INSTACART
MARKET BASKET ANALYSIS

CAPSTONE PROJECT 1 - FINAL REPORT

RAVITEJA BODLA

raviteja.bodla@gmail.com

+1 647-674-2021

OVERVIEW:

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 behavior, 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 behavior.

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 behaviors 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:

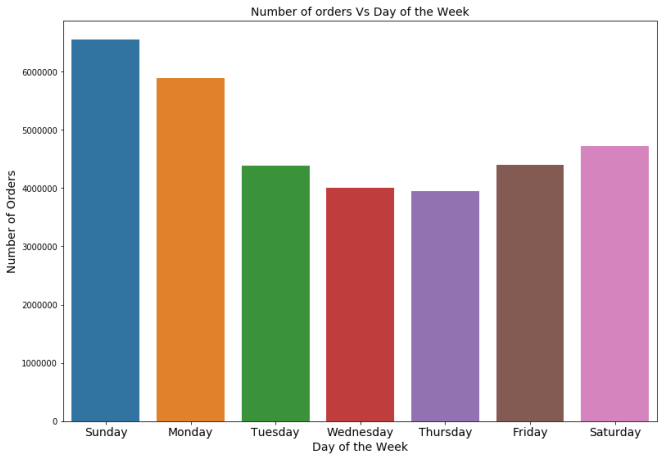
S.no	File	Shape	Attributes
1	aisles.csv	134 x 2	1) aisle_id 2) aisle
2	departments.csv	21 x 2	1) department_id 2) department
3	orders.csv	3.42 million x 7	1) order_id 2) user_id 3) eval_set 4) order_number 5) order_dow 6) order_hour_of_day 7) days_since_prior_order
4	order_products_prior.csv	3.24 million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

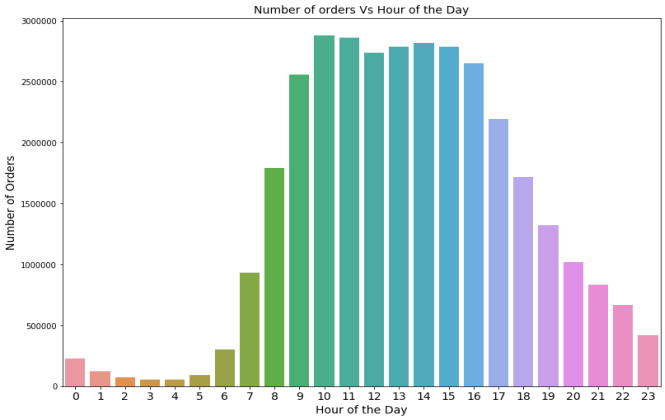
1) Number of orders per day of the week:

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



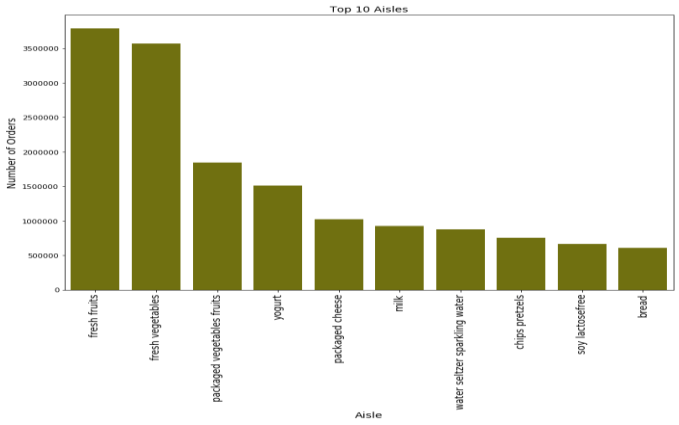
2) Number of orders per hour of the day:

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



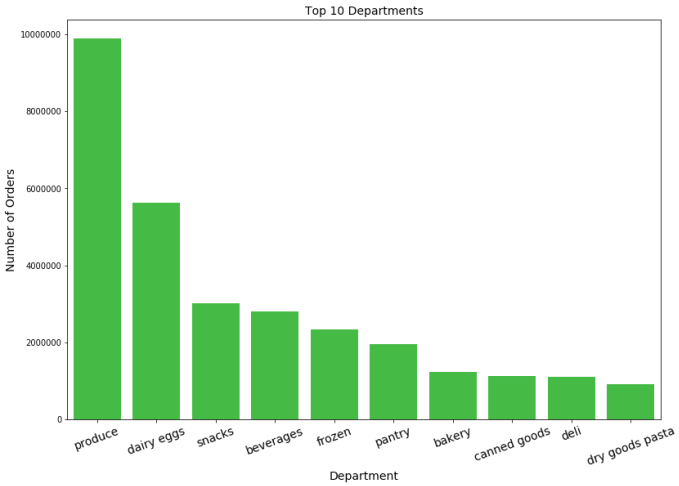
3) Top 10 Aisles:

