instacart

Market Basket Analysis

By: Mario Samalo

Outline

- Objective
- Data
- Data Exploration
- Feature Extraction
- Models Summary
- LGB Model Discussion
- Recommendations

Objective

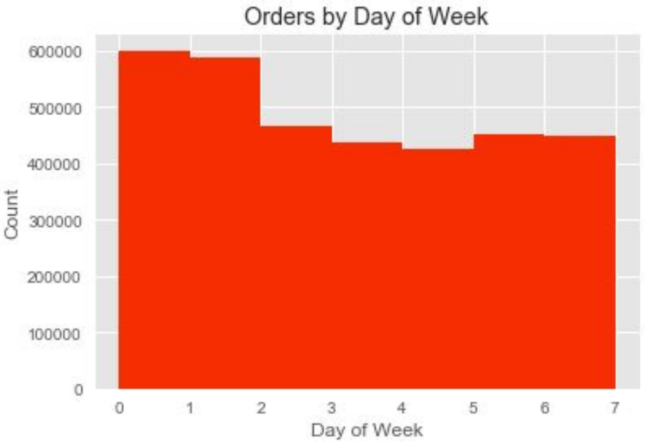
Find the significant factors that affect the probability of a customer returning and ordering again and provide optimal recommendations to customers to increase overall shopping and browsing experience, increase revenue from sales, and increase overall customer satisfaction.

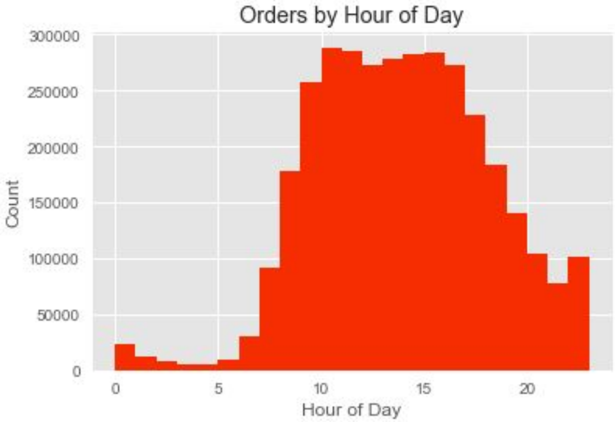
Data

- The data can be obtained from kaggle, https://www.kaggle.com/c/instacart-market- basket-analysis/data.
- The dataset is a relational set of files describing Instacart 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 data was comprises 6 different tables: Orders, Prior Order Data, Products, Departments, Aisles, Order Product Train

Data Exploration (#1)

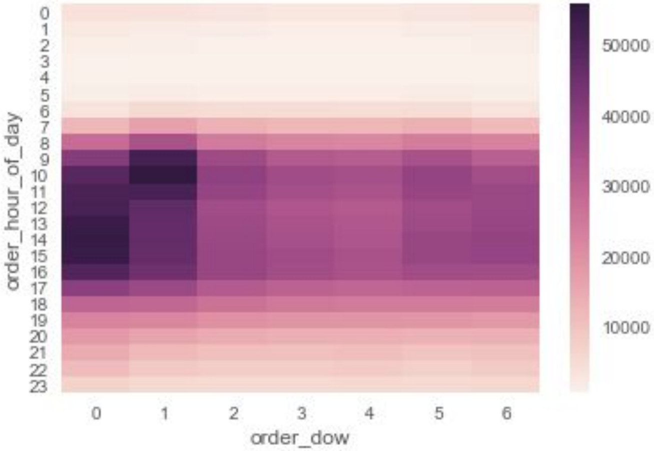
- Day 0 and Day 1 have the most number of orders, but it is unclear what days Day 0 and Day 1 represent (Left)
- Peak hours are between 9AM and 5PM (Right)



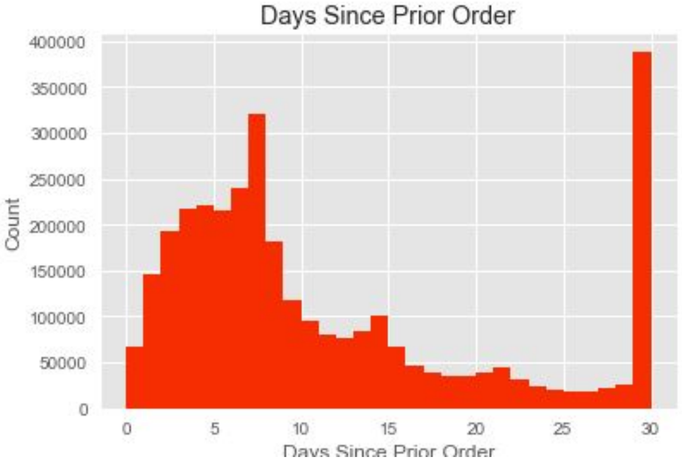


Data Exploration (#2)

The peak day and hours combination occurs on day 0 and day 1 between 9AM and 5PM.

