Integrated Program in Business Analytics

InstaCart Market Basket Analysis.

IPBA_B13_Group_F

Varun Bathija |

Agenda

I. Business Problem and Objectives
II. Executive Summary
III. Data Overview
IV. Data Preparation
V. Exploratory Data Analysis
VI. Model Development and Validations
VII. Businesss Recommendations and Conclusions

I. Business Problem and Objectives

- Instacart is an American company that operates a grocery delivery and pick-up service. The company offers its services via a website and mobile app. It has provided an anonymized dataset that contains a sample of over 3 million grocery orders from more than 200,000 Instacart users.
- Using the data, the objective is to build a model that is capable of predicting which previously purchased products will be in a user's next order. This model would be based on different factors like product details, order details and their relationship with user's order data.
- One of the use case identified for the dataset is predicting which product would the user purchase again. Upselling new products by advertising can be undertaken to existing customers.

II. Use Case Summary

How Machine Learning and Artificial Intelligence is revolutionizing grocery delivery service.

- **Dynamic pricing** - The dynamic pricing concept revolves around using ML and AI in shopping to determine the best pricing strategy for different products. For this, algorithms analyze data from different sources, such as historical sales, competitor prices, stock levels, and special occasions. One of the tactics of dynamic pricing is cross selling a discounted item (e.g., buns) with a complimentary product (hot dogs) at a full price. This strategy helps reduce food waste by lowering the prices of goods nearing their expiration date.
- **Improving inventory management** - It has become a trend to use AI in supermarkets and grocery stores to manage inventory at warehouses, shops and malls. It is important to keep the flow of supply and demand of products moving throughout the lifecycle of a company. By analyzing the historic data and analyzing the data and features of various products we can manage the inventory effectively of a company without incurring additional storage, delivery or any other overhead costs.
- **Enriching In-Store Experience -** AI-powered technology is committed to offering customers with locating products in nearby stores in real-time on their smartphones by sending notifications. It improves customer experience and aids them in deciding what they should buy in their budget. Moreover, these technology help grocers understand the product their customers love and bring them to their store. Next time, they suggest similar products to improve their shopping experience and keep them engaged.

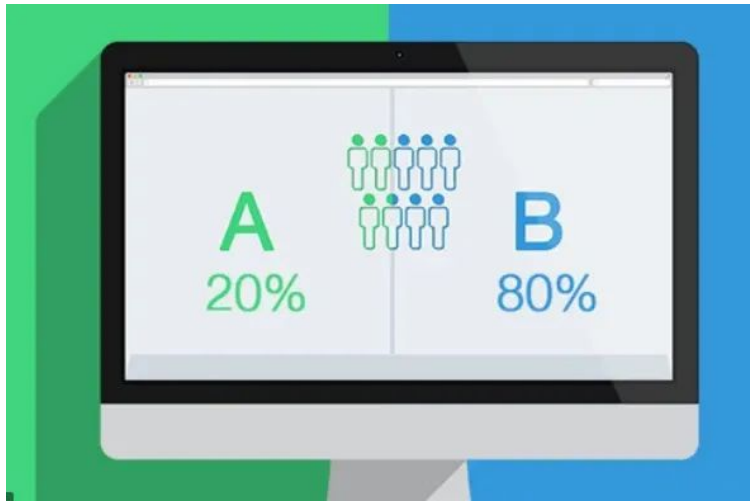
Use case of how a real-time company is using ML and AI in the grocery domain.

Hungryroot's AI-Powered Personalized Grocery Service

- Hungryroot is different from other online grocers because it uses machine learning and predictive modeling to fill a customer's cart based on their needs and objectives.
- The smart algorithm predicts a weekly basket of fresh groceries and simple recipes that best meet their needs.
- While customers have the option to edit their order or change their preferences at any time, Hungryroot is finding that the predictive technology is accurate. Most customers trust the company to select the majority of their weekly groceries for them, and 72% of all items purchased are chosen by the algorithm. As customers continue to engage with the service over time, the predictive model becomes more and more accurate.

How Data analytics and ML solutions can be used to help instacart.

1. **Advertisement of products** is Instacart's business model to drive grocery stores sales. So maximizing the customer conversion rate through advertisement and targeting customers to right grocery items or stores would generate revenue for Instacart. About 30% of all purchases made on Instacart go to advertised products.
2. **Boosting the traffic to a specific store for visibility** also increases Instacart revenue options, as grocery store owners are willing to pay if there is lot of new customer traffic on their page. They get sizable portion of revenue.
3. They also need an estimate by how much percent ML can improve operational efficiency which throws light on various **A/B testing** that can be done.

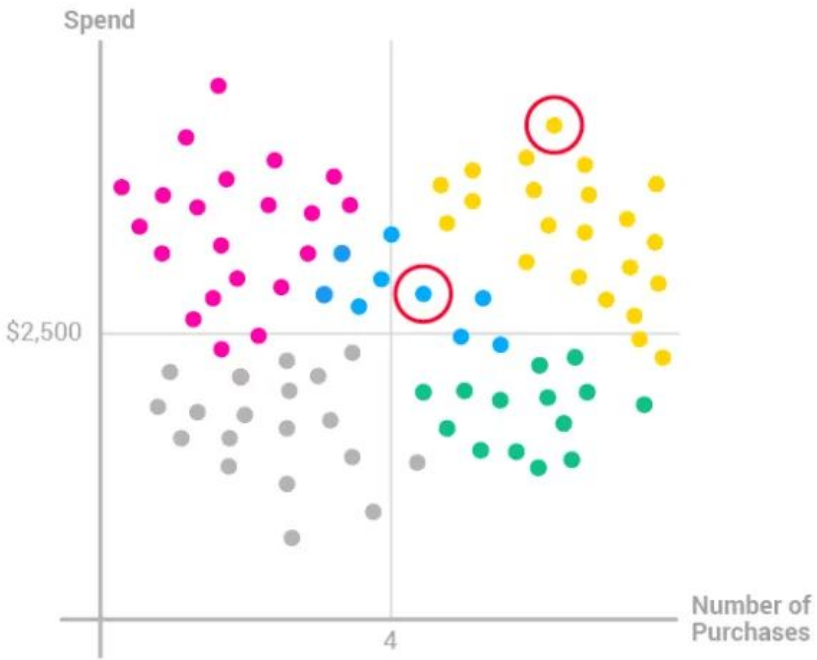


4. They also need a **forecast of how much a customer spends** on their service, to forecast demand which is essential in having an equilibrium in supply/demand.

5. It would also need to know **how many customers are repeat customers** which purchase on regular basis, with which they can develop smart algorithms to predict their needs ahead of time.

6. **Recommending items dynamically to the user** (Which he/she might be interested) would also increase conversion.

7. They also **need to identify customers based on their behavior**. There are 3 types of customers, one that will go ahead and shop. Two which walked away because of poor delivery options and three were just here to browse and not buy.



Clustering customers based on Spent amount.

Image: Jeremy Stan, Instacart

III. Data Overview.

- The dataset consists of around 3,00,00,000 records of grocery orders by approximately 200,000 Instacart users and product information specific to each individual order.

Table descriptions

The kaggle dataset that has been provided different CSV files which we will be using to conduct all our analysis and finally build the model .

- Aisles.csv - This file has the id and description of the aisle.
- Departments.csv - This file has the department name and id.
- Order_products__prior.csv - This file has the details of the order in which a user has added a particular product.
- Orders.csv - This file provides us with all the information regarding which day , hour, after how long the order has been reordered and then product details, aisles details etc.
- Orders_products__train.csv - This table will be used later to understand and predict reordered items or the next items to be predicted.
- Products.cvs - This contains the product name and the department and aisle to which that particular product belongs.

```
aisles.head(10)
```

	aisle_id	aisle
0	1	prepared soups salads
1	2	specialty cheeses
2	3	energy granola bars
3	4	instant foods
4	5	marinades meat preparation
5	6	other
6	7	packaged meat
7	8	bakery desserts
8	9	pasta sauce
9	10	kitchen supplies

	department_id	department
0	1	frozen
1	2	other
2	3	bakery
3	4	produce
4	5	alcohol
5	6	international
6	7	beverages
7	8	pets
8	9	dry goods pasta
9	10	bulk

```
order_products__prior.head(10)
```

	order_id	product_id	add_to_cart_order	reordered
0	2	33120	1.0	1.0
1	2	28985	2.0	1.0
2	2	9327	3.0	0.0
3	2	45918	4.0	1.0
4	2	30035	5.0	0.0
5	2	17794	6.0	1.0
6	2	40141	7.0	1.0
7	2	1819	8.0	1.0
8	2	43668	9.0	0.0
9	3	33754	1.0	1.0

```
order_products__train.head(10)
```

	order_id	product_id	add_to_cart_order	reordered
0	1	49302	1	1
1	1	11109	2	1
2	1	10246	3	0
3	1	49683	4	0
4	1	43633	5	1
5	1	13176	6	0
6	1	47209	7	0

```
products.head(10)
```

	product_id	product_name	aisle_id	department_id
0	1	Chocolate Sandwich Cookies	61	19
1	2	All-Seasons Salt	104	13
2	3	Robust Golden Unsweetened Oolong Tea	94	7
3	4	Smart Ones Classic Favorites Mini Rigatoni Wit...	38	1
4	5	Green Chile Anytime Sauce	5	13
5	6	Dry Nose Oil	11	11

```
orders.head(10)
```

	order_id	user_id	eval_set	order_number	order_dow	order_hour_of_day	days_since_prior_order
0	2539329	1.0	prior	1.0	2.0	8.0	NaN
1	2398795	1.0	prior	2.0	3.0	7.0	15.0
2	473747	1.0	prior	3.0	3.0	12.0	21.0
3	2254736	1.0	prior	4.0	4.0	7.0	29.0
4	431534	1.0	prior	5.0	4.0	15.0	28.0
5	3367565	1.0	prior	6.0	2.0	7.0	19.0
6	550135	1.0	prior	7.0	1.0	9.0	20.0
7	3108588	1.0	prior	8.0	1.0	14.0	14.0
8	2295261	1.0	prior	9.0	1.0	16.0	0.0
9	2550362	1.0	prior	10.0	4.0	8.0	30.0

