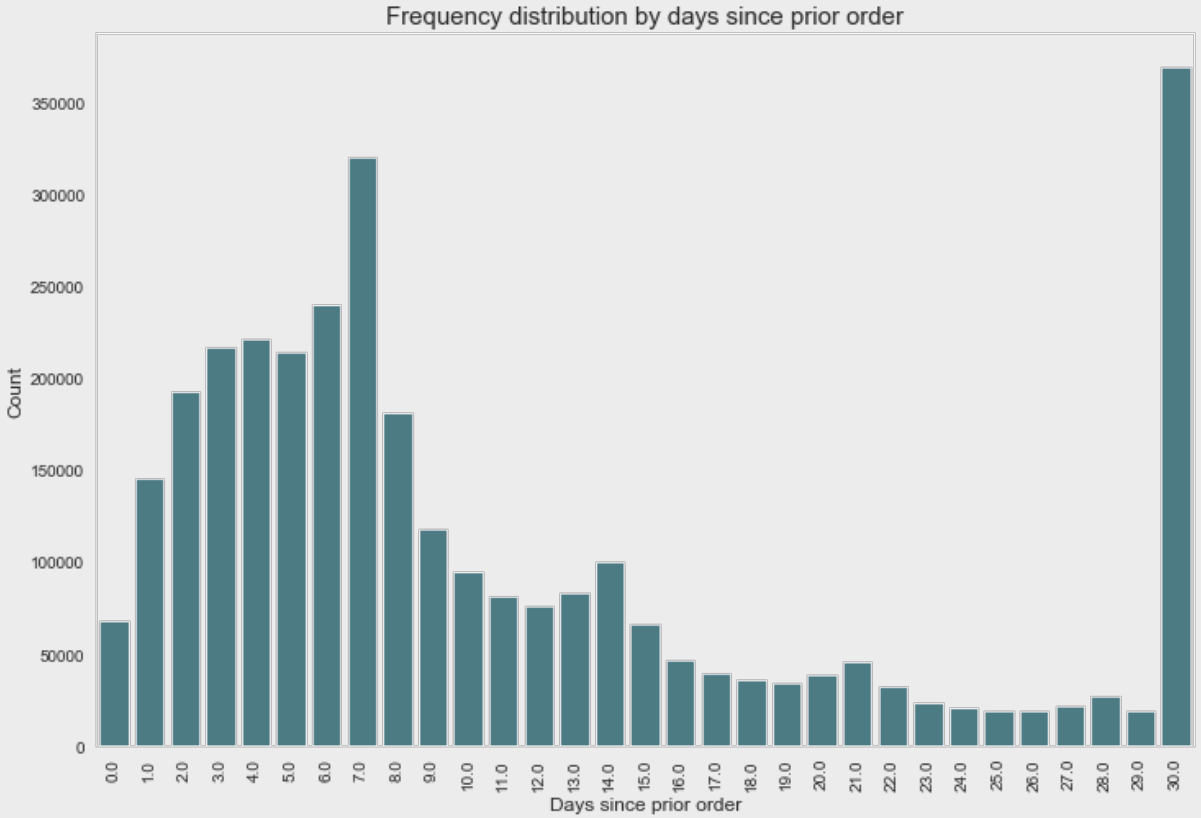


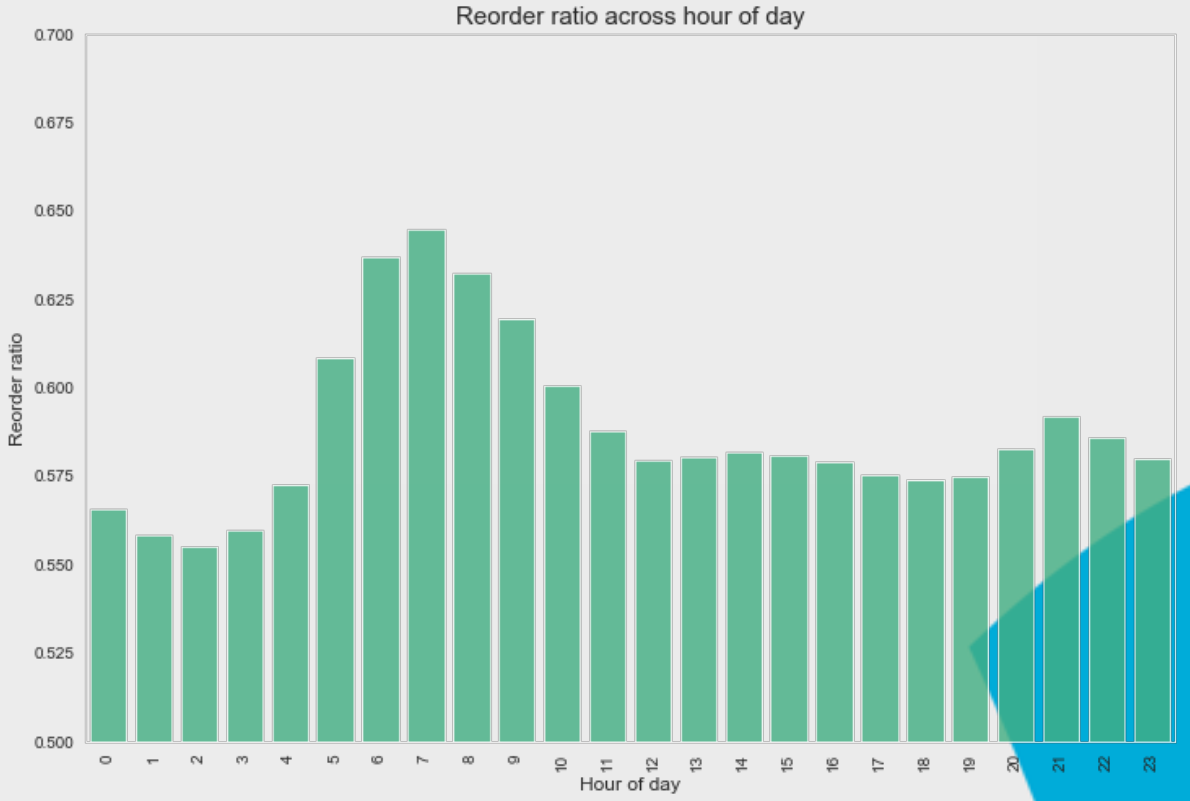
Sunday and **Monday** have relatively more orders during a week
Most of orders occur during the time range 9-17 of a day

