Igor Mandel
imandel@telmar.com

Causal models in estimation of the advertising ROI

Abstract

A new approach to causal modeling was recently proposed and published. Its main differences from more traditional counterfactual framework (potential outcomes, structural graphs and equations) are lying in several features: it doesn't require any assumptions of "potentials", working only with existing data; it works successfully with data where covariates are correlated among themselves, but practically not correlated with dependent variable (thus sharply eliminating the multicollinearity problem), it replaces regression-like paradigm to the concept of intrinsic probability, etc. It was actively used in media planning by Telmar and showed very promising results. In this presentation these results are discussed in combination with another challenging question: how orientation of the media planning not just to people, but to people with specific (highest or lowest marketing value) may change the way the ROI is measured and traced? Questions like that are undoubtedly very important from both practical and theoretical points of view.

Key words: causal modeling, intrinsic probabilities, media planning, Telmar, ROI

1. Causal models

The general model for separation of the causal effects from the random ones have been conceptually proposed in (Mandel, 2013) and mathematically developed in (Lipovetsky and Mandel, 2015). It is distinct from the actively developing area of the causal modeling like in (Pearl 2009) and the concept of potential outcomes, presumably lying in its base (Rubin, 2006) in several aspects:

a. It is focused on **one dependent variable**, not many, and does need for that reason acyclic graphs presentation (although could be generalized in this direction as well)

b. It assumes that certain events are in reality "caused" by two types of causes: **random**, i.e. those for which we could not find the association with any measurable characteristic (covariate) of the data and **covariate-specific,** i.e. those determined by that covariate with certain probability.

c. It doesn't mean the **physical or behavioral "causes" as forces** (the only real, not "potential" actors in the game), but merely the fact, that each part of the universe, constrained by covariate's value, has its own, different from others, probabilities to generate the outcome.

d. The event occurs whether random or specific causes make it occur. If they worked simultaneously – **it would not produce any different result**, it would still occur just once.

e. This latter assumption makes a main point of departure from the traditional regression and structural equations models (a part of the causal machinery): the **causes need not "accumulate" to make the outcome**; each works separately

The basic equation of this type of causal analytics is following from the formula of probabilities summation:

$$S = S_{causal} + S_{random} - S_{causal}S_{random} \qquad (1)$$

where S is a probability of the event in question, decomposed to the respective components. If one has K covariates and each is associated with its particular (hidden) probability ***p*** of the generating ***Y*** (the outcome), and also there is a general probability of the random occurrence of ***Y, r,*** then the main equation to estimate ***p*** and ***r*** parameters, as shown in (Lipovetsky and Mandel 2015) is:

$$S_i = 1-(1-r)\prod_{k=1}^{K}(1-p_k)^{x_{ik}} \qquad (2)$$

where Si is a frequency of ***Y*** in the particular cell of the design matrix (each ***i*** stands for particular combination of values of ***K*** covariates) and ***Xik*** is a value of covariate in row ***I*** of this matrix.

The estimation of the parameters in (2) and its problems were discussed in the mentioned articles. Here I want to focus on very specific case: when all covariates **are independent,** i.e. the whole data set is just a description of the data, broken into K groups.

For this case, direct general regression-like estimation based on design matrix is not an option, because the number of parameters is K+1 (K – for specific ***p*** and 1 for random ***r***) is higher than number of rows in a matrix K, therefore, the new way of estimation should be created.

One of the approaches is to add some noise in the data, which artificially creates more rows in design matrix and allows to run the regression. However accuracy for this approach is actually not that high. Another, much more solid approach is based on the fact that in each of K groups now only two "forces" are in play: random and group-specific. It means, for each group equation (1) works directly, with ***r*** common for all groups:

$$S_j = p_j + r - p_j r; \quad j = 1, K \qquad (3)$$

If one may correctly estimate ***r*** – the estimations of ***p*** will be obtained automatically from (3).

The simplest way to do this is to test different values of r from 0 to 1, while simultaneously varying different values of p, to find the combination where deviations of theoretical S from (3) from observed S values is minimal. For example, if K=2 if one varies r, p1 and p2, each with step 0.01 – there is just one combination of all of three parameters, what makes the squared (or absolute) error minimal. The total number of combinations though is 100^3 = 1m, which is large for such a simple task. I ran this type of simulation and the results were excellent, as expected: if data is artificially generated with certain p and r, the recovery of these parameters with this procedure was very precise. But it is clear that it **will not work for a large K.**

I found that if one systematically changes the p values in the generation process by replacing with random values and run several times – the results remain surprisingly good. This is of critical importance, for it eliminates the problem of dimensionality.