IV. Data Preparation

Missing Values.

After initially exploring the data we found that there were missing values in the column days_since_prior_order. This column depicts the number of days before a particular product has been ordered by a user.

We can found around 6% of values to be missing.
In order to deal with the missing values, we have imputed 0 for them.

```
% of data which has missing values
order_id                  0.000000
user_id                   0.000000
eval_set                  0.000000
order_number              0.000000
order_dow                 0.000000
order_hour_of_day         0.000000
days_since_prior_order    0.060276
dtype: float64

% of unique users vs total data :  0.06027594185817766
```

Filtering data for model creation.

In order to train our model effectively we filtered out our dataset during the training phase to take into account users having 30 items associated with them. Taking all users with less product orders was skewing our model and hence this step was necessary.

V. Exploratory Data Analysis

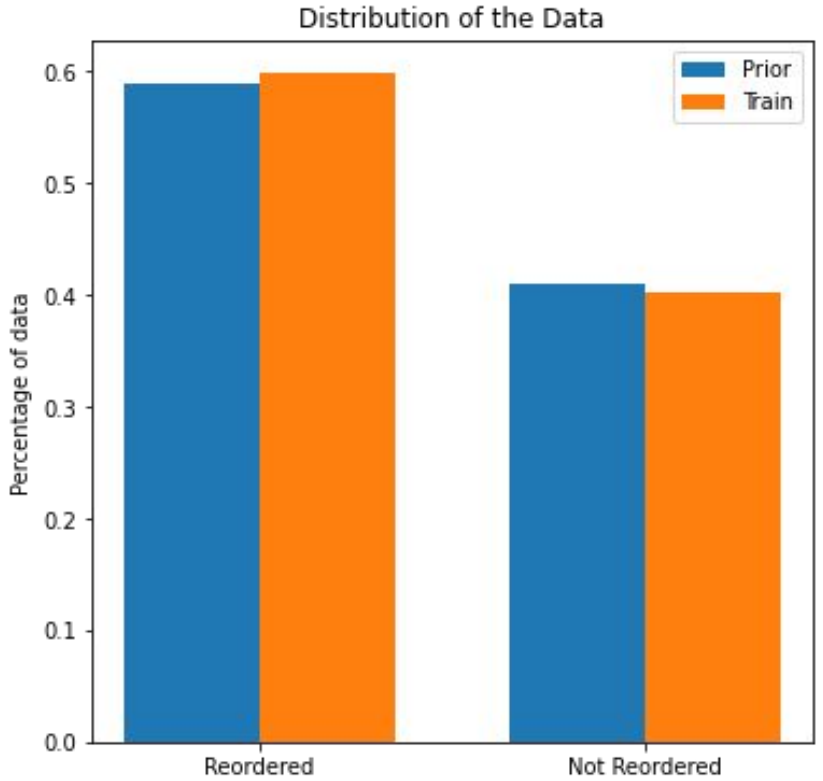
- EDA has been performed by a combination of Python, R, Tableau and MS excel.

- **Univariate Analysis**

Distribution of Target Variable:

Analysis:

- The distribution of the target variable ‘reordered’ is almost equal in both prior and train set.
- We have orders with 60% of reordered products.

