SESUG Paper 252-2019

Market Basket Analysis on Instacart

Aravind Dhanabal, Oklahoma State University

ABSTRACT

Market Basket Analysis is a technique used by retailers to understand customer behavior while purchasing from their stores. In the process of online shopping, you have probably seen a section called “suggestions for you” or “customers who bought this item also bought” in which Market Basket Analysis plays an important role. The implementation of this analysis was aided by the initiation of electronic point of sales systems. Store owners used handwritten records and digital records of the customer transactions which were generated by point of sales system. This was effectively used to analyze ample amount of data to know about customer purchasing behavior and pattern.

In this paper we are going to understand and help Instacart to make use of their customer transaction data and focus on descriptive analysis on the customer purchase patterns, items which are bought together and units that are highly purchased from the store to facilitate reordering and maintaining adequate product stock. Also, to identify the clusters and subgroups of customers possessing similar purchase behavior and to visualize the data to provide productive recommendations which focus on improving the revenue and customer experience through segmentation and prediction models.

Our dataset has variables which focused on the orders as well as the time of the orders. So, order related and time related features were created in order to predict whether a product will be reordered or not. This was fully used in the modelling stages and will be used further for the future research of this project.

This paper will enable Instacart to enhance the user experience by suggesting the next likely product to purchase to the customer during the order process. Further, this paper will outline a marketing strategy for Instacart and similar retailers including sending personalized communications to customers reminding them to order again, by highlighting the predicted products in those communications.

Introduction

In modern world, with the advancements coming in the Internet of Things, the necessity for innovation is on high demand. According to Stastica, the Statistic portal, 1.8 billion people worldwide purchased goods online in 2018, which will continue to increase in the following years. More research works and innovations are in progress in order to meet people's demand and their satisfaction. One of the key techniques used to understand the customer purchasing behavior is the analysis on their transaction details.

Understanding the transaction is a must to any form of business and its effect will lead to increase in sales. Especially in a retail store, it can be achieved by understanding the purchasing pattern of customer and related products which were sold together. This enables impulse buying from customers and also to understand their usual purchasing pattern and their effects towards the retail market.

To understand well enough about the market basket, I have decided to analyze the Instacart transactional data. Instacart is an e-commerce website that allows users to shop for groceries from a local grocery store online, and then sends an Instacart personal shopper to pick up and deliver the orders made by users the same day. These processes allow retailers to conduct analysis on purchase iterations by users but understanding the customer purchasing patterns and behaviors can become tedious and challenging.

With the provided dataset, I could come up with the best possible models for the below mentioned business objectives.

Model 1: To predict the next likely product, the customer would purchase during the ordering process

Model 2: To develop a model which can predict whether a product will be reordered or not

Some of the key data points required for the above-mentioned models were:

Model 1: Transactional details of the products, order number, date and time of the transactions, aisles and departments where the product belongs

Model 2: Reordered details, time and day details of the products ordered, transactional details, departments and aisles details

Data Description

In this section, we are going to discuss in detail about the dataset provided by Instacart. A total of 6 datasets were provided which gave information on customer transactional and purchasing order details.

Section 1: Aisles

This dataset provides information on the aisles such as aisle ID and aisle names, through which the products were organized.

Variables	Description
Aisle ID	Labels the ID of the aisles
Aisle name	Mentions the aisle name in the retail stores

Table 3.1 – Details of Aisles data set

Section 2: Department

This dataset provides information on the departments such as department names and department Id.

Variables	Description
Department ID	Labels the ID of the departments
Department name	Mentions the department name in the retail stores

Table 3.2 – Details of departments data set

Section 3: Order_Products_prior

This dataset gives information on the orders, products, and reordered products

Section 4: Order_Products_train

This dataset is same as order_products_prior and it is a trained dataset.

Variables	Description
Order ID	Labels the ID of the order made by customer
Product ID	Labels the ID of the products purchased by customers
Add to cart order	Sequence of the order placed in the cart
Reordered	Denotes whether the products are reordered or not

Table 3.3 – Details of order_prior_train data set

Section 5: Orders

This dataset has information about the customer orders like order ID, order number, week day of the order, hour of the order, user ID and days since prior order.

