Capstone Project: Instacart Market Basket Analysis

ANUSHREE SRINIVAS

MENTOR: SRDJAN SANTIC

Problem to be solved and Motivation

- Instacart, a grocery ordering and delivery app, aims to make it easy to fill your refrigerator and pantry with your personal favorites and staples when you need them
- 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
- The objective of this capstone is to address two research questions:
  - Predict whether a product will be reordered or not in the future by the customer
  - Predict which department will the next product ordered belong to
- The ability to identify which products the customers are likely to purchase again, and automatically adding those to cart through obtained predictions or provide a seamless interface for doing so will enhance their user experience

Client

- Instacart is looking to use this analysis to better serve their customers.

- The data science team at Instacart will be the client for which the conducted data analysis as part of the capstone project will be beneficial.

Feature Engineering for prediction of Reordered products

- **Order related features**
  - Order_id
  - Order_number
  - Average_days_between_orders
  - Nb_orders(Number of orders)
  - Average_basket
- Total items
- Aisle
- Department
- Product
- User_id
- **Time related features**
  - Order_hour_of_day
  - Order_dow(day of week)
  - Days_since_prior_order
  - Days_since_ratio

Feature Engineering for Department prediction

- **Order related features**
  - Order_id
  - Order_number
  - Average_days_between_orders
  - Nb_orders(Number of orders)
  - Average_basket
  - Orders
  - Reorders
  - Reordered rate
- Total items
- User_id
- **Time related features**
  - Order_hour_of_day
  - Order_dow(day of week)
  - Days_since_prior_order
  - Days_since_ratio