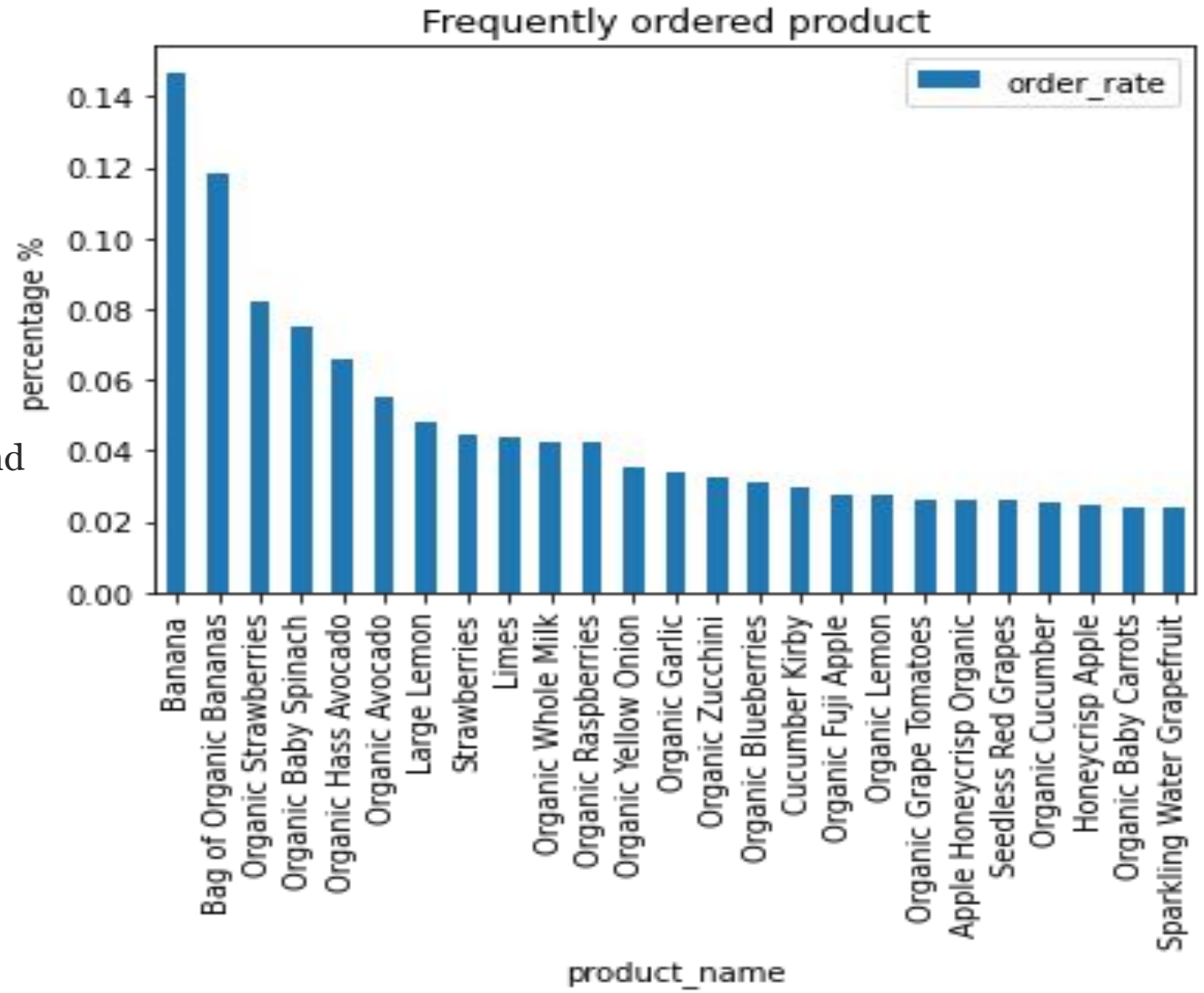


Frequently of ordered products

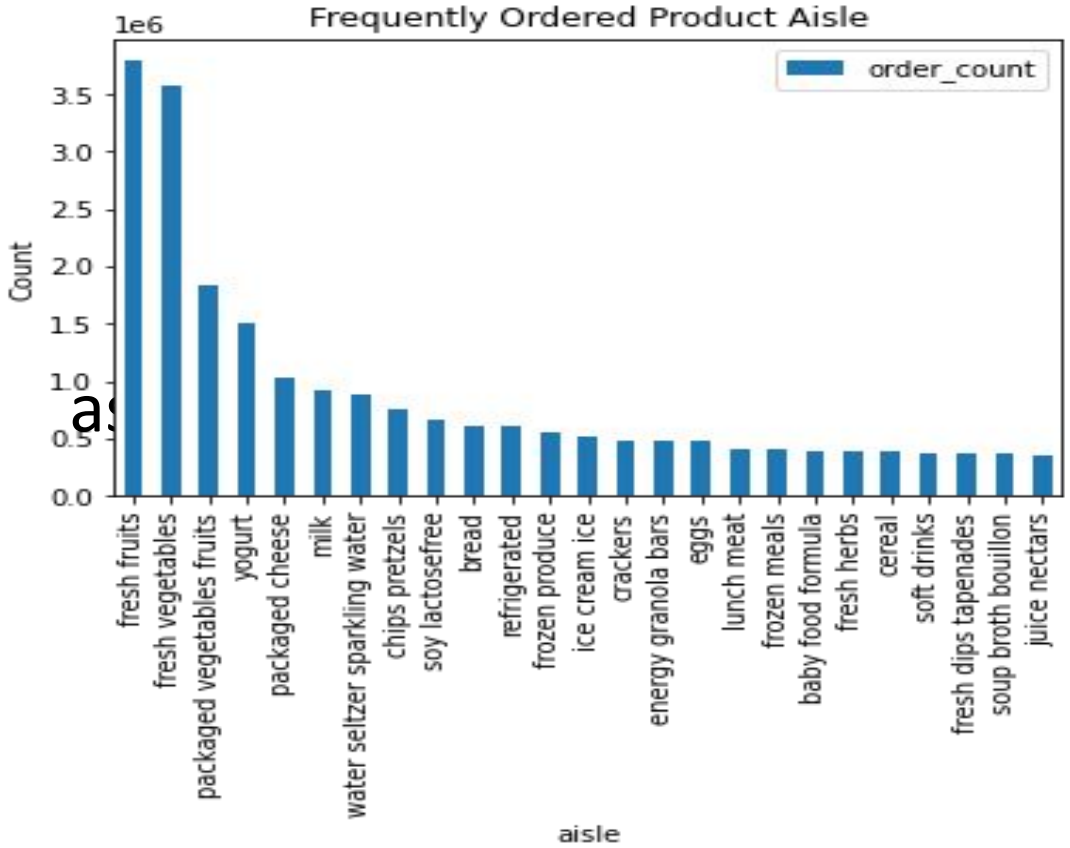
Analysis:

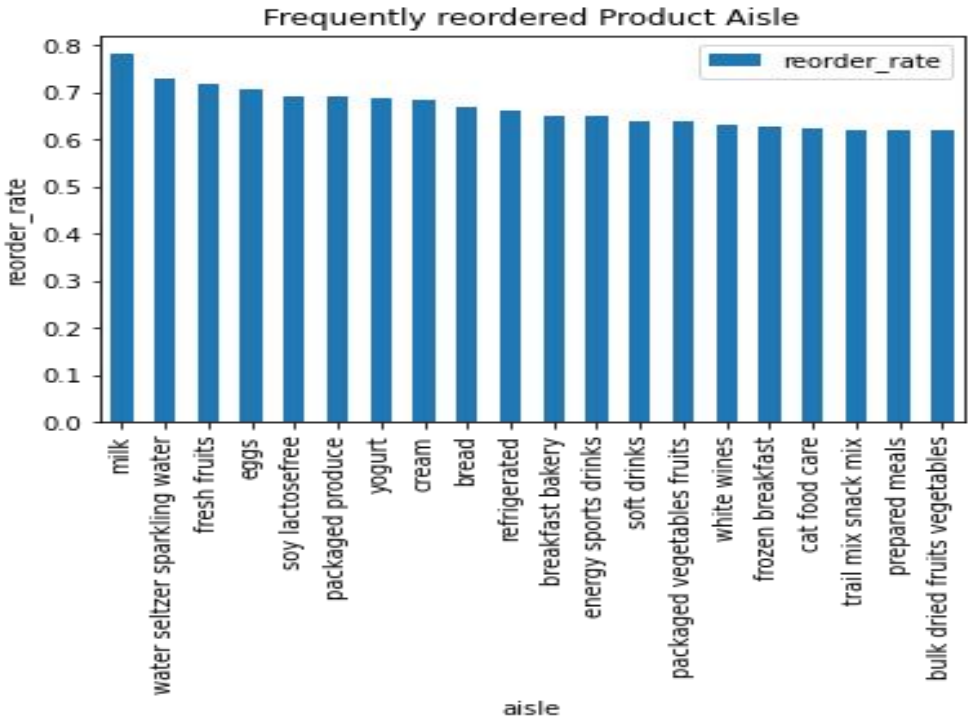
- It can be seen that most of the products which are ordered are organic foods / fresh fruits (especially Bananas)
- Bananas have highest order rate.
- Top 5 frequently ordered products are organic in nature

- The least ordered product identified were high end syrups and alcohols



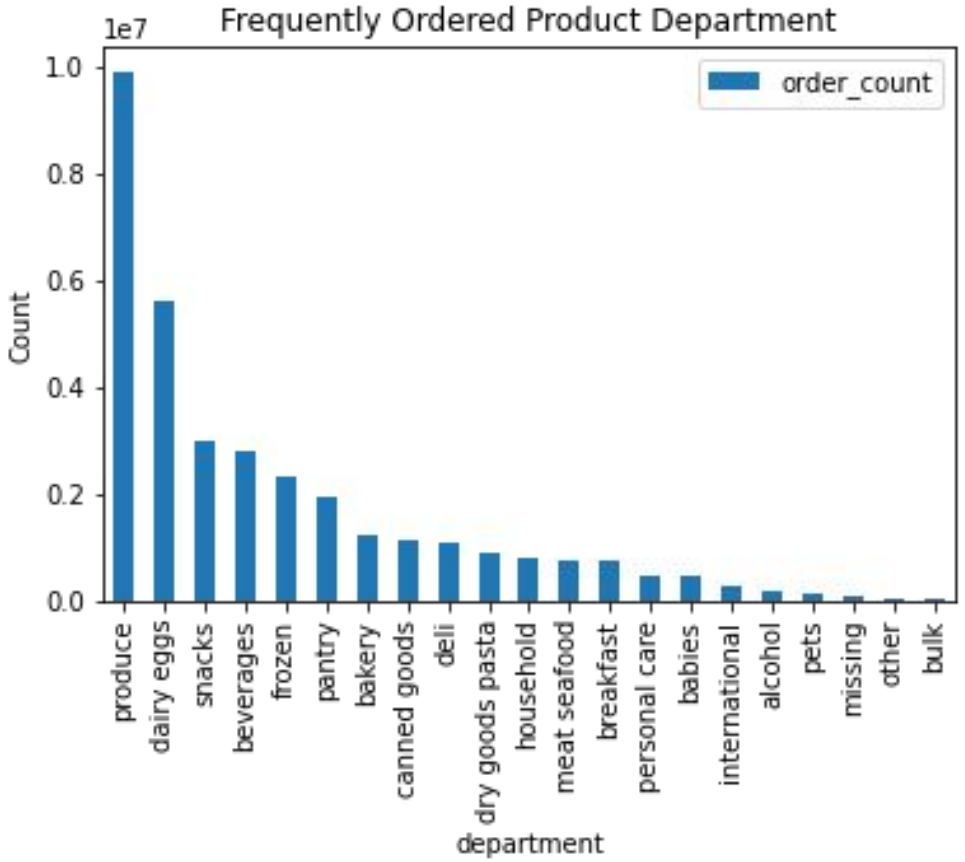
Aisle Analysis for ordered and reordered products

