Moving from big data and machine learning to smart data and causal modelling: a simple example from consumer research and marketing

Norman Fenton, 23 March 2015

Company ABC provides a service to a large number of organisations of different sizes in various countries. The company is concerned at the high rate of customer 'churn' – i.e. of organisations who fail to renew the service when their annual subscription ends. Since a lot of data is collected on each customer they decide to use a big data machine learning approach to discover which factors/combination of factors can be used to identify those customers most likely to be lost. Their customer database looks like this:

	"RISK FACTORS"				
Customer number	**SME**	**Outside EU/USA**	**No Discount offered**	**Recorded complaint**	***Customer Lost***
1	N	N	N	N	***N***
2	N	N	Y	N	***Y***
3	N	N	N	N	***N***
4	N	N	N	N	***N***
5	Y	N	N	N	***N***
6	N	N	Y	N	***N***
7	N	N	N	Y	***Y***
8	N	N	N	N	***Y***
9	Y	N	N	N	***N***
10	N	N	N	N	***N***
11	N	N	N	N	***N***
12	N	Y	N	N	***N***
9999	Y	N	N	Y	***Y***
10000	N	Y	N	N	***N***

N= "No", Y= "Yes"

A total of 1893 out of the 10,000 customers are lost (i.e. 18.9%). By grouping the customers according to the different 'profiles' of the 'risk factors' we also discover the information in Table 1 :

Table 1: Risk Factor Profiles

"RISK FACTOR PROFILE"						
SME	**Outside EU/USA**	**No Discount offered**	**Recorded complaint**	***Number of customers with this profile***	**Total with this profile lost**	**% of those with this profile lost**
N	N	N	N	3213	270	8.4
N	N	N	Y	357	96	27.0
N	N	Y	N	2142	523	24.4
N	Y	N	N	567	108	19.1
Y	N	N	N	1377	226	16.4
N	N	Y	Y	238	76	31.8
N	Y	N	Y	63	19	30.2
N	Y	Y	N	378	108	28.6
Y	Y	N	N	243	58	23.8
Y	N	Y	N	918	253	27.6
Y	N	N	Y	153	45	29.4
N	Y	Y	Y	42	14	33.1
Y	N	Y	Y	102	33	32.8
Y	Y	N	Y	27	9	31.7
Y	Y	Y	N	162	49	30.5
Y	Y	Y	Y	18	6	33.7

This information is useful. It is clear that even a single 'risk factor' in the profile leads to an increased probability of the customer being lost.

Throwing the best data mining and machine learning techniques at the data will result in models that all essentially provide the following information:

- Each of the four risk factors increases the chance of a lost customer.
- In order of impact (from most to least important) they are:
  1. Recorded complaint
  2. No discount offered
  3. Outside EU/USA
  4. SME

A full sensitivity analysis of these risk factors on lost customers is shown in Figure 1.

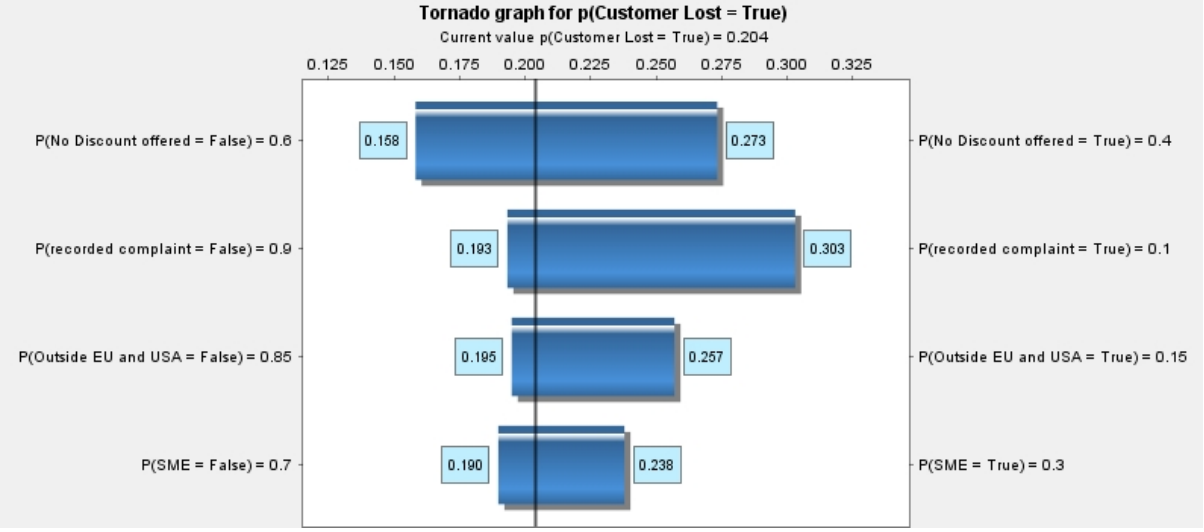


Figure 1 Tornado graph showing sensitivity analysis of risk factors on lost customers

The models (such as logistic regression, learnt BN, k-nearest neighbours etc, which in causal structural terms you can think of as Figure 2) will all provide you with a means of predicting the probability a customer is lost based on their 'profile'. Moreover, these predictions will be much more accurate than just tossing a coin. But they will still not be very 'accurate'.

