ESOMAR | Events

Demystifying Machine Learning

Accelerating innovation by exploiting what is already known

Sjoerd Koornstra • Wim Hamaekers

ESOMAR
Office address:
Atlas Arena, Azië Gebouw
Hoogoorddreef 5
1101 BA Amsterdam
Phone: +31 20 664 21 41
Fax: +31 20 664 29 22

Email: customerservice@esomar.org
Website: www.esomar.org

Publication Date: September 2018
ESOMAR Publication Series Congress 2018
ISBN 92-831-0301-7

Copyright

Published by ESOMAR, Amsterdam,
The Netherlands
Edited by: Deborah S. Fellows

About ESOMAR

ESOMAR is the global voice of the data research and insights community, representing a network of 35,000 data professionals.

With more than 4,900 members from over 130 countries, ESOMAR's aim is to promote the value of market and opinion research in illuminating real issues and bringing about effective decision-making.

To facilitate this ongoing dialogue, ESOMAR creates and manages a comprehensive programme of industry specific and thematic events, publications and communications, as well as actively advocating self-regulation and the worldwide code of practice.

ESOMAR was founded in 1948.

About ESOMAR Membership

ESOMAR is open to everyone, all over the world, who believes that high quality research improves the way businesses make decisions. Our members are active in a wide range of industries and come from a variety of professional backgrounds, including research, marketing, advertising and media.

Membership benefits include the right to be listed in the ESOMAR Directories of Research Organisations and to use the ESOMAR Membership mark, plus access to a range of publications (either free of charge or with discount) and registration to all standard events, including the Annual Congress, at preferential Members' rates.

Members have the opportunity to attend and speak at conferences or take part in workshops. At all events the emphasis is on exchanging ideas, learning about latest developments and best practice and networking with other professionals in marketing, advertising and research. CONGRESS is our flagship event, attracting over 1,000 people, with a full programme of original papers and keynote speakers, plus a highly successful trade exhibition. Full details on latest membership are available online at www.esomar.org.

Contact us

ESOMAR

ESOMAR Office:
Atlas Arena, Azië Gebouw
Hoogoorddreef 5
1101 BA Amsterdam
The Netherlands
Tel.: +31 20 589 7800

Email: customerservice@esomar.org
Website: www.esomar.org

Demystifying Machine Learning

Accelerating innovation by exploiting what is already known

Sjoerd Koornstra • Wim Hamaekers

Introduction

Research departments are under pressure. They are expected to deliver faster, cheaper and more impactful insights than ever before. Instead of doing more and faster research, insight departments are also able to revisit existing data sources. Often, companies have plenty of valuable data sources at their disposable, without being aware of their full potential. Moreover, relevant databases are often publicly available via APIs or sold via data brokers. Yet, the biggest hurdle lies in making sense of this abundance of data. Companies are struggling to connect different sources due to different structures, missing values and other complexities. In the last years enormous advancements have been made in machine learning and data science. Although demystifying is needed in order to better understand this discipline in context. With a creative and pragmatic mind-set, the problem can be solved by borrowing techniques from this field of data science. We show a case – in the beverage industry – where we exploited existing data sources to uncover a hidden layer of insights.