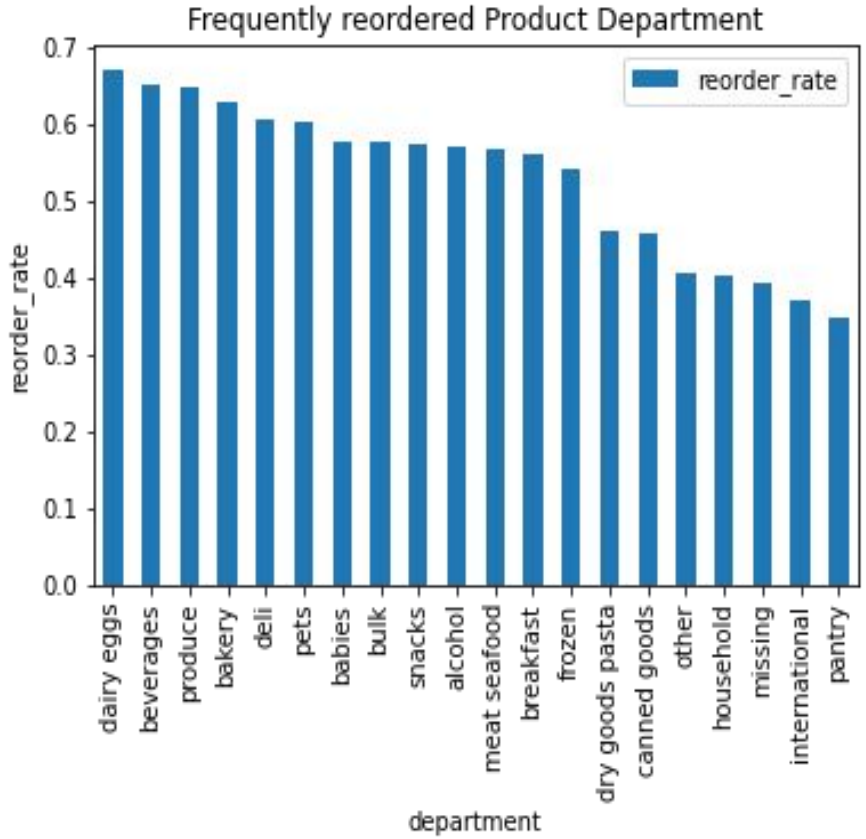




- Milk, sparkling water, fruits, eggs, yogurt are most common aisles the product is reordered from, as they are items which are daily consumed.
- Also these are the products that lasts only few days , thus high reorder rate.
- As we can see, most products are ordered from Fresh Fruits and Fresh Vegetables aisles.
- Other frequently ordered items are from Yogurt , Packaged Vegetables and packaged cheese aisles.
- Least frequently ordered items are from Air fresheners, Baby accessories, Baby bath body care etc. aisles.

