6/21/2020

Anu Kashyap

Springboard

Instacart Market Basket Analysis

Springboard Capstone Project # 2

Contents

[Background 2](#_Toc43645654)

[Problem Statement & Objective 2](#_Toc43645655)

[Data 2](#_Toc43645656)

[Exploratory Data Analysis 3](#_Toc43645657)

[1. Orders by Order Number 3](#_Toc43645658)

[2. Order Frequency 3](#_Toc43645659)

[3. Order Distribution by Days Since Prior Order 3](#_Toc43645660)

[4. Order Distribution by Order Size 4](#_Toc43645661)

[5. Products by Department 4](#_Toc43645662)

[6. Top 20 Products by # Orders 5](#_Toc43645663)

[7. Product by Order Day of Week 6](#_Toc43645664)

[8. Product by Order Hour of Day 7](#_Toc43645665)

[9. Product Reorder Ratio 7](#_Toc43645666)

[10. Reorder Ratio by Add to Cart Order 8](#_Toc43645667)

[11. Reorder Ratio by Order Day of Week and Hour of Day 8](#_Toc43645668)

[Independent Variables 9](#_Toc43645669)

[Comparison Between Prediction Models 10](#_Toc43645670)

[Variable Importance 11](#_Toc43645671)

[Insights 12](#_Toc43645672)

# Background

Instacart operates an online grocery delivery and pick-up service. Orders are fulfilled and delivered by an Instacart personal shopper, who picks, packs, and delivers the order within the customer's designated time frame

Currently they 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

Through a competition, Instacart is challenging the Kaggle community to use this anonymized data on customer orders over time to predict which previously purchased products will be in a user’s next order.

# Problem Statement & Objective

Problem statement: Which products will an Instacart consumer purchase again?

The overall objective is to predict products that a user will buy again, try for the first time or add to cart next during a session

* Instacart currently uses XGBoost, word2vec and Annoy in production on similar data to sort items for users to “buy again”
* This data, and the algorithms trained upon it, are enabling Instacart to revolutionize how consumers discover and purchase groceries
* This helps Instacart make the right product recommendations to the customer, thereby, making the shopping experience more convenient for consumer

# Data

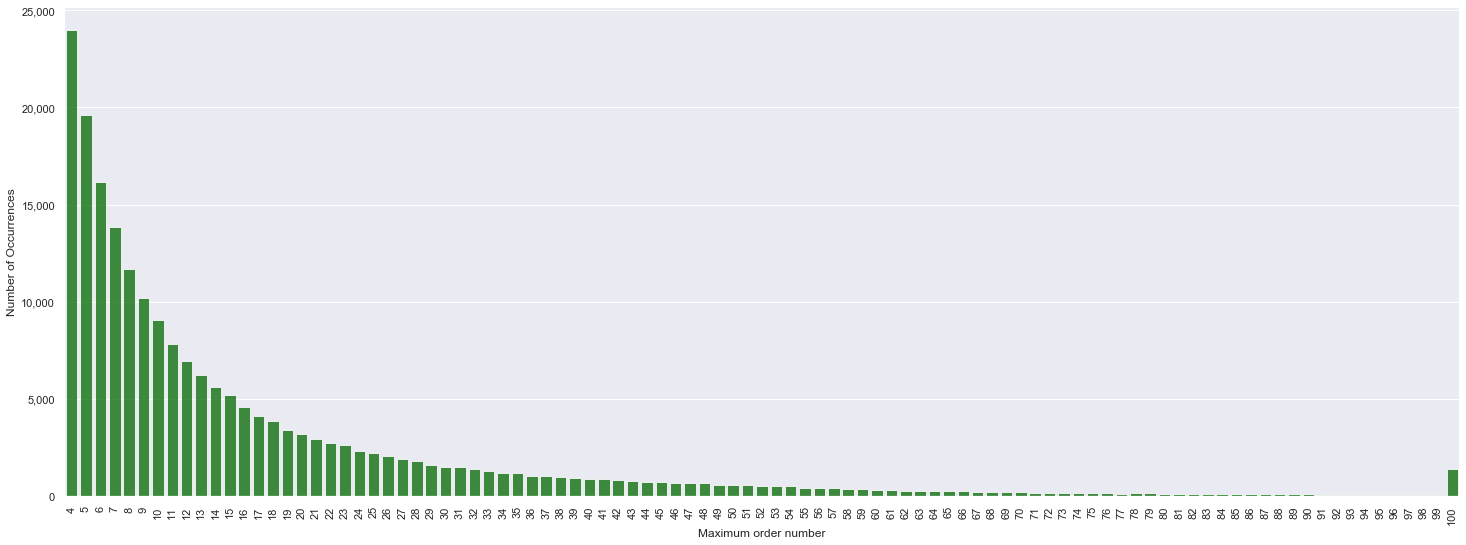
The dataset for this competition is a relational set of files describing customers' orders over time. The dataset is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. The datasets are:

1. aisles.csv: contains the aisle\_id and aisle\_name of a product
2. departments.csv: contains the department\_id and department\_name of a product
3. order\_products\_\_prior.csv: contains previous order contents for all customers. 'reordered' indicates that the customer has a previous order that contains the product
4. order\_products\_\_train.csv:
5. orders.csv: Contains information about which set (prior, train, test) an order belongs. Need to predict reordered items only for the test set orders. 'order\_dow' is the day of week
6. products.csv: contains mapping between product, aisle and department

# Exploratory Data Analysis

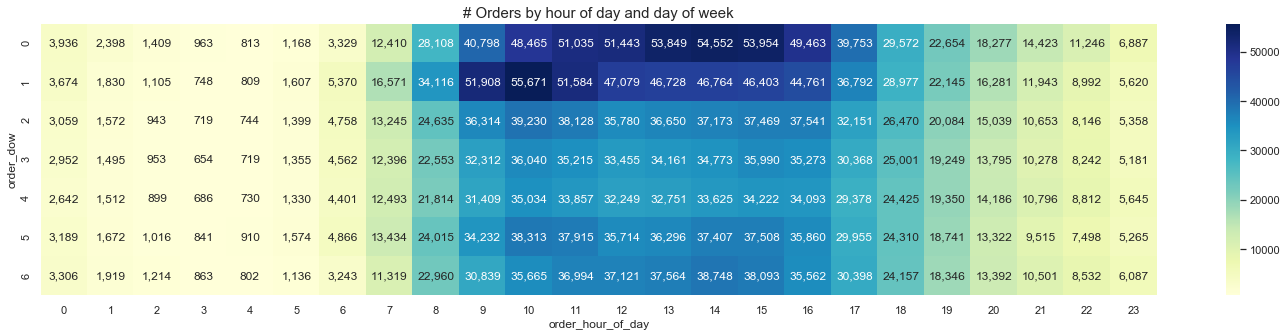
## Orders by Order Number

* All users have at least 4 orders and at most 100 orders



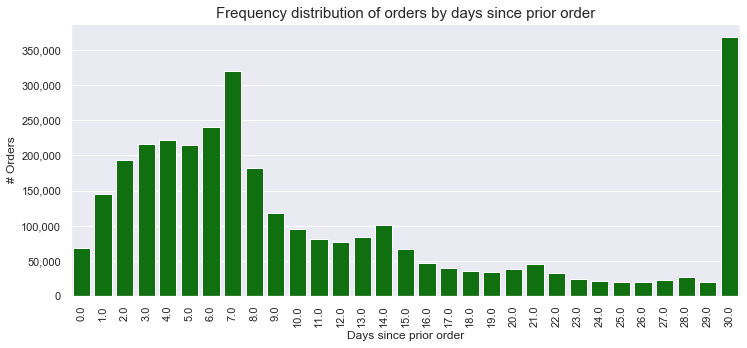
## Order Frequency

* Saturday afternoon and Sunday mornings/afternoons have high order frequencies



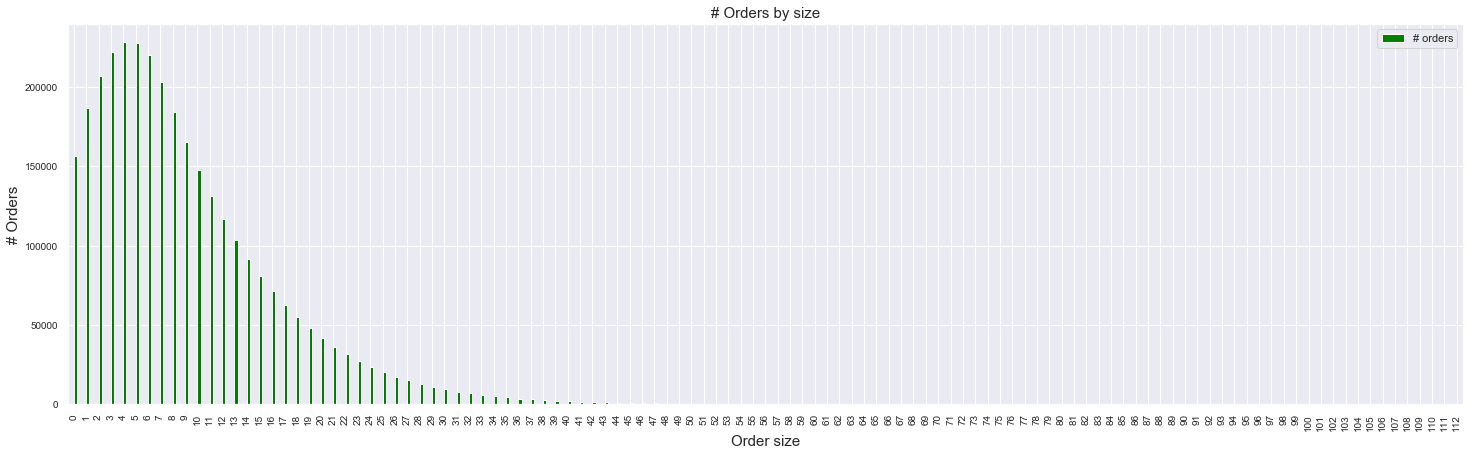
## Order Distribution by Days Since Prior Order

* Customers order once in every week (peak at 7 days) or once in a month (peak at 30 days)
* Also observed smaller peaks at 14, 21 and 28 days (weekly intervals)