Insights and Exploratory Analysis - Understanding the orders



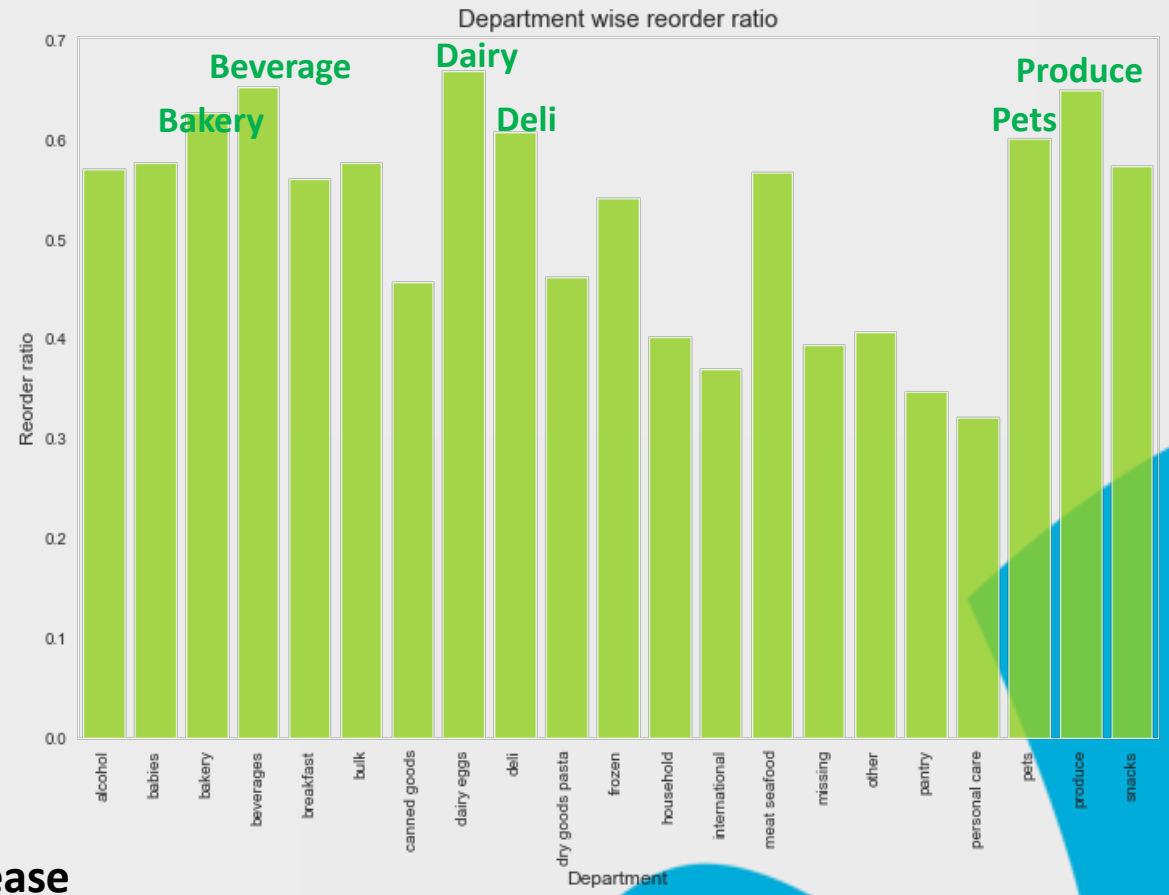
Insights and Exploratory Analysis - Understanding the reorders