Department Analysis for ordered and reordered products

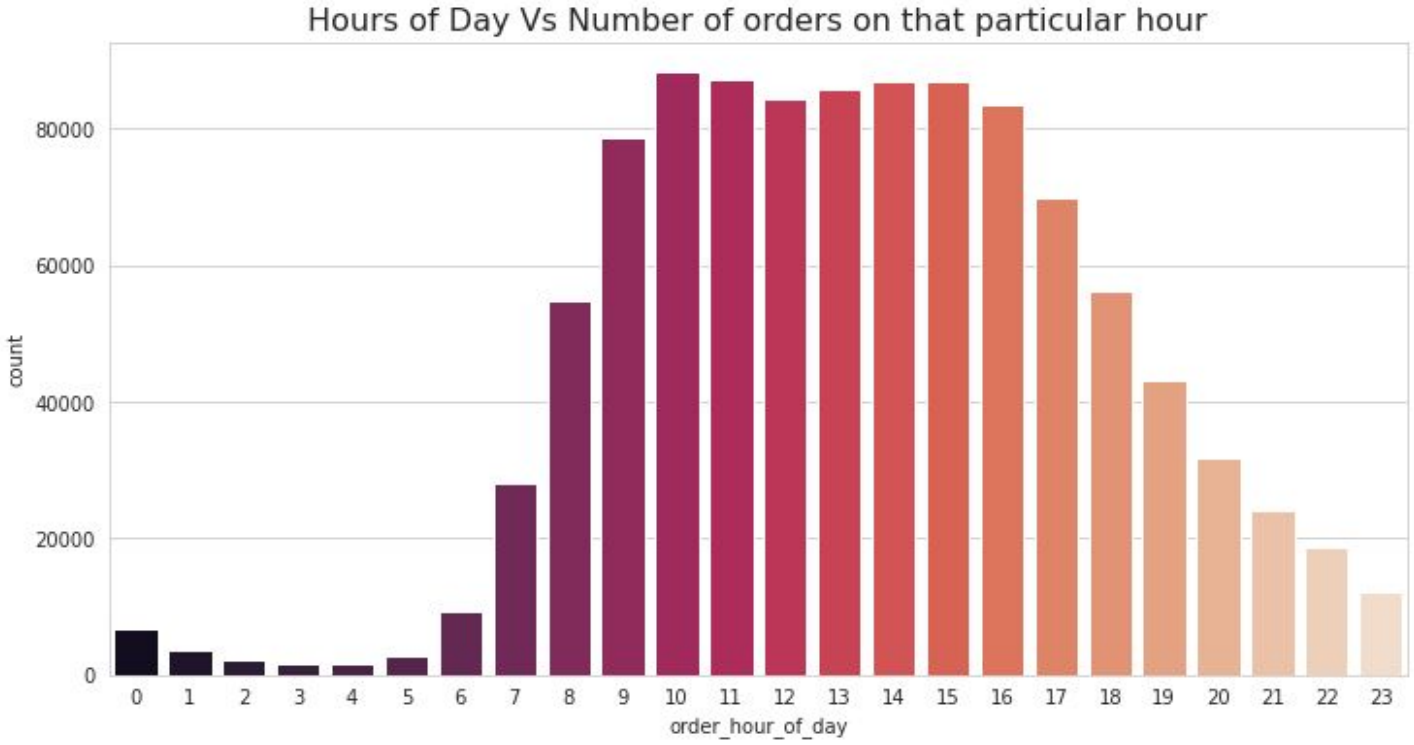


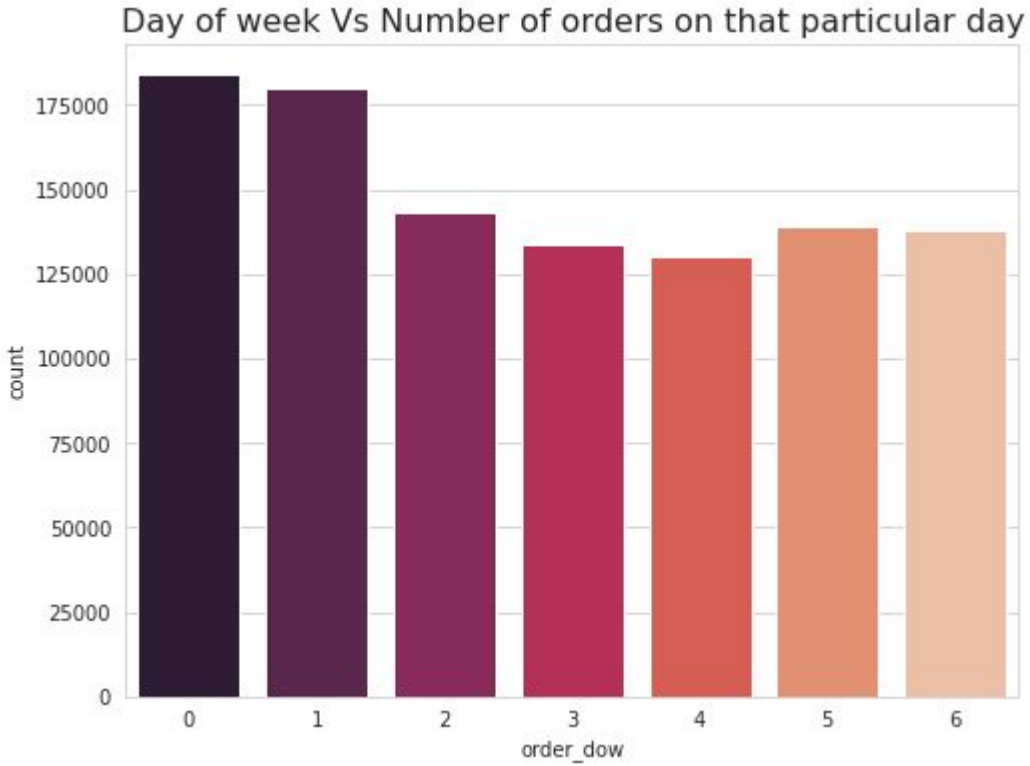


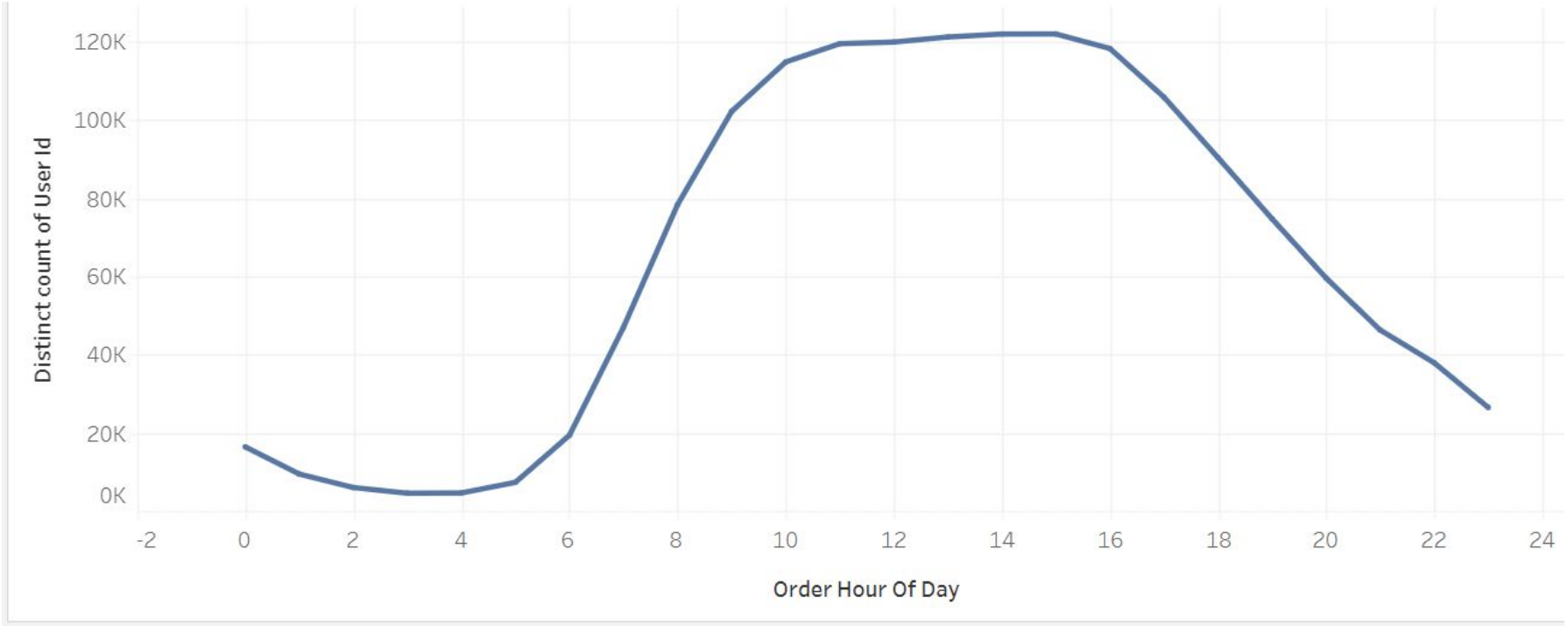
- There are total of 21 departments.
- The most ordered products are from the **produce department** having fruits, vegetables, herbs etc.
- Lowest order rates are for personal care, babies, international and alcohol departments.
- We see highest reorder rate for dairy eggs and beverages.

Analysis of User Behaviour on time, day of the week.

- It can be seen that day 0 and day 1 have the higher number of orders. Day 0 has been taken as sunday day 1 monday and so on.
- For the overall data we can see that during the start of the 9 am onwards upto 6 pm the orders are maximum. During mid night and late night times the orders are minimum.

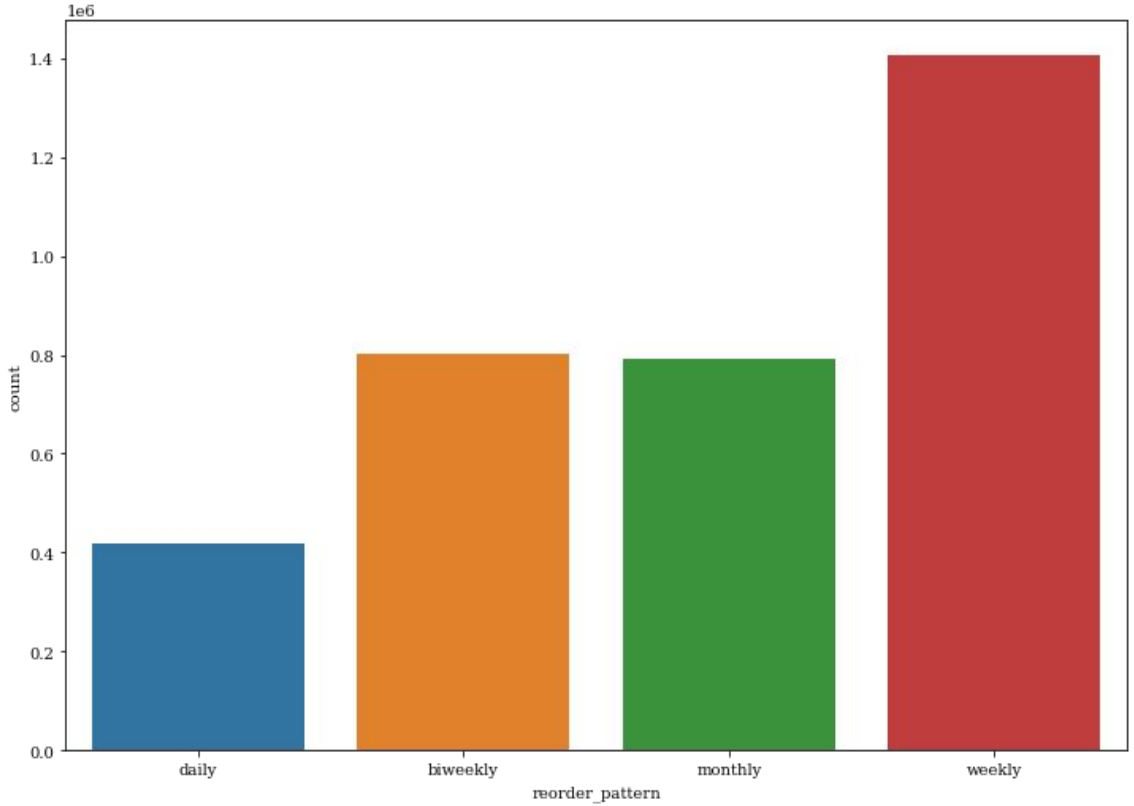






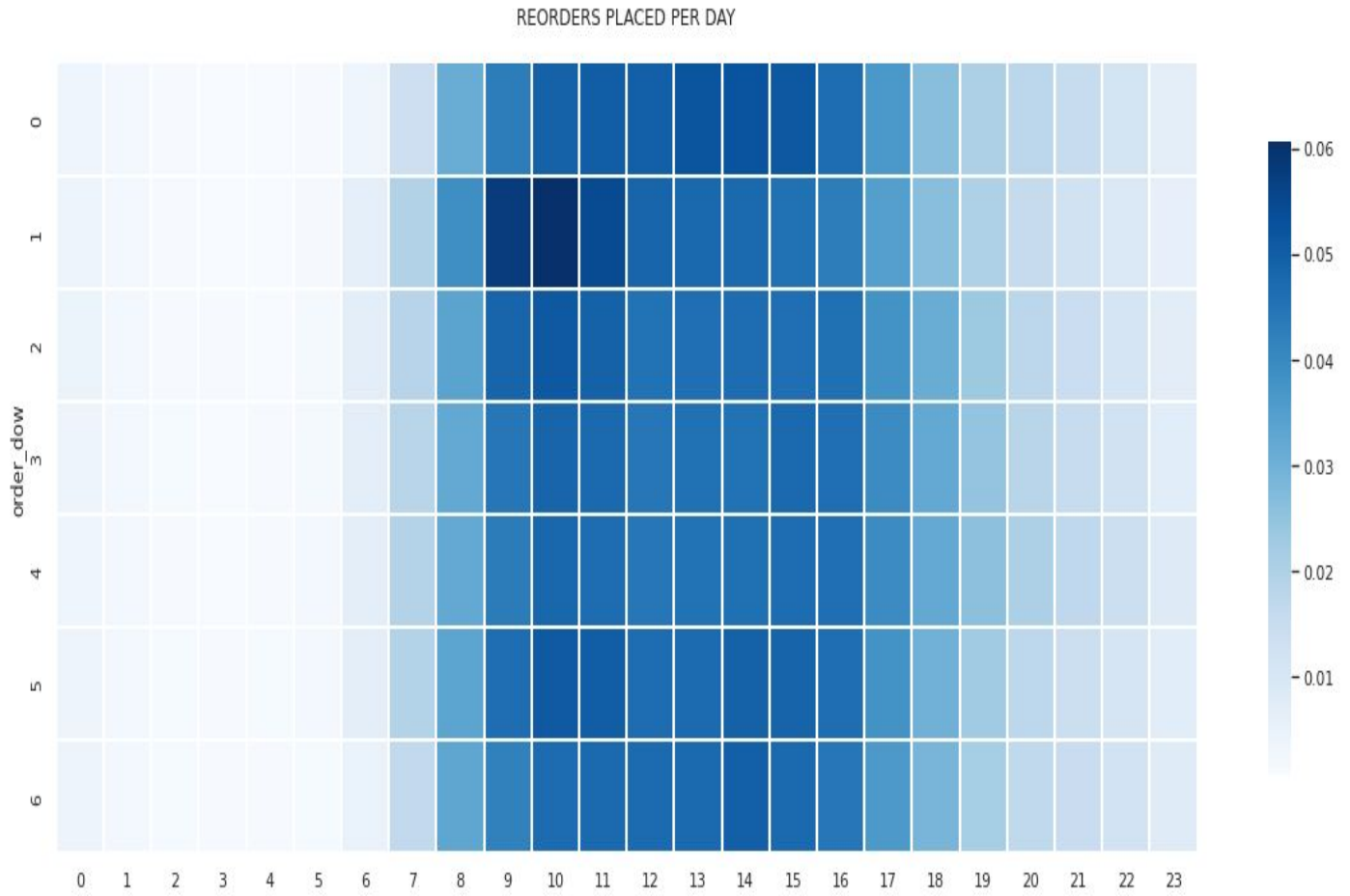
From the graph we can see that the number of users are most active from 10 am to 5 pm.
It would be more advisable to broadcast advertisements during this particular time.

