Business insight

What's the difference in terms of ROI between the predictive models built by the people from the "*T.J. Watson Research Center*" (with average AUC of 85.21% - the winner of the KDD2009 cup) and a predictive model built with TIM (with average AUC of 82.55) for the KDD2009 cup. What does these 3 percents difference in accuracy represent in terms of euros?

Note: The "*T.J. Watson Research Center*" in Yorktown is a research center that is 100% financed by IBM. Since this center is 100% financed by IBM, it would by logical to think that all these researchers used the SPSS software (recently acquired by IBM) to create all their predictive models for the KDD2009 competition. This is actually not the case. There was no model created with SPSS amongst the 3x1200=3000 individual models created by the IBM team (they used the classical "ensemble technique" that requires building hundreds of different models. The final prediction is the "average" of the individual prediction of the 1200 individual models). For a complete list of the software packages used by the IBM team, see:
http://www.business-insight.com/downloads/IBM_kdd2009_paper.pdf
And now, two small remarks:

- Don't you think that it's a little bit "Strange" that SPSS was not used at all? Personally, I see it as a public confession that SPSS is a worthless datamining software.
- The time and the computing power required to develop these 3600 individual models make the IBM solution inconceivable in the "real world". The TIMi solution is the only solution amongst the top winners of the KDD2009 cup that can easily be used in any "real world" or "industrial-world" context.

To answer the above question, let's first make some assumptions:

1. In this document, we will study the ROI of a predictive model. We will assume that this predictive model is used for cross-selling purposes. For the KDD2009 cup, we didn't know which product was offered in "cross-selling". The only thing that we knew is that, one of the 3 problems was a cross-selling problem. To fix our idea, let's assume, for example, that the objective is to "predict" which customers (amongst all the customers that currently have a "*mobile phone subscription*") will buy an "*internet connection subscription*" (in addition to their "*mobile phone subscription*").

2. Amongst all the customers that currently have a "*mobile phone subscription*", there are already around 2% of people that currently already have an "*internet connection subscription*" (this number is extracted from the "appetency/cross-selling" problem from the KDD2009 challenge dataset and represents the reality of the challenge). In other word, the size of the "target population" (the "buyers") is (currently) 2% of the global population.

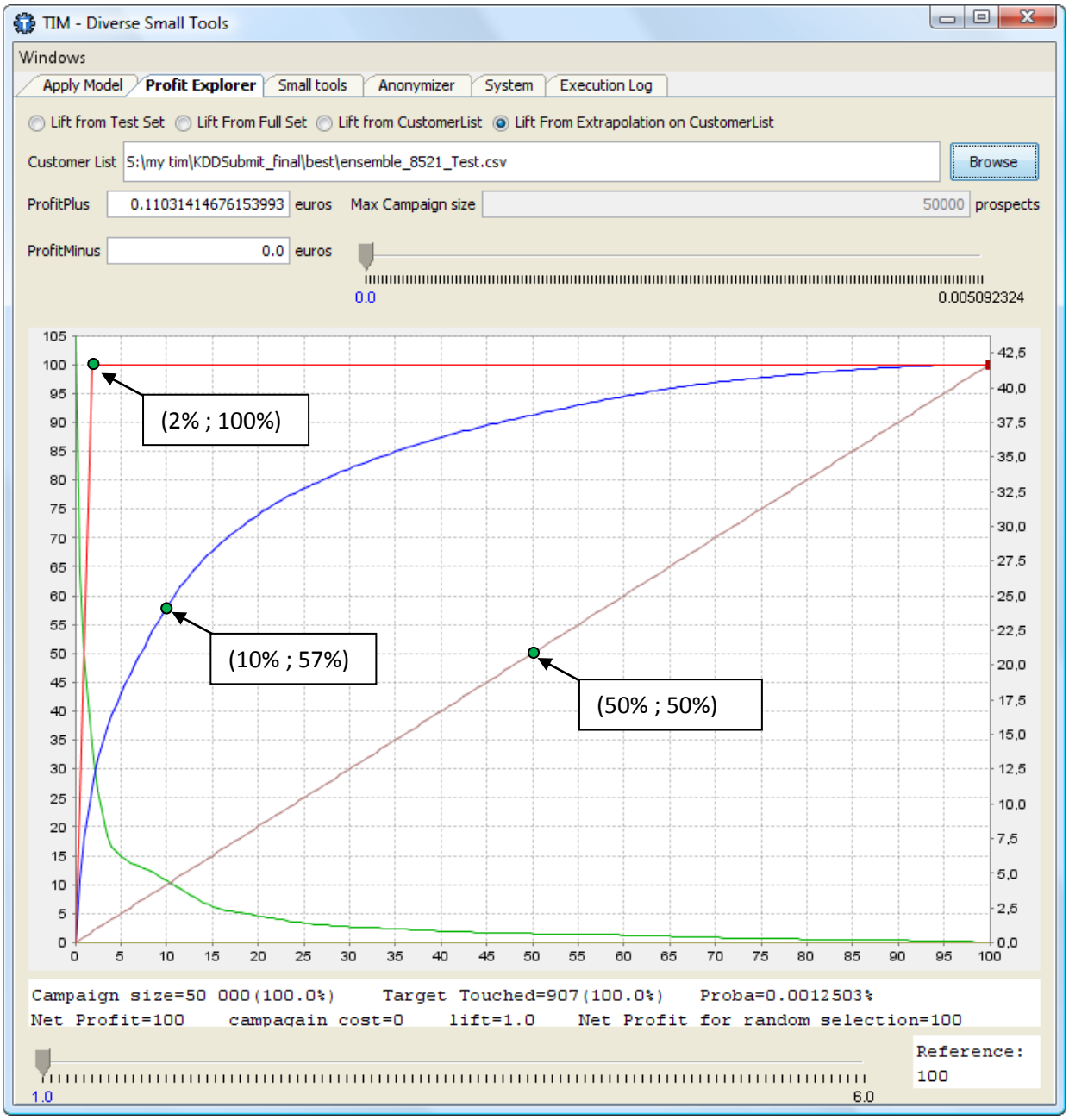
3. If we manage to successfully contact an individual that really wants to subscribe to "Orange" to obtain an "*internet connection*", we will assume that this individual won't forget about the offer and really do the subscription.

