Milestone Report

Capstone Project 1

Raviteja Bodla
Data Science Career Track
Springboard

Introduction:

Maintaining logistics is a huge part of any retail business. With the advent of online shopping and concepts like “same day delivery”, retail business has become highly competitive. Delay in delivery of the customer orders due lack of stocks will have a drastic impact on the business revenue and customer satisfaction as the customer can switch to another retailer.

By carefully studying customer buying behaviour, a near accurate recommendations can be provided to the retailers regarding what stocks to be maintained in order to make the customer experience good. In addition to the retailer recommendation, a customer recommendation system can be built to suggest appropriate products based on the customers buying behaviour.

Potential Clients and Applications:

All the retailers of various fields like food, grocery, fashion etc., can benefit from such a model. The potential clients can benefit by buying and maintaining stocks beforehand, can predict which customers to target for new product launches based on previous buying behaviours and sales. Additionally, this model can also be used in preparing food combos, menu card and display menu design.

Data:

The data used for this project is obtained from Kaggle competition ‘Instacart Market Basket Analysis’. The dataset is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. For each user, we provide between 4 and 100 of their orders, with the sequence of products purchased in each order. The data contains 6 comma separated files (csv) files:

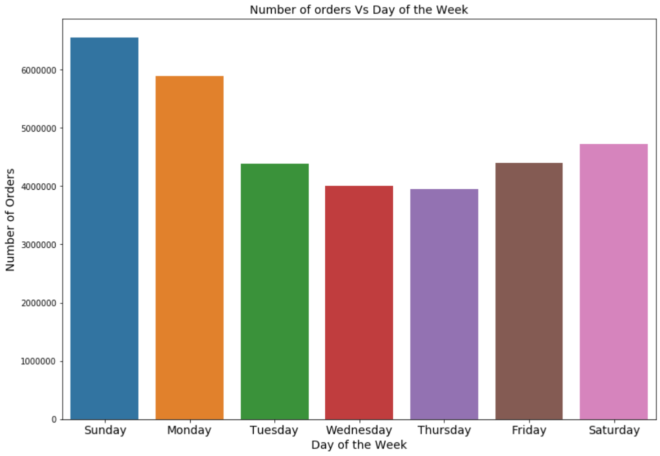
S.no	File	Shape	Attribues
1	aisles.csv	134 x 2	1) aisle_id 2) aisle
2	departments.csv	21 x 2	1) department_id 2) department
3	orders.csv	3.42 million x 7	1) order_id 2) user_id 3) eval_set 4) order_number 5) order_dow 6) order_hour_of_day 7) days_since_prior_order
4	order_products_prior.csv	3.24 million x 4	1) order_id 2) product_id 3) add_to_cart_order 4) reordered
5	order_products_train.csv	1.38 million x 4	1) order_id 2) product_id

			3) add_to_cart_order 4) reordered
6	products.csv	4970 x 4	1) product_id 2) product_name 3) aisle_id 4) department_id

Exploratory Data Analysis:

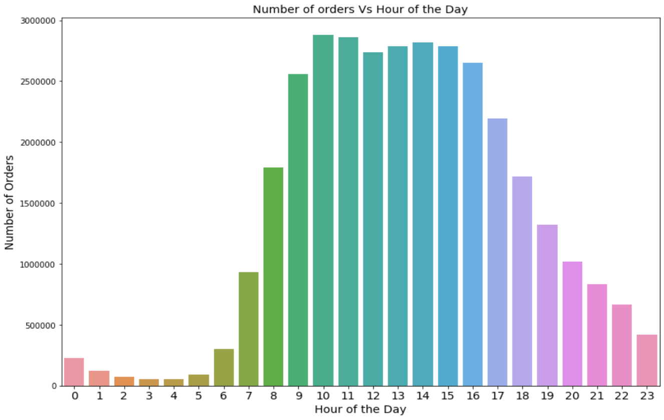
1)**Number of orders per day of the week:**

-Number of orders are high on Sunday, which gradually decreased till Thursday and again increases towards Friday and Saturday.