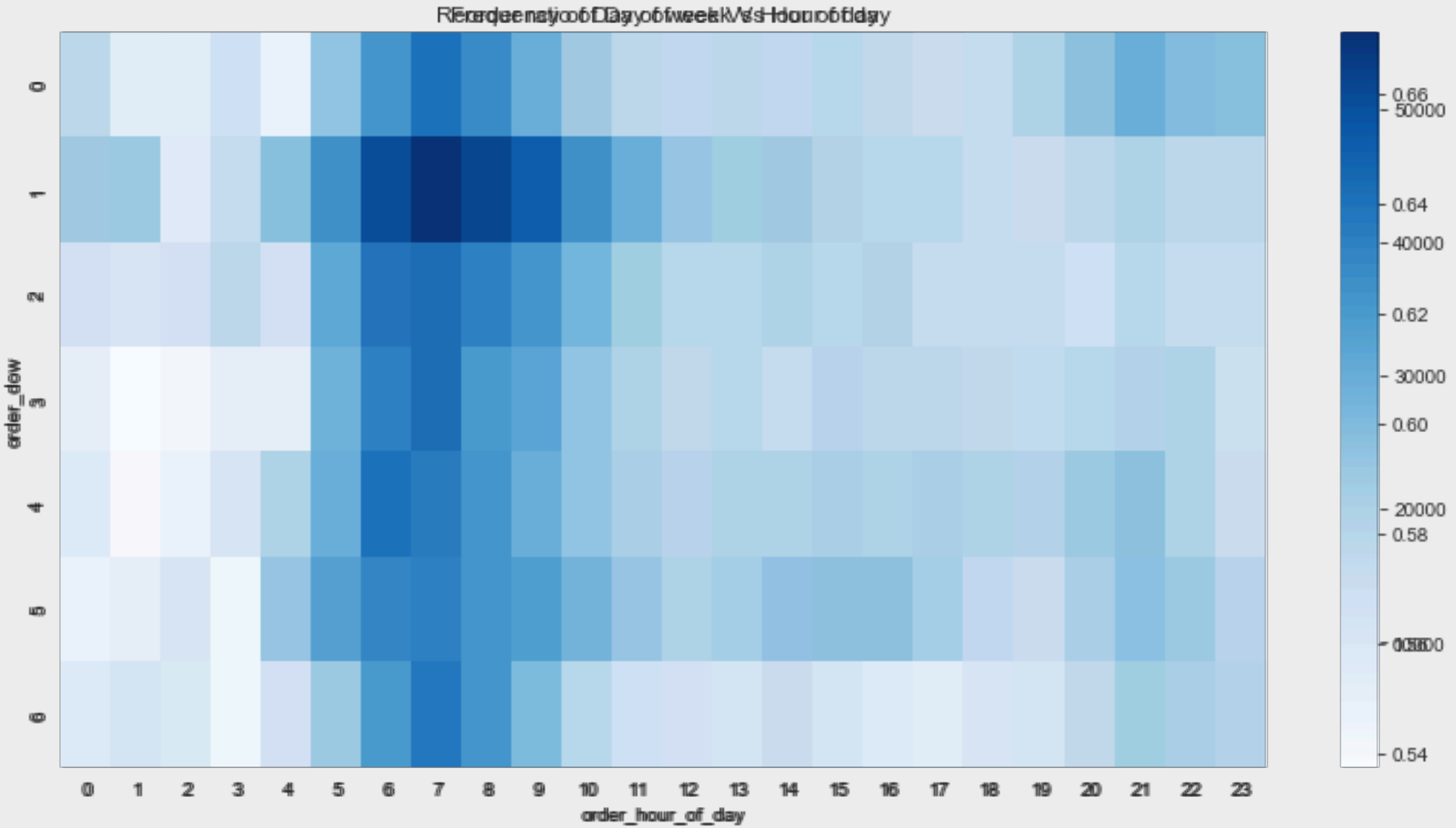
The average times user buy a product is 3.31
The maximum times product is 98