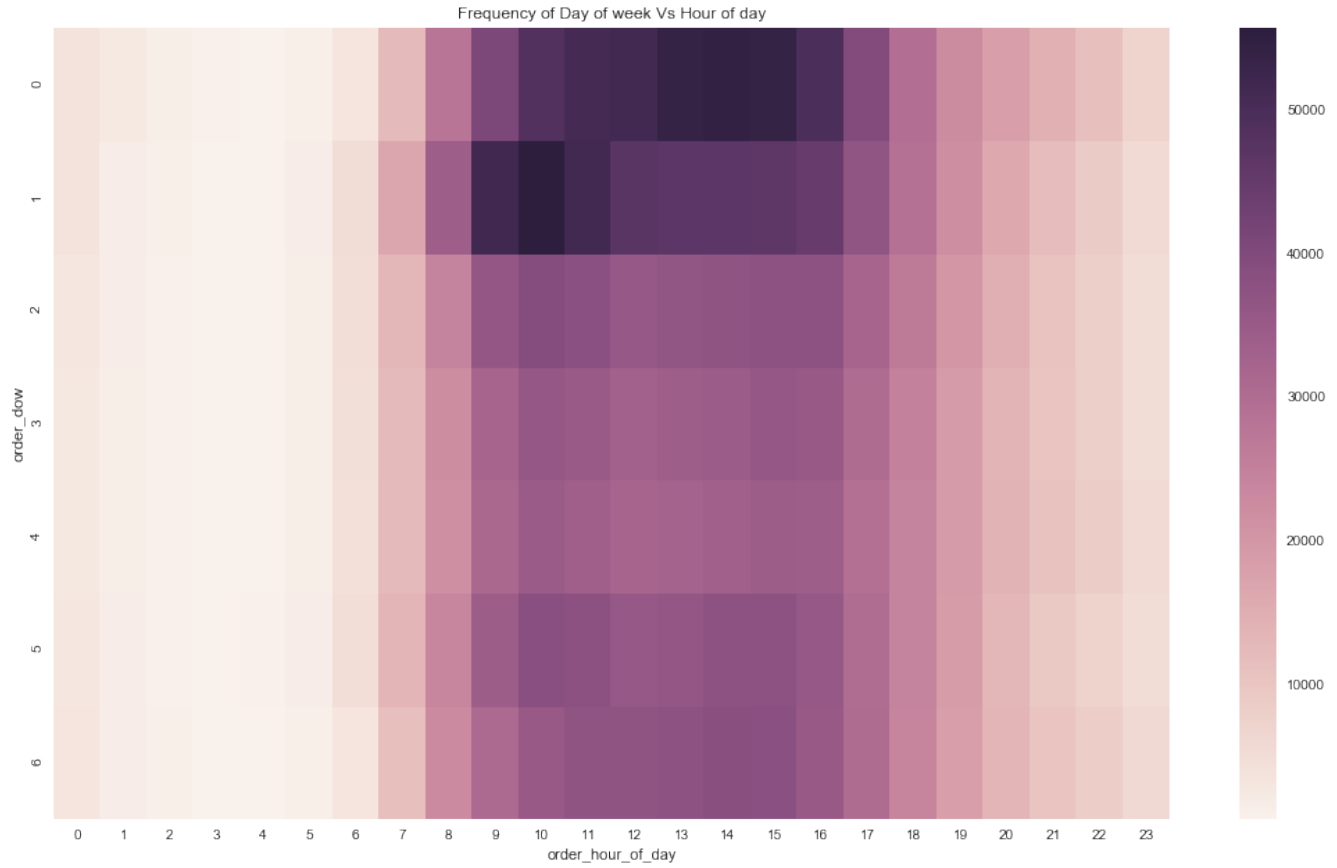
Exploratory Data Analysis

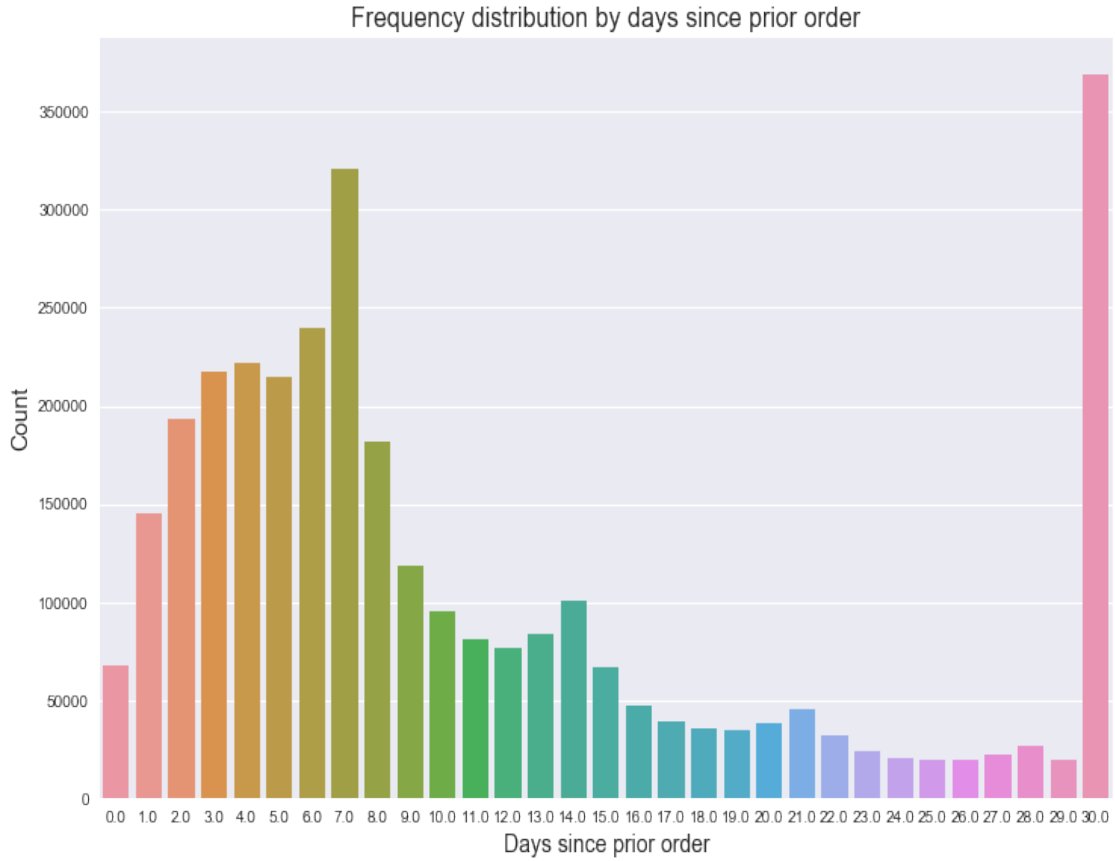
- The number of orders is maximum on Sunday followed by Monday.
- Thursday has the least number of orders.
- Most orders on Sunday are placed between 2-3pm.
- On Mondays, most orders are placed between 9-11AM.
- Weekends, peak orders are in the afternoon from 2-4pm.
- Whereas in the weekdays, it's in the morning from 10AM-12PM.