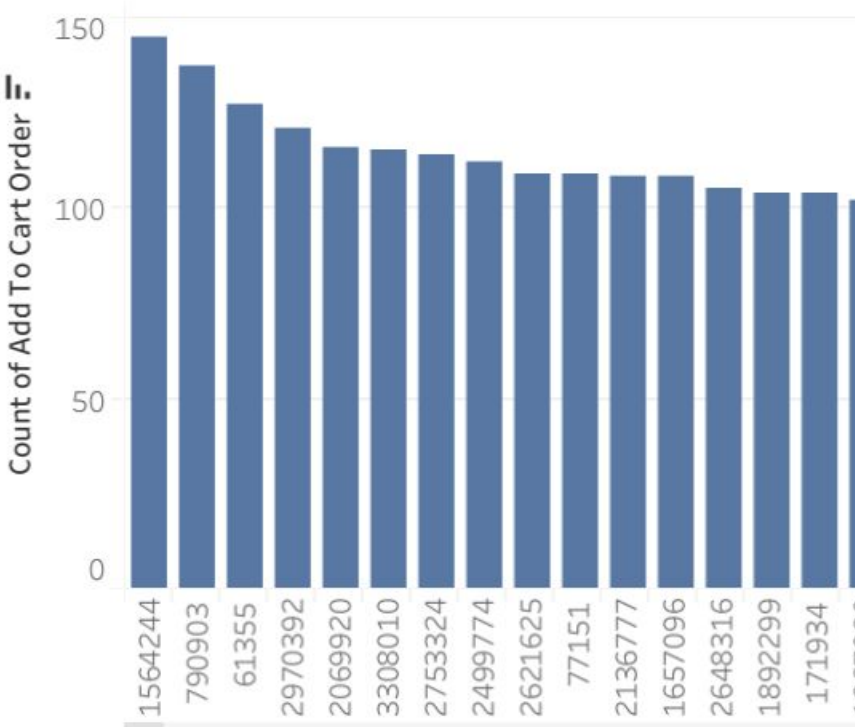
These graph shows us the number of days after which a particular product has been reordered. It can be seen that there is a spike for reorders at the 7 day (one week mark) a small spike on the 14 day (2 week mark) and a massive spike at the 30 day mark. These are the days approximately when the orders will be reordered.



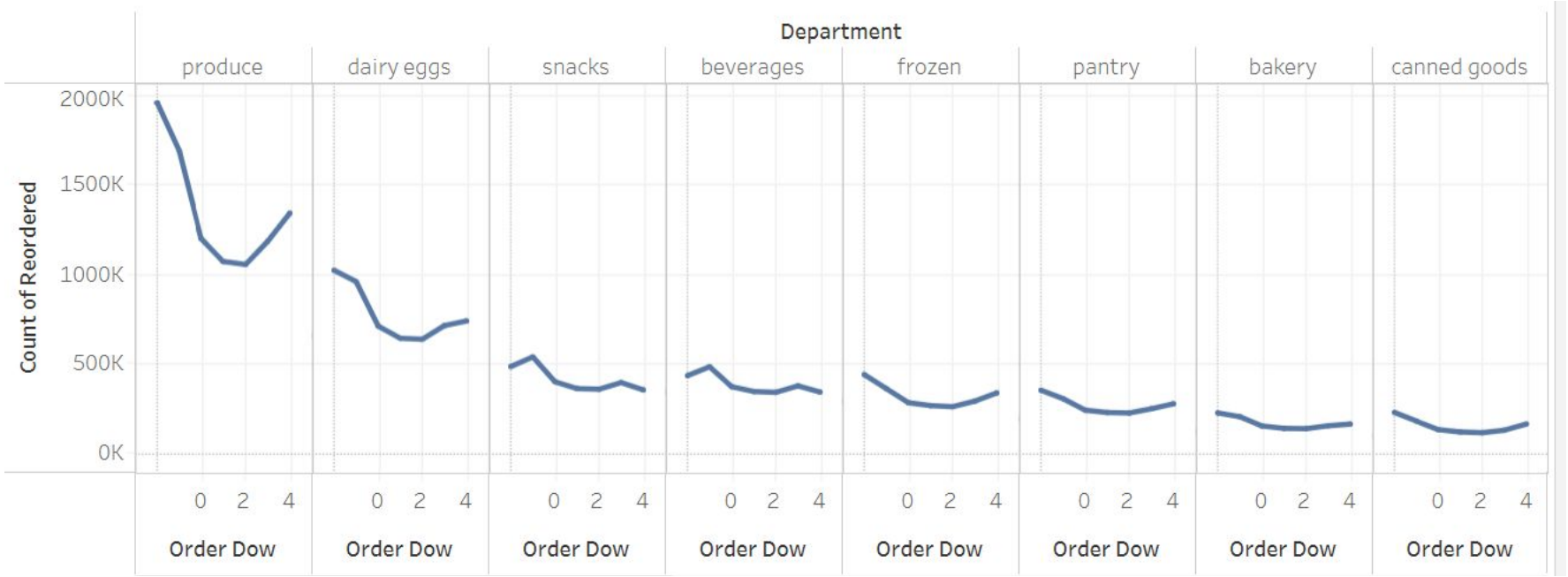
Reorder rate of orders placed every day

- Plot describes reorder rate of every day w.r.t orders placed on that day.
- We can see that of all orders that were placed on any day, most reorders were placed from 8 AM — 4 PM, on any given day





The maximum items placed in a single order are 145 items and the minimum order placed and the minimum is 1.



From the above graph we can see that the reorder for each department tends to decrease past day 0 except for Produce and dairy eggs.

EDA Summary Report:

- The dataset has no missing values except for in orders.csv. Only **6%** of values are missing from days_since_prior_order column.
- The target variable is '**reordered**' column and its distribution is similar in both prior and train set.
- For every user we have around 4–100 order details.
- We observed that on any day - any hour most frequently ordered product is **Banana and Bag of organic bananas** and
  frequently ordered aisle is **fresh fruits and vegetables** and department is **produce**.
- Bananas have highest order rate and top 5 frequently ordered products are **organic** in nature.
- **Milk, sparkling water, fruits, eggs, yogurt** are most common aisles the product is reordered from, as they are items which are daily consumed, and one rarely switches from their usual meal plan. Also these are the products that lasts only few days , thus high reorder rate.
- We observed high reorder rate in **organic foods** and daily consumed items and low orders in **personal care department.**

- Most shopping is done on **Sundays and Mondays**. Also least orders were placed on **Thursday**. People tend to restock their supplies on Sundays.Also we can see that of all orders that were placed on any day, most reorders were placed from **8 AM — 4 PM**, on any given day.
- Most people restock after a **week or a month**. It seems, some people prefer buy a week / month supplies at once.
- There are 3 products (Protein Granola Apple Crisp, Unpeeled Apricot Halves in Heavy Syrup , Single Barrel Kentucky Straight Bourbon Whiskey) which were never ordered. May be their alternatives were ordered.

VI. Model Development and Evaluation

After having researched various algorithms we explored 4 algorithms and compared the accuracy of them to decide which model would be best suited for our analysis.

The 4 algorithms were -

1. Linear regression
2. Decision Tree
3. Random Forest
4. Light GBM.

We Split the dataset into training and test data and checked the accuracy based on various metrics.

1. RMSE
2. MSE
3. MAPE

Linear regression

Linear regression is a statistical method used to analyze the relationship between two continuous variables, where one variable (the dependent variable) is predicted based on the other variable (the independent variable) using a straight line. The line is determined by finding the best-fit line that minimizes the difference between the predicted and actual values of the dependent variable. This method is commonly used in data analysis and is a simple yet powerful tool for making predictions and understanding relationships between variables.

