Project 2: predicting products purchase in the next order

Introduction

The dataset for this project is a relational set of files describing customers' orders over time. The goal of the project is to predict which products will be in a user's next order. The dataset is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. For each user, we provide between 4 and 100 of their orders, with the sequence of products purchased in each order. We also provide the week and hour of day the order was placed, and a relative measure of time between orders.

Dataset Description

Each entity (customer, product, order, aisle, etc.) has an associated unique id. Most of the files and variable names should be self-explanatory.

Aisles.csv

Aisles id: contain id of the products

Aisles name: name of the products (soap, coffee, red wine etc.)

Department.csv

Department id: contains id of department

Department name: meat sea food, alcohol, pets etc

Order_product_prior.csv

These files specify which products were purchased in each order. order_products__prior.csv contains previous order contents for all customers. 'reordered' indicates that the customer has a previous order that contains the product. Note that some orders will have no reordered items. You may predict an explicit 'None' value for orders with no reordered items.

Order_id: order id

Product_id: product id

Add_to_cart_order: no. of time product add to cart

Reorder: product previously ordered

Order.csv

order_id: order id

user_id: user id

eval_set: tells the order is prior or test

order_number: total number of order the products by user

order_dow: no. of time order the product in the day of week

order_hour_of_day: order the product on the specific duration of the day

days_since_prior_order: order the product prior since day

product.csv

product_id: contain the product id

product_name: contain name of the product

aisle_id: contain the aisle id which is the same as the aisles.csv

department_id: contain the department id in which the product belongs to

sample_submission.csv

your output will be the following format

order_id:17

product_id:39276

product_name: Banana

Phase 1: pre-processing of the dataset

In phase 1 you need to do data analysis and data cleaning. For an example you can find out below facts from the data

- What do people buy?
- At what hour of day? (8-18)
- At which day of week? (0-1)
- When do customer order?
- When they order again? (1 week)
- How many prior orders? (3)
- How many items they generally order? (5)
- More often sold product? (top 10) i.e. Bananas
- How often do people order the same items again? 59%
- Which product put first? i.e. towel
- Try new things after 30 days

Data cleaning

It is possible that data issues may exist i.e. Null values N.A. There may exist some columns which you do not need at all i.e. outlet identifies etc. You also need to product five number summary.

Note: There are so much to do in data cleaning, I just give some examples and hints.

Phase 2:

Some techniques

You must implement at least three techniques you are free to apply any technique taught during course work.

- Linear regression
- Decision tree model
- Association rule mining
- Clustering
- Word2vec
- Neural Network etc.

Instructions: Submit one comprehensive report explaining the analysis done on the predicting next order of the user and the different techniques used along with the python code only.