Variables	Description
Order ID	Labels the ID of the order made by customers
User ID	Labels the ID of the users who made the purchase
Order number	Denotes the order number made by the customer
Order_dow	Denotes the day of the week, the order made by the customer
Order hour of day	Denotes the hour of the day, the order made by the customer
Days since prior order	Denotes the number of days since last order

Table 3.4 – Details of orders data set

Section 6: Products

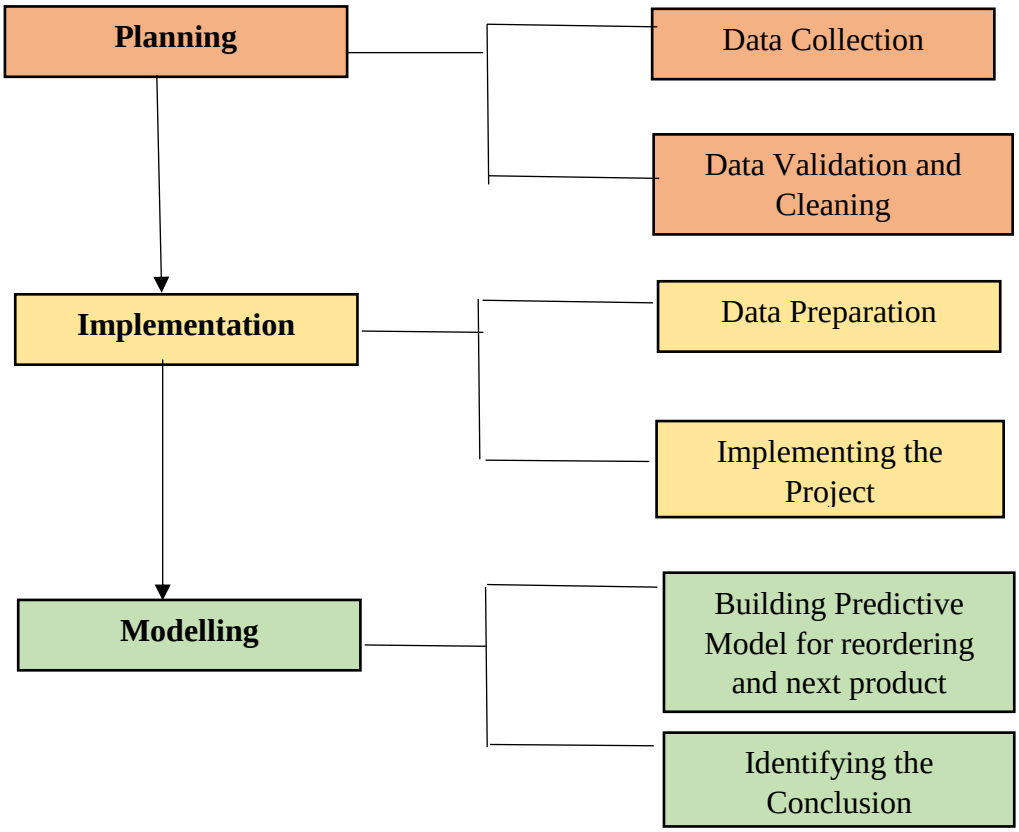
This dataset gives information on the products such as product name, product ID, aisle and departments, which were sold to the customer.

Variables	Description
Product ID	Labels the ID of the products purchased by customers
Product Name	Denotes the product name purchased by the customer
Aisle ID	Labels the ID of the aisles
Departments ID	Labels the ID of the departments

Table 3.5 – Details of products data set

Methods

This project focus on predicting the next products the customer tends to purchase and also whether a product is reordered or not by the customers in their next purchase.



Figures 4.1 – Methods adopted for the purpose of analysis

Data Collection

The datasets were provided by Instacart Technology Company and was taken from Kaggle to perform the analysis. The datasets provided by Instacart had complete information of over 3 million grocery orders from more than 200,000 Instacart users. Both product data and customer data from Instacart includes 50,000 unique products, week and the time of purchase, different product aisle and departments. Understanding the data, dairy products, fruits and vegetables were purchased the most across all the departments and people tends to purchase and reorders 60% of their previous orders mostly on Sunday and Monday