One may increase the accuracy of the estimation by the following:

- varying r not from 0 to 1, but from 0 to observed frequency of Y, but with smaller steps
- varying p also not from 0 to 1, but from 0 to 5r or something like that. The reasoning is that particular groups, most likely will not be very different from the baseline r (at least in adverting area, where I applied all that)

Many experiments show that the proposed simulation techniques produce very good estimations for different situations. If data, indeed, had been generated in such a way that Y=1 if and only if one of the causes (either random or specific) had worked out – we can make a good estimation of the coefficients. But what if data is made differently?

In one experiment, Y was just random, i.e. not related in anyway with both X values and causal coefficients. However, the algorithm found certain values of r and then, respectively, estimated p – and observed S were again almost the same as theoretical ones. These values were, of course, **very different** from the ones used in generation. It shows that the similarity of predicted and observed frequencies to each other **cannot serve as a usual goodness of fit statistics like R^2**. When data is not generated, one does not know, if either p and r are correct or not, since any ones "fit" the data. It raises the fundamental question: can we say, **what mechanism generated the data in reality?**

A general answer to the question about the data generated mechanism, I believe, is negative – there could be so many mechanisms out of all possible. But it is in the same venue, as fictitious foundations of the regression or causal models – no one knows what exactly happened behind the scene. But what we could do is to find another type of criteria, which distinguishes data which has assumed relations between Y and X via causal model and those which does do not.

It is based on the following observation: if in the data set, generated with "correct outcome" Y is replaced by a random variable, having the same frequency as in the original data set, the R^2, as mentioned, is not changed. What changed are these two statistics:

- Average error, calculated as the difference of the estimated and used in generation causal coefficients in proper model (almost 0) and in one with random Y (very high);
- Correlation between these two sets of coefficients – almost 1 if data set is "correct" and practically zero if Y is random.

- Both of these statistics change respectively as just part of Y is replaced by random values (say, 30% is random, 70% - properly designed)

These features allow to introduce something that could be called "**model relevance criteria**", a new type of statistics, which differs both from traditional goodness of fit and machine learning's errors of testing.

2. Application of causal models in estimation of the advertising ROI

These modifications of the main causal model (2), together with other considerations, led the creation of a special system of large marketing data analysis at Telmar, called **Telmar Audience Effect (TAE),** which was recently launched.

The main purpose of the system is to select just a few most informative variables out of many potentially influencing the results and use them to make the model of the target (independent) variable). For modeling itself CHAID analysis is used, as a practical tool providing convenient marketing solutions. Then the causal analysis is applied to obtained segments. Let's consider it on one example.

The purpose was to identify the segments of population, intensively consuming any **Bourbon** whiskey (8.6% of the population). MRI data provides such information (all calculations below do not have any commercial value and made only for demonstrational purposes). I selected for targeting all **1,258 Demographic candidate variables** to be considered.

The first thing TAE does is dramatically reduces the number of variables. It applies an algorithm to find optimal selection among two criteria, in this case: Index, i.e. how concentrated the Bourbon consumption is within a group is, and Target population, i.e. how many customers this group has (see Fig.1.).

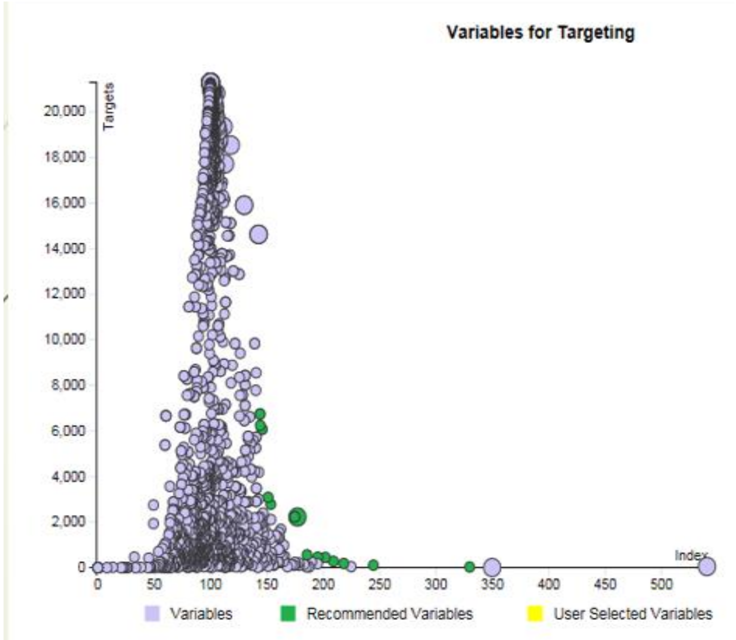

Fig.1. Pareto-like variables selection for two criteria

Ideally, if some variable has high values for both of these – it is very good for marketing campaign. The set of criteria may vary.