Insights and Exploratory Analysis - Understanding the reorders



Reorder ratio decrease as Add to cart order position increase

Insights and Exploratory Analysis - Understanding the reorders



Feature Engineering
& Clustering Analysis

Feature Engineering and Clustering Analysis

Use Merge and Group by agg function on order_product data to create new features:

Products

- product_total
- product_reorder
- product_first_order
- product_second_order
- product_third_order
- product_reorder_pro
- product_triorder_pro
- product_reorder_ratio
- product_reorder_times

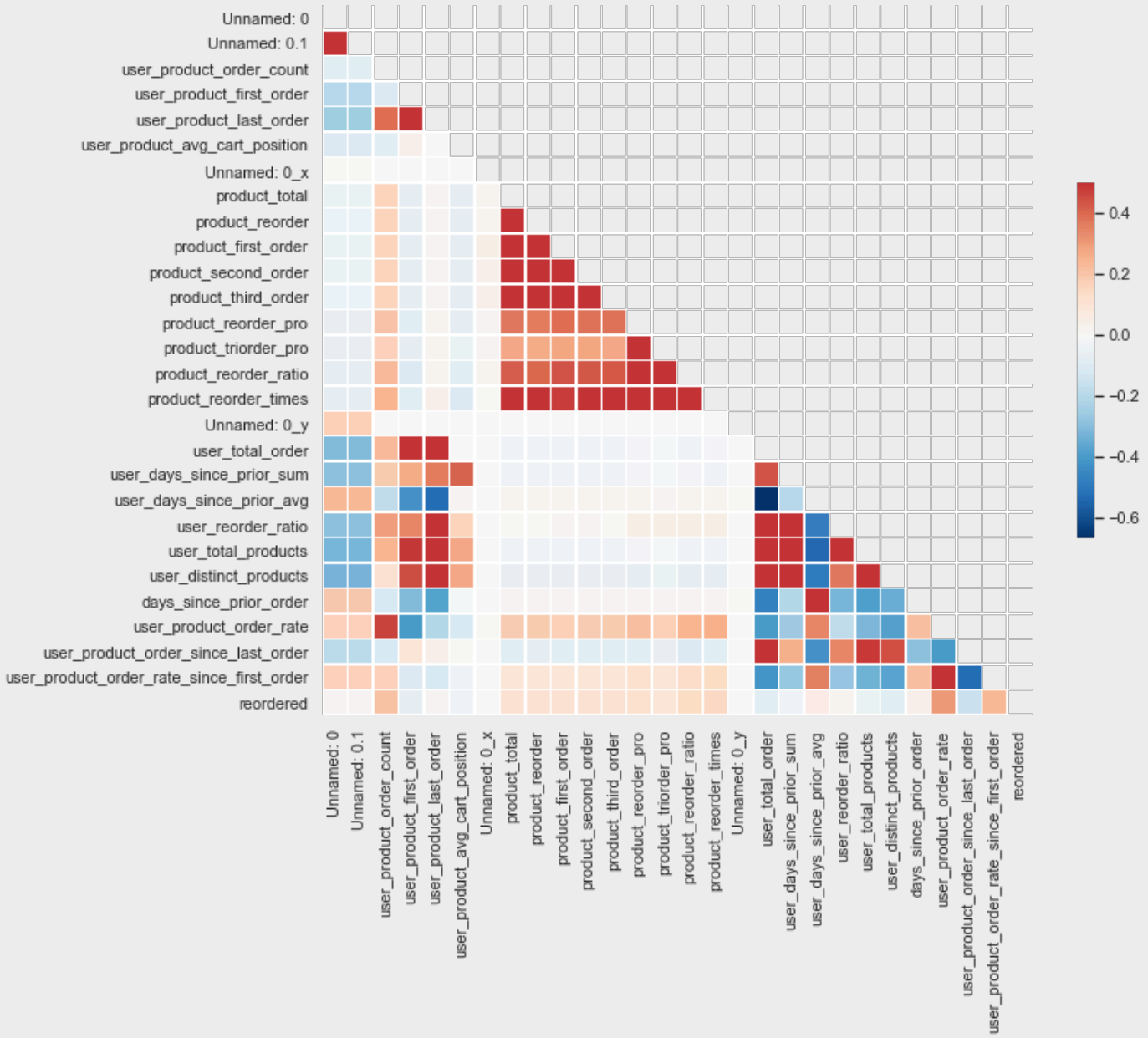
Users

- user_total_order
- user_days_since_prior_sum
- user_days_since_prior_avg
- user_reorder_ratio
- user_total_products
- user_distinct_products