Exploratory Data Analysis

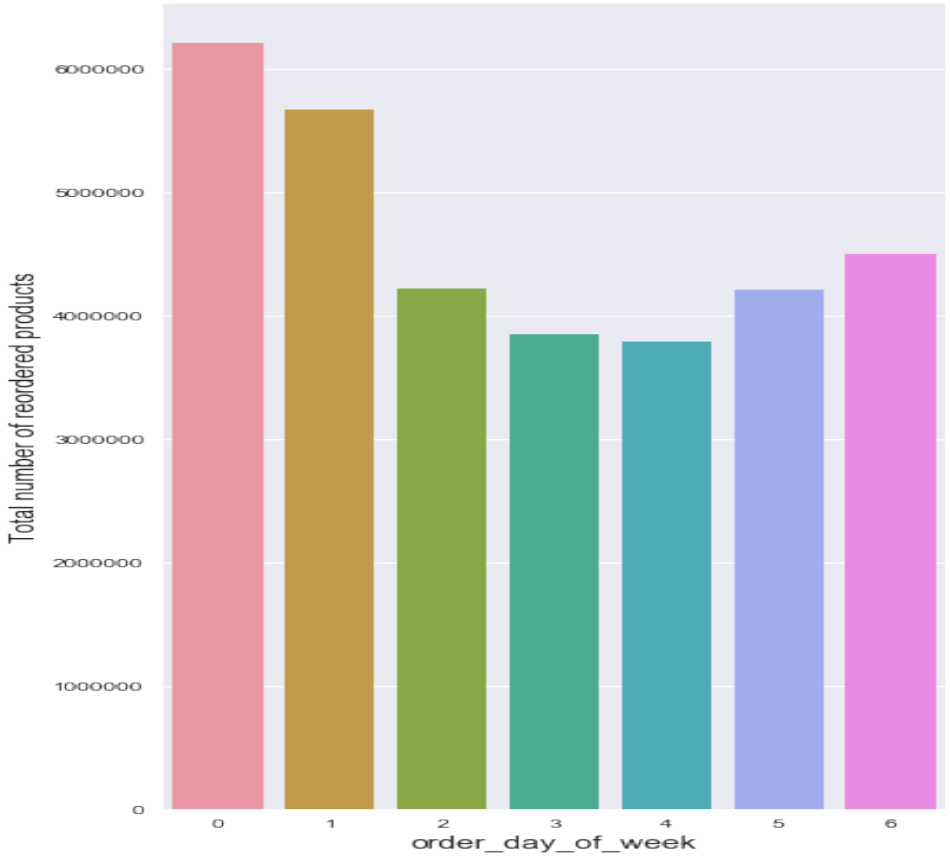
- Customers generally order weekly. And there's a monthly peak as well.
- Most ordered products are fruits like bananas, strawberries and organic products.
- The fresh food and fresh vegetables aisles are the most frequently visited.
- Department wise frequency is most for produce and dairy eggs.
- Most products are reordered on Sunday followed by Monday and Saturday
- Most products are reordered from 10-11AM followed by 1-3pm.