Data Preparation

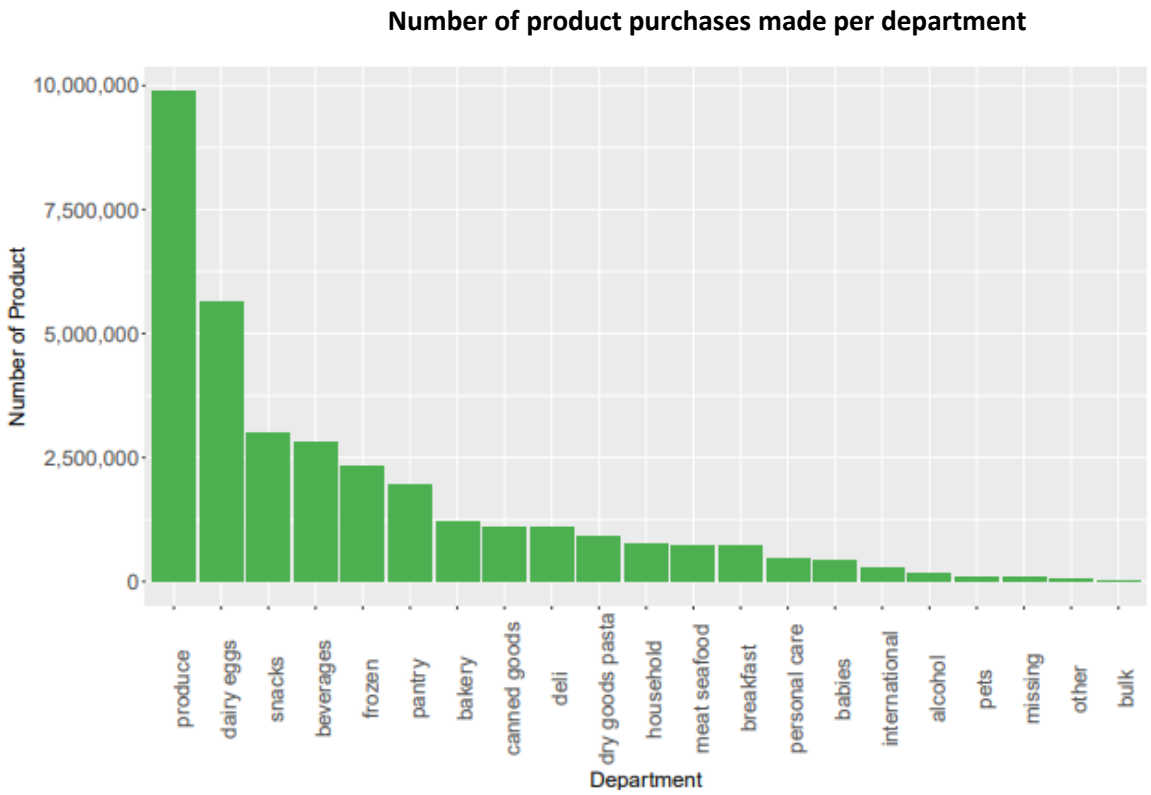
Information about customer purchases and transaction details are delivered to us through six different datasets. Order and Product dataset form the base of the complete transactions and were merged to a single dataset through the common Product and Order ID variables accordingly. Later, aisles and departments datasets were merged with the order and product combined dataset through aisle ID and department ID to form a master dataset to commence the analysis. SAS Studio was used for preparing the data and other manipulation operations to proceed further for the analysis.

Data Cleaning and Manipulation

There were no null or empty values for the variables like aisle, departments, Order_product_prior, order_product_train and products datasets. Orders dataset has some null values in days since prior order variable and only 5% of the values were found to be missing and this has been rejected since the count is very low to be a significant issue. All the datasets were merged using SAS Studio and SAS Enterprise Guide.

Summary statistics

Below attached are the charts which will give a brief description about the summary statistics of the final dataset.