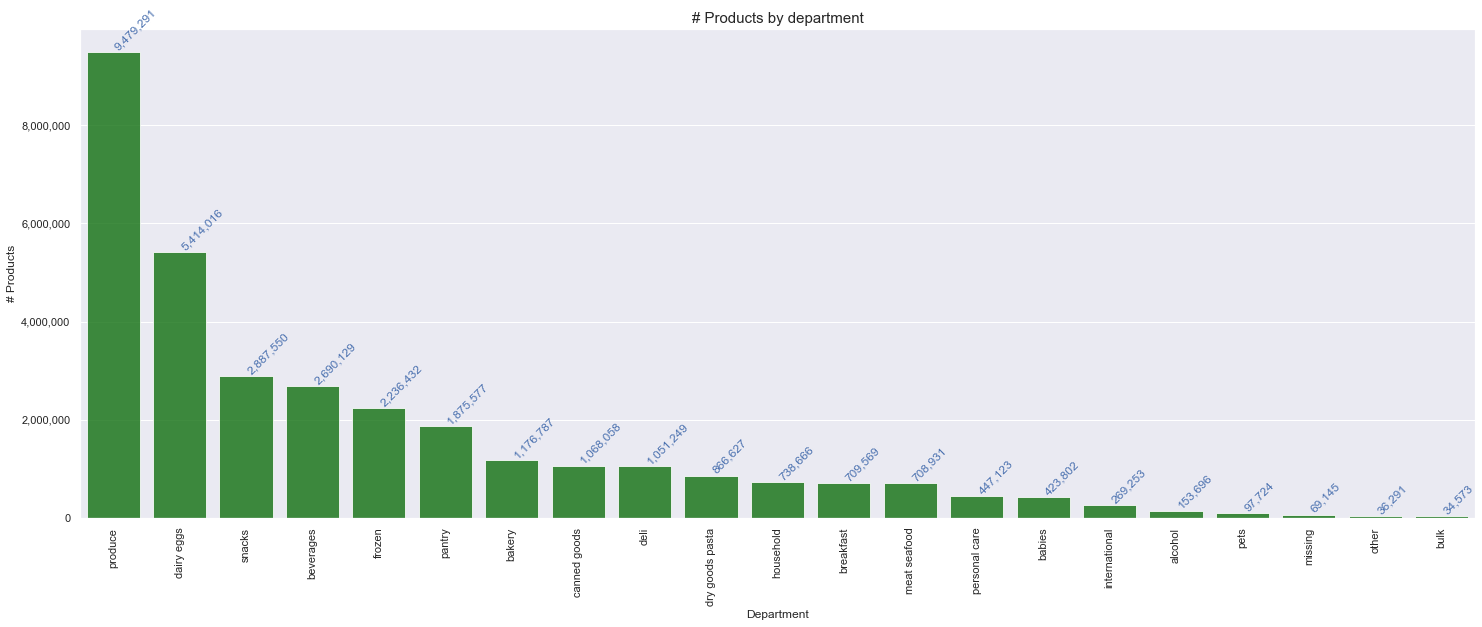
## Order Distribution by Order Size

* Maximum order size is 5, with most orders with 3-7 products



## Products by Department

* Most products are ordered from the Produce department (Fruits and vegetables)



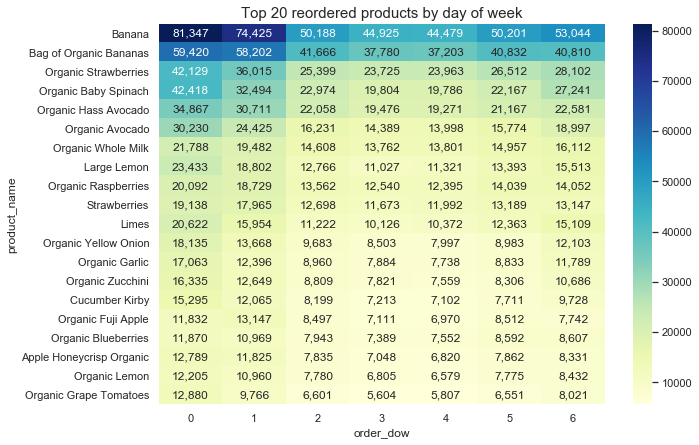
## Top 20 Products by # Orders

* Most of the top 20 ordered products are either fruits and vegetables

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **#** | **Aisle** | **Department** | **Product** | **# Orders** | **Reorder Ratio** |
| 1 | fresh fruits | produce | Banana | 472,565 | 84.4% |
| 2 | fresh fruits | produce | Bag of Organic Bananas | 379,450 | 83.3% |
| 3 | fresh fruits | produce | Organic Strawberries | 264,683 | 77.8% |
| 4 | packaged vegetables fruits | produce | Organic Baby Spinach | 241,921 | 77.3% |
| 5 | fresh fruits | produce | Organic Hass Avocado | 213,584 | 79.7% |
| 6 | fresh fruits | produce | Organic Avocado | 176,815 | 75.8% |