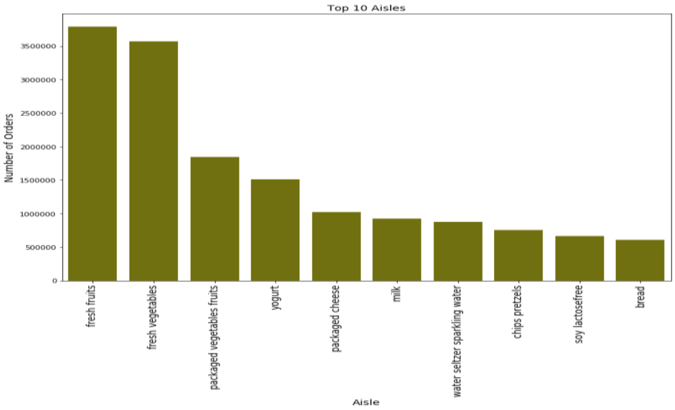
2) Number of orders per hour of the day:

-Number of orders increased from 6 AM until 10 Am and stayed high until 4PM and gradually decreased towards the night



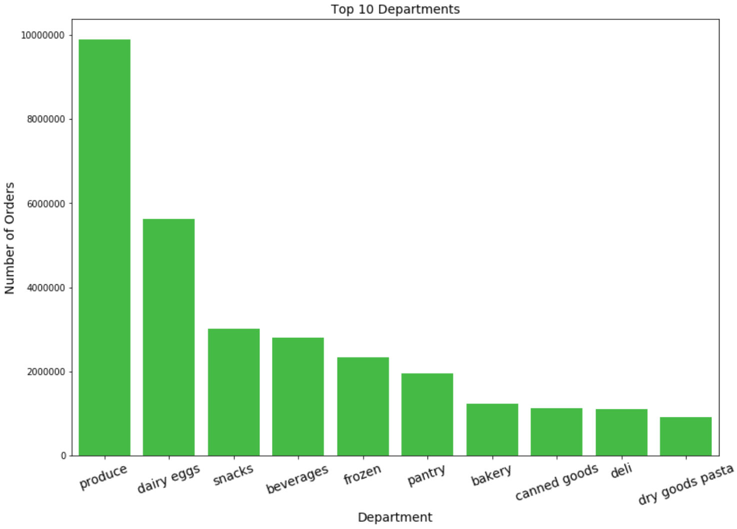
3)Top 10 Aisles:

-'Fresh fruits' is the top Aisle with highest orders followed by 'Fresh vegetables' and 'packaged vegetables fruits'.



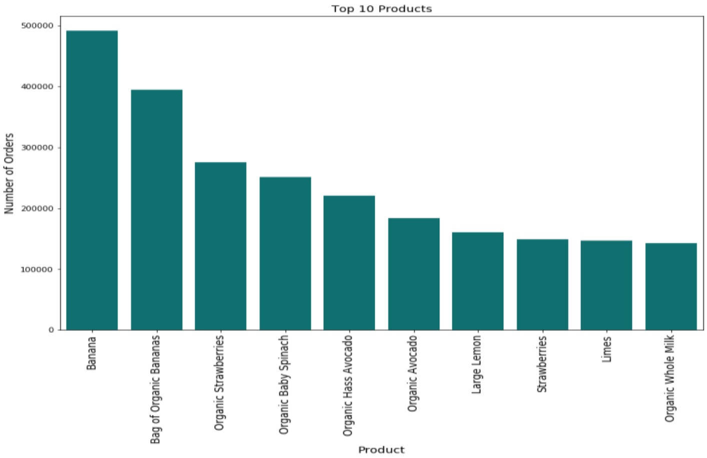
4)Top 10 Departments:

-Produce is the top department with highest number of sales followed by dairy eggs and snacks.