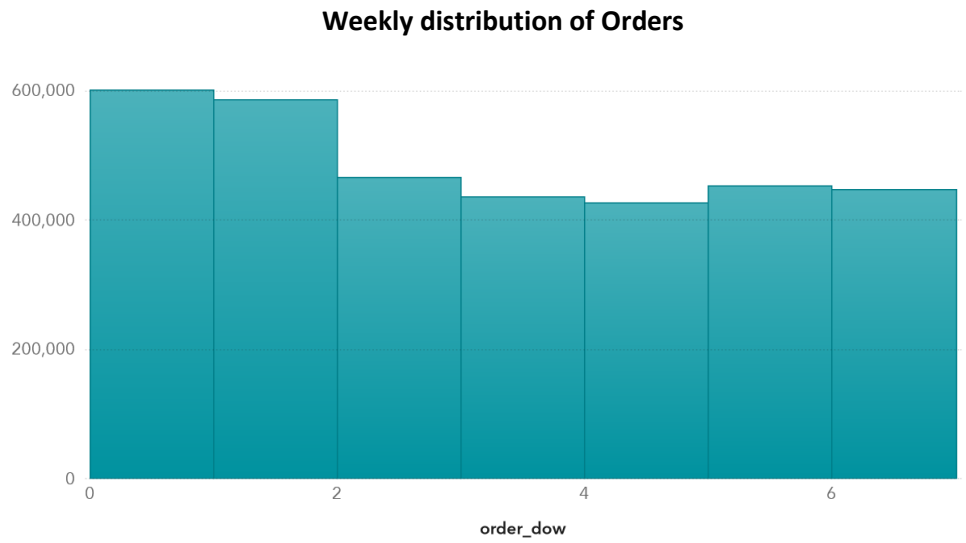


Figures 4.2 – Number of product purchases made per department

Basic product based results:

- Fruits and vegetables (Produce) and dairy products were sold the most across departments
- Shakes, beverages and frozen food were also popular across departments

Figures 4.3 – Hourly distribution of Orders



Figures 4.4 – Weekly distribution of Orders

Basic Customer based results:

- Customers usually prefer to purchase between 10am and 4pm across all days
- Sunday and Monday were the most preferred days by the customer to order the groceries
- Mostly, customers typically reorders 50% of the previous ordered products

Sampling Strategy and Data Partitioning

Master dataset merging all the required datasets were compiled separately for analysis. It is found that the master data was balanced with 41% No and 59% Yes values. So it is not required to perform any of the sampling techniques and hence can move forward to data partitioning. For further analysis, master dataset were partitioned with 70% - 30% split for training and validation data.

Modeling Techniques

Model 1: To predict the next likely product, the customer would purchase during the ordering process

Using associative rule methodology in SAS Enterprise Miner, the next product sequence the customer purchases can be interpreted using Lift, support and confidence.

Support – It is the percentage of transactions that comprise all of the items in a dataset. The more the support value the more frequently the product occurs. High support values are preferred for ample amount of future transactions.

Confidence - It is the probability that a transaction that contains the items on the left hand side of the also contains the item on the right hand side. The more the confidence value, the greater the likelihood that the product on the right hand side will be purchased.

Lift - Lift is nothing but the ratio of Confidence to Expected Confidence. It is the probability of all of the products in a rule occurring together by the product of the probabilities of the items on the left and right hand side occurring as if there was no association between them.

Some of the association rules which were interpreted from our methodology are mentioned below:

	Rules
Rule 1	Organic Strawberries & Organic Hass Avocado ==> Organic Baby Spinach & Bag of Organic Bananas
Rule 2	Organic Hass Avocado & Organic Baby Spinach ==> Organic Strawberries & Bag of Organic Bananas
Rule 3	Organic Yellow Onion & Organic Baby Spinach ==> Organic Garlic
Rule 4	Organic Strawberries & Organic Garlic ==> Organic Yellow Onion
Rule 5	Organic Garlic & Bag of Organic Bananas ==> Organic Yellow Onion
Rule 6	Organic Yellow Onion & Organic Strawberries ==> Organic Garlic
Rule 7	Organic Yellow Onion & Bag of Organic Bananas ==> Organic Garlic
Rule 8	Organic Garlic & Organic Baby Spinach ==> Organic Yellow Onion
Rule 9	Organic Hass Avocado & Bag of Organic Bananas ==> Organic Lemon
Rule 10	Organic Lemon & Bag of Organic Bananas ==> Organic Hass Avocado
Rule 11	Organic Strawberries & Organic Lemon ==> Organic Hass Avocado
Rule 12	Organic Strawberries & Organic Hass Avocado ==> Organic Lemon
Rule 13	Organic Strawberries & Organic Baby Spinach & Bag of Organic Bananas ==> Organic Hass Avocado
Rule 14	Organic Strawberries & Organic Hass Avocado ==> Organic Raspberries
Rule 15	Organic Yellow Onion & Bag of Organic Bananas ==> Organic Hass Avocado
Rule 16	Organic Hass Avocado & Organic Baby Spinach ==> Organic Yellow Onion
Rule 17	Organic Garlic & Bag of Organic Bananas ==> Organic Hass Avocado
Rule 18	Organic Ginger Root ==> Organic Garlic
Rule 19	Organic Italian Parsley Bunch ==> Organic Garlic

