# Acquire Valued Shoppers Challenge

## Predict which shoppers will become repeat buyers

Consumer brands often offer discounts to attract new shoppers to buy their products. The most valuable customers are those who return after this initial incented purchase.  With enough purchase history, it is possible to predict which shoppers, when presented an offer, will buy a new item. However, identifying the shopper who will become a loyal buyer -- prior to the initial purchase -- is a more challenging task.

The Acquire Valued Shoppers Challenge asks participants to predict which shoppers are most likely to repeat purchase. To aid with algorithmic development, we have provided complete, basket-level, pre-offer shopping history for a large set of shoppers who were targeted for an acquisition campaign. The incentive offered to that shopper and their post-incentive behavior is also provided.

This challenge provides almost 350 million rows of completely anonymised transactional data from over 300,000 shoppers. It is one of the largest problems run on Kaggle to date.

## Submission File

Submissions are evaluated on [area under the ROC curve](http://en.wikipedia.org/wiki/Receiver_operating_characteristic) between the predicted probability that a customer repeat-purchased and the observed purchase outcomes.

For each customer (id) in testHistory.csv, predict a probability that the customer repeat-purchased the product from the promotion they received. Your submission file must have a header and should look like the following:

id,repeatProbability

12262064,0

12277270,0

12332190,0

...

**Data source:** [**https://www.kaggle.com/c/acquire-valued-shoppers-challenge/data**](https://www.kaggle.com/c/acquire-valued-shoppers-challenge/data)

## Data Description

**Warning: this data set require about 22GB of space.**

This data captures the process of offering incentives (a.k.a. coupons) to a large number of customers and forecasting those who will become loyal to the product. Let's say 100 customers are offered a discount to purchase two bottles of water. Of the 100 customers, 60 choose to redeem the offer. These 60 customers are the focus of this competition. You are asked to predict which of the 60 will return (during or after the promotional period) to purchase the same item again.

To create this prediction, you are given a minimum of a year of shopping history prior to each customer's incentive, as well as the purchase histories of many other shoppers (some of whom will have received the same offer). The transaction history contains all items purchased, not just items related to the offer. Only one offer per customer is included in the data. The training set is comprised of offers issued before 2013-05-01. The test set is offers issued on or after 2013-05-01.

## Files

You are provided four relational files:

|  |  |  |  |
| --- | --- | --- | --- |
| **file** | **about** | **index** | **columns** |
| **transactions.csv** | contains transaction history for all customers for a period of at least 1 year prior to their offered incentive. | * 'id' * 'chain' * 'dept' * 'category' * 'company' * 'brand' * 'date' * 'productsize’ * 'productmeasure' * 'purchasequantity' * 'purchaseamount' | 349655790 |
| **trainHistory.csv** | contains the incentive offered to each customer and information about the behavioral response to the offer. | * 'id', * 'chain', * 'offer', * 'market', * 'repeattrips', * 'repeater', * 'offerdate' | 160057 |
| **testHistory.csv** | contains the incentive offered to each customer but does not include their response. (you are predicting the repeater column for each id in this file) | * 'id' * 'chain', * 'offer' * ‘market' * ‘offerdate' | 151484 |
| **offers.csv** | contains information about the offers. | * 'offer' * 'category' * 'quantity' * 'company' * 'offervalue' * 'brand' | 37 |

* **transactions.csv**
* Only ‘productmeasure’ has null number,Total: 10296541(2.94%)
* A negative value in productquantity and purchaseamount indicates a return.
  + Purchasequantity: 21518
  + Purchaseamount: 808321

|  |  |  |  |
| --- | --- | --- | --- |
| **trainHistory.csv** | | | |
| id | Unique | 160057 |  |
| Null | 0 |  |
| chain | Unique | 130 |  |
| Null | 0 |  |
| offer | Unique | 23 |  |
| Null | 0 |  |
| market | Unique | 34 |  |
| Null | 0 |  |
| repeattrips | Unique | 56 |  |
| Null | 0 |  |
| repeattrips | Unique | 2 |  |
| Null | 0 |  |
| offerdate | Unique | 56 |  |
| Null | 0 |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **testHistory.csv** | | | |
| id | Unique | 151484 |  |
| Null | 0 |  |
| chain | Unique | 131 |  |
| Null | 0 |  |
| offer | Unique | 29 |  |
| Null | 0 |  |
| market | Unique | 34 |  |
| Null | 0 |  |
| offerdate | Unique | 89 |  |
| Null | 0 |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **offers.csv** | | | |
| Offer | Unique | 37 |  |
| Null | 0 |  |
| category | Unique | 20 |  |
| Null | 0 |  |
| quantity | Unique | 2 |  |
| Null | 0 |  |
| company | Unique | 18 |  |
| Null | 0 |  |
| offervalue | Unique | 7 |  |
| Null | 0 |  |
| brand | Unique | 19 |  |
| Null | 0 |  |

## Fields

All of the fields are anonymized and categorized to protect customer and sales information. The specific meanings of the fields will not be provided. Part of the challenge of this competition is learning the taxonomy of items in a data-driven way.

|  |  |  |
| --- | --- | --- |
| ***history*** | | |
| **id** | A unique id representing a customer |  |
| **chain** | An integer representing a store chain |  |
| **offer** | An id representing a certain offer |  |
| **market** | An id representing a geographical region |  |
| **repeattrips** | The number of times the customer made a repeat purchase |  |
| **repeater** | A boolean, equal to repeattrips > 0 |  |
| **offerdate** | The date a customer received the offer |  |

|  |  |  |
| --- | --- | --- |
| ***Transactions*** | | |
| **id** | see above |  |
| **chain** | see above |  |
| **dept** | An aggregate grouping of the Category (e.g. water) |  |
| **category** | The product category (e.g. sparkling water) |  |
| **company** | An id of the company that sells the item |  |
| **brand** | An id of the brand to which the item belongs |  |
| **date** | The date of purchase |  |
| **productsize** | The amount of the product purchase (e.g. 16 oz of water) |  |
| **productmeasure** | The units of the product purchase (e.g. ounces) |  |
| **purchasequantity** | The number of units purchased |  |
| **purchaseamount** | The dollar amount of the purchase |  |

|  |  |  |
| --- | --- | --- |
| ***offers*** | | |
| **offer** | see above |  |
| **category** | see above |  |
| **quantity** | - The number of units one must purchase to get the discount |  |
| **company** | see above |  |
| **offervalue** | The dollar value of the offer |  |
| **brand** | see above |  |

* The transactions file can be joined to the history file by (id,chain).
* The history file can be joined to the offers file by (offer).
* The transactions file can be joined to the offers file by (category, brand, company).
* A negative value in productquantity and purchaseamount indicates a return.

## Workflow

## File Structure

* Bdse08\_t1
  + Data
  + ETL
  + Statistics
  + Web
  + DB

## Idea

## Data Cleaning / Integration / Transformation

## Describe Statistic

## Analysis Method

## Analysis Result

## Other

## Member

## Reference