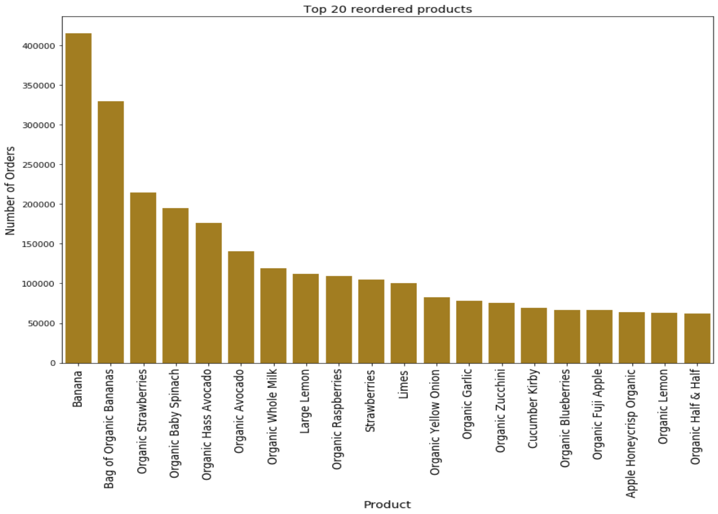
5)Top 10 Products:

-'Banana' is the top product with highest number of orders followed by bag of organic bananas and organic strawberries.



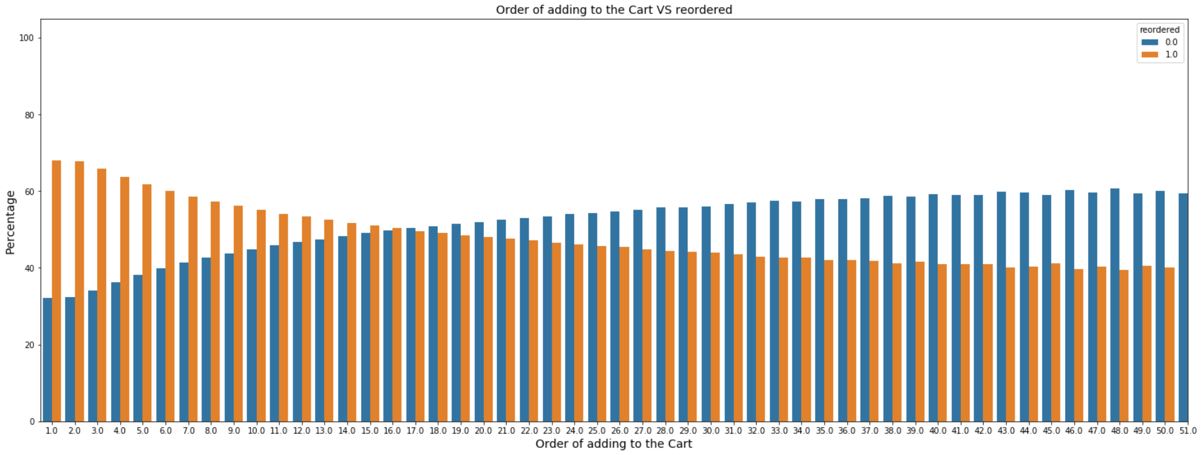
6)Top 20 reordered Products:

-'Banana' is the top reordered product with highest number of orders followed by bag of organic bananas and organic strawberries

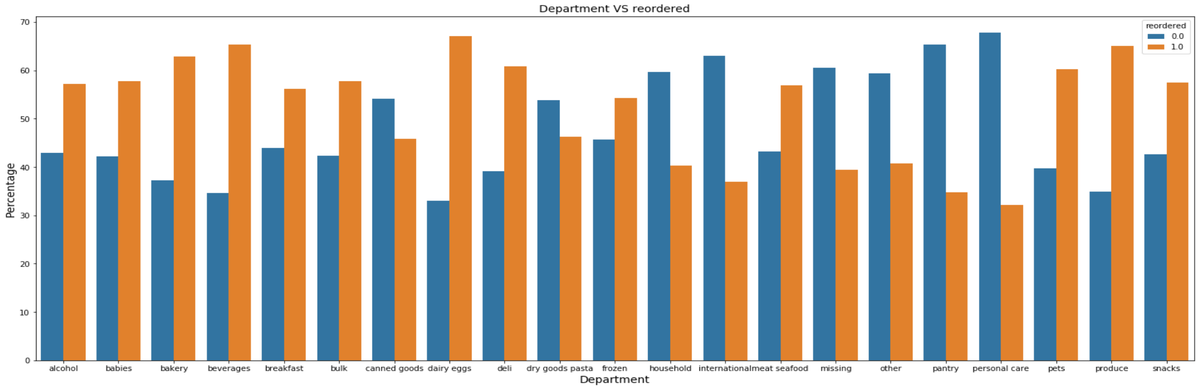


7) Added to the Cart sequence Vs reordered percentage:

