Machine Learning Engineer Nanodegree

Capstone Proposal

Pradeep Sai
July 10th, 2017

Proposal

Domain Background

If only there was a way to predict what the future orders would be then It would save lot of time and increase efficiency in the way It's delivered. Both retailer and customer would win in this way.

Instacart which delivers groceries through online orders has open sourced their 3 million orders of data https://tech.instacart.com/3-million-instacart-orders-open-sourced-d40d29ead6f2 . This makes the problem wide open and for everyone to provide interesting insights.

This project is derived from a Kaggle competition called
"Instacart Market Basket Analysis"
https://www.kaggle.com/c/instacart-market-basket-analysis

Problem Statement

The aim is to predict what the customers will be ordering next. Current data set contains 4-100 recent orders of users, our goal would be to derive some relation between what the user has ordered before and what they might order in future.

Datasets and Inputs

The data sets are obtained from https://www.kaggle.com/c/instacart-market-basket-analysis/data , Instacart has open sourced this data. We have 5 sets of files to train upon and should produce a submission file based on analysis made on these 5 files.

1. Aisles - Contains where in the storage products are located,
   Aisle id - Number value separate each Aisle
   Aisle - String value describing each Aisle (processed, soups, salads…)
2. Departments - A bit higher level categorization.
   Departmentid:- Number value to separate each department.
   Department(frozen, dry):- String value describing each department
3. order_products - Contains information of which product is bought in each ordered.
   order_id:- Numerical value indicating order number
   product_id:- Numerical value describing each product purchased separately
   add_to_cart_order:- Order in which products are added to cart from the app.
   reordered :- 1 indicates that the product was bought before.
4. orders: - Mostly self-explanatory
5. products: - Description of each product.

Solution Statement

Depending on time of the order and how frequently the products are reordered we can derive a relation to what products might be ordered again. We can use the order history details as a feature to our ML model and train it to predict what products user is more likely to order in the future.

Benchmark Model

Since this is a Kaggle hosted competition I can compare my F1 score calculated with the rest of the users on leaderboard and get a relative idea on how much better my model performs or how much it can be improved further.

Evaluation Metrics

Mean F score is used to provide the accuracy of the model. F score considers precision and recall and provide a good insight on how the model performs. Say if 60% of the people don't order anything and our model predicts None for all orders we would still be correct 60% of the time in-order to avoid this false interpretation we make use of precision and recall deriving a meaningful result
$F1 = 2 \cdot (\text{precision} \cdot \text{recall}) / (\text{precision} + \text{recall})$

Project Design

Our goal would be to predict what the user would be ordering next

1. We can join order_products__prior and order_product by user id and product id,

From this we can derive

- Each user ordering product

Product related details

- Number of orders for each product
- Minimum order number for a product
- Maximum orders placed for a product

User information details.

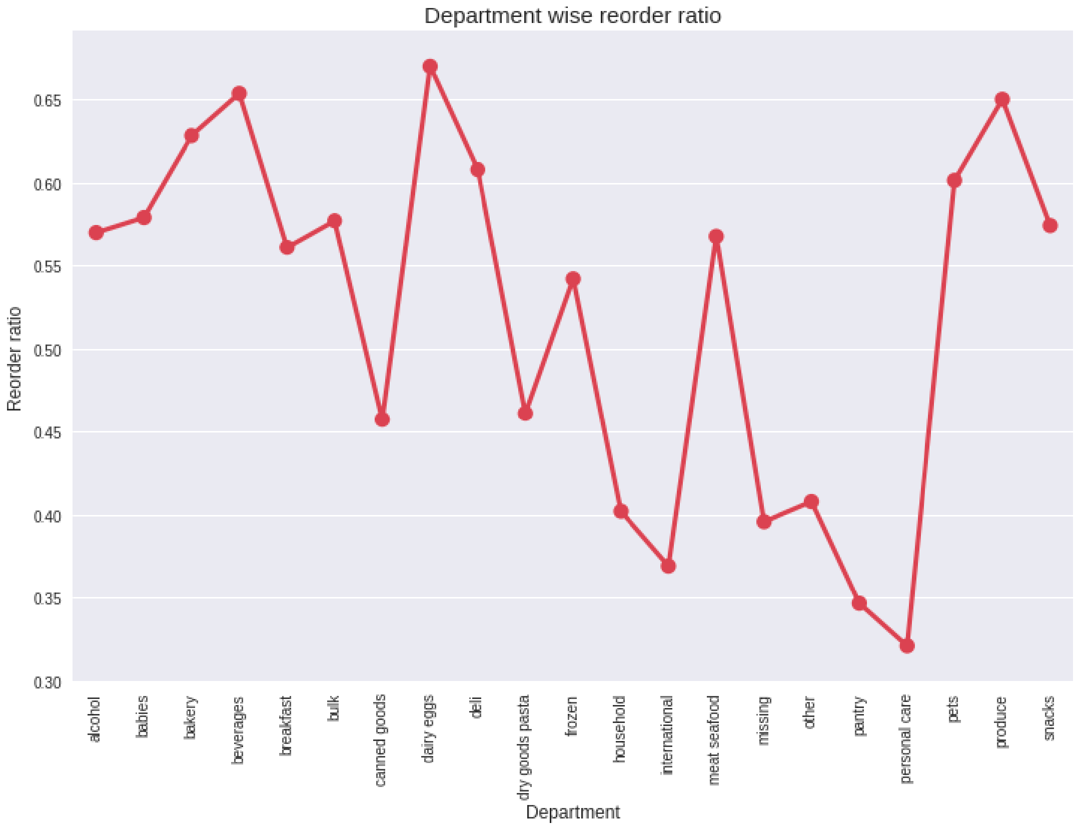
- Total number of second orders placed given first order
- Total days passed since first order
- Average time between orders

And when we use these features and from our train set we can see which orders are being re-ordered, if we find any new items for the first time we can drop it for now. A probability can be associated with each product being reordered again or not.

It would be a Multi class classification problem on the top, but internally we would be asking questions like whether product would be reordered or not based on features mentioned above.

I propose to use XGboost and some other techniques and compare performance of each model.

Below visualization shows re-order ratio based on each department, as we can observe dairy eggs are reordered the most. It makes sense since it would be most used item in everyday life. So there seems to be some sought of correlation between re-ordering and product category.


Department wise reorder ratio
Reorder ratio
alcohol
babies
bakery
beverages
breakfast
bulk
canned goods
dairy eggs
deli
dry goods pasta
frozen
household
international
meat seafood
missing
other
pantry
personal care
pets
produce
snacks
Department