Good Models out of an assembly line or advances in targeting

How to adapt processes and structures from manufacturing industry
to improve the benefit of predictive targeting

Author: Andrea Ahlemeyer-Stubbe

in: Transactions on Machine Learning and Data Mining, Volume 7, Number 2,
ibai publishing, Leipzig, October 2014

Abstract:

To detect fast changes in customer behavior or to react in as focused a manner as possible, predictive modelling must be done in good quality to get effective predictions of customer behavior and it has to be done fast to be relevant under business aspects. Modeling speed is of great importance in industry as time is a crucial factor. This necessity requires a different technical set up for model development to fulfill both needs: quality and development speed. Today most companies like to develop their models individually with the help of specialists.

But for a lot of companies, this way takes too long: even though the models are excellent, the time to develop them sometimes kills the advantages of a better prediction. This article describes the general structure and ideas how to implement industry-focused model production that will help to react quickly to changing behavior. We will discuss the key success factors and the pitfalls of this assembly line model product.

1 Marketing in the Web 2.0

Customer profiles are really valuable if they contain more information than just customer's past behavior, but also include present reactions and give reliable predictions about their future conduct.

A company's existing customer data may, for example, provide information about purchase and payment history, address, age, gender, etc. For all customers that can be identified this data is stored and used for targeting purposes. Even if there is a technical solution enabling you to identify customer who are already known when they visit your website, you never know all the people visiting your site. But there are plenty of reasons to learn more about them and to target them as well and to do it in the same personalized way as you do with those you can identify. But the only information available is what you can filter out of the log file.

So you know which and how many pages they have seen and for how long, where the came from and much more.

To use this information each individually brings obvious advantages: With the knowledge of the typical click patterns of customers reacting similarly and the knowledge of the positioning and placement of advertising for a product which appeals to most customers you will create a big benefit. With information such as entry and exit page, or click behavior, the structuring of web sites can be continually improved and their representation can be optimized on the web. Knowledge of the origin, points to the channels, which should best be addressed by the specific target groups.

All this information put together gives a multifaceted view of customers and target groups. To measure success, in order to recognize and to identify trends, changes and new needs of the target groups early, and seize enterprise opportunities, consider off- and online market study as well as on the analysis of the volume of data existing in the enterprise reporting, data mining and market observations. On-line measures of interested parties use the clicking advice or page impressions consulted as yardsticks [2].

But nothing changed in the last years as rapidly as the use of the Internet. 10 years ago, Web 1.0 offered static websites, Web 1.5 offered dynamic websites and since 2005 more users expect interactive Websites (Web 2.0). The World Wide Web develops itself constantly in large steps and Web 3.0 will start soon.

The Internet medium penetrates the market with its various application possibilities and offers world-wide access – across all society and ethical layers. Currently 79.9% of the Germans are already on the Web [8].

Ever more humans use the Internet ever more extensively (for search, E-Mail, forums, blogs, podcasting, on-line one of plays,…) – briefly: The Internet is a fast, global, highly competitive marketplace. Above all, the Web is used increasingly by many customers and enterprises as a starting point for the search for services and products.

The expenditure for on-line marketing today already takes more than 19.2% [13] of the total expenditure for advertising. This will continue to rise in all industries, because this market place opens various marketing possibilities – and demands completely new marketing strategies. Medium-spreading, target group-specific, relevantly – today must be tailored advertising best personally to each customer. For it is necessary to improve customer loyalty, facilitate the acquisition of new customers and improve the response or the conversion rate – all at very low advertising cost [3].

The current challenges facing new methodologies and technologies are, for example, the analysis of log files [14], information on the origin of the visitor, what browser he uses, which and how many pages he has viewed. The advantages are obvious: the person who knows the typical click behavior of their customers can determine with this knowledge, the positioning and placement of advertising for a product that appeals the most to its customers. With information such as exit and entry side, the structuring of web sites can be continuously improved and the representation can be optimized on the web [4].

Of exceptional value for strategic planning are reliable predictions about future developments in the behavior and needs of customers. The development of predictive behavioral targeting ensures that such predictions are placed on a statistically validated foundation.

2 Predictive Behavioral Targeting

For a company to receive exciting customer profiles, improve the relevance of its online offerings and optimize its long term online marketing ROI, it would not only need information about the historical and current habits of its customers, but also about their future conduct [5], [10]. [11]. So it is important to discover patterns in customer behavior, for this we need to identify a specific user. This makes the predictive behavioral targeting.

Methods such as descriptive statistics, click-stream analysis [6], discriminant analysis, regression methods [3], decision trees, neural networks, case-based reasoning (CBR) [12], cluster analysis [11] and time series analysis are used.