Users and Products

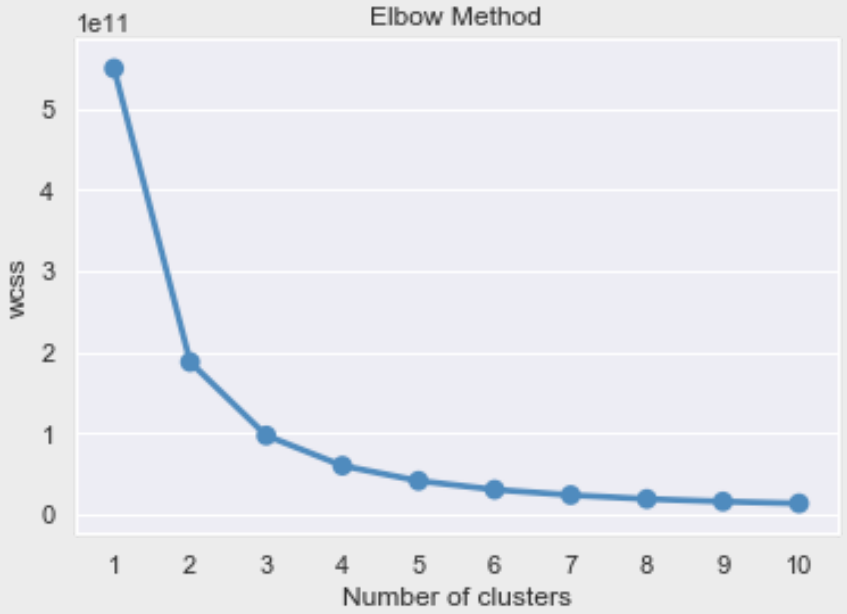
- user_product_order_count
- user_product_first_order
- user_product_last_order
- user_product_avg_cart_position
- user_product_order_rate
- user_product_order_since_last_order
- user_product_order_rate_since_first_order

Feature Engineering and Clustering Analysis

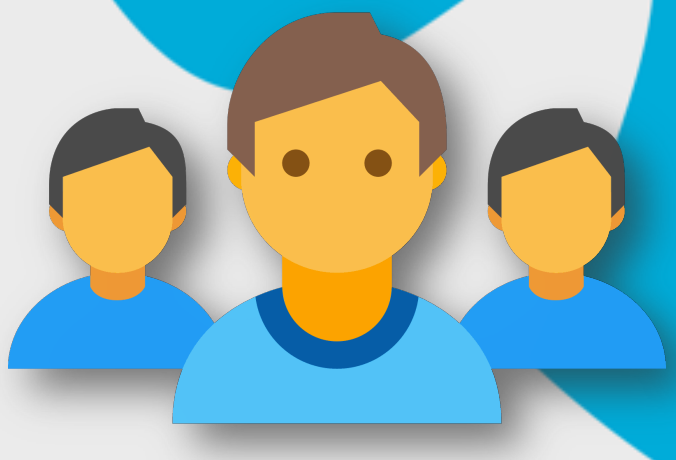


Feature Engineering and Clustering Analysis

Optimal Cluster Number: 4



- user_total_order
- user_days_since_prior_avg
- user_reorder_ratio
- user_total_products
- user_distinct_products



Cluster	Measure 1	Measure 2	Measure 3	Measure 4	Measure 5
Most Active	37.28	11.44	0.73	759.99	199.17
Active	31.05	12.98	0.64	393.18	135.84
Regular	19.82	15.31	0.53	176.51	79.30
Less Active	8.31	16.42	0.42	52.30	31.24