-There is a clear pattern that items that were added to the cart first are reordered most. -The pattern continued until item 16 and after that the reorder percentage gradually decreased.



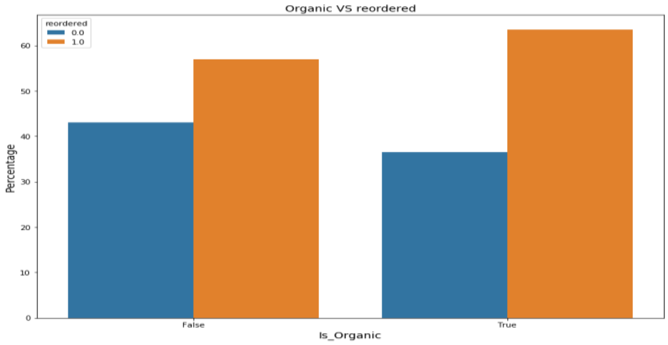
8) Department Vs Reordered Percentage:



-'Dairy eggs' is the department with highest reordered rate followed by 'produce' and 'beverages'.

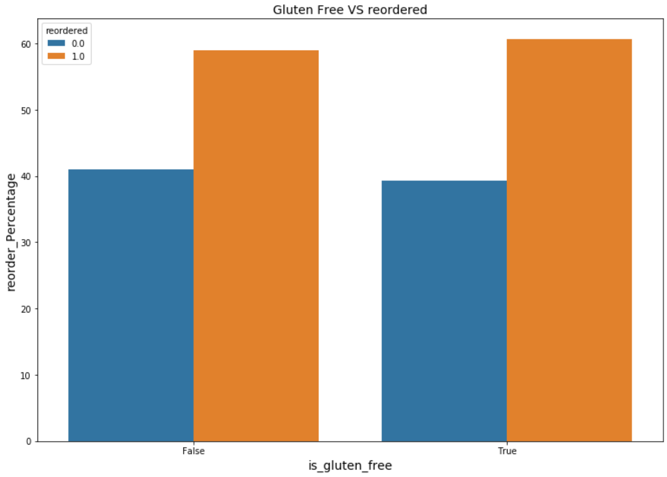
9) Organic Foods Vs Reordered:

-Reorder rate is more in Organic foods compared to other foods.



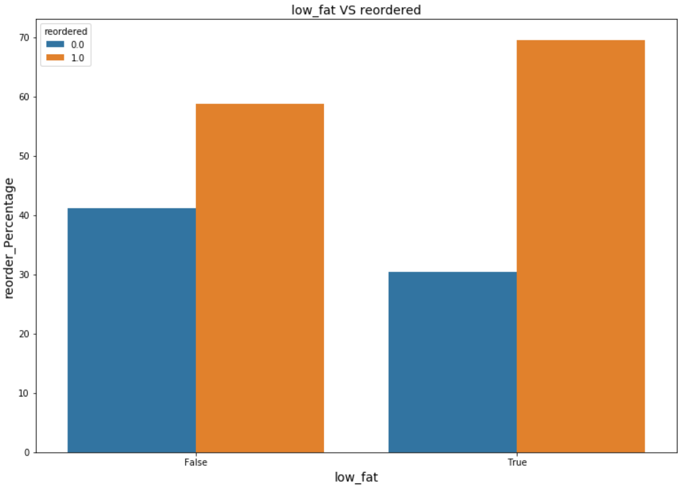
10) Gluten free VS reordered

-Reorder rate is more in Gluten free foods compared to other foods.



11) Low fat VS reordered

-Reorder rate is more in low fat foods compared to other foods.