Based on analysis of user profiles and user structures (such a sage, lifestyle, peer group affiliations, browsing behavior) predictive models are created for future behavior [13]. For example, the decision on where to place banners which users should be shown, is based on the sites he visited or on the basis of what he's doing on these sites.

Previously contextual advertising based on the content of a website, identified content which is best suited for a display. The predictive behavioral targeting based on the users actual and past behavior, the right person for a quote with the ability to identify user profiles, which opens the way for predictive behavioral targeting relevant advertising.

Predictive targeting makes a lot and that information gain immensely in value when they are in the right place as quickly as possible in the right form ready. Fully automatic Predictive Targeting and real-time modelling of on-line behavior create for it the conditions.

3 Fully automatic Predictive Targeting and real-time modelling of on-line behavior

Using the necessary algorithms for analysis in real time opens up the new fully automated predictive targeting, individual and lasting forms of communication and offers a head start by modeling the real-time online behavior.

This means an evolution step comparably from the handicraft to the production – inclusive reductions at the complexity. The classical way to build complex forecast models by hand, is in the role of „Master Workshop". But even a large staff of analysts (craftsmen) cannot cope with the huge amounts of data and the high number of models – this needs a fully automated „assembly line" for prediction models.

3.1 Function

The core of the fully automatic predictive targeting system is the construction of prediction models. It includes all functionality to build with a team of analysts complex forecast models by hand („Master Workshop") but also in a second module („assembly line") it builds very fast fully automated, simple, click-based predictive models, automatically backs up its quality and makes it available for use.

In the „assembly line" all models are calculated, which is a relatively simple task in the field of predictions (predictive modeling), for example, only the models whose target variable is a dichotomous structure (clicked vs. not clicked, purchased vs. not bought, visits vs. not visited, etc.) [9]. These prediction models cover, for example, a large portion of the orders for banner sand optimization behavior targeting. Special analysis such as cluster analysis, are performed by the analysis team in the „Master Workshop".

It is decided by an administrative process whether a forecast model goes in the workshop or in the assembly line to be manufactured. Each model receives a clear ID and is archived. The substantial elements the assembly line contains are shown in Figure 1.

Modules of the Automatic Targeting Systems

- **Control**
  - Administration
  - Model Management
  - Management of the variables
- **Model Development**
  - Selection of the Target Variable =1 and Selection of the Target Variable =0
  - Model-specific random sampling to obtain a data set
  - Supply of a Learning Data Set and a Test Data Set
  - Model development
  - Model validation
  - Quality Assurance Monitor
  - Release Process
  - Handing over Process
- **Model Examination**
  - Model-specific random sampling to obtain a data set
  - Model Examination
  - Quality Assurance Monitor
  - Confirmation Process
- **Model Archive**
- **Variable Processing and Supply**

Fig. 1. The Elements of an Assembly Line

3.2 Architecture

The areas of „Master Workshop“ (working range of the analysis team) and „assembly line“ (automatic targeting) are in the existing architecture of a company included in a way that the environment and ist benefits can be used as far as possible. This includes especially all tasks around data preparations.

In general: All the developed models will be passed as a code / script and archived in an archive, including its metadata and also about their use. The application of models based on the scripts is the final step of the calculation of variables at the end of a session or a slot. For each active model the prediction is calculated as relevant forecast value per unique client (UC) and is stored in a separate variable and made available to the Target Builder.

In the Target Builder (an instrument for targeted delivery of content), these predictive capabilities of profiles are used to provide target audiences fror online campaigns and make them ready to be marked. So every user with fitting profile is addressed.