Recommendation Engine Modeling & Diagnostics

Recommendation Engine Modeling and Diagnostics

01 Logistic Regression

02 Lasso Regression

03 XGBoost Model

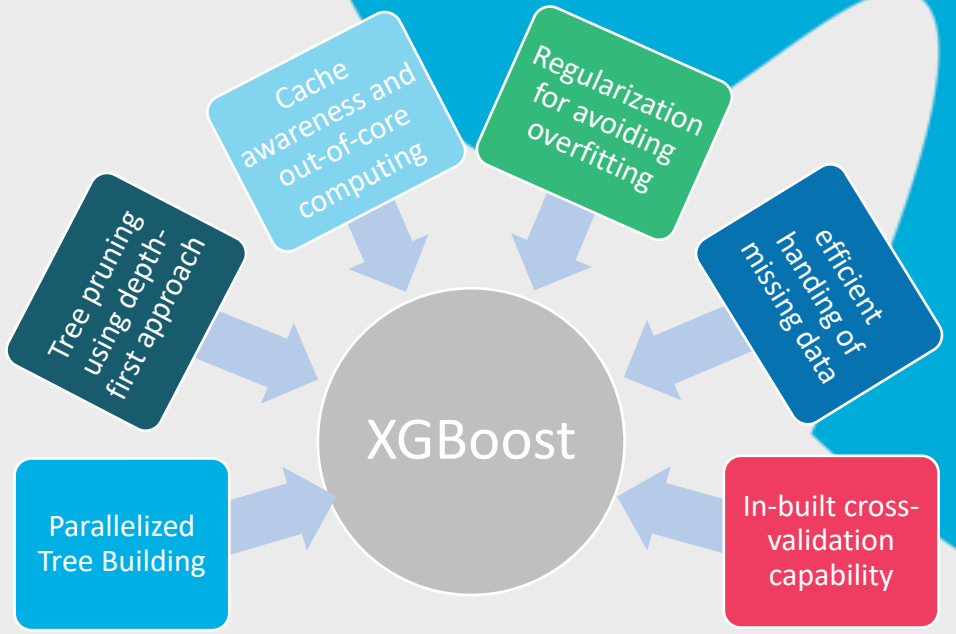
04 Optimize threshold

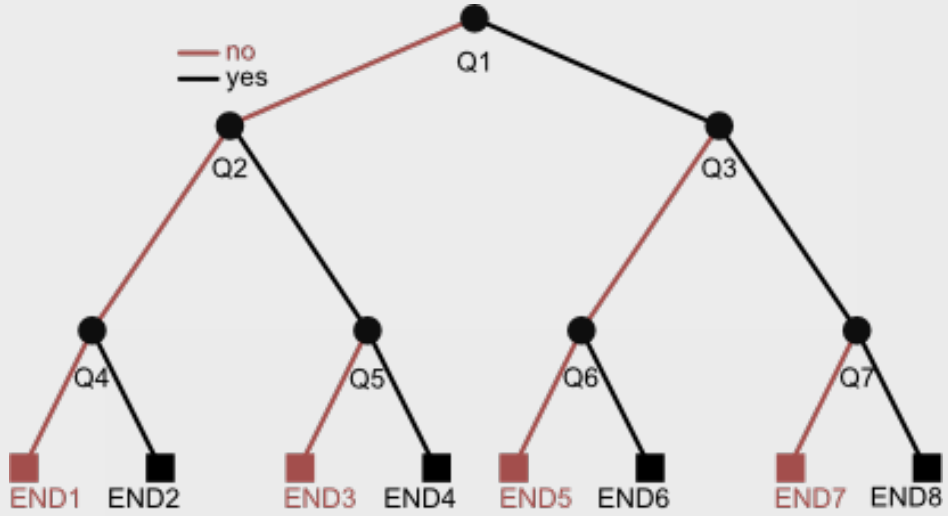
05 Diagnostics

Recommendation Engine Modeling and Diagnostics

XGBoost Parameters :

objective	reg:logistic	
eval_metric	logloss	
eta	0.1	Step size shrinkage used in update to prevents overfitting
max_depth	6	Maximum depth of a tree
min_child_weight	10	
gamma	0.70	Minimum loss reduction required to make a further partition on a leaf node of the tree.
subsample	0.76	randomly sample before train to prevent overfitting
colsample_bytree	0.95	subsample ratio of columns when constructing each tree
alpha	2e-05	L1 regularization term on weights.
lambda	10	L2 regularization term on weights.