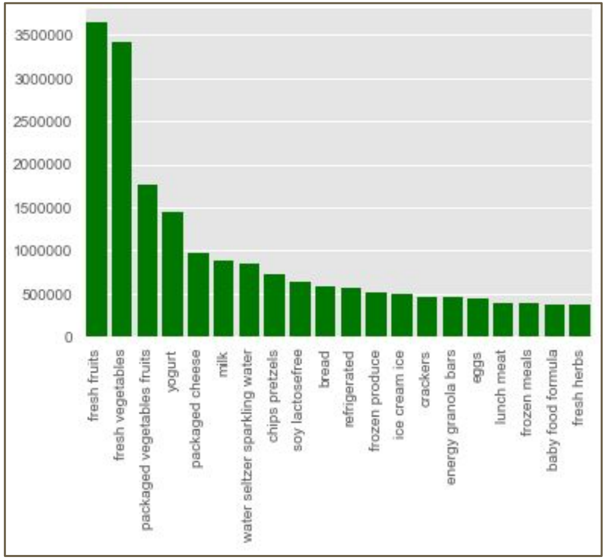


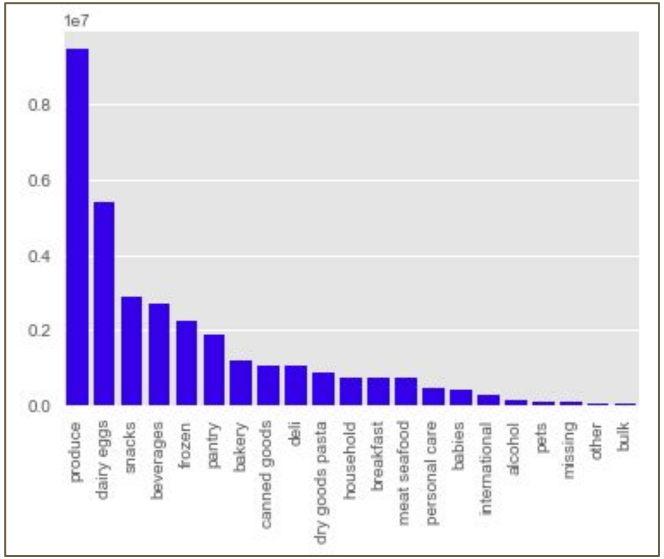
A lot of customers put another order after a week or a month which makes sense because people tend to reorder after a week or a month



Data Exploration (#3)

- The top products (Green) and aisles are those of fruits and vegetables
- Produce department dominates orders which is consistent with fruits and vegetables as the top products (Blue)

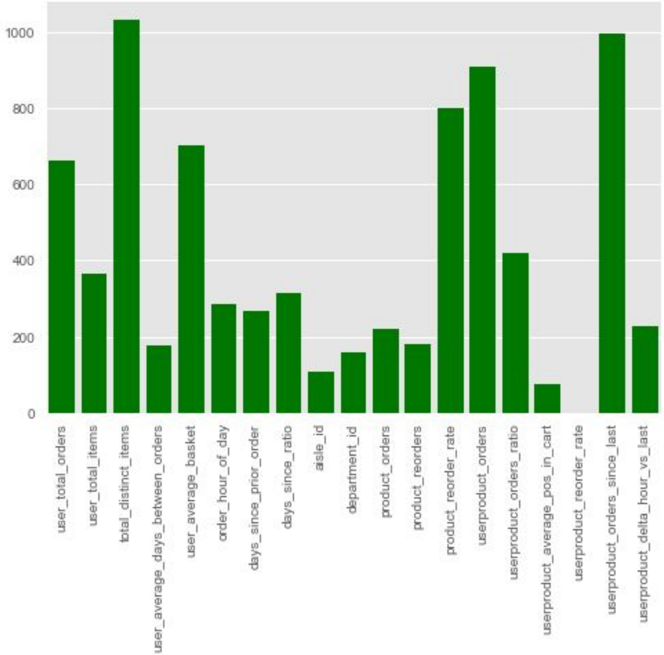




Features

The most important features turns out to be (we picked those with score greater than 100) :