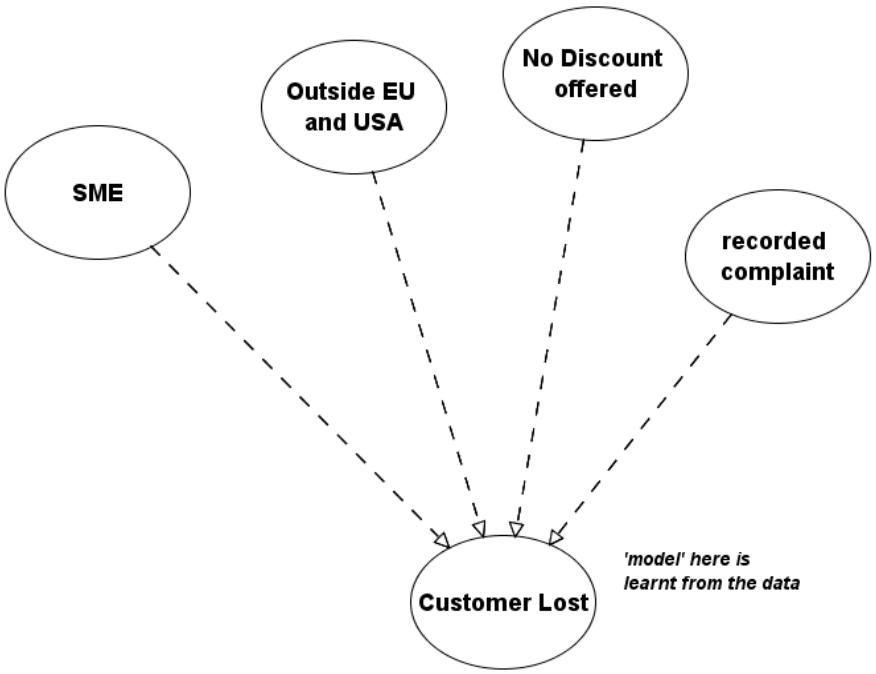


Figure 2 Model learnt from data

There are many standard ways to measure 'accuracy' of these models and the particular measure you use depends on what the objectives for the model are. For example, if the objective is to identify the '5% of customers most at risk of being lost' then the accuracy can be measured as the proportion of those customers who actually ***were*** lost out of the 5% that the model said were most

at risk. For example, the model might suggest that the 5% most at risk are those with profile YYYY, NYYY, YNYY.

Unfortunately, irrespective of how you measure accuracy (and there are many ways to do so) the accuracy of all these models will actually never be particularly good. In this example all of the state-of-the-art machine learning techniques will produce results that are about 32% accurate in identifying the 5% of customers most at risk.

So the data alone provides limited predictive analytics. Moreover, even though it does provide reasonable information on customers at risk of being lost it provides only very limited information on how to improve things: the only factor directly within our control is to offer a discount to more customers. We have learnt that customers who have registered a complaint are the most likely to be lost, so although we cannot stop them registering a complaint we can plan to improve service to reduce the number of future complaints. However, the other factors (whether the customer is an SME and whether they are based outside EU/USA) are beyond our control (unless we wish to take the irrational step of barring such customers).

Yet it turns out that the data is hiding some crucial 'knowledge' – easily obtained from people within the organisation, but not in the data. Specifically, the customer support team tries to call a reasonable number of its existing customers shortly before their contracts are due for renewal. For example, over 80% of customers who have recorded a complaint will get a call. It is known that getting a personal call can turn a dissatisfied customer into a satisfied one. It therefore turns out that the support team is already using some of the key 'risk factors' to decide which customers to call.