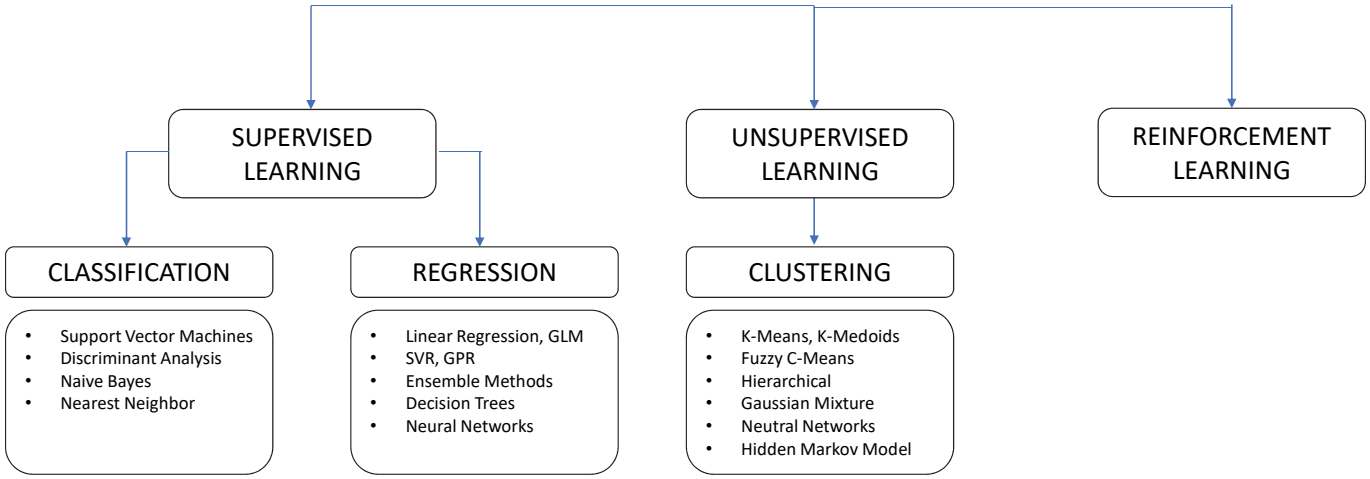
Machine learning: Explanation

Machine learning has become a hype and is used for a various number of things not necessarily correctly. When you ask people to define machine learning you usually receive the following answer: Advanced analytical tools which have the ability to learn by so-called self- learning algorithms. But what does self-learning mean? We can distinguish within machine learning three different types: Supervised, Unsupervised and reinforcement learning. Within Supervised and Unsupervised learning no self learning is taking place. In both cases, the learning comes from the amount of data available: the more data available, the more 'learning' the system is. e. The only type where self-learning takes place is Reinforcement, also applied in robots. It reflects roughly only 1% of all machine learning applications. Figure 1 shows an overview of the main machine learning tools where half consist of statistical techniques. Also for these statistical techniques counts; the more data, the higher the predictive value and better estimated results. We think that machine learning needs a proper definition. In our view machine learning is an algorithm which:

- Is not assuming certain distributions of the variables /features (no statistical testing)
- Is applied to generate a relationship between input and output but you are not interested in the coefficients
- Is usually using three sub sets to validate the model:
  - Training: to generate the model
  - Validation: to optimize the model
  - Test: to verify the model

This distinguishes real machine learning tools from statistical applications. The application of machine learning within market research can be seen in automation of research (bot applications, drop out predictions), Text Analytics and Mining, Social Media Analytics, Image Mining and Predictive Modelling .

Figure 1. Machine learning algorithms



Machine learning: Current barriers

The knowledge about machine learning within Insights Department from companies is limited. Insights manager do not know the possibilities of machine learning and have difficulties in understanding the results. The majority of these analyses are performed with machine learning tools by data scientists. Data scientists are needed to build the analysis models. Data scientists can analyze but are usually less strong in the interpretation and the link with the business context. Insights managers are not able to steer and challenge these data scientists. McKinsey has also recognized this and according to them there is a huge need for analytics translators, the link between data science and the business. This is described by McKinsey in the article "Analytics translator: The new must-have role" for the *Harvard Business Review*.

According to McKinsey translators work closely with the business leaders to understand the business issues and prioritization. Translators then use their knowledge of AI and analytics to translate the business issue into a briefing for the data professionals. These data professionals produce models and solutions. The translators then ensure that insights are distilled from these models and solutions that the business can understand and the business can act on.

Domain knowledge is the most important skill that a translator must have. They must be experts in their sector / sector, discipline and company to place the value of AI and analytics in the business context. They do not necessarily have to be able to build quantitative models themselves. They have to know which model variants are available (e.g. deep learning versus logistic regression) and for which business problem they can be applied. Translators must be able to interpret the results, discover potential model errors, such as overfitting, and challenge the data scientists.