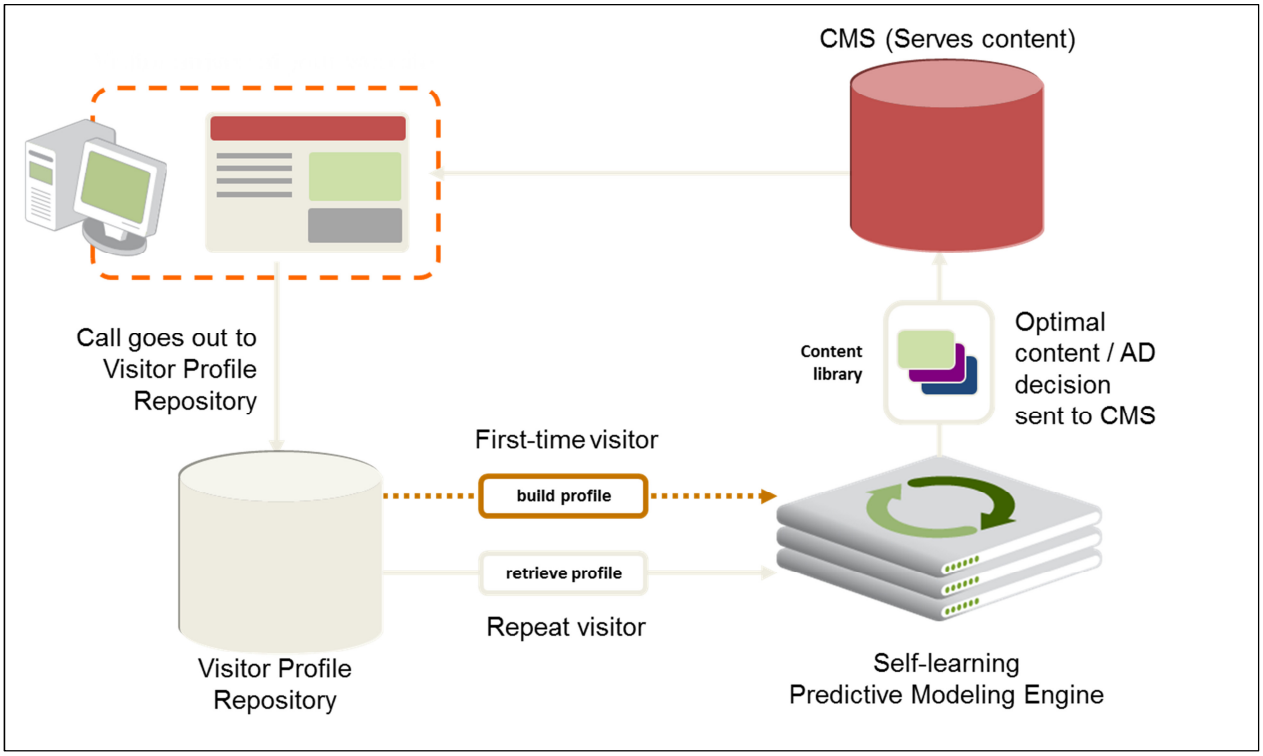

Fig. 2. Automatic Predictive Targeting

3.3 Data flows and database

The bases for all analyses are the behavior data of the UC on the respective websites.

Ideally the behavior data can be enriched by information from questioning or login data. One can differentiate between standard/general behavior and/or interest-conditioned information. These profile and behavior data per UC are computed relative to different time windows. Both are formed from large numbers of variables. The variables should reflect different views on the UC.

In the context of the modelling in the assembly line, only an examined subset of variables is used, in order to grant stability, robustness an performance essential automation. That can be seen somehow as the standard dataset. The size and content of it is the outcome of experience and domain knowledge obtainable from the analyses team. Using the Pareto principle the dataset should contain roughly 20 % of all potential data to be the basis of 80% of all easy prediction problems. To ensure that data preparation fits the time constraints of the Assembly line, it is important to use optimization techniques that help to deliver the data as fast as possible. After the structure of standard dataset is defined and adjusted, it will be created fully automatically by the systems.

As in other industrial processes it is important to identify and separate different process steps, such as data handling and calculations so that they can be standardized and optimized to deliver an outcome of reliable quality and with a more or less small standard error. About the data preparation including all necessary (automatically done) transformations, it lead us to the fact we should choose robust methods that will help to generate data that fits most (modeling) situations but might not be the optimal one for every single case as you would get, when you do the analytics by hand through an expert in your company. Where you try to optimize the modeling result by doing a constant looping out of

data preparation, analytics and measurement, especially under time constraints, this looping has to be at a minimum.

An essential part of the daily profile building step is data preparation. It is needed both for the modeling / verification as well as for the application of existing scores. All of has to be done in a time-critical area.

So that modeling or deployment can take place in the current session, it is indispensable that the session data can be accessed at any time in the session.

To train und validate a model, very fast sampling is mandatory. Please keep in mmd that the relation between those who act (target=1) and those who do not act (target=0) is very unbalanced. It is very likely that you have just 500 UC acting and 5 million UC not acting. As part of the modeling issue, it makes a big difference in calculation time whether you have to calculate slightly more than 5 million records or just a couple of hundred. So if you start to build an assembly line it is required to work out and test a sampling strategy that best suits your individual situation

3.4 Modelling Aspects

Similar to the data preparation, the assembly line needs for its modelling engine a data mining algorithm that will deliver good and stable models without any interaction with an expert. The algorithm should be fast und the time needed tot the deployment of models should be as small as possible, especially if it is planned to use the assembly line tot nearly real time forecasting. If it is under business reason not so important to be real time, for example it is enough if you use the data of the last finished session as the newest one to be included in the forecasting, or as a base forecast that is shaped by the content dick on or searched for, then the time for modeling und deployment might not be so time critical. Based on many tests simulating the situation like an assembly line, out of all the algorithms, the family of decision trees wins most, and delivers fast very good results.