| 7 | fresh fruits | produce | Large Lemon | 152,657 | 69.6% |
| 8 | fresh fruits | produce | Strawberries | 142,951 | 69.8% |
| 9 | fresh fruits | produce | Limes | 140,627 | 68.1% |
| 10 | milk | dairy eggs | Organic Whole Milk | 137,905 | 83.0% |
| 11 | packaged vegetables fruits | produce | Organic Raspberries | 137,057 | 76.9% |
| 12 | fresh vegetables | produce | Organic Yellow Onion | 113,426 | 69.7% |
| 13 | fresh vegetables | produce | Organic Garlic | 109,778 | 68.0% |
| 14 | fresh vegetables | produce | Organic Zucchini | 104,823 | 68.8% |
| 15 | packaged vegetables fruits | produce | Organic Blueberries | 100,060 | 62.9% |
| 16 | fresh vegetables | produce | Cucumber Kirby | 97,315 | 69.2% |
| 17 | fresh fruits | produce | Organic Fuji Apple | 89,632 | 71.2% |
| 18 | fresh fruits | produce | Organic Lemon | 87,746 | 69.0% |
| 19 | fresh fruits | produce | Apple Honeycrisp Organic | 85,020 | 73.5% |
| 20 | packaged vegetables fruits | produce | Organic Grape Tomatoes | 84,255 | 65.6% |

