**Name:** Suraj Kumar Padhy

**Email address:** skp309skp@gmail.com

**Contact number:** 7978440416

**Anydesk address:**

**Years of Work Experience:** 1.2

**Date:** 20th Oct 2020

## Self Case Study -1: Instacart market basket analysis

"After you have completed the document, please submit it in the classroom in the pdf format."

Please check this video before you get started: https://www.youtube.com/watch?time_continue=1&v=LBGU1_JO3kg

## Overview

*** Write an overview of the case study that you are working on. ***(MINIMUM 200 words)*** ***

1. Instacart is an online grocery shopping app. Back in 2017 instacart announced a dataset release, which is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 Instacart users.

2. In this Kaggle competition we use this instacart dataset to perform market basket analysis, means based on user baying pattern we need to find which products will get reordered in the next order.

3. The goal of this competition was basically we need to predict which all products will be reordered in the next order given user prior purchase history(  set of previous orders, products in that previous orders).

4. The given dataset comes with 6 different csv files

    i. Order.csv:-provides metadata for each order such as(order id, user id, eval set, order number, order day  of week, order hour of day, day since prior order)

    ii. eval set(prior ,train, test):- denotes which set order belongs to.

        1. If a user is train user its prior orders are present in order_product_prior.csv and the last order present in order _product_train.csv.

        2. If the user is a test user its prior orders present in order_product_prior.csv and we need to predict the reordered products for this order id.

    iii. order_product_prior.csv, order_product_train.csv:- These files specify which products were purchased in each order (order id, product id, add to cart number, reordered.

    iv. Product.csv (product id, product name, aisle id, department id), aisle.csv (aisle id, aisle), departments.csv (department id, department):- aisle and department are only used to classify products.

5. In this competition's they use F1 score as performance metric which makes sure our models have both high precision and high recall.
6. Finally we need to predict grocery reorders given a user's purchase history.

## Research-Papers/Solutions/Architectures/Kernels

*** Mention the urls of existing research-papers/solutions/kernels on your problem statement and in your own words write a detailed summary for each one of them. If needed you can include images or explain with your own diagrams. it is mandatory to write a brief description about that paper. Without understanding of the resource please don't mention it***

1. https://www.kdd.org/kdd2016/papers/files/adf0160-liuA.pdf in the paper we can understand how to do featurization for E-Commerce such do some feature engineering on count/ratio features, aggregation features, recent activity features, etc. From this paper we can conclude there is no features that generated is a strong indicator of class labels, so we need hundreds of features.

2. https://kimiyoung.github.io/papers/fang-ijcai-2015.pdf in this paper is for E-Commerce Repeat Buyer Prediction, we can use the way of fraternization of this problem to our problem we can featurize instacart data such that some of the new features belong to user related features, product related, user and product related, and some repeat features related to user and product.

3. https://medium.com/kaggle-blog/instacart-mark et-basket-analysis-feda2700cded in this blog the author used XGBoost to create two gradient boosted tree models one for predicting reorders (which previously purchased products will be in the next order) and second model is for (will the user's next order contain any previously purchased products).

4. https://www.kaggle.com/c/instacart-market-basket-analysis/discussion/38097 in this discussion the approach was to fit a variety of models to the prior data and use the internal representations from these models as features to second-level models.

5. https://www.kaggle.com/paulantoine/light-gbm-benchmark-0-3692 hear in the kernal they created some features  related to user, product, product and user and trained a lightGBM model, which is a gradient boosting framework that uses tree based learning algorithms in which the tree grows depth wise first.

---

## First Cut Approach

*** Explain in steps about how you want to approach this problem and the initial experiments that you want to do. ***(MINIMUM 200 words)*** ***

*** When you are doing the basic EDA and building the First Cut Approach you should not refer any blogs or papers ***

1. Form the research we can assume that the product which has been ordered many times in history have the high chance that it will be reordered in the next order, so we need to find top ordered products and we can generally recommend these products so that the user can consider the product again.

2. Next, we need to understand in which hour of the day the users purchase items and what is the interval between two consecutive orders this will give insights about how the customer is behaving.

3. We also need to know that from which aisles and departments many products are getting orders because of which we can say that the products from these aisles and departments might have the higher probability of getting reordered.

4. Next we need to investigate how many products were present in each order, how often, how many times they were ordered and average percentage of reordered products present in an order.

5. Then using the prior data for user, we need to featurize the data in such way that we can pose this problem as a binary predication problem - given a user , a product and his purchase history predict whether or not given product will be reordered in the next order.

6. From the research done on this problem we came to know that there is no such important features that strongly related to output so we need create many features and check which is working well.

7. To featurize the data, we need to find the features such that the features will preserve much information about the user, product and also need some features that will get the information about how a given user is related to a given product.

8. After going through the above mentioned links I gathered some features I will train a gbdt model with those features. The features are mentioned below.

   **a.** User related features such as :-

   i. no of orders

   ii. AVG of percentage of reorder product in an order

   iii. Users average days between orders

   iv. User average order size

   v. Total no of product ordered by the user

   vi. Total distinct items ordered by the user

   vii. Time of day the user visits

   viii. no of first purchased items by the user

   **b.** Product related:

   I. Mean, median, max or min of the product purchased per user

   II. Mean, median, max or min position of the product in cart

III. Mean, median, max or min of no of times the product reordered per user

IV. Average of consecutive order contain the product per user

V. Average no of probability of this product reordered in last 30 OR 50 percent of the orders per user

VI. No of times the product present in last 5 orders per user

VII. Median of the hour of day the product is ordered

VIII. Median of the day of week the product is ordered

IX. Median of the time of between the product reordered per user.

**c.** product and user related:-

I. percentage of the product purchased by the user

II. day since the item last purchased

III. streaks

IV. average position of the product in cart

V. difference between current and average position of the product in cart

VI. average hour of day this product ordered by the user

VII. difference between current and average hour of day this product ordered by the user

VIII. average day of week this product reordered

IX. difference between current and average day of week this product reordered

X. aisle , departments:-distance(bow) from most frequent aisle, departments for that user

---

**Notes when you build your final notebook**:

1. You should not train any model either it can be a ML model or DL model or Countvectorizer or even simple StandardScalar
2. You should not read train data files
3. The function1 takes only one argument “X” (a single data points i.e 1*d feature) and the inside the function you will preprocess data point similar to the process you did while you featurize your train data
    a. Ex: consider you are doing taxi demand prediction case study (problem definition: given a time and location predict the number of pickups that can happen)
    b. so in your final notebook, you need to pass only those two values

c. 
```
def final(X):
        preprocess data i.e data cleaning, filling missing values etc
        compute features based on this X
        use pre trained model
        return predicted outputs
final([time, location])
```

d. in the instructions, we have mentioned two functions one with original values and one without it

e. final([time, location])   # in this function you need to return the predictions, no need to compute the metric

f. final(set of [time, location] values, corresponding Y values)  # when you pass the Y values, we can compute the error metric(Y, y_predict)

4. After you have preprocessed the data point you will featurize it, with the help of trained vectorizers or methods you have followed for your train data
5. Assume this function is  like you are productionizing the best model you have built, you need to measure the time for predicting and report the time. Make sure you keep the time as low as possible
6. Check this live session: https://www.appliedaicourse.com/lecture/11/applied-machine-learning-online-course/4148/hands-on-live-session-deploy-an-ml-model-using-apis-on-aws/5/module-5-feature-engineering-productionization-and-deployment-of-ml-models