NYU | STERN

TERM PROBLEM STATEMENT AND DATA DESCRIPTIONS

To help you gain experience working with real-world data sets, this project will require that you mine actual transaction data to solve a problem of interest to your client, an online retailer, that is interested in understanding how to better structure the advertising and cross-selling opportunities it directs at different customers. *In particular, the firm wishes to determine which types of promotional campaigns (discounts based on a specific promo code) are most effective for which types of customers.*[1]

Context overview

Imagine that you have been approached by Gganeph, Inc., an established online retailer. Gganeph frequently conducts marketing campaigns for its diverse product range. The retailer has asked you to assist the firm in optimizing its discount marketing process using the power of machine learning.

Discount marketing and promo code usage are very widely used promotional techniques to attract new customers and to retain & reinforce loyalty of existing customers. You have almost certainly experienced these directly. For example, a clothing retailer that you have used in the past might conduct a President's Day sale, and send you an email with the code: "PREZ2022." If you elect to purchase an item on sale from the retailer, you will be asked at checkout to enter any promo codes. In order to get your discount you would need to enter "PREZ2022" at which point the discount will be applied to your order. The measurement of a consumer's propensity towards coupon usage and the prediction of the redemption behavior are crucial parameters in assessing the effectiveness of a marketing campaign.

Gganeph contacts its customers about promotions via a number of channels including: email, pop-up notifications, etc. Some campaigns include promo code discounts that are offered for a specific product, while others might apply to a range of products or even be site-wide.

The marketing and finance departments at Gganeph have determined that the campaigns can have widely varying uptake by customers. Finance is concerned that Gganeph is missing opportunities to attract clients for long periods of time during the less-effective campaigns. In order to address this, the COO would like to be able to predict how effective Gganeph campaigns are likely to be, for different customer segments. One way to do this is to examine which customers are most likely to respond to which types of promotions, when delivered over which channels.

Specific Business need: Predicting Promo Code Uptake

Gganeph would like to be able to better predict the probability that a customer will redeem the promotional code (and buy the promoted product) for each coupon and customer combination in newer prospective campaigns.

The retailer has asked you to design the data mining task, mine the data, and describe your results. (See the *Term Project Instructions* documents for the details of the deliverable and documentation.)

[1] This data and problem are from real-world promotional campaigns. The data is real-world data. However, the specific domain of the problem, and some of the details of the data have been modified for confidentiality purposes.

Data Overview

The IT department at Gganeph has agreed to provide you with information on previous campaigns, customers and transactions. The firm has operations in both the US and EU, and as a result, some of the data that Gganeph collects may not be shared with your team. However, you will have access to a variety of data sets related to Gganeph request.

The data available in this problem is delivered in six data files:

- `code-redeemed-train.csv`
- `promotional_campaigns-train.csv`
- `all-transactions-train.csv`
- `products-train.csv`
- `demog_train.csv`
- `promo-cd-2-prod-mapping-train.csv`

These are available in an archive file named `train.zip`.

The tables in the Data Dictionary section below provide the definitions of the variables in the different data sets, and the schema details how the different files relate to each other. If you are unfamiliar with joining data files, please read the article: *How-to-Join-tables-in Python .pdf*, which is included in the Brightspace Term Project folder.)

The basics of a transaction flow is as follows:

1. Customers receive promo codes under various campaigns
2. A customer with a promo code may redeem it for any valid product for that promo code provided the purchase date is within the campaign's valid window (`promo-cd-2-prod-mapping-train`, `promotional_campaigns-train`)
3. If a customer redeems the code for a valid item under the campaign, this will be recorded (`code-redeemed`), along with the discount that the promo code offered (`dscnt_pcode` in `all-transactions`).
4. If a customer purchases an item but does not use the promo code, that will also be recorded.

Data Dictionary

code-redeemed-train.csv (redemption behavior)

Variable name (column name, feature, etc.)	Description
redemption_id	Unique id for promo code-customer combination
c_id	Unique id for promo campaign
pcode_id	Unique id for promo code
cust_id	Unique id for customer account
Redeemed (TARGET!)	Indicator of whether the promo code was redeemed on a purchase

promotional_campaigns-train.csv (historical promotional campaign chracteristics)

Variable name (column name, feature, etc.)	Description
c_id	Unique id for promotional campaign
c_type	Anonymized campaign Type (X/Y) used internally by Gganeph
start_date	First day of the campaign
end_date	Last day of the campaign
start_year	Year of **start_date**
end_year	Year of **end_date**

all-transactions-train.csv (transaction data for duration of campaigns)

Variable name (column name, feature, etc.)	Description
date	Date of transaction
cust_id	Unique id for customer account
prod_id	Unique id for the product purchased
qty	Number of units of item purchased
price	Price at which item was listed
dscnt_store	Discount that was given on the website without promo code
dscnt_pcode	Discount that was available on the website if promo code was used
year	Year of **date**

products-train.csv (product data for all products sold by Gganeph during campaigns)

Variable name (column name, feature, etc.)	Description
prod_id	Unique id for product
manuf	Unique id for product manufacturer / supplier
brand_type	Indicator of Gganeph's internal brand classification
category	Broad product category

demog_train.csv (selected customer demographics)

Variable name (column name, feature, etc.)	Description
cust_id	Unique id for customer account
age_cohort	Customer's broad age group
marital_status	Marital status of customer (if known)
home_is_rental	Indicator of whether customer owns or rent's home
n_family	Number of family members in customer's household (if known)
no_of_children	Number of children in customer's household (if known)
income_range	Customer's broad income range (if known)

promo-cd-2-prod-mapping-train.csv (junction table mapping promos to products)

Variable name (column name, feature, etc.)	Description
c_id	Unique id for promotional campaign
prod_id	Unique id for product

The database schema is given on the next page.

Gganeph Working Table Schema