Table of Contents

Who We Are

Valassis is the leader in marketing technology and consumer engagement. We work with over 60,000 companies and brands in a wide array of industries, partnering to anticipate consumer intent, inspire action, and create demand, while saving consumers money. We own and use predictive intelligence to anticipate purchase intent, deliver value, optimize campaigns, and drive results by inspiring consumers to act. Every day, Valassis converts billions of signals into meaningful engagement across all channels to drive exponential growth.

How Do We Do This?

We examine content across 80 billion web pages each day and derive meaning from words and sentences to determine likely purchase intent. Our page-level analysis enables a deep understanding of the shopper's browsing behavior across 1400 interest topics. For example, we know the difference between someone interested in running shoes, running for office or running a small business. We can differentiate between recent in-market shopping patterns and long-term interests over time. The in-depth understanding of the shoppers' online activity is only part of their story. To get a holistic view, marketers must connect the shoppers' online activity to their real-world actions. The Valassis Consumer Graph does this by synthesizing accurate, real-world data across 540K stores, 900K points of interest, and 110MM households together with browsing and buying data.

The Challenge: Predicting Rare Events

Background

In digital advertising, a "conversion" refers to the event when the shopper clicks on the ad and performs a valuable action such as signup, registration, or make a purchase. Since "conversion" is a measurable event, it represents a reasonable proxy for the number of customers acquired during the ad campaign. Increasingly, brands and agencies looking to put a value on the Return on Advertising Spend (ROAS), require marketers such as us to optimize the ad spend such that customer acquisition is maximized.

In order to wisely spend the limited marketing dollars, we need to identify the shoppers who are more likely to respond to our ad and convert. While the number of devices to target is nearly one billion, the number of conversion events range from just a few hundreds to few thousands during the period of the ad campaign. In other words, these conversion events are extremely rare.

Example Challenge

For example, one objective might be to predict the shoppers who are likely to convert with very little false alarm. An added impediment is that the information we know about the shopper is incomplete which means the data is sparse.

The Ask

The dataset is already divided into two groups (refer to Table 1) for your convenience, training and validation. Use the training data to build a model to predict shoppers who are likely to convert. Then, use the validation data to evaluate the performance of your model. Feel free to use any appropriate metric to evaluate the models on the training and validation sets.

For various reasons, marketers will have to understand the profile of the shoppers who converted. What is your take on the profile of the converters? What other insights can you gain from the data and the model you have built.

Data Description

There are two zip files that hold the same data, albeit, in a different format.

- valassis_dataset.zip: Preferable for Python programmers.
- valassis_tallskinny_dataset.zip: Preferable for R programmers.

Python Programmers

Download the valassis_dataset.zip file and unzip it, you should see three files.

Table 1: Description of attached files

File Name	Description
training.csv	Training Data: Contains the samples needed to train the model Record count: 96,406 Positive samples: 1,465 Negative samples: 94,941
validation.csv	Validation Data: Contains the samples for validating the model. Record count: 80,008 Positive samples: 620 Negative samples: 793,88

interest_topics.csv	Contains the topic label and topic description.

Interest Topics Dataset

Each row in this dataset represents one of the interest topics.
# of record in the file: 1,411

Table 2: Schema for interest_topics.csv

Column Name	Data Type	Description
topic_id	Integer	Numerical identifier of the topic
topic_name	String	Interest topic

Training and Validation Dataset

Each row in this dataset represents a shopper's long-term and short-term interests in one of the 1,411 categories that we have.

Table 3: Schema for files training.csv and validation.csv

Column Name	Data Type	Description
userID	Integer	User identifier. Unique for each row in this csv file.
inAudience	Boolean	This column represents whether or not the shopper has converted in the past. TRUE: Shopper has converted FALSE: Shopper has not converted
ltiFeatures	Dictionary Key: String Value: Double	Long Term Interests. The dictionary's keys are one of interest topics. The value represents the proportional interest the user has in that topic. For example, if the *ltiFeatures* of one of the users is {'34': 0.7, '41':0.3}, it can be inferred as follows. The keys represent the *topic_id* in the interest_topics.csv file.

		Topic 34 is */Arts & Entertainment/Movies* Topic 41 is */Games/Computer & Video Games* This user has 70% interest in */Arts & Entertainment/Movies* and 30% interest in */Games/Computer & Video Games*.
stiFeatures	Dictionary Key: String Value: Double	Short Term Interests Description is same as above, however, for shopper's short term interest.

R Programmers

Download the valassis_tallskinny_dataset.zip file and unzip it, you should see three files.

Table 1: Description of attached files

File Name	Description
training_tallskinny.csv	Training Data: Contains the samples needed to train the model Record count: 96,406 Positive samples: 1,465 Negative samples: 94,941
Validation_tallskinny.csv	Validation Data: Contains the samples for validating the model. Record count: 80,008 Positive samples: 620 Negative samples: 793,88
interest_topics.csv	Contains the topic label and topic description.

Interest Topics Dataset

Each row in this dataset represents one of the interest topics.
# of record in the file: 1,411

Table 2: Schema for interest_topics.csv

Column Name	Data Type	Description
topic_id	Integer	Numerical identifier of the topic
topic_name	String	Interest topic

Training and Validation Dataset

Each row in this dataset represents a shopper's long-term and short-term interests in one of the 1,411 categories that we have.

Table 3: Schema for files training_tallskinny.csv and validation_tallskinny.csv

Column Name	Data Type	Description
userID	Integer	User identifier.
inAudience	Boolean	This column represents whether or not the shopper has converted in the past. TRUE: Shopper has converted FALSE: Shopper has not converted
topic_id	Integer	Integer representing the topic id
ltiFeatures	Double	Level of long-term interest in the topic represented by the topic_id column.
stiFeatures	Double	Level of short-term interest in the topic represented by the topic_id column.

Python Code Snippet to Inspect Data

```
def load(file):

    import pandas as pd
    import ast
    df = pd.read_csv(file)

    # convert the column values from literal string to dictionary
    df['ltiFeatures'] = df['ltiFeatures'].apply(ast.literal_eval)
    df['stiFeatures'] = df['stiFeatures'].apply(ast.literal_eval)

    return df

# load all the data
training = load("training.csv")
validation = load("validation.csv")
interest_topics = pd.read_csv("interest_topics.csv")

# inspect the data
interest_topics.head()

training.head()

validation.head()
```