Instacart: Market Basket Analysis

12/30/2020

—

Nusrat Islam
Springboard DS Career Track

Problem Statement

Develop a machine learning model to predict the future behaviour or buying pattern of customers using Instacart data of 3 million grocery orders from over 200,000 users.

Context

The Instacart data include orders of 200,000 Instacart users with each user having between 4 and 100 orders. Instacart indicates each order in the data as prior, train or test.

Prior orders describe the past behaviour of a user while train and test orders regard the future behaviour that we need to predict. We want to predict which previously purchased products i.e. prior orders will be in a user's next order (train and test orders). The setting of the Instacart problem is described in the figure below.

	order_id	user_id	eval_set	order_number	order_dow	order_hour_of_day	days_since_prior_order
0	2539329	1	prior	1	2	8	NaN
1	2398795	1	prior	2	3	7	15.0
2	473747	1	prior	3	3	12	21.0
3	2254736	1	prior	4	4	7	29.0
4	431534	1	prior	5	4	15	28.0

This is a classification problem because we need to predict whether each pair of user and product is a reorder or not. This is indicated by the value of the reordered variable, i.e. reordered=1 or reordered=0 (see figure below).

	order_id	product_id	add_to_cart_order	reordered
0	2	33120	1	1
1	2	28985	2	1
2	2	9327	3	0
3	2	45918	4	1
4	2	30035	5	0

Criteria for success

The success of the project will be based on the mean F1 score of 0.3 or more.

Constraints

- We do not have information on how often an user cancels or returns a product:
    - Once a order is placed what is the likelihood that it will be cancelled before fulfilling
    - After an order has been completed the chance of the product being returned
- New users sign up all the time and knowing how to define the retention of an user it might be easy to understand if they will reorder in the future

Stakeholders

This information could be valuable for Instacart for suggesting users the 'frequently bought together' items and help boost sales. They could also keep stock piled up if needed.

Data Sources

Anonymized transactional data of 3 Million Instacart Orders,

Data source: Dataset

Deliverables

A GitHub repo containing the work you complete for each step of the project,

including:

■ A slide deck

■ A project report