Conclusion:

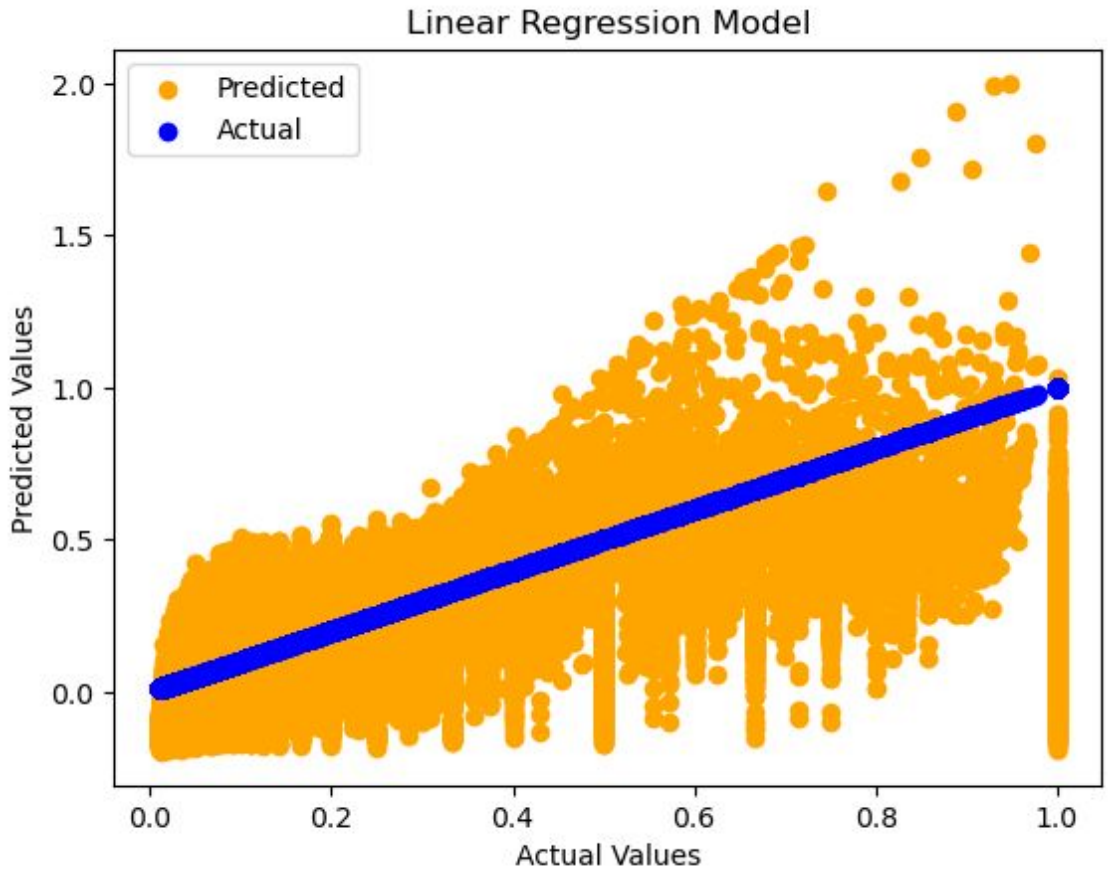
The image shows the plot of the actual vs the predicted values for the LInear Regression Model. The model shows linear and positive relationship but the relationship is weak since the R-squared value is not high.The higher the R-squared value, the more accurately the regression equation models your data. Moreover, data shows some negative predictions as well

The metrics associated with this model are -

RMSE - 0.241877

MSE - 0.058504

MAPE - 1.265804



Decision Tree Model

A decision tree model is a popular machine learning algorithm used for both classification and regression tasks. It is a tree-like model where each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label or a numerical value.

The decision tree algorithm starts with the entire dataset and recursively splits the data into smaller subsets based on the most significant attribute until a stopping criterion is met. The
splitting process continues until the tree reaches a predetermined
depth or no more significant attributes are left to split on.

Conclusion:
The image shows the plot of the actual vs the predicted values for the Decision Tree Model. The model shows

linear and positive relationship and it is strong,

since the R-squared value is high. There seems to be pattern from

the model but data is widespread showing higher chances of

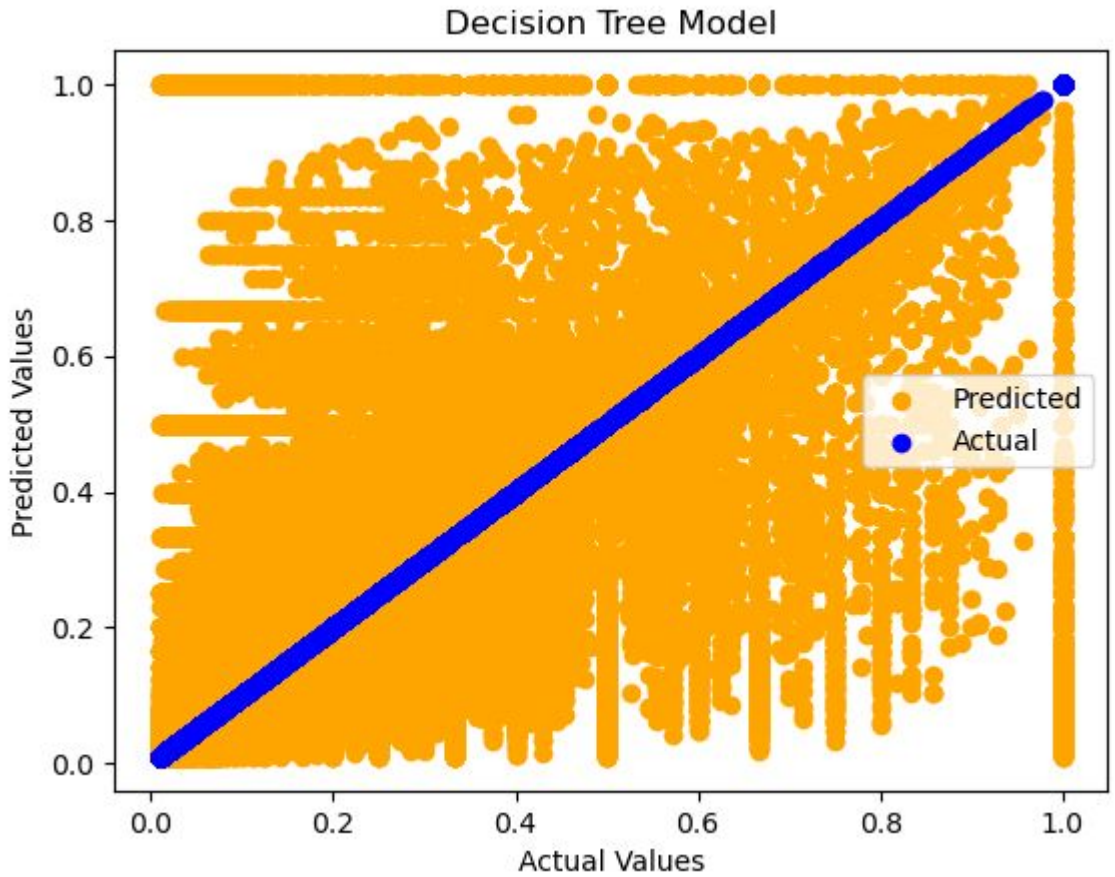
wrong predictions

The metrics associated with this model are -

RMSE - 0.315096

MSE - 0.099285

MAPE - 1.173060



Random Forest Model

Random Forest is a popular machine learning algorithm used for both classification and regression tasks. It is an ensemble method that combines multiple decision trees to create a more robust and accurate model. Random Forest works by creating a set of decision trees, each of which is built on a random subset of the original data and a random subset of the attributes.

The algorithm then uses these trees to make predictions by taking a vote among the individual tree predictions.
One of the main advantages of Random Forest is that it can handle high-dimensional data and is less prone to overfitting than a single decision tree. Additionally, it can provide a measure of feature importance, which can be useful in identifying the most important variables in the dataset.

Conclusion:
The image shows the plot of the actual vs the predicted values for the Random Forest Model. The model shows