Recommendation Engine Modeling and Diagnostics

The shape of train data: (1038515, 4)
The shape of test data: (346102, 2)
The shape of prior data: (32434489, 4)

Data Balance

{**0**: **7852735**, **1**: **621926**}

Threshold – 0.5

	False (0)	True (1)
Negative	6255140	27000
Positive	461230	36359

Accuracy: 92.799%
F-1 Score: 0.1296

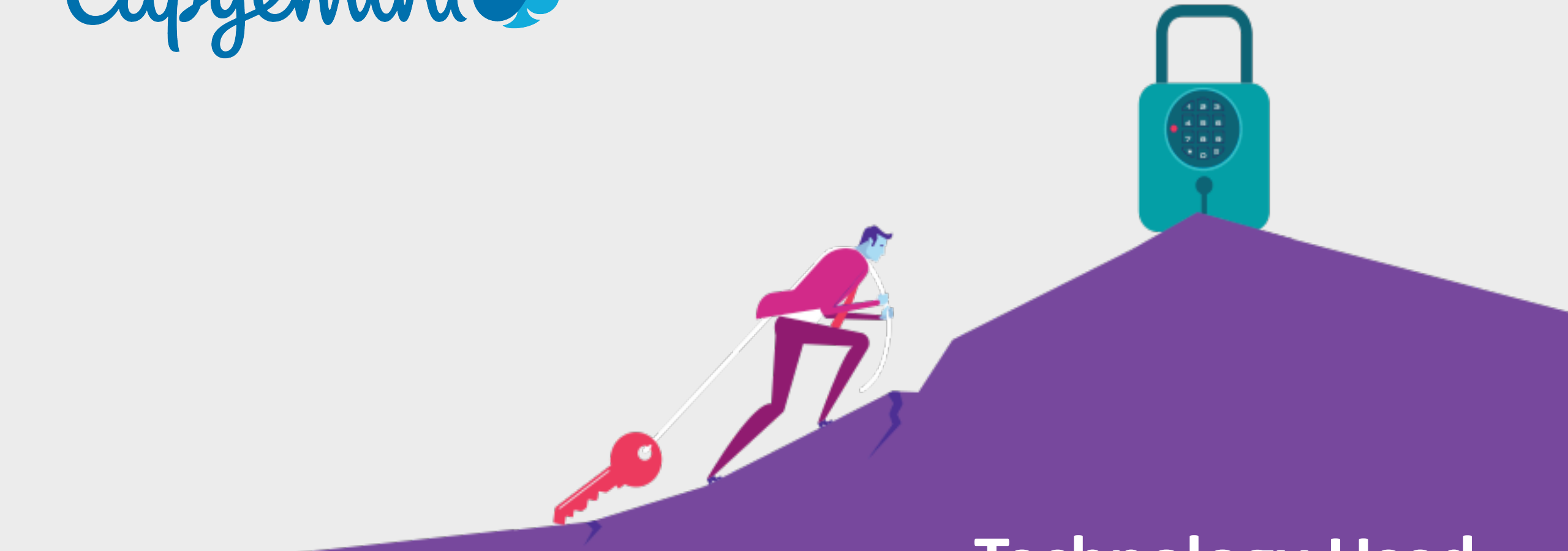
Optimal Threshold for ROC Curve: Best Threshold=0.073372
Optimal Threshold for Precision-Recall Curve: Best Threshold=0.184227
Optimal Threshold Tuning on F-1 Score: Best Threshold= 0.190