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



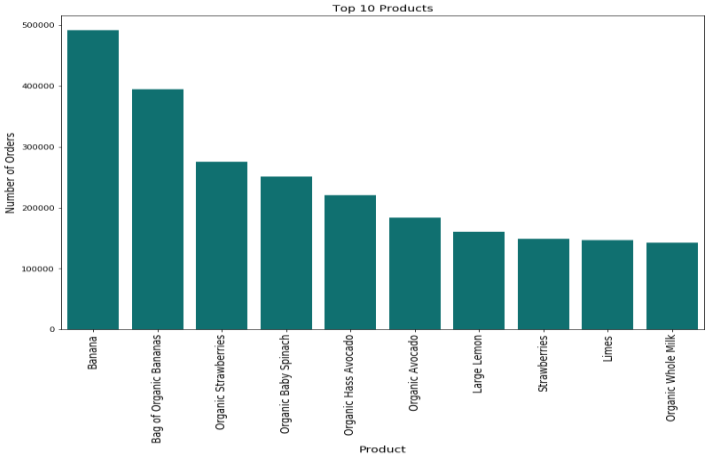
4) Top 10 Departments:

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



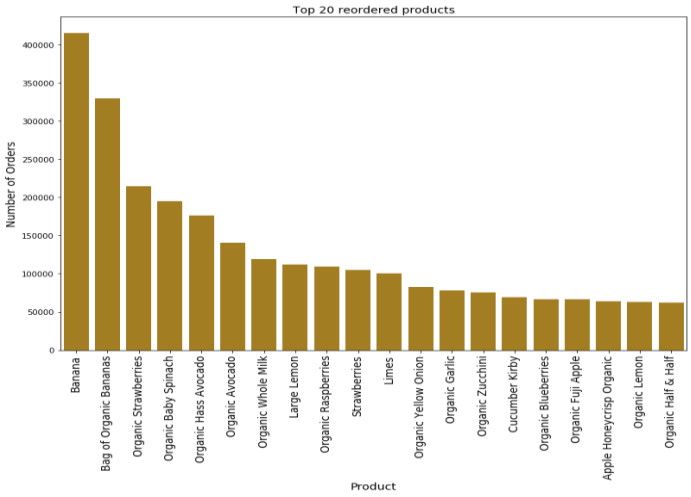
5) Top 10 Products:

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



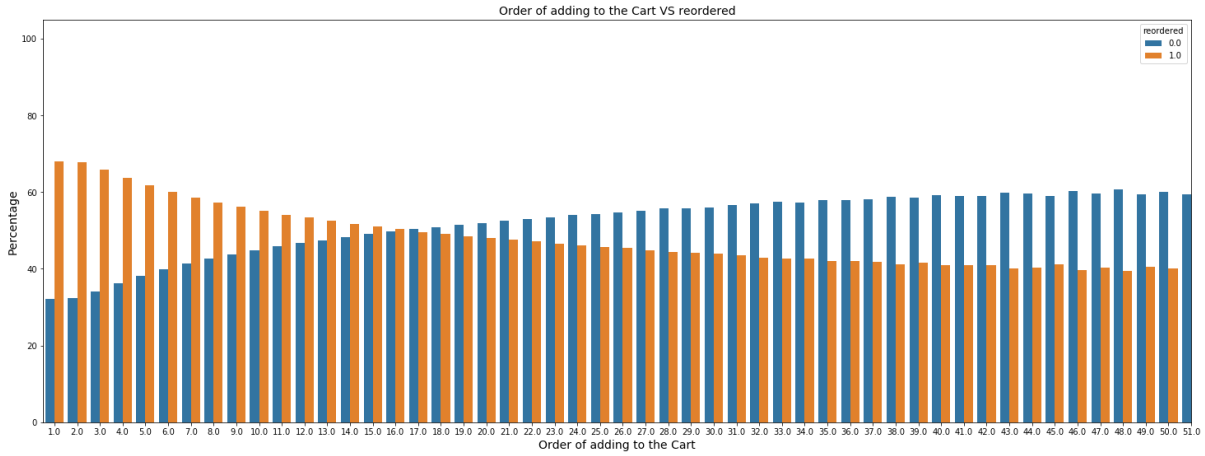
6) Top 20 reordered Products:

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

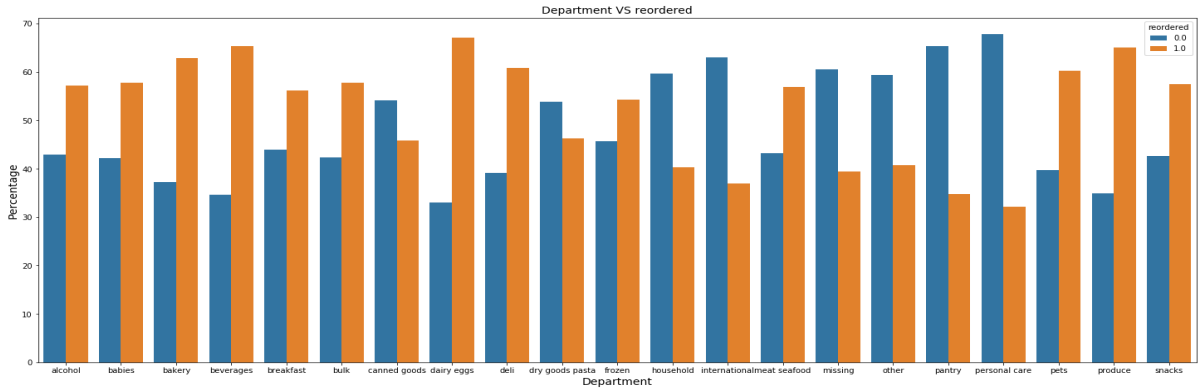


7) Added to the Cart sequence Vs reordered percentage:

-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.



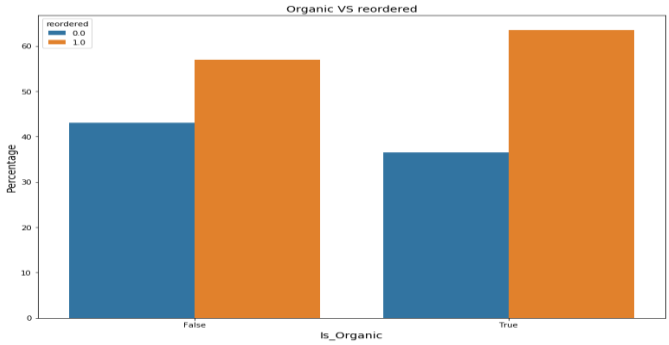
8) Department Vs Reordered Percentage:



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

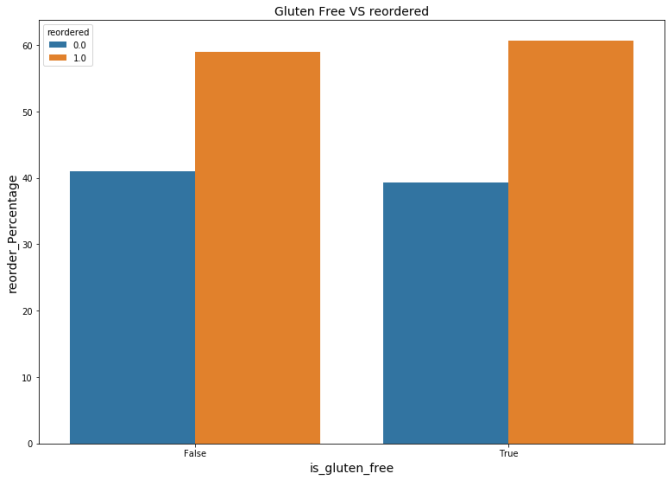
9) Organic Foods Vs Reordered:

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



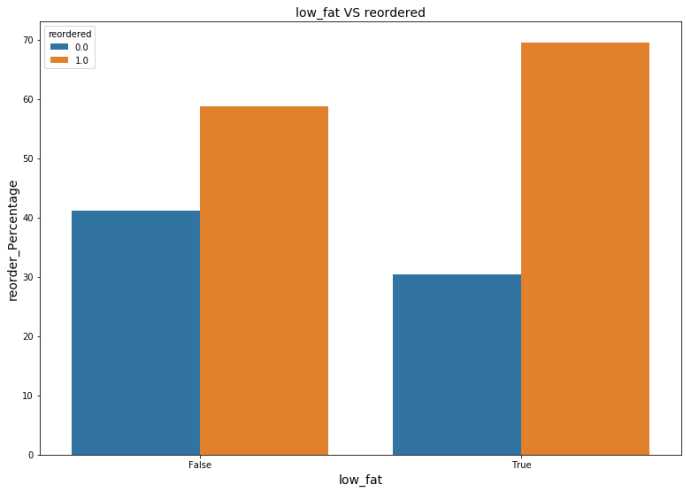
10) Gluten free VS reordered

-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.