Second, it calculates gain chart of the CHAID model in two versions – traditional and causal-specific (Fig.2). It allows to make decision based on either logic, as a function of business objectives: if one is going to target only these people who would buy because they belong to specific groups – she should look at the green cumulated (causal related) values; if the goal is to target everyone (thus, potentially, losing money on these who will buy anyway) – she should follow the blue dots.

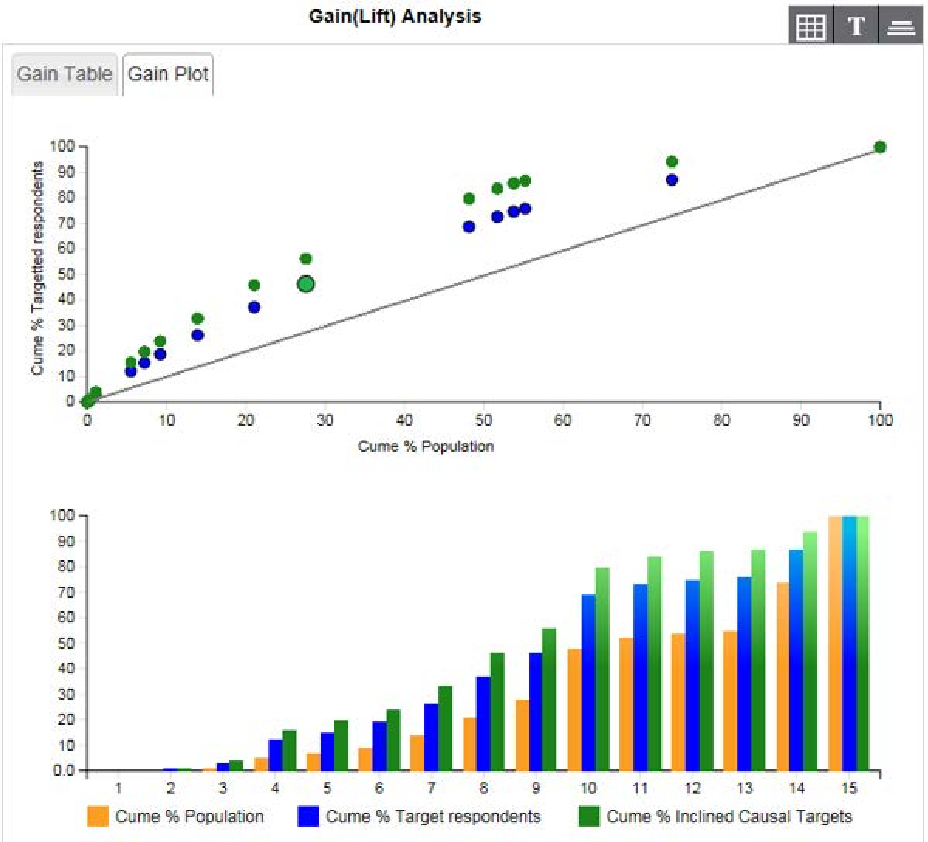

Fig.2. Two gain charts for traditional and causal specific targets

The proposed combination of these two ideas – goal oriented selection of the variables and further narrowing of the targeting by specifying causally determined individuals within groups – seems very promising. The next step could be to make causal modeling inside the decision tree logic.

References

Mandel, I. (2013), Fusion and causal analysis in big marketing data sets. *Proceedings of JSM - Section on Statistics in Marketing*. Montreal, Canada: American Statistical Association.

S. Lipovetsky and I. Mandel (2015) Modeling Probability of Causal and Random Impacts. *Journal of Modern Applied Statistical Methods 2015, Vol. 14, No. 1, 180-195.*

Pearl, J. (2009), *Causality: Models, reasoning, and inference*. Cambridge: *Cambridge University Press.*

Rubin, D. B. (2006). *Matched sampling for causal effects*. *Cambridge, MA: Cambridge University Press.*