Inferential Statistics:

As per the visual EDA report, it was observed that items that were added first into the cart have high reorder rate and organic, low-fat and gluten-free foods have high reorder rate. This report presents the results of inferential statistics performed to check the significance of above EDA results.

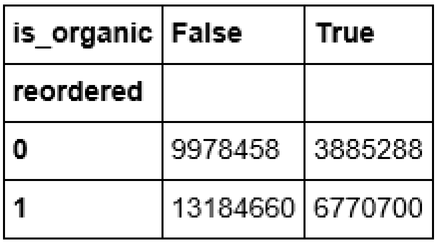
Added to the cart order and Reorder rate:

This test was performed to check if there is any relationship between the added to cart order and reorder rate. To perform the test, we have used Z-test as the reordered variable is a binary variable and added to the cart order is a multivariate feature. There were 6 datasets originally (aisles.csv, departments.csv, orders.csv, order_products_prior.csv, order_products_train.csv and products.csv) which were merged into one dataset 'total1.csv'. From the merged dataset, required columns ('order_id', 'add_to_cart_order', 'reordered') were taken and Z-test was performed at 0.05 significance level.

Results:
1) There is a significant relationship between added to cart order and reorder rate.
2) The p-value obtained was almost equal to 0.

Organic food and Reorder rate:

This test was performed to check if there is any relationship between the product name with 'Organic' in it and reorder rate. A new feature (is_organic) was created which has 'True' if 'Organic' was present in the 'product_name' feature. A contingency table was created with reordered as index and is_organic as columns.



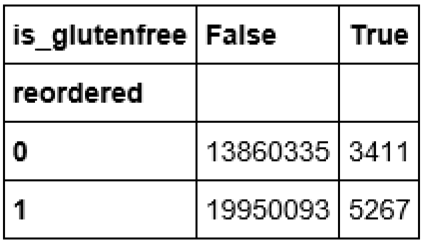
is_organic	False	True
reordered		
0	9978458	3885288
1	13184660	6770700

To perform the test, we have used Fisher exact test at 0.05 significance level as the both the features reordered and is_organic are binary features.

Results:
1) There is a significant relationship between the product name with 'Organic' in it and reorder rate.
2) The p-value obtained was almost equal to 0.

Gluten free food and Reorder rate:

This test was performed to check if there is any relationship between the product name with ‘Gluten’ and ‘Free’ in it and reorder rate. A new feature (is_ glutenfree) was created which has ‘True’ if ‘Gluten’ and ‘Free’ were present in the ‘product_name’ feature. A contingency table was created with reordered as index and is_ glutenfree as columns.



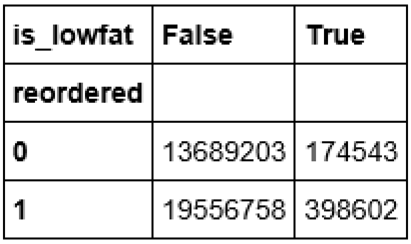
is_glutenfree	False	True
reordered		
0	13860335	3411
1	19950093	5267

To perform the test, we have used Fisher exact test at 0.05 significance level as the both the features reordered and is_ glutenfree are binary features.

Results:
1) There is a significant relationship between the product name with ‘Gluten’ and ‘Free’ in it and reorder rate.
2) The p-value obtained was almost equal to 0.001.

Low fat food and Reorder rate:

This test was performed to check if there is any relationship between the product name with ‘Low’ and ‘Fat’ in it and reorder rate. A new feature (is_ lowfat) was created which has ‘True’ if ‘Low’ and ‘Fat’ were present in the ‘product_name’ feature. A contingency table was created with reordered as index and is_ lowfat as columns.



is_lowfat	False	True
reordered		
0	13689203	174543
1	19556758	398602

To perform the test, we have used Fisher exact test at 0.05 significance level as the both the features reordered and is_ lowfat are binary features.

Results:
1) There is a significant relationship between the product name with ‘Low’ and ‘Fat’ in it and reorder rate.
2) The p-value obtained was almost equal to 0.