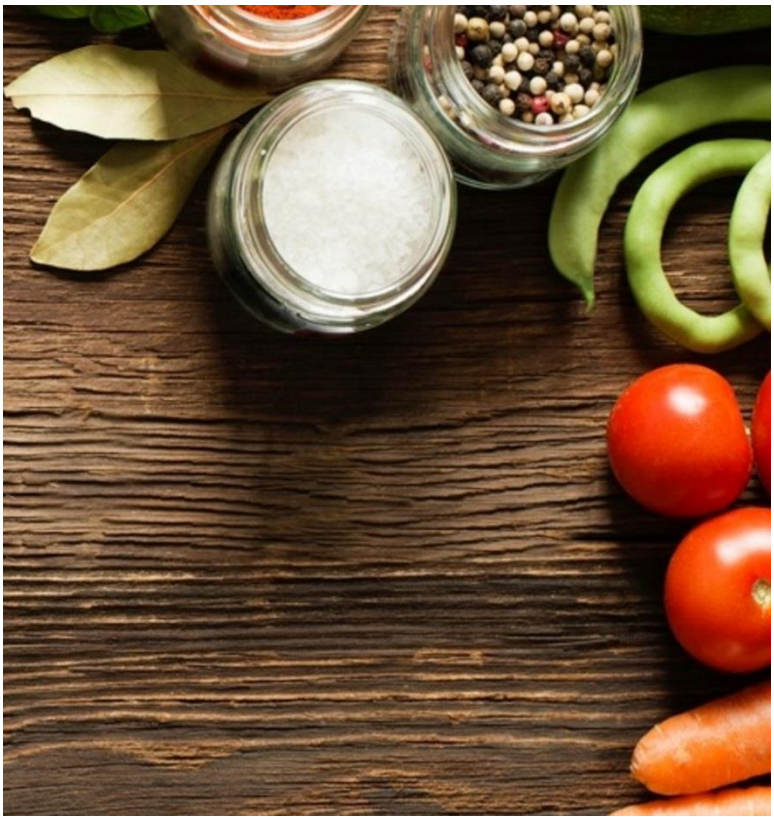

Instacart Market Basket Analysis

PREDICT NEXT BASKET

Prajakta Gujarathi | Capstone Project | 10/5/2017

Definition

PROJECT OVERVIEW:

Instacart, a grocery ordering and delivery app, aims to make it easy to fill customer's refrigerator and pantry with their personal favorites and staples when they need them. After selecting products through the Instacart app, personal shoppers review order and do the in-store shopping and delivery for customers. But achieving this simplicity cost effectively at scale requires an enormous investment in engineering and data science.

Knowing customer's next order will be helpful for Instacart business in following way:

- Optimize algorithm which enroute Instacart shoppers, for timely and efficient delivery
- Balancing supply and demand of customers. This includes estimating Instacart's capacity to fulfill orders to create optimal customer experience.
- Provide customers with recommendation by analyzing their previous purchase history.

PROBLEM STATMENT

In this project goal is to use previous transactional data of customer to develop models that predict which products a user will buy again. Task involve are following:

1. Download data from Recently, Instacart open source data from 3 Million Instacart Orders, Open Sourced.
2. Explore, visualize and analyze important dimensions
3. Preprocess data by adding or reducing dimension as required.
4. Train a different models using preprocessed data and identify best model
5. Predict next carts products for test users.
6. Use test data to analyze final model performance.

Metrics:

F1 is a common metric for binary classifiers; it takes into account both true positives and true negatives with equal weight.

$$PRE = \frac{TP}{TP + FP}$$

$$REC = TPR = \frac{TP}{P} = \frac{TP}{FN + TP}$$

$$F_1 = 2 \cdot \frac{PRE \cdot REC}{PRE + REC}$$

TP = will be product which are actually present in next order.

FP = will be product that incorrectly predicated to be in next order.

TN = will be product which are not present in next orders.

FN = will be products which are supposed to be in next order be predicated as not.

Project Design:

- EXPLORATORY DATA ANALYST:

  Perform EDA on data in order to find out:

  1. Important features
  2. How the data distribution
  3. Handling missing data
  4. Calculated statistics relevant to the problem

- PREPROCESSING AND DATA PREPARATION:

  In this step I merge data from Orders.csv, Orders_product_prior.csv, order_products_train.csv, aisle.csv and department.csv.

  Create few new features that I feel are important like reorder ratio, user order ratio, average numbers orders in a cart etc.

- ## MODEL CREATION:

  In this step I applied various model on preprocessed data like Logistic regression, Adaboosting, Xgboost and light gradient boosting etc

  Using Grid Search tuned the model parameters.

- ## MODEL EVALUATION:

  For model evaluation I am going to used accuracy and f1 score used for binary classification.

- ## PREDICT TEST DATA VALUES:

  Final best model is then used to predict next orders of Test data. Prediction are saved in format order_id followed by Product_ids present in that order.