FEATURE ENGINEERING:

Product Features:

1) **prior_orders:** Number of prior orders of the product.
2) **prior_reorders:** Number of reorders of the product based on prior orders.
3) **prior_reorder_rate:** Reorder rate of the product based on prior orders calculated as number of prior reorders divided by number of prior orders of the product.
4) **is_gluten_free:** If the product is gluten free of not, takes the value of 1 if the product name has 'gluten' and 'free' in it and 0 if not.
5) **is_organic:** If the product is organic of not, takes the value of 1 if the product name has 'organic' in it and 0 if not.
6) **is_low_fat:** If the product is low fat of not, takes the value of 1 if the product name has 'low and 'fat' in it and 0 if not.
7) **prior_avg_add_to_cart:** Average added to the cart rank for each product based on prior orders.
8) **product_avg_orderdow:** Average order day of week of the product.
9) **product_avg_orderhour:** Average order hour of the product.
10) **is_top_100_reordered:** Binary feature, 1 if the product is among the top 100 reordered products.
11) **is_aisle_top_30_reordered:** Binary feature, 1 if the aisle is among the top 30 reordered.
12) **is_department_top_10_reordered:** Binary feature, 1 if the department is among the top 10 reordered.

User Features:

1) **mean_days_between_orders:** Average days between all orders.
2) **number_of_orders:** Total number of orders of the users.
3) **prior_all_products:** All the previously ordered products by the user.
4) **jaccard_similarity:** Similarity of the present order with the previous order calculated as cardinality of the intersection of sets divided by the cardinality of the union of the sample sets.
5) **mean_order_similarity:** Average order similarity score of all the orders of the user.
6) **prior_total_items:** Number of products ordered by the user.
7) **total_distinct_products:** Number of unique products ordered by the user.
8) **average_products_per_basket:** Average number of items per order for user.
9) **user_avg_orderdow:** Average order day of week of the user.
10) **user_avg_orderhour:** Average order hour of the product.

User X Product Features:

1) **user_product:** product id + user_id * 100000
2) **UP_orders:** Total number of orders based on user_product.
3) **UP_mean_pos_in_cart:** Average position in the cart based on user_product.

Other features:

1) 6 binary features for each day of the week
2) 23 binary features for each hour of the day.

MODEL:

The data was prepared with above mentioned features and the data was split into train and test sets in 70:30 ratio. This is a classification task and several models were tested to predict which product will the user reorder in the next order. The models are evaluated based on F1 score. The F1 score is the harmonic average of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

$$F_1 = \left(\frac{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}}{2} \right)^{-1} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$

Initially, the models were predicting all 0's due to class imbalance. Hence the data was split with stratified splitting so that both the 1's and 0's are equally represented in train and test sets. Below are the accuracy score, recall score, precision score and f1 score for different models.

S.No	Model	Accuracy Score	Recall Score	Precision Score	F1 Score
1	Logistic Regression	0.90	8.04	0.66	0.16
2	Decision tree	0.82	0.16	0.14	0.15
3	Random Forest	0.90	0.005	0.34	0.01
4	Light gradient boosting model (Lightgbm) (Threshold 0.1)	0.61	0.65	0.15	0.25