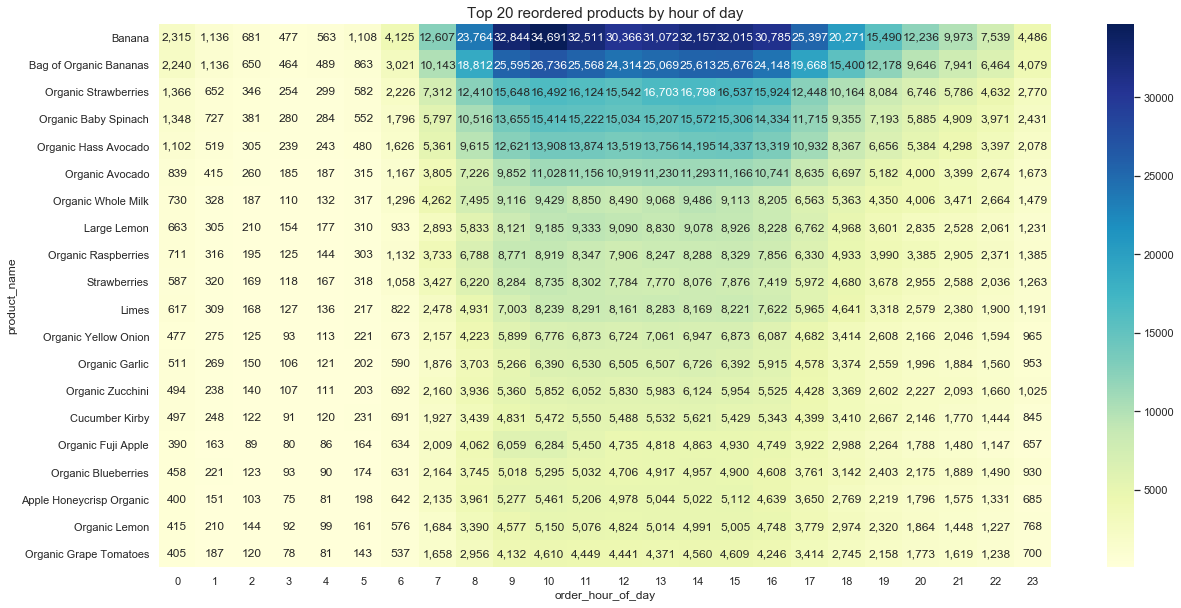
## Product by Order Day of Week

* Banana is the most ordered and reordered product and mostly ordered on Saturday and Sunday



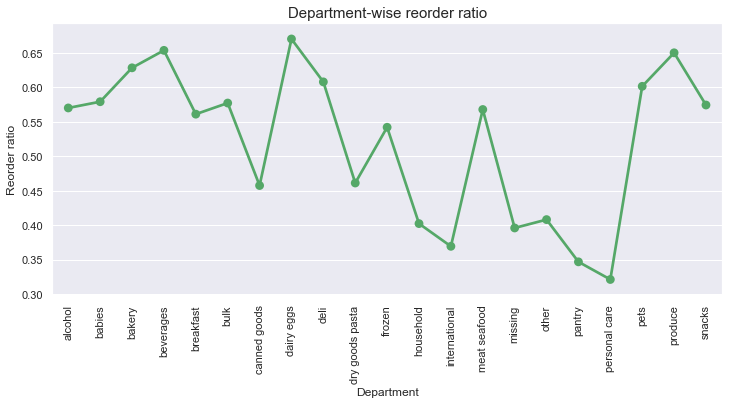
## Product by Order Hour of Day

* Banana is the most ordered and reordered product and mostly ordered during afternoon hours



