84.51°

TARTAN DATA SCIENCE CUP

Data Dictionarv

Data and Problem Overview

Understanding if a customer is likely to buy a particular product is an important factor that drives decisions at Kroger. Students are asked to use customers purchase histories (spanning 205 days) to predict which customers will purchase eggs in the following week. What factors drive an increase or decrease in purchase likelihood? How can 84.51° use that information to make better decisions? What actions can be taken to increase the probability of households buying eggs in their next visit?

This dataset contains household level transactions over 205 days from a group of 967 households who are frequent shoppers at Kroger. It contains all of each household's grocery purchases. For certain households, demographic information is provided.

Time Period

Students will be provided training data for days 500 – 704 and will be asked to forecast if each of the 967 households will purchase at least 1 unit of eggs in days 705 – 711. Since this is an out-of-sample prediction, so no further information will be provided about days 705-711.

Response Variable

Products are classified as eggs by `commodity_desc="EGGS"`. The response variable will be a binary classification variable – does a household have quantity purchased > 0 for eggs during days 705 – 711?

Please submit a two-column .csv file that contains the `household_key` and `probability` of purchasing eggs for each of the 967 households in the training dataset. A template for your submission can be found at:
www.stat.cmu.edu/tartandatasciencecup/episodeII/team_name.csv

IMPORTANT: When submitting your response, your predictions must match the exact format of the team_name.csv file:

- All predictions in a single .csv file
- Same number of rows and same column names
- Your team name should be the file name

The Brier score will be used to evaluate your predicted probabilities.

TARTAN DATA SCIENCE CUP: DATASET DETAILS

This dataset contains household level transactions over 205 days from a group of 967 households who are frequent shoppers at Kroger. Each row of the dataset is essentially the same as what would be found on a grocery store receipt.

The columns correspond to information about each product sold and demographic information for a portion of households. (Due to nature of the data, the demographic information is not available for all households.) The columns / variables in the dataset are described below:

`transaction_data_8451`

Variable	Description
household_key	Uniquely identifies each household
BASKET_ID	Uniquely identifies a purchase occasion
DAY	Day when transaction occurred
PRODUCT_ID	Uniquely identifies each product
QUANTITY	Number of products purchased during the trip
BASE_SPEND_AMT	Dollar amount of products(s) without any discounts
LOY_CARD_DISC	Discount applied due to retailer's loyalty card program
COUPON_DISC	Discount applied due to coupon
NET_SPEND_AMT	Final dollar amount household pays
DEPARTMENT	Groups similar products together
COMMODITY_DESC	Groups similar products together at a lower level
SUB_COMMODITY_DESC	Groups similar products together at the lowest level
AGE_DESC	Estimated age range
MARITAL_STATUS_CODE	Marital Status (A - Married, B - Single, U - Unknown)
INCOME_DESC	Household Income
HOMEOWNER_DESC	Homeowner, renter, etc.
HH_COMP_DESC	Household composition
HOUSEHOLD_SIZE_DESC	Size of household up to 5+
KID_CATEGORY_DESC	Number of children present up to 3+

It may be important to calculate the price per product. To do so, use this formula:

- `PRICE_PER_PRODUCT = NET_SPEND_AMT / QUANTITY`