Threshold – 0.19

	False (0)	True (1)
Negative	5823556	458584
Positive	276118	221471

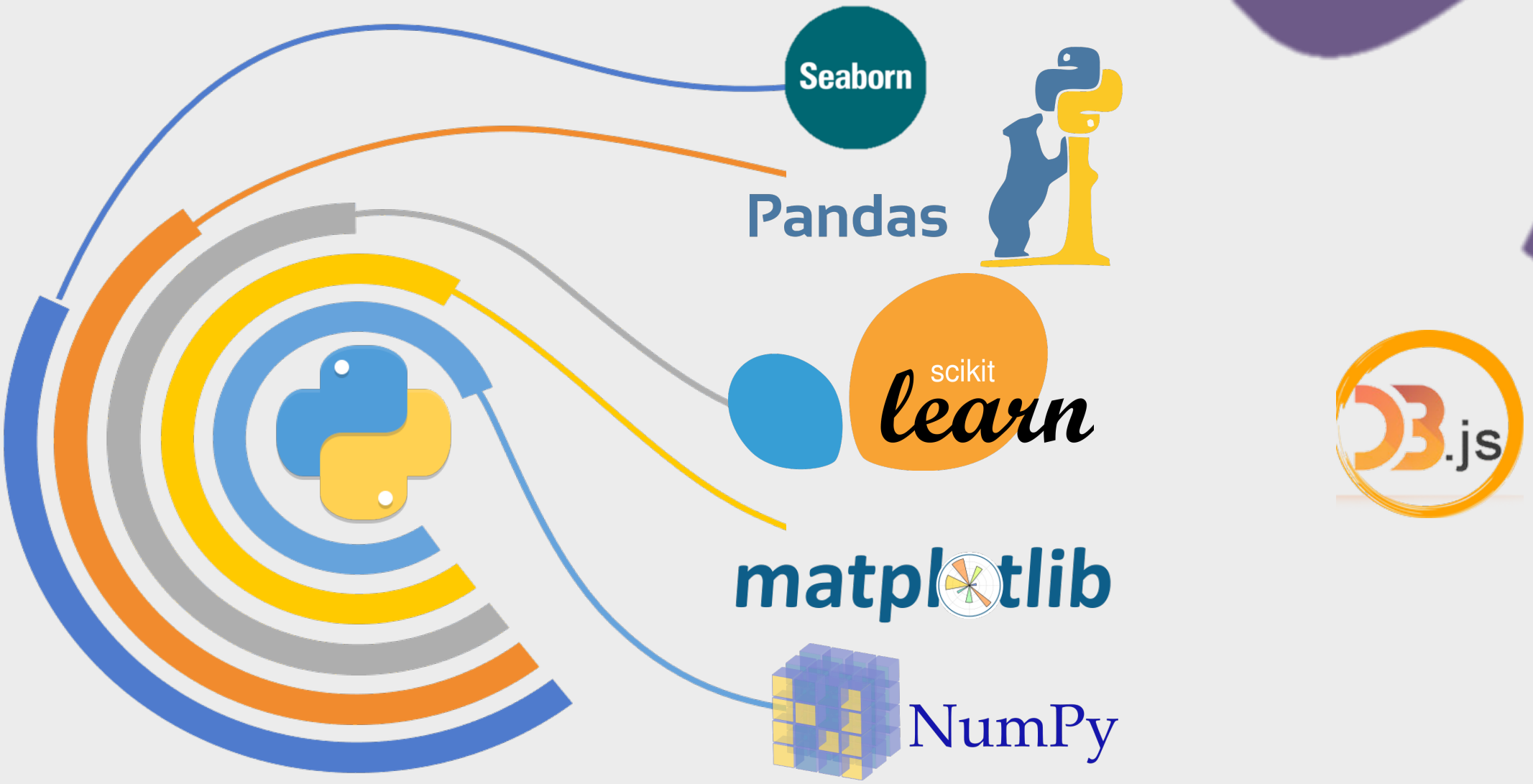
Accuracy: 89.163%
F-1 Score: 0.3761

Capgemini



Technology Used & Future Work

Technology Used & Future Work



1 Improve the Model

2

More data
Clicked, Liked

Feedback

3

User Clustering

4

Matrix factorization algorithms
- Hybrid model
- Deep Learning Method
- Add Content Based model

5

Test Model in real life scenarios
-Click through rate
-conversion rate
-user rating score

Capgemini


Capgemini

Capgemini


Capgemini

Brand in detail

Color palette

This page displays our primary and secondary brand colors and their respective color breakdowns.

Our primary color palette consists of two colors, Capgemini Blue and Vibrant Blue.

Our secondary color palette consists of three colors; Deep Purple, Tech Red and Zest Green.

We use Gray as the background color for cut-out images. Furthermore, white space should feature heavily within our identity.

We also have an additional color palette to be used for infographics. Please refer to these the Infographics section further in the document.

CAPGEMINI BLUE
Pantone 7551C
CMYK C100 – M40 – Y10 – K3
RGB R0 – G112 – B175

DEEP PURPLE
Pantone 2091C
CMYK C89 – M100 – Y0 – K60
RGB R41 – G35 – B83

TECH RED
Pantone 7100
CMYK C0 – M90 – Y50 – K0
RGB R235 – G48 – B76

VIBRANT BLUE
Pantone 2195C
CMYK C90 – M0 – Y5 – K0
RGB R0 – G171 – B219

ZEST GREEN
Pantone 2303C
CMYK C10 – M0 – Y100 – K0
RGB R140 – G230 – B22

GRAY
Pantone Cool Gray 7
CMYK C0 – M0 – Y0 – K50
RGB R136 – G206 – B236

WHITE
CMYK C0 – M0 – Y0 – K0
RGB R255 – G255 – B255