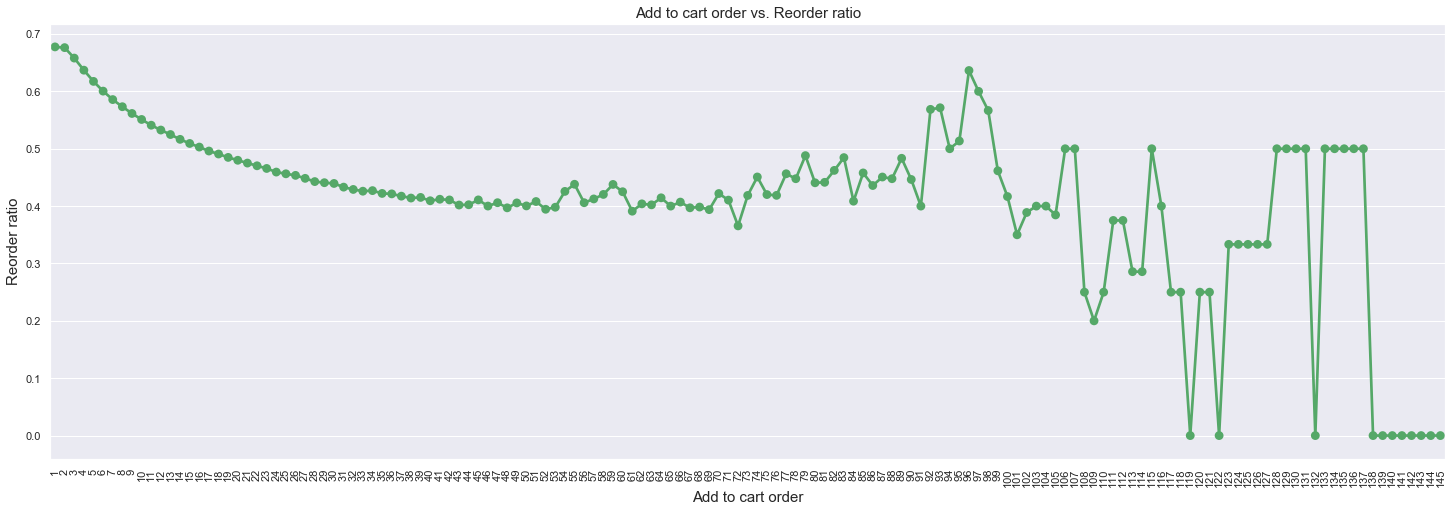
## Product Reorder Ratio

* Dairy eggs department has highest reorder ratio and Personal care has the lowest



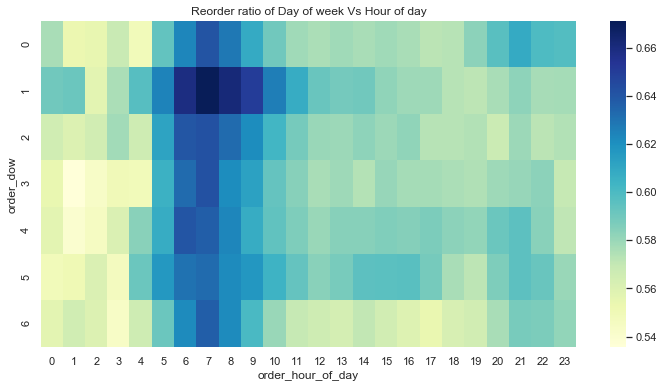
## Reorder Ratio by Add to Cart Order

* The products that are added to the cart initially are more likely to be reordered again compared to the ones added later which makes sense as we tend to first order all the products we buy frequently and then search for new products



## Reorder Ratio by Order Day of Week and Hour of Day

* Most of the reordered or frequently ordered products are ordered in the early morning hours



# Independent Variables

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Variable name** | **Description** | **Type** |
| 1 | user\_tot\_orders | Number of total orders for a user (user level) | Numeric |
| 2 | user\_tot\_prods | Total number of items a user bought (user level) | Numeric |
| 3 | user\_tot\_dist\_prods | Number of distinct products a user bought (user level) | Numeric |
| 4 | user\_avg\_days\_bet\_orders | Average number of days between orders for a user (user level) | Numeric |
| 5 | user\_avg\_order\_size | Average number of items in a user’s order (user level) | Numeric |
| 6 | order\_hour\_of\_day | Hour of the day for the order for which we need to predict products (order level) | Numeric |
| 7 | days\_since\_prior\_order | # of days since the prior order for the order for which we need to predict products (order level) | Numeric |
| 8 | days\_since\_ratio | days\_since\_prior\_order/user\_avg\_days\_bet\_orders (order level) | Numeric |
| 9 | Department | Department name of the product (order level) | Categorical |
| 10 | prod\_ordered | # of times the product has been ordered (product level) | Numeric |
| 11 | prod\_reordered | # of times the product has been reordered (product level) | Numeric |
| 12 | prod\_reorder\_rate | prod\_reordered / prod\_ordered | Numeric |
| 13 | uxp\_tot\_orders | Total number of orders at user X product level | Numeric |
| 14 | uxp\_order\_rate | uxp\_tot\_orders / user\_tot\_orders | Numeric |
| 15 | uxp\_avg\_pos\_in\_cart | Average position in cart at user X product level | Numeric |
| 16 | uxp\_reorder\_rate | uxp\_reordered/user\_reorderd\_prods | Numeric |
| 17 | uxp\_orders\_since\_last\_order | # orders since the user ordered this product (Total # of orders for a user – order number of the product when it was ordered by the user) | Numeric |
| 18 | uxp\_delta\_hour\_vs\_last | order\_hour\_of\_day - | Numeric |

