

Analytics Project 2026

Dynamic pricing strategies have become an essential part of online business. Large online shops with a wide range of products rely on the automatic adjustment of product prices. The focus here is not on price personalization but rather price optimization at the product level. The goal is thus not to provide customer-specific and personalized prices. On the contrary, regular price adjustments lead to a price per product in line with the market that will maximize parameters such as revenue and gross margin.

In the project we will consider that it is not only the price but also the interaction between different product attributes that influences revenue.

Scenario

A mail-order pharmacy uses a dynamic pricing strategy in the form of daily automatic price adjustments for its online shop. To evaluate its success, prices, revenue figures and different product attributes are recorded. User behavior expressed in actions such as clicks on products, assigning shopping baskets and purchases is also recorded.

The objective is to create a model to predict whether a product is bought and in which quantity.

The attributes of all the products occurring in the learning period that do not change with time are listed in the "*items.csv*" file.

The information that changes with time for the learning period is in the "*train.csv*" file.

The key to linking information that changes with time and information that does not is the product number under the "pid" attribute.

A single data set from the "*train.csv*" file contains information about the action of a user regarding a particular product and other product information that changes with time (e.g. the competitor's price). The "*click*," "*basket*," and "*order*" columns provide information about the type of action. Only one value per line in these columns can be "1", the others are "0". If a product has only been clicked, the value of the "*click*" column is "1". If the product was added to the shopping basket, the value in the "*basket*" column is "1". If the product was purchased, the value in the "*order*" column is "1". This does not provide information as to the number of units of the product purchased or added to the shopping basket.

Variables

Items.csv

Variable name	Description	Value range
pid	Product number	Natural number
manufacturer	Manufacturer (anonymized)	Positive whole number
group	Product group	String consisting of capital letters and numbers
content	Package content	Positive floating-point number or string in the format numberXnumber, e.g. 5X10
unit	Unit	String of capital letters
pharmForm	Dosage form	Three-digit string of capital letters
genericProduct	Generic flag	{0, 1}
salesIndex	Dispensing regulation code	Natural number
category	Main shop category	Natural number
campainIndex	Action label	{A, B, C}
rrp	Reference price	Positive decimal number

Train.csv

Variable name	Description	Value range
lineID	Key for unique identification of user action	Natural number
day	Day in the observed period	Natural number
pid	Product number	Natural number
adFlag	Advertising flag indicating if the product is part of a marketing campaign	{1, 0}
availability	Availability status	{1, 2, 3, 4}
competitorPrice	Lowest competitors price	Positive decimal number
click	Click flag	{0, 1}
basket	Basket flag	{0, 1}
order	Order flag	{0, 1}
price	Product price	Positive decimal number
revenue	Revenue	Positive decimal number

There are missing values which must be treated adequately. Perhaps not all attributes contribute to the classification. You will have to calculate new attributes.

Try out at least three different classification algorithms and compare them.

What is the business aspect of the problem?

Present your findings with the help of data story telling in a paper and an on-site presentation.

You can find the datasets on moodle. Train consists of over 2.5 million records.

Source: Data Mining Cup 2017 (adapted)