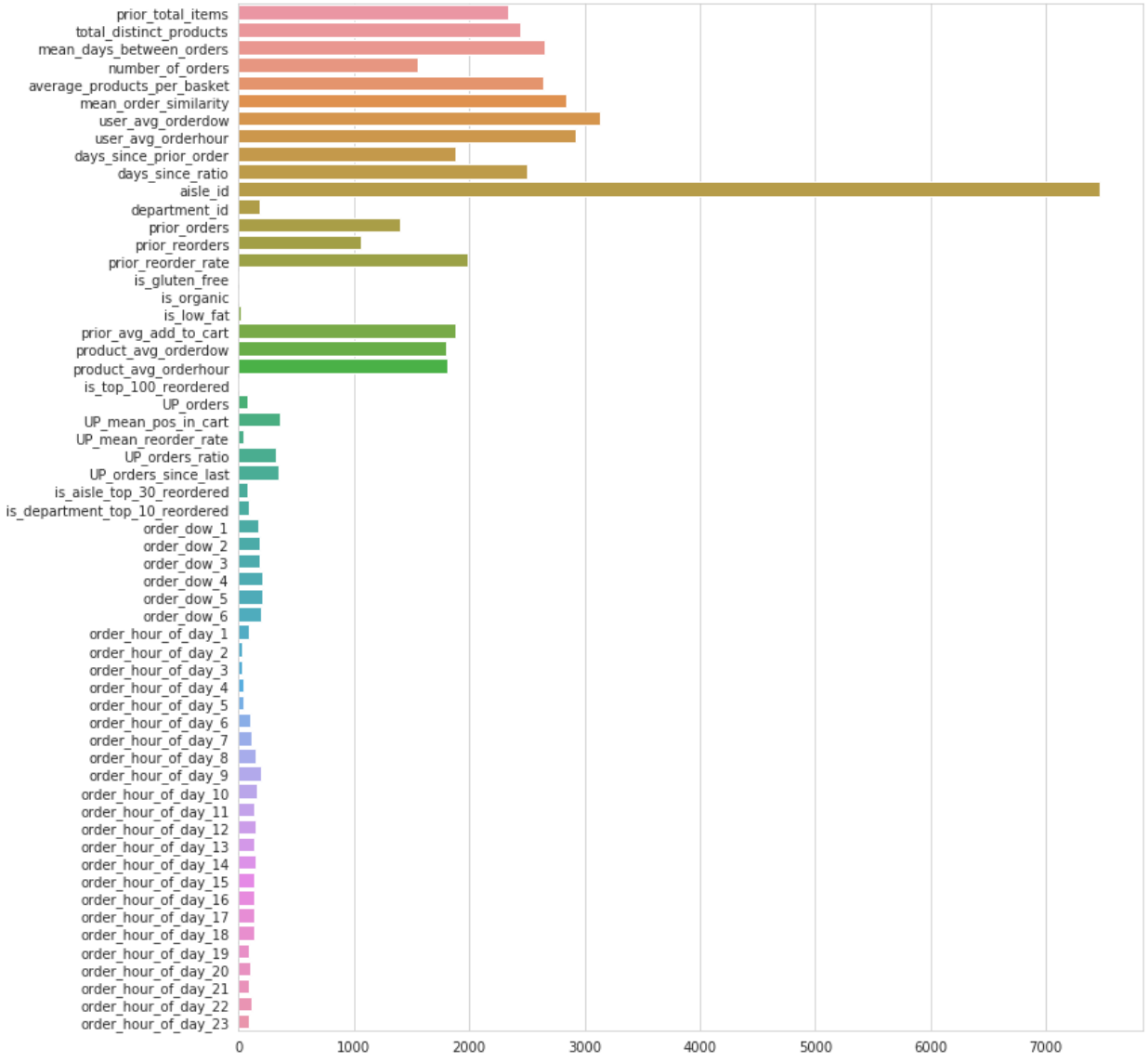
As the Light gradient boosting model achieved best F1 score among all the models, F1 score was checked for different probability threshold values.

S.No	Probability Treshold	Accuracy Score	Recall Score	Precision Score	F1 Score
1	0.10	0.622	0.661	0.158	0.255
2	0.11	0.672	0.594	0.168	0.262
3	0.12	0.716	0.530	0.179	0.267
4	0.14	0.751	0.470	0.189	0.269
5	0.15	0.803	0.416	0.199	0.266
6	0.16	0.823	0.365	0.209	0.260
7	0.17	0.839	0.318	0.232	0.252
8	0.18	0.852	0.276	0.243	0.242

9	0.19	0.861	0.211	0.252	0.211
10	0.20	0.869	0.217	0.261	0.217

Probability threshold of 0.4 provided the best f1 score.

FEATURE IMPORTANCE:



CONCLUSION:

This report highlighted the processes of data wrangling, inferential statistics, data visualization, feature engineering and predictive modelling performed on the Instacart basket dataset. All the results and insights gained as part of these processes were also highlighted. With these insights, a

Light Gradient Boosting Classifier was built to predict which product will the user reorder in the next order and score of 0.26 was obtained.