How does the above relate to Insights Departments? Insights managers play a role that is somewhat similar to translators. Together with mostly marketing the business issues are discussed as well as which (consumer) information is needed to make a decision. Insights managers then often brief agencies and ensure that the correct method is applied. They then ensure that research results are translated into information that is necessary for taking business decisions. Insights managers are the connectors between the business issues and the understanding of consumers. This domain knowledge is most important for a translator role. It seems logical, therefore, that Insights managers are trained in data science knowledge in order to fulfil this translator's role

Case study

An important research question for many beverage companies is which flavours to launch in which markets. In practice, brands usually launch one basic variant, together with two or three flavours, in a given market. Market research is then executed to identify the flavours with the most potential. But this comprehensive research costs time and money. Moreover, some markets are too small/underdeveloped to allocate research budget to, and are therefore neglected research-wise. Another problem is that consumer research often identifies liquids with enormous consumer rejection – which is often considered a waste of money and resources. Thus, the goal of this project was to develop a tool which would enable to screen certain flavours upfront and check their potential in a given market.

We started to look for data sources already available or available at data brokers with whom we have a long-lasting subscription agreement. We discovered the following data sources

1. *Other beverage data sources*
   We looked at categories which are the front runners concerning new flavours. Leading categories are in our view chewing gum, tea and carbonated soft drinks (CSDs). One of our data brokers had a carbonated soft drinks database. This database contains 517 flavour combinations on category level (not on brand level), from 92 countries and two consecutive years. The flavour could be one flavour, e.g. apple or lime, but also a flavour combination e.g. apple / lime. We put the following assumptions:
   - Higher CSD flavour volumes suggest higher potential for beverage category flavours.
   - Steeper growth (vs decline) over the two consecutive years suggest trending flavours.

2. *Beverage flavour launches*
   We had a subscription to new product launch service. This supplier compiled a database for us consisting of the amount of new flavour introductions in a given country. This concerns 50 new flavour introductions in 73 countries. The assumption is that with more flavour introductions of that certain flavour in a given country, the respective flavour will be more receptive and have a higher volume potential.

3. *Company Sales Volume*
   This data source contains the sales volume from the different flavours in nearly 100 markets in two consecutive years. This database indicates the success and relative importance of in market flavours.

4. *Motivational Positioning Platforms*
   The company has developed so called motivational positioning platforms. This was based upon comprehensive desk research from qualitative motivational studies, which were conducted in several markets globally. 10 Motivational positioning platforms are detected. Also an assessment was made the of the role different flavours are playing on these motivational positioning platforms.

5. *Syndicated consumer flavour appeal / preference studies*
   One of our data brokers had executed two consumer flavour appeal / preference studies – one study for our category, and one study for an adjacent category. The studies were executed in 25 countries and around 50 flavours were evaluated. The evaluation was reported as the preference share of each flavour per market. This indicates the potential of a flavour as declared by consumers. This measurement was even fine tuned within different category types. For our category this was distinguished in sweet, tart/sour, unique and spicy category types.

So we had different datasources and we want to assess the potential of each from these 517 flavours per market.
How did we continue? We first structured the data in five consecutive steps:

Step 1. Categorizing flavours
We had two categorizing approaches. One from Haystack which distinguishes 20 main flavour categories, and another was used within the company and consists of 12 flavour categories. As a first step, each of the 517 flavour combinations from the CSD data set were classified in 20 main categories. This was done by sensory experts. A flavour combination could be assigned to more than one category (e.g. “Apple/Lime” would fall into both the Apple category and the Citrus category). The same approach was used for the 12 company flavour categories. Each of the 517 flavour combinations from the CSD data set were classified by the same sensory experts in 12 main categories. Why did we use two categorization approaches? The one from Haystack is more granular and the people from Haystack are used to work with this. The Company approach is known by the company and created commitment.

Step 2. Allocating flavours to motivational positioning platforms
Each of the 517 flavour combinations were assigned to the motivational platforms A flavour combination could be assigned to more than one positioning platform. This allocation procedure used as input the categorization of flavours as described in step 1.

Step 3. Calculating new flavour introductions
For each of the 517 flavours, we looked at how many new introductions of that flavour happened in each country. For flavour combinations (e.g. Apple/Lime), we calculated the mean of new introductions of each flavour separately.

Step 4. Adding flavour appeal / preference to the database
Each of the 517 flavours was given a score on eight different parameters, based on results from the two appeal / preference studies. The eight different parameters came from two studies and four sub categories per study. Each parameter ranged from 0% to 100%. According to the preference shares of flavour type (per country), the parameters were then weighted and added to come to one category / appeal measure and one adjacent category appeal / preference measure per country.