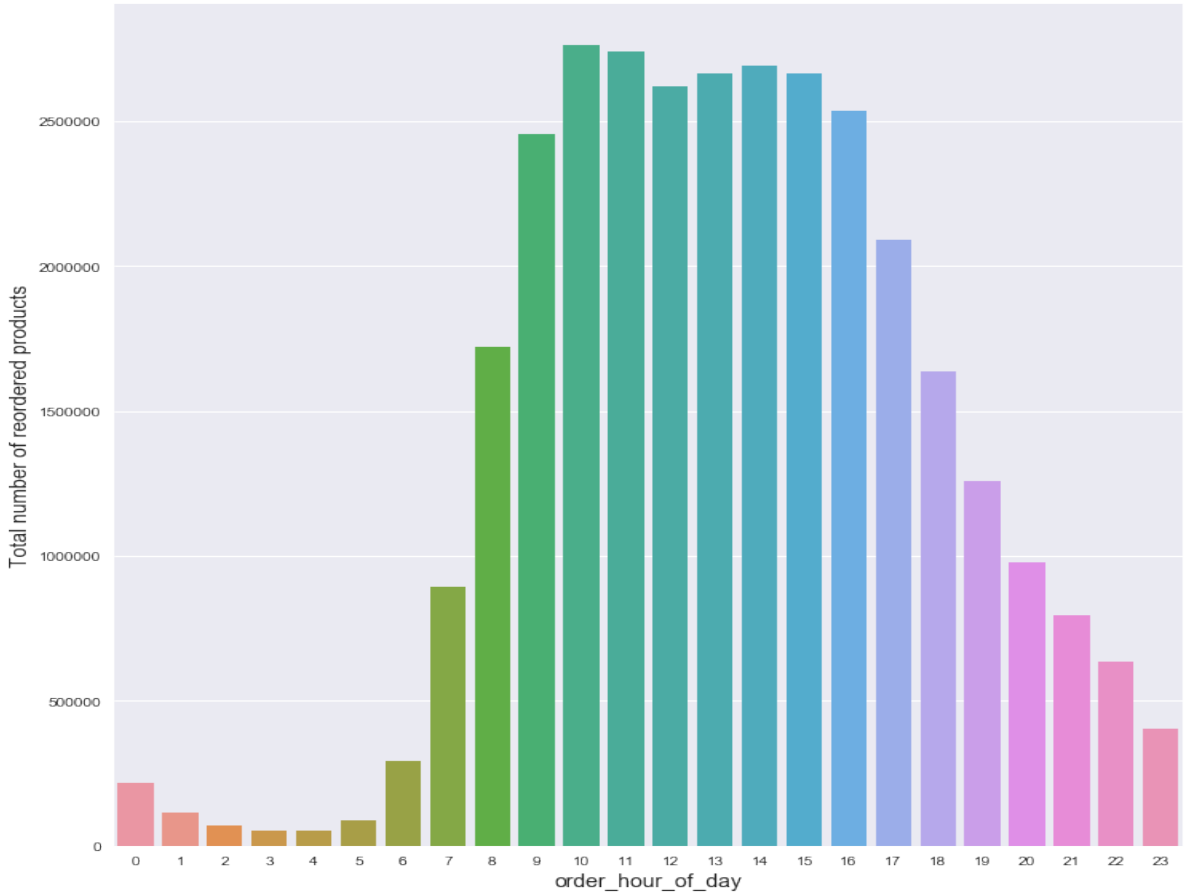
Visualizing Instacart User Behavior





Visualizing Instacart User Behavior





Algorithms and Results

- Research question 1: Predict a product will be reordered or not

Models	Accuracy Score
Logistic Regression	59.7%
AdaBoost Classifier	65.6%
RandomForest Classifier	66.6%
Gradient Boosting Classifier	**67.1%**

Algorithms and Results

- Research question 2: Predict the department from which a product will be ordered

Models	Log loss score
Random Forest Classifier	**2.342**
Gradient Boosting Classifier	2.344
Adaboost Classifier	2.979

Recommendations for the Client

- These analyses can be used to run promotional and marketing campaigns targeting specific customers during peak time.
- The insights generated can be used to provide a seamless interface to enhance customer's user experience by knowing about the customer's reordered products and automatically adding those to cart.
- Personalized communications can be sent to customers' preferences, reminding them to order again.
- To improve customer satisfaction by timely delivery and reduce wait time, the shopper base can be increased by hiring new shoppers who can especially work around the peak time.

Future Research

- **Try non-linear models:** The models that were used in here were all linear models. Non-linear models could be implemented to see if better results can be achieved.
- For better predictions, market basket algorithms such as apriori can be implemented.
- For predicting whether a product is reordered or not, algorithms that predict binomial categories better can be used.
- For predicting a multi-category variable like department, other multi-nomial algorithms can be applied.
- **New features:** New features could be created to help us generalize better on the test dataset thereby achieving better results.