Table 4.5 – Association rules

Model 2: To predict whether a product will be reordered or not

Various modeling techniques like logistic regression, Decision tree, Gradient boost classifier were adopted to predict whether a product will be reordered or not. The best model is accessed and validated with the help of model comparison node comparing all the results of the different models used.

Product Models

Model	Accuracy	Misclassification
Logistic Regression	63%	37%
Decision tree	68%	32%
Gradient Boosting	59%	41%

Table 4.6 – Accuracy, misclassification for models

As can be seen from the table above, accuracy was the highest for Decision tree model. However, Logistic regression model has the high sensitivity value which means 83% of the reordered products have been accurately captured by the model.

Recommendations

Based on the associative rule methodology and models to predict the reorder of products, some of the recommendations have been made:

- It will be productive to run promotional and marketing campaigns with the help of the associative rules. Based on the prediction of the next product, customers can be given additional offers by bundling the products together for a lesser price and customize the products based on the association rules
- Based on the reordering model, personalized communications can be very lucrative by reminding the customers to reorder the products or can be added to the cart automatically based on the customer preferences
- We would recommend Instacart to add the products directly to the customer's cart or to provide a suggestion list when they make their purchase in order to enhance the customer experience
- By knowing the rate of products reordered, Instacart can make use of the reordered data to analyze the inventory stocks by ensuring the replenishments and proper scheduling of the products to increase internal productivity

References

- *Journal article* - A.A. Raorane, R.V. Kulkarni, B.D. Jitkar, **Association Rule – Extracting Knowledge Using Market Basket Analysis**, Research Journal of Recent Sciences, 1 (2) (2012), pp. 19-27
- *Journal article* - A. Herman, L.E. Forcum, Joo Harry. **Using Market Basket Analysis in Management Research,** Journal of Management, 39 (7) (2013), pp. 1799-1824
- *Website* - Megaputer blog, An introduction to market basket analysis. Retrieved from https://www.megaputer.com/introduction-to-market-basket-analysis/
- *Website* - Margaret Rouse, Basic understanding of Market basket analysis. Retrieved from https://searchcustomerexperience.techtarget.com/definition/market-basket-analysis
- The Instacart Online Grocery Shopping Dataset 2017, Accessed from https://www.instacart.com/datasets/grocery-shopping-2017

Acknowledgement

I would like to thank Dr. Miriam Mcgaugh, Clinical Assistant Professor in Marketing, for her guidance and support in the due course of this project. We would also like to acknowledge the guidance and insight provided by **Dr. Goutam Chakraborty**, **SAS Professor of Marketing Analytics and Director of MS in Business Analytics and Data Science.**