Step 5. Adding CSD volume data + Company Category volume data
The CSD volumes and their absolute growth were added to the dataset. Also the Company volume data was added to the dataset. Volumes of specific flavours were summed up per country.

In total, the complete database consists of 49 features.

- Country
- Flavour
- 20 flavour segments (a)
- 12 flavour segments (b)
- 10 motivational positioning platforms
- CSD volume + growth
- Category volume + growth
- New flavour introductions
- Category Appeal / preference + Adjacent Category Appeal

Every country has 517 lines (rows) in the database given the 517 flavours (one row per flavour).

Eventually, we like to predict the potential for each flavour, in each country. However, the dataset consists of missing variables, which would make it impossible to give a potential score for each flavour in each country. That is why a method was developed to estimate these missing variables. Not all CSD flavours are on the market in all countries. However, we want to predict the potential CSD volume of a given flavour if it were on the market in a certain country. Additionally, we want to make an estimation of the evolution this flavour would make, as if it would be on the market in that country. These estimations can then be included for the final flavour potential score.

Not all countries were present in the ‘new introductions dataset’. However, we like to calculate the potential in all countries. This is why we need to make estimations in the countries which are not present in this dataset.

To impute the mentioned missing values, we then employed a predictive algorithm (machine learning method) to extrapolate our data to unknown markets. This predictive tool is a methodology that provides a solution to complex incomplete data problems. This machine learning method searches for patterns in the complete dataset (over all the included parameters: country, positioning platforms, flavour categories, flavour appeal, etc.) and predicts the missing value, based on similarities with other observations.

Table 1. Data file description

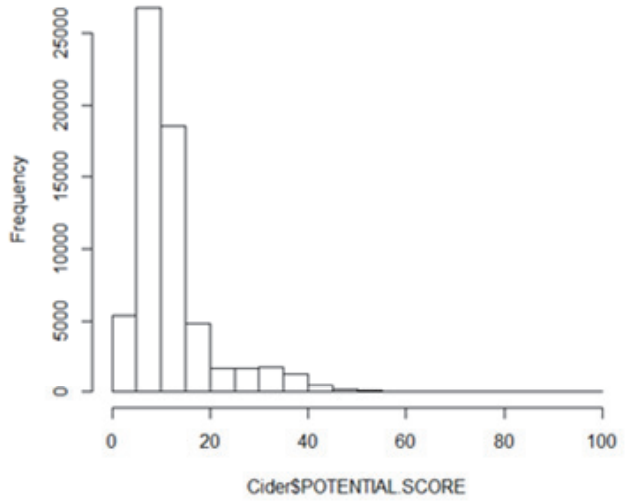
		20 Flavour Segments (a)			12 Flavour Segments (b)			10 Positioning Platforms			CSD	CSD	Category	Category	Flavour	Category	Adj. Category
Country	Flavour	Segment 1a	Segment 2a	Etc.	Segment 1b	Segment 2b	Etc.	Platform 1	Platform 2	Etc.	Category	Growth	Volume	Growth	Introductions	Appeal	Appeal
Afghanistan	Acai	0	1		1	1		0	0		0.44	0.04	1.1	-0.07	7	0.4356	0.078
Afghanistan	Aloe	0	1		1	1		0	0		0.55	0.29	4.58	0	16	0.638	0.2646
Afghanistan	Aloe Vera/ Blackberry	0	1		0	1		1	0		72	0	0.26	0.03	21	0.3168	0
Afghanistan	Aloe Vera/ Lemon	0	1		0	1		0	0		16.04	-3.68	24.64	0.02	21	0.4356	0.0778