linear and positive relationship and it is strong,

since the R-squared value is high. The actual values are

associated with high changes in predicted values. But, data is

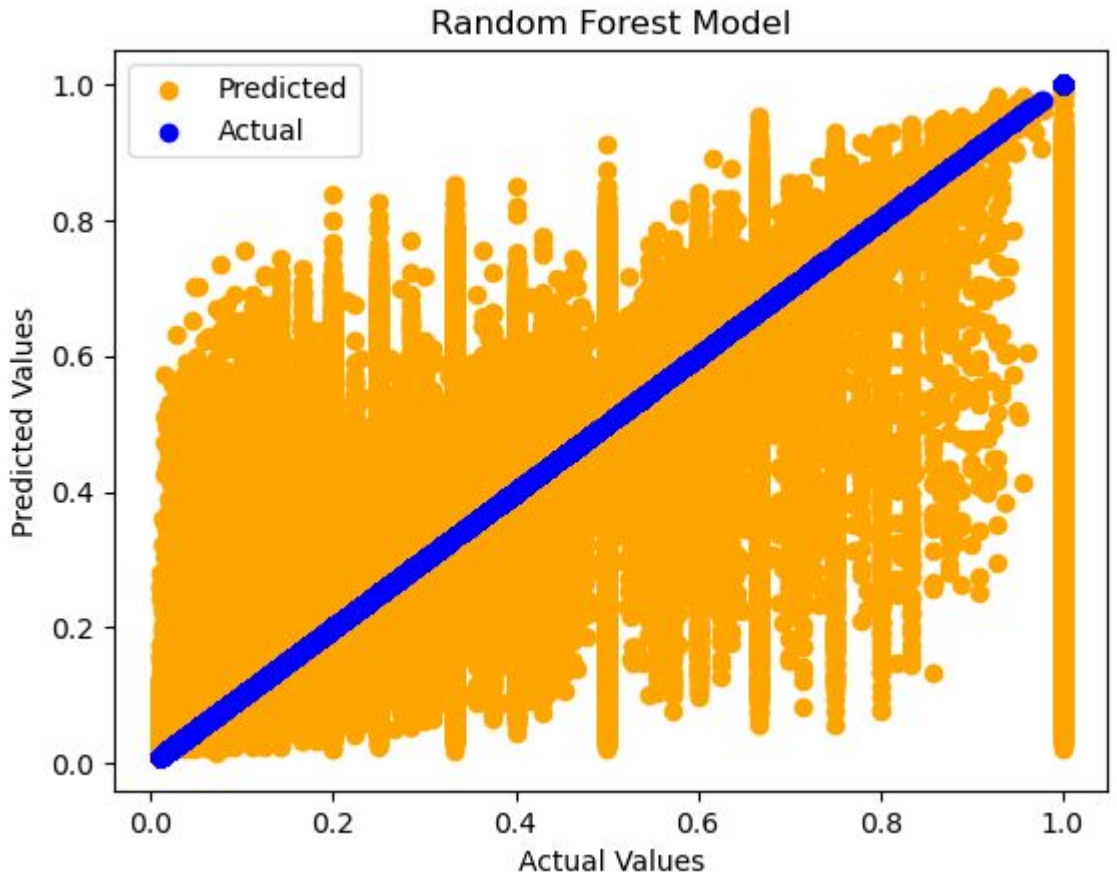
widespread for this model.

The metrics associated with this model are -

RMSE - 0.223143

MSE - 0.049793

MAPE - 1.002330



Light GBM

Light GBM (Gradient Boosting Machine) is a machine learning algorithm used for both classification and regression tasks. It is a type of boosting algorithm that works by iteratively building weak decision trees and combining them into a strong predictor.

Light GBM differs from other boosting algorithms in that it uses a gradient-based approach to select the best split points, making it faster and more efficient for large datasets.
It also uses a leaf-wise approach rather than a level-wise approach, which further improves its speed and reduces memory usage.

Conclusion:
The image shows the plot of the actual vs the predicted values for the LightGBM Model. The model shows

linear and positive relationship and it is strong,

since the R-squared value is high. The data points are

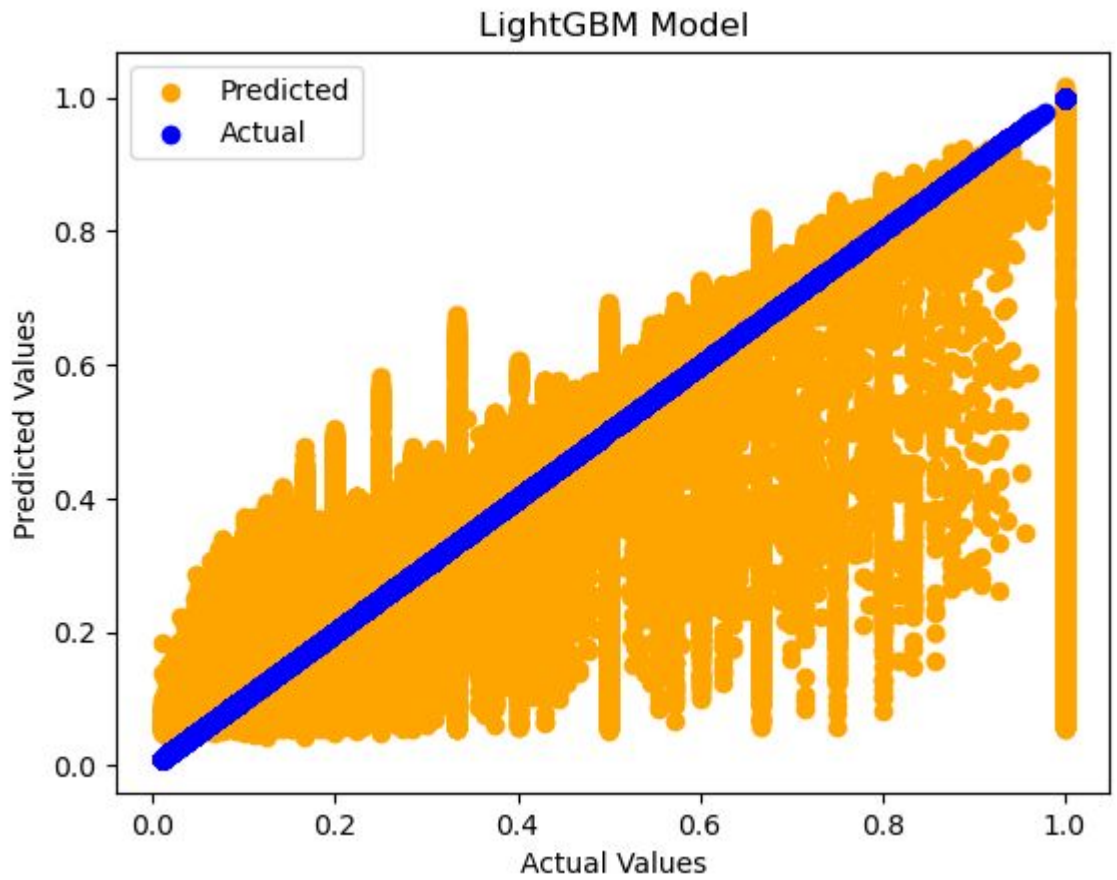
clustered more tightly than other models.

The metrics associated with this model are -

RMSE - 0.217481

MSE - 0.047298

MAPE - 0.899919



Metrics to analyze the accuracy.

RMSE: Root Mean Square Error

RMSE is calculated as the square root of the average of the squared differences between the predicted and actual values. It is expressed in the same units as the predicted and actual values.
RMSE is a measure of the difference between the predicted values and the actual values, with lower values indicating better accuracy. It is commonly used in regression analysis to evaluate the performance of models that make continuous predictions.

MSE: Mean Square Error

MSE is calculated as the average of the squared differences between the predicted and actual values. It is expressed in the square of the units of the predicted and actual values.
MSE is a measure of the difference between the predicted values and the actual values, with lower values indicating better accuracy. It is commonly used in regression analysis to evaluate the performance of models that make continuous predictions.

MAPE: Mean Absolute Percentage Error

MAPE is calculated as the average of the absolute percentage differences between the predicted and actual values. It is expressed as a percentage.
MAPE is a measure of the percentage difference between the predicted values and the actual values, with lower values indicating better accuracy. It is commonly used in forecasting and time series analysis to evaluate the performance of models that make continuous predictions.

Comparison of metrics

	Algorithm	RMSE	MSE	MAPE
1	Decision Tree	0.315096	0.099285	1.173060
2	Random Forest	0.223143	0.049793	1.002330
3	Linear Regression	0.241877	0.058504	1.265804
4	Light GBM	0.217481	0.047298	0.899919

The above table represent the measure of model by analysing above table we can say that RMSE metrix of light GBM model gives the most accurate result. because lower the value more accuracy will be there.

Hence we recommend to move forward with Light GBM model.

Final Predictions

Products	Department	Category
Organic Whole Milk	Dairy	High
Low fat Chocolate Milk	Dairy	High
One day one Banana Pack of 4	Produce	High
Natural Artesian Water 12 fl oz	Beverages	High
Baby Spinach	Produce	High
Chocolate Whole Grain Graham Snacks Bunny Grahams	Snacks	Mid
Organic Whole Strawberries	Frozen	Mid
Light French Silk Slow Churned Ice Cream	Frozen	Mid
Spaghetti Protein Plus Pasta	Dry Goods	Mid
Brut Sparkling Wine	Alcohol	Mid
Black Beans	Canned Goods	Low
XL Emerald White Seedless Grapes Preserved	Canned Goods	Low
Authentic South River Miso Paste	Canned Goods	Low
Organic Vegetable Quinoa Soup	Canned Goods	Low
Free & Clear Bleach	Household Items	Low

VII. Business Recommendations and Conclusions.

From the models created and after analysis of their accuracy we have come to the conclusion that the best algorithm for our dataset would be to use the LightGBM model.

The model is based on the previously ordered items the customer has made. The predictions made by the model can be used in many ways as discussed in the beginning, mainly -

1. Advertisement
2. Dynamic Pricing
3. Inventory Management

Based on the final prediction list we have also observed that the most in-demand products belong to the perishable items (dairy & produce) along with small quantity frequently consumed beverages. Various offers can be made to customers to target these items.
Further based on the analysis on the types of orders being sold at particular days and times of the days, target advertisements could be done for the customers.

सिद्धिमूल प्रबन्धनम्
भा. प्र. सं. इन्दौर
IIM INDORE

Thank You

jigsaw
A UNEXT COMPANY