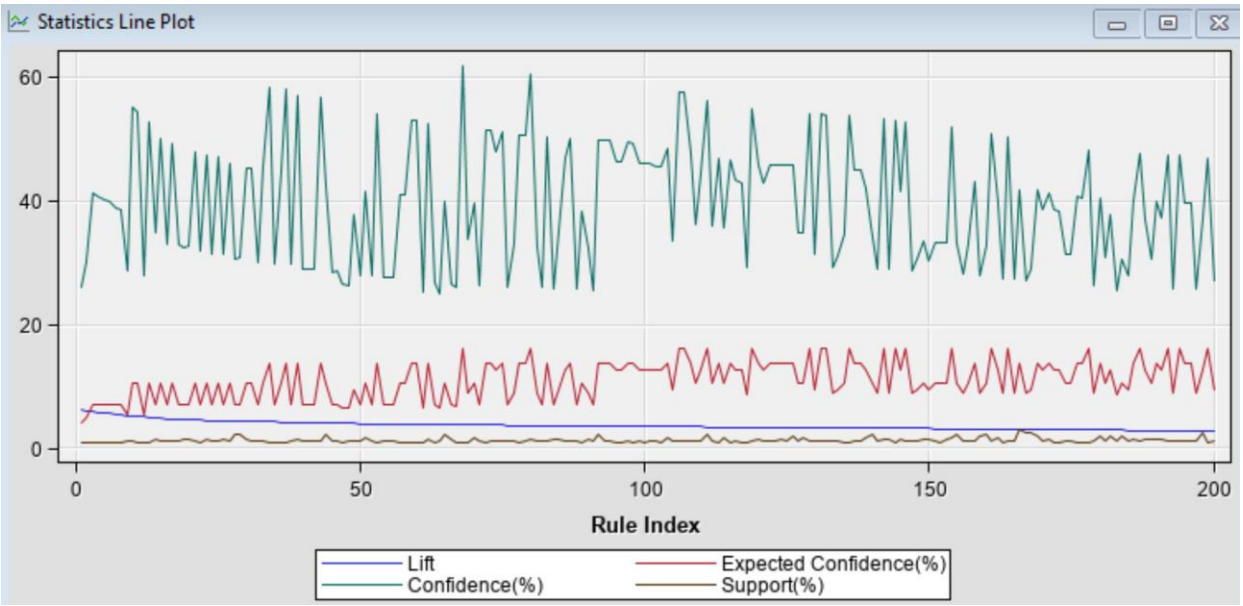
Appendix

SAS Code for merging datasets

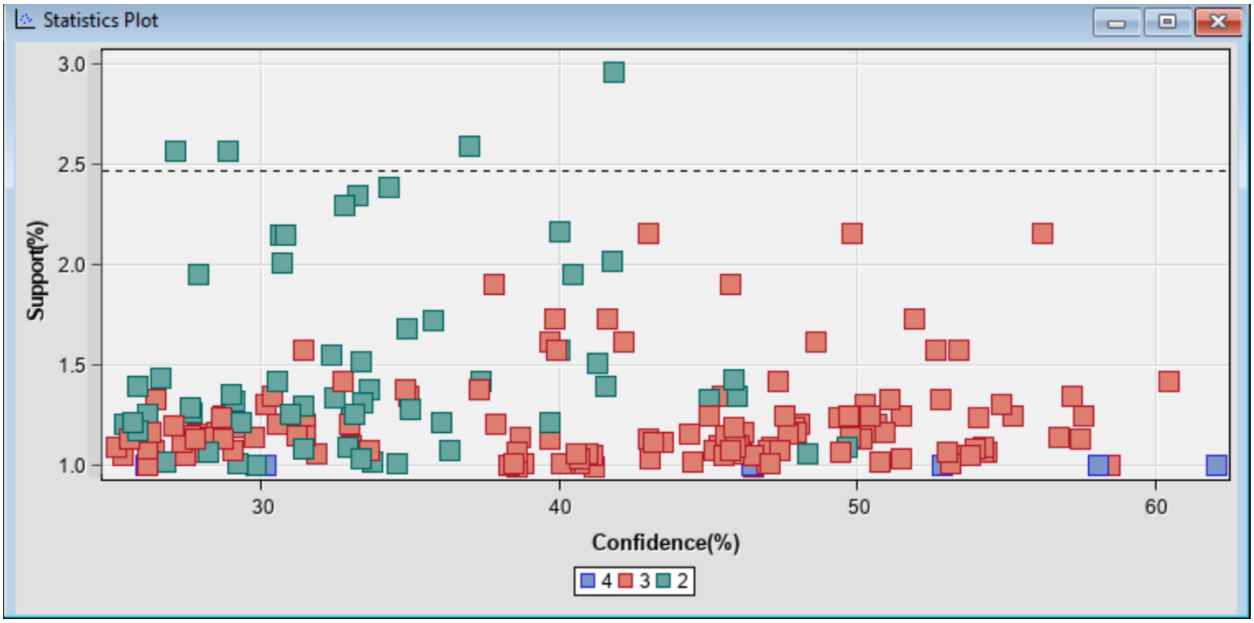
```
/*To merge products, aisles and department datasets*/

proc sql;

create table Products_aisles_departments as

(select a.product_ID, a.product_name, a.aisle_id, a.department_id, b.aisle, c.department

from products a, aisles b, departments c

where a.aisle_id=b.aisle_id and a.department_id=c.department_id);

quit;


/*To merge orders and products*/

proc sql;

create table orders_final as

(select a.order_ID, a.user_id, a.order_num, a.order_dow, a.order_hour_of_day,
a.days_since_prior_order, b.product_id, b.reordered, b.add_to_cart

from orders a, orders_products_prior b

where a.order_id=b.order_id);

quit;


/*To merge orders_final and products_aisles_departments*/

proc sql;

create table masterdataset as

(select a.order_ID, a.user_id, a.order_num, a.order_dow, a.order_hour_of_day,
a.days_since_prior_order, a.product_id, a.reordered, a.add_to_cart, b.product_name, b.aisle_id,
b.department_id, b.aisle, b.department

from orders_final a, products_aisles_departments b

where a.product_id=b.product_id);

quit;
```