PATTERN RECOGNITION **COMPLETED MISSING VALUES**

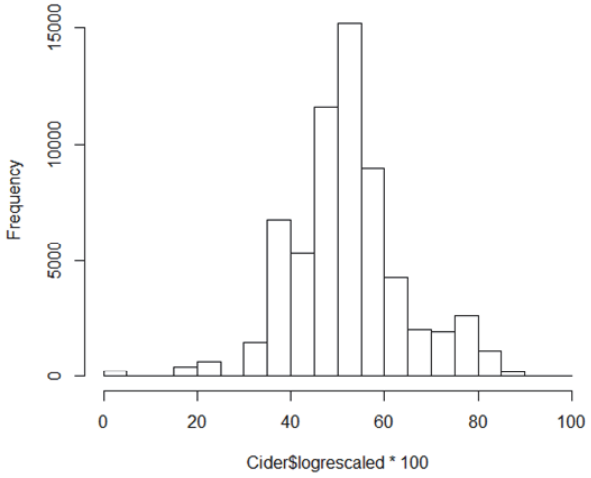
The final flavour potential is composed of five scores. All scores are calibrated from 0 to 100. The CSD score (average of volume and of growth), the category score, the launch introductions sore, the category appeal / preference score and the adjacent category appeal / preference score. Notice that the two flavour segmentations and the motivational positioning platforms are only exploited to complete the missing values. The 42 variables have been indispensable for pattern recognition by applying the machine learning tool. The final score is a weighted average of the five described scores. Finally, our potential flavour score was transformed by means of a logarithmic transformation to come to a more natural distribution. Why? Because our original distribution is skewed. Our mean and median do not coincide, which can cause problems in the interpretation of our potential score. And more importantly our stakeholders want easy to understand figures. Logarithmic transformation shortens distances between large numbers and enlarges distances between small numbers. The average and median are both 50%. More natural distribution around the mean, which makes our categories more natural and more interpretable, does not change ranks between flavours. Best flavours remain best flavours. Worst flavours remain worst flavours.

Figure 2. Logarithmic data transformation

Skewed distribution before transformation:



Normal distribution after transformation:



We moulded our database into a user-friendly tool. With this tool, stakeholders could easily estimate the probability of success for certain flavours in any given market and could decide to prioritize the potential flavour portfolio. We give them also the probability to simulate with the weights of the different factors.

Final conclusion

The application yielded an abundance of insights. In total, the application was able to give indicative insights on almost 50,000 products. To collect similar results with traditional research methods would cost easily over the 100 millions of euros and years of fieldwork. The tool proved an enormous asset in prioritizing R&D and was heavily used by business managers. This case shows that with a creative mind-set, research departments could greatly benefit from connecting existing data sources. By exploiting what already is known, companies can become more agile entities that are better equipped to meet today's business standards.

Final thoughts

We have been talking about the role of the analytics translator. We already see the following applications areas for a translator within Insights Departments.

1. Companies usually have a lot of data sources available while they are not aware of the potential. In addition, more and more datasets are available via APIs or data brokers. This is e.g. our case study.

2. In addition, more text data becomes available; existing research reports as well as social media that can be analyzed with text analytics. Text analytics mainly uses machine learning tools.

3. Media departments also need data science support. Working with programmatic buying and audience targeting via DMPs (Data Management Platforms) is mainly done with machine learning tools. A company needs people who can judge the work of media agencies. This role can be fulfilled by a CMI translator.

As mentioned many companies require the Insights Departments to deliver cheaper and faster and generate more impact. More impact means that it is better to be able to feed the business leader on time with 80% of the required information than with 99% but take more time, while the business decision has already been taken. The application areas 1 and 2 mentioned above can play a crucial role in this. First use and analyze the knowledge that is already available within a company. This is facilitated if you have a good data infrastructure structure and knowledge management system. An analytic translator can play a proactive role in this field, also in the often difficult and political cooperation with ICT.

What does this mean for market research agencies? Many companies will outsource the data scientist activities. On the one hand because the amount of work does not justify a data scientist function, on the other hand because flexibly the data analytics can be hired. Agencies can adopt this capability and employ data scientists as a service for their clients.

References

DriveSens, Haystack International, 2018

Henke, N., Levine, J., McInerney, P. (2018), You Don't Have to Be a Data Scientist to Fill This Must-Have Analytics Role, *Harvard Business Review*

Roberts, D. E., Mustard, T. D., The Taste Signature Revealed, St. Albans, Ecademy Press

Data transformation statistics:
https://en.wikipedia.org/wiki/Data_transformation_(statistics)#Reasons_for_transforming_data

Introducing to Machine Learning: https://nl.mathworks.com/campaigns/offers/machine-learning-with-matlab.html

Ng, A., Machine Learning – Stanford University

.

The Authors

Sjoerd Koornstra is Partner, The House of Insights, Netherlands

Wim Hamaekers is Managing Partner, haystack International, Belgium