Although the telephone calls are not included in the database, the support team know that a call is especially important for an 'at risk' customer – they know from experience that about 70% of the most dissatisfied customers will renew their contracts with the call compared to just 20% without the call. When this additional 'expert judgment' is incorporated into a causal model we get the results shown in Table 2:

Table 2: Data with expert judgment added

"RISK FACTOR PROFILE"					CALL MADE	NO CALL MADE
SME	Outside EU/USA	No Discount offered	Recorded complaint	% of those with this profile lost	% of those with this profile lost	% of those with this profile lost
N	N	N	N	8.4	**5.8**	**11.5**
N	N	N	Y	27.0	**25.1**	**34.6**
N	N	Y	N	24.4	**23.1**	**28.4**
N	Y	N	N	19.1	**18.5**	**20.2**
Y	N	N	N	16.4	**15.9**	**17.4**
N	N	Y	Y	31.8	**28.3**	**53.6**
N	Y	N	Y	30.2	**27.2**	**46.0**
N	Y	Y	N	28.6	**26.1**	**39.4**
Y	Y	N	N	23.8	**22.7**	**27.3**
Y	N	Y	N	27.6	**25.4**	**36.1**
Y	N	N	Y	29.4	**26.7**	**42.3**
N	Y	Y	Y	33.1	**28.9**	**62.0**
Y	N	Y	Y	32.8	**28.7**	**59.7**
Y	Y	N	Y	31.7	**28.1**	**52.7**
Y	Y	Y	N	30.5	**27.4**	**47.1**
Y	Y	Y	Y	33.7	**29.2**	**66.6**

So, whereas the 'machine learnt' model looks like Figure 2, the correct causal model looks like Figure 3:

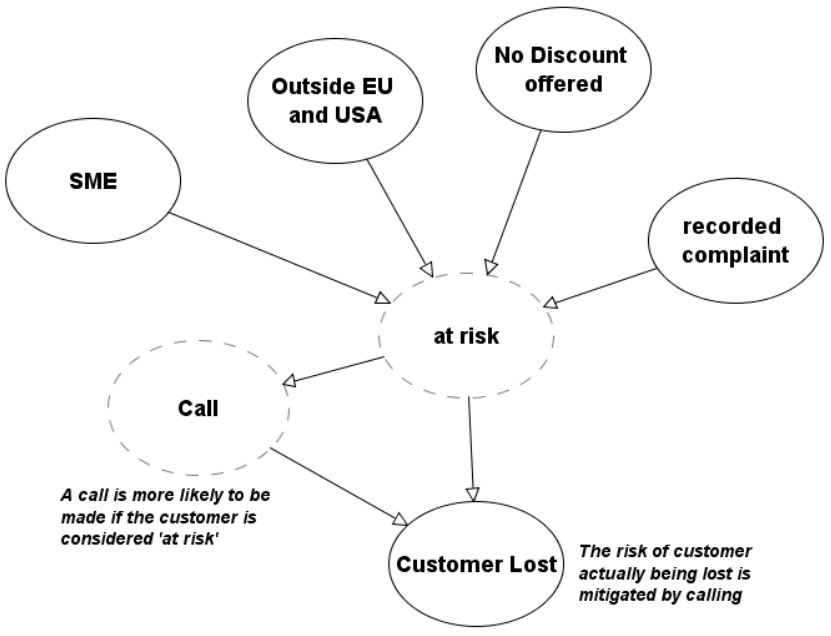


Figure 3 Causal model incorporating expert judgment

When we run the model in AgenaRisk the marginal probabilities are shown in Figure 4.

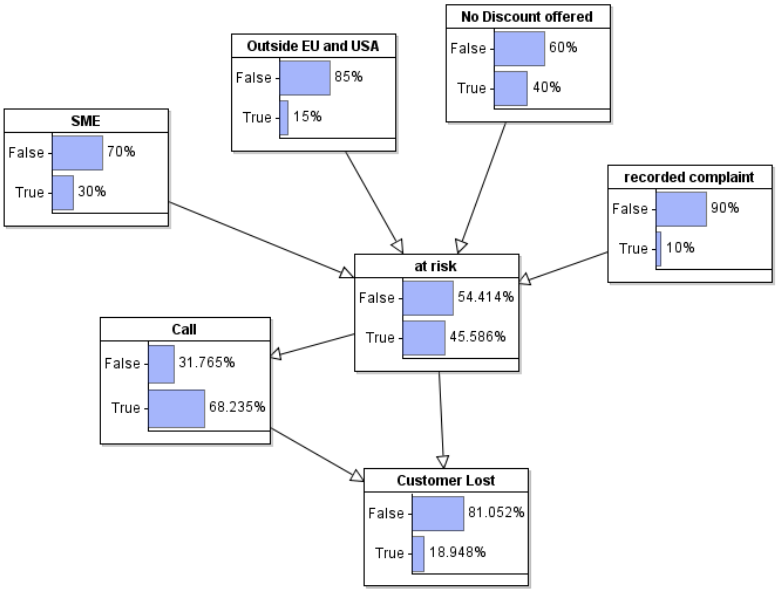


Figure 4 Model with marginal probabilities

For example, this shows there is an 18.9% chance a customer is lost (which matches the data). Figure 5 shows the revised probabilities when all risk factors are True. Note that in this case the chance a customer is lost increases to 33.6%