# Comparison Between Prediction Models

|  |  |  |  |
| --- | --- | --- | --- |
|  | **XGBoost** | **Light GBM** | **Random Forest** |
| Dependent variable | Predicts whether a product will be ordered in the user’s next ordered | Predicts whether a product will be ordered in the user’s next ordered | Predicts whether a product will be ordered in the user’s next ordered |
| Hyperparameter Tuning Method | Randomized Search | Randomized Search | Grid Search |
| Optimal Hyperparameters | objective: ‘binary:logistic’  n\_estimators: 50  max\_depth: 5  learning\_rate: 0.2 | boosting\_type: 'gbdt’  objective: 'binary’  metric: {'binary\_logloss’}  num\_leaves: 96  max\_depth: 10  feature\_fraction: 0.9  bagging\_fraction: 0.95  bagging\_freq: 5 | max\_dept: 7  n\_estimators: 70  max\_features: ‘auto’ |
| Accuracy (Train set) | 90.1% | - | 90.7% |
| Accuracy (Hold out set on Kaggle) | 37.4% | 37.8% | 17.4% |

# Variable Importance

* uxp\_order\_rate and uxp\_orders\_since\_last\_order are the top two predictors

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Variable name** | **Importance** | **Description** |
| 1 | uxp\_order\_rate | 44.7% | uxp\_tot\_orders / user\_tot\_orders |
| 2 | uxp\_orders\_since\_last\_order | 17.9% | # orders since the user ordered this product (Total # of orders for a user – order number of the product when it was ordered by the user) |
| 3 | uxp\_reorder\_rate | 9.9% | uxp\_reordered/user\_reorderd\_prods |
| 4 | uxp\_tot\_orders | 7.3% | Total number of orders at user X product level |
| 5 | prod\_reorder\_rate | 4.9% | prod\_reordered / prod\_ordered |
| 6 | department | 3.8% | Department name of the product (order level) |
| 7 | days\_since\_prior\_order | 2.6% | # of days since the prior order for the order for which we need to predict products (order level) |
| 8 | days\_since\_ratio | 1.4% | days\_since\_prior\_order/user\_avg\_days\_bet\_orders (order level) |
| 9 | prod\_reordered | 1.3% | # of times the product has been reordered (product level) |
| 10 | uxp\_delta\_hour\_vs\_last | 1.2% | order\_hour\_of\_day - |
| 11 | user\_tot\_prods | 0.8% | Total number of items a user bought (user level) |
| 12 | user\_avg\_days\_bet\_orders | 0.7% | Average number of days between orders for a user (user level) |
| 13 | user\_avg\_order\_size | 0.7% | Average number of items in a user’s order (user level) |
| 14 | user\_tot\_dist\_prods | 0.7% | Number of distinct products a user bought (user level) |
| 15 | user\_tot\_orders | 0.7% | Number of total orders for a user (user level) |
| 16 | prod\_ordered | 0.6% | # of times the product has been ordered (product level) |
| 17 | uxp\_avg\_pos\_in\_cart | 0.6% | Average position in cart at user X product level |
| 18 | order\_hour\_of\_day | 0.2% | Hour of the day for the order for which we need to predict products (order level) |

# Insights

* Out of the three classification models used, XGBoost, Light GBM and Random Forest, Light GBM classification model gives the best accuracy at 37.8% on the Kaggle hold out set
* Light GBM hyperparameters were determined using Randomized Search and the best hyperparameters are:
  + Boosting\_type: 'gbdt’
  + Objective: 'binary’
  + Metric: {'binary\_logloss’}
  + Num\_leaves: 96
  + Max\_depth: 10
  + Feature\_fraction: 0.9
  + Bagging\_fraction: 0.95
  + Bagging\_freq: 5
* The top 5 variables as per the variable importance are:
  1. uxp\_order\_rate (44.7%)
  2. uxp\_orders\_since\_last\_order (17.9%)
  3. uxp\_reorder\_rate (9.9%)
  4. uxp\_tot\_orders (7.3%)
  5. prod\_reorder\_rate (4.9%)