Instacart Market Basket Analysis Proposal

1. Introduction:

Instacard is an American techonology company that gets order through grocery shopping app, pick up ordered items from local grocery stores and deliver it to the customers. The purpose of this competition is, by getting the data of past shopping record of customer, they would like to develop model that predicts which of the product ,a customer , will buy again. By do so, Instacart would like to assist customer’s next order with their favorites and staples.

1. Objectives:

Instacart have provided the dataset of customer’s order through time. The dataset is divided into 3 entities, Aisles, departments, order products. Aisles and Department entities contain the list of id and the name of product matches with id. The order\_product file has customer id and their past ordered product id and how many they have purchased and indication of reordered. The products file has all the item listed with their product id, name, aisle id, and department id. Orders file is important file that contains the order id, user id, order number, order dow, order hour of dat and days since prior order which is all the record of transaction for each customers

1. Outline

The purpose of this project is to predict what products a customer will like to purchase again. First,I will combine order products and order products to add aisle id and department id to order products file on order id. Then I will match the order products file and orders file with order\_id in order to find out which of the product for each customer bought and reodered and see their pattern or shopping. Each of the column will be graphed in order to find out preference. Also the relationship of data and indication of reordered.

1. Delivery :

The order file has 3 sets, prior, train and test. These file will be devided into 2 set train and test for prediction. For training data, the train set will be devided into 70/30% for another train and test set. Since this data seems to have labeled data and predict outcome, I will be using supervised Learning such as logistic regression, SVM or XGBoost. The right model will be further investigated.