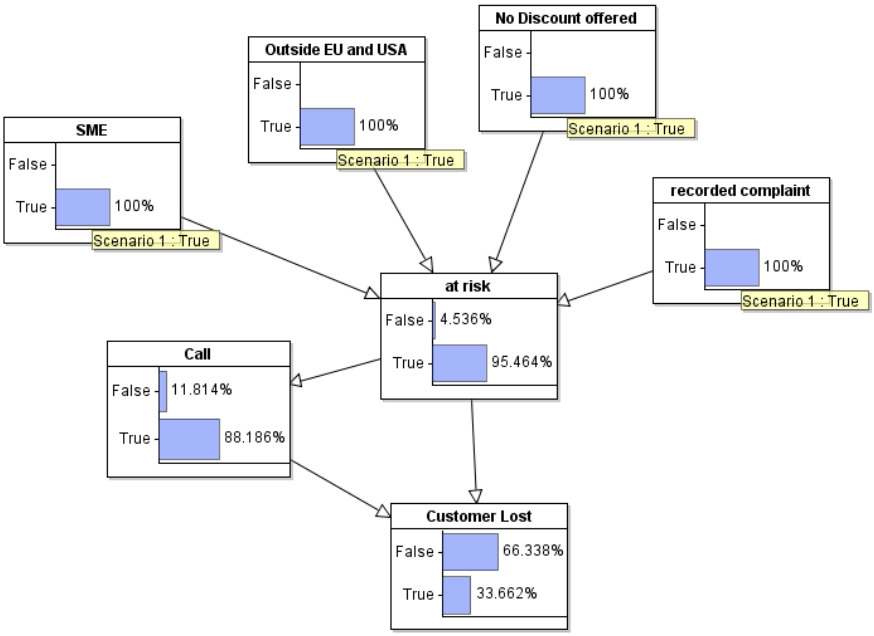
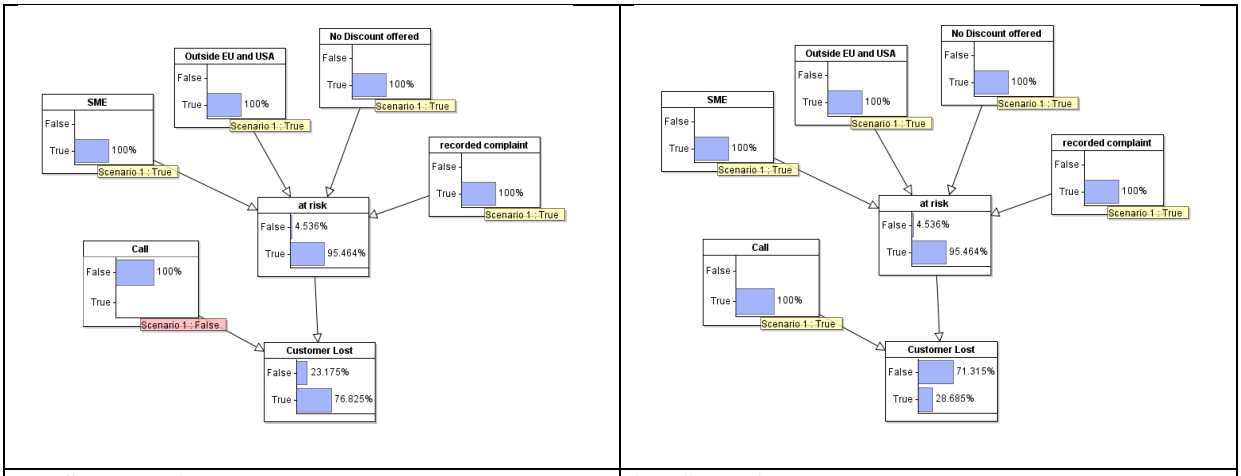


Figure 5 All risk factors True

However, note that in this case the probability of making a call has increased to 88%.

Figure 6 shows[1] the different impact of making and not making the call in this 'high risk' scenario. When the call is ***not*** made the probability the customer is lost is 76.8%, but this drops to 28.6% when the call is made. This crucial difference was 'lost' in the pure machine learnt model.



a. Call is not made | b. Call is made

Figure 6 Comparing case when call is or is not made

[1] The action of making the call or not requires us to break the link from 'at risk' to 'call' – this is essentially Pearl's do-calculus