Association rules:

Statistics Line Plot



Statistics Plot

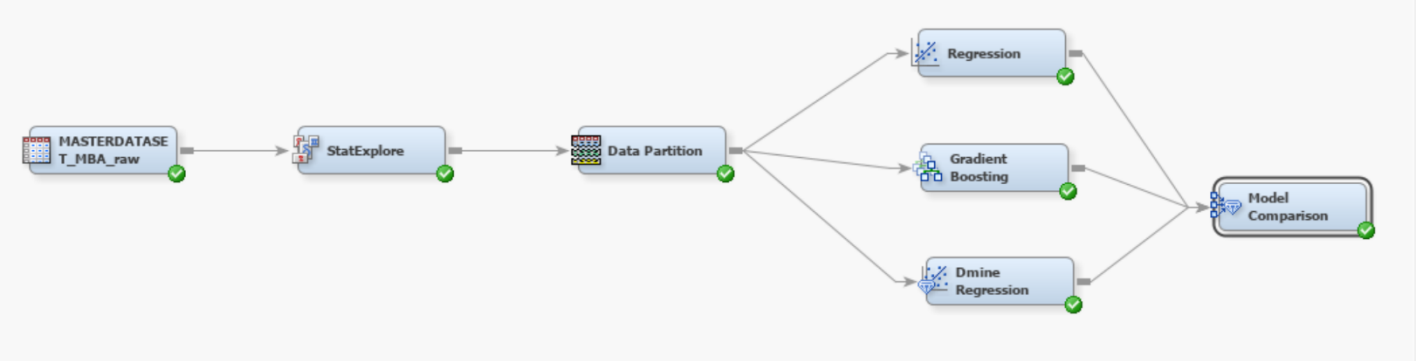


Rule Description

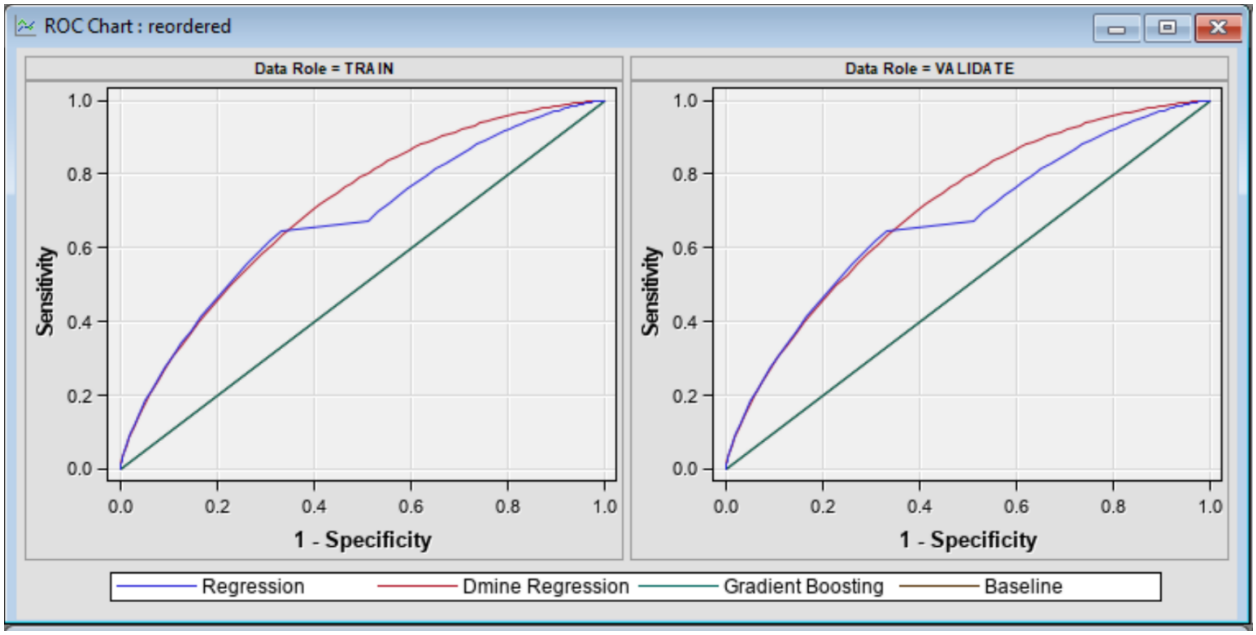
Rule Description

Map	Rule
RULE1	Organic Strawberries & Organic Hass Avocado ==> Organic Baby Spi...
RULE2	Organic Hass Avocado & Organic Baby Spinach ==> Organic Strawb...
RULE3	Organic Yellow Onion & Organic Baby Spinach ==> Organic Garlic
RULE4	Organic Strawberries & Organic Garlic ==> Organic Yellow Onion
RULE5	Organic Garlic & Bag of Organic Bananas ==> Organic Yellow Onion
RULE6	Organic Yellow Onion & Organic Strawberries ==> Organic Garlic
RULE7	Organic Yellow Onion & Bag of Organic Bananas ==> Organic Garlic
RULE8	Organic Garlic & Organic Baby Spinach ==> Organic Yellow Onion
RULE9	Organic Hass Avocado & Bag of Organic Bananas ==> Organic Lemon
RULE10	Organic Lemon & Bag of Organic Bananas ==> Organic Hass Avocado
RULE11	Organic Strawberries & Organic Lemon ==> Organic Hass Avocado
RULE12	Organic Strawberries & Organic Hass Avocado ==> Organic Lemon
RULE13	Organic Strawberries & Organic Baby Spinach & Bag of Organic Bana...
RULE14	Organic Strawberries & Organic Hass Avocado ==> Organic Raspberr...
RULE15	Organic Yellow Onion & Bag of Organic Bananas ==> Organic Hass A...
RULE16	Organic Hass Avocado & Organic Baby Spinach ==> Organic Yellow ...