- User total orders
- User total items
- Total distinct items
- User average basket
- Order hour of day
- Days since prior order
- Days since ratio
- Product reorder rate
- User-product orders
- User-product orders ratio
- User-product orders since last
- User-product delta hour vs last



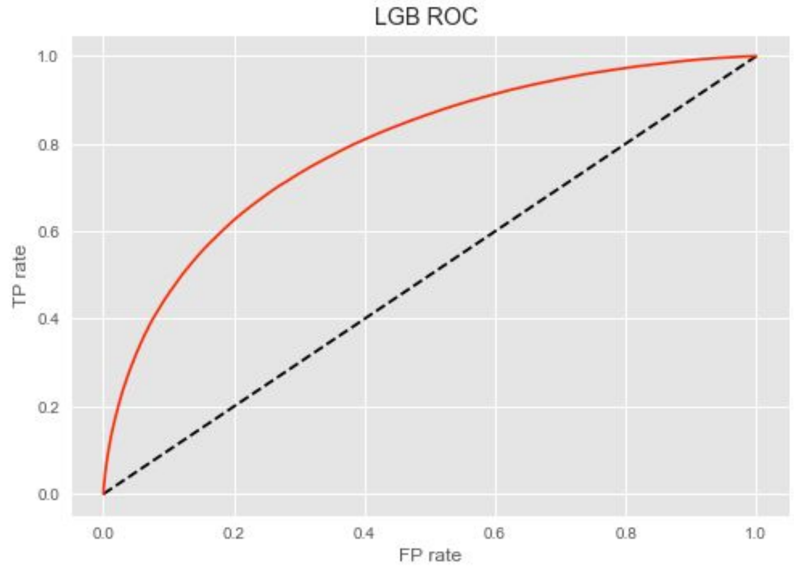
Models Summary

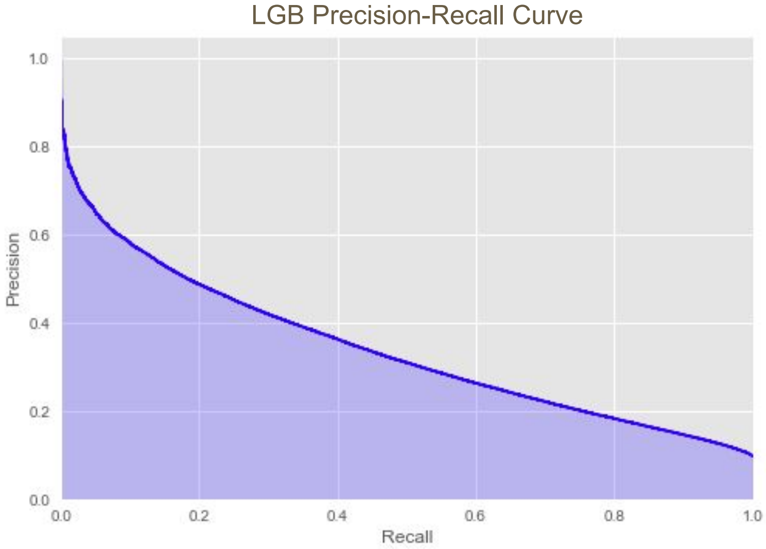
We tried 3 models and fine tuned their parameters using 3 folds cross validations. The best model (based on F-1 & AUC score) turns out to be Light Gradient Boosting and the overall results are summarized below

Models	Threshold	F-1 Score	AUC
Logistic Regression	0.22	0.29	0.61
Decision Tree	0.21	0.34	0.65
Light Gradient Boost	0.2	0.38	0.67

LGB Model Discussion

- After training the model, we found the best combination of parameters.
- Using that combination we were able to achieve an AUC up to 67%
- This model is the most accurate compared to the other two models





Recommendations

- Important metrics to look at are total orders, total items, distinct items, average basket size,how long has it been since the customer places order, reorder date, and so on

- Keep reminding the customer to reorder their groceries on say a weekly basis since it is shown that the longer it takes one to re-order his or her groceries, the less likely he or she is going to reorder

- Recommend products the customer has high reorder rate since products that have high reorder rate tends to increase probability of them getting reordered

- Include related products as recommendations or “products you may like” as it is shown that customers that order a lot of products tend to reorder more

Thank You !