The fully automated quality control is the next step in the process to deal with; the task here is to define the small border between not "quite as good" and "good enough under business reason". If your expectations for the modelling quality are too high you will end up with lots of models transferred back from assembly line to workshop to be redone by the experts, so you will lose time and it will cost you more money to produce the models. If your quality is too low you will lose business, e.g. lower click rates.

In quality control we are looking after new (freshly developed) models as well as after models that have been in duty for a while, so the chosen quality control process and it measurements must be able to do a constant quality control on all active models to notice in time when a model needs refreshment.

3.5 Challenges and critical success factors

The complexity of the process of fully automated predictive targeting and modeling of real-time online conduct presents some statistical challenges: During sampling the minimum reasonable number of events = 1 (e.g. clicks), stratification, sampling routines has to be fixed sensitively. Forecasting methods are judged in terms of prediction quality, stability, development, etc. Performance, durability, run-time behavior, parameterization, and automation of error detection must be considered in the selection of quality assurance methods.

Similarly, critical success factors should not be ignored: Are uncontrollable optimization steps I algorithms influencing the ad server that originally predicted massive click behavior? Are tags are missing in the banner? Are there so few clicks that it takes too long to get a critical mass for modeling?

Results

Fully automatic targeting and predictive modeling of real-time online behavior are at the very beginning of their development and are valuable tools provided critical success factors and requirements are carefully observed in the implementation and the environment. All being well we can obtain fully automated predictive targeting even based literally on the last click in real time, and these predictions can be flexible and up to date. This allows rapid response to changes and trends in the rapidly changing online marketplace.

References

1. Alberto Messina et all: A generalised cross-modal clustering method applied to multimedia news semantic indexing and retrieval, Juan Quemada et all (Ed.), Proceedings of the 181h International Conference on World Wide Web, WWW 2009, Madrid, 2009

2. Alice Klever: Behavioural Targeting: An Online Analysis for Efficient Media Planning?, Diplomica, 2009

3. Andrea Ahlemeyer-Stubbe: Behavioral Targeting: Which Method produces the most Robust Prediction? A Confrontation between Decisions Trees, Neural Networks and Regressions, Petra Perner (Ed.): Advances in Data Mining. 9th Industrial Conference, JCDM 2008, Ibai Publishing, 2009

4. Andrea Ahlemeyer-Stubbe: Predictive Targeting: Buzzword or Reality - The potential of Automatic Behavioral Targeted Advertising in Online Marketing, Petra Perner (Ed.), Advances in Data Mining. 9th Industrial Conference, ICDM 2008, Ibai Publishing, 2009

5. David S. Evans:, The Online Advertising Industry: Economics, Evolution, and Privacy, The Journal of Economic Perspectives, Volume 23, Number 3, Summer 2009 , pp. 37-60(24), 2009

6. Deepak Agarwal et all: Spatio-temporal models for estimating click-through rate, Juan Quemada et all (Ed.), Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, 2009

7. Foster J. Provost et all: Data acquisition and cost-effective predictive modeling: targeting offers for electronic commerce, Maria L. Cmi et all (Ed)., Proceedings of the 9th International Conference on Electronic Commerce: The Wireless World of Electronic Commerce, 2007, Minneapolis, 2007

8. Internet World Stats 2011, http://www.intemetworldstats.comlstats4.htm, 25.September 2011

9. Joseph Reisinger, Marius Pasca: Bootstrapped extraction of class attributes. Juan Quemada et al (Ed.), Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, 2009

10. Jun Yan, et al: How much can behavioral targeting help online advertising?, Juan Quemada et all (Ed.), Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, 2009

11. OVK Online-Report 2011/01, Online-Vermarkterkreis(OVK) im Bundesverband Digitale Wirtschaft (BVDW) e-V., (Hrsg.), Düsseldorf, 2011

12. P. Perner, Case-based reasoning and the statistical challenges, Journal Quality and Reliability Engineering International Volume 24 Issue 6 (October 2008), p 705-720.

13. Petra Perner, G. Fiss: Intelligent E-marketing with Web Mining, Personalization, and User-Adapted Interfaces, Petra Perner (Ed.), Advances in Data Mining, Applications in E-Commerce, Medicine, and Knowledge Management [Industrial Conference on Data Mining, Leipzig, Germany, June 2002]. Lecture Notes in Computer Science, Springer 2002

14. M. Reichle, P. Pemer, K.-D. Althoff, Data Preparation of Web Log Files for Marketing Aspects Analyses, In: Petra Perner (Ed.): Advances in Data Mining, Applications in Medicine, Web Mining, Marketing, Image and Signal Mining, Ines 4065, Springer 2006,p. 131-145.