RULE17	Organic Garlic & Bag of Organic Bananas ==> Organic Hass Avocado
RULE18	Organic Ginger Root ==> Organic Garlic
RULE19	Organic Italian Parsley Bunch ==> Organic Garlic
RULE20	Organic Hass Avocado & Bag of Organic Bananas ==> Organic Rasp...
RULE21	Organic Yellow Onion & Organic Strawberries ==> Organic Hass Avo...
RULE22	Organic Hass Avocado & Organic Baby Spinach ==> Organic Garlic
RULE23	Organic Raspberries & Bag of Organic Bananas ==> Organic Hass Av...
RULE24	Organic Strawberries & Organic Hass Avocado ==> Organic Yellow O...
RULE25	Organic Zucchini & Bag of Organic Bananas ==> Organic Hass Avoca...
RULE26	Organic Strawberries & Bag of Organic Bananas ==> Organic Raspbe...
RULE27	Organic Strawberries & Organic Garlic ==> Organic Hass Avocado
RULE28	Organic Garlic ==> Organic Yellow Onion
RULE29	Organic Yellow Onion ==> Organic Garlic
RULE30	Organic Strawberries & Organic Raspberries ==> Organic Hass Avoc...
RULE31	Organic Yellow Onion & Organic Baby Spinach ==> Organic Hass Av...
RULE32	Organic Hass Avocado & Bag of Organic Bananas ==> Organic Yello...
RULE33	Limes & Bag of Organic Bananas ==> Organic Hass Avocado

Model Comparison



ROC Chart


ROC Chart : reordered
Data Role = TRAIN
Data Role = VALIDATE
Sensitivity
1 - Specificity
0.0
0.2
0.4
0.6
0.8
1.0
Regression
Dmine Regression
Gradient Boosting
Baseline