   In reality, this is not always the case. Even if people want to buy, even if they have a really nice offer, sometime they forget. With TIM, you can easily detect&target these "missing" individuals (this is called a "response model"), to reach a near-optimal conversion rate.

If you don't want to make any "response model" (or if your datamining software does not allow you to easily&efficiently create these models), you should multiply all the ROI given in this document by a pessimistic security factor (let's say "0.5") to account for the fact that some people forget to subscribe even if they want to.

4. For the time being, we will assume that there are only around 50.000 customers that have a "Orange" mobile phone subscription. In reality, "Orange" has a lot more customers and we will have to "adjust" later-on the results obtained so far to take into account the real-number of "Orange" customers.

Let's start by introducing the notion of "lift curve". The lift curve represents the quality of the predictive model. The best predictive model for the KDD2009 (obtained by the IBM team) had a "lift curve" (with AUC=85%) that is very close to the lift curve represented in blue in the screenshot below:



When we “apply” a predictive model to a list of “*potential buyers*” (in our case, the “*potential buyers*” are the people that have a “*mobile phone subscription*” at “Orange”), we can sort all these people from the best (with the highest probability of purchase of the “*internet connection subscription*”) to the worse (with the lowest probability of purchase, estimated by the predictive model). We obtain an ordering (in technical terms: a "ranking") where the first clients are the best one. The quality of this ranking is illustrated on the Lift Curve (the blue curve on the graph above). The estimated probability of purchase is illustrated with the green curve above.

For example, in the graph above, we see that, if we contact the top 10% of this list of “ranked client list” (this ranked-list is constructed using the “*probability of purchase*” estimated by the predictive model), we will find, inside our database 57% of the buyers (we will find 57% of the people that will actually purchase the “*internet connection subscription*”). Thus:

- The horizontal axis represents the list of all the “*potential buyers*” (from 0% to 100%) (i.e. all the the people that are known by “Orange”)
- The vertical axis represents the % of Target found by the predictive model (from 0% to 100%).

The “perfect predictive model” is able to “find” all the buyers (i.e. all the people that will buy an “*internet connection subscription*”) (i.e. 100% of the “target population”) by selecting only 2% of the global population. The “perfect predictive model” is illustrated by a red line that goes through the point (2% ; 100%). It’s not possible to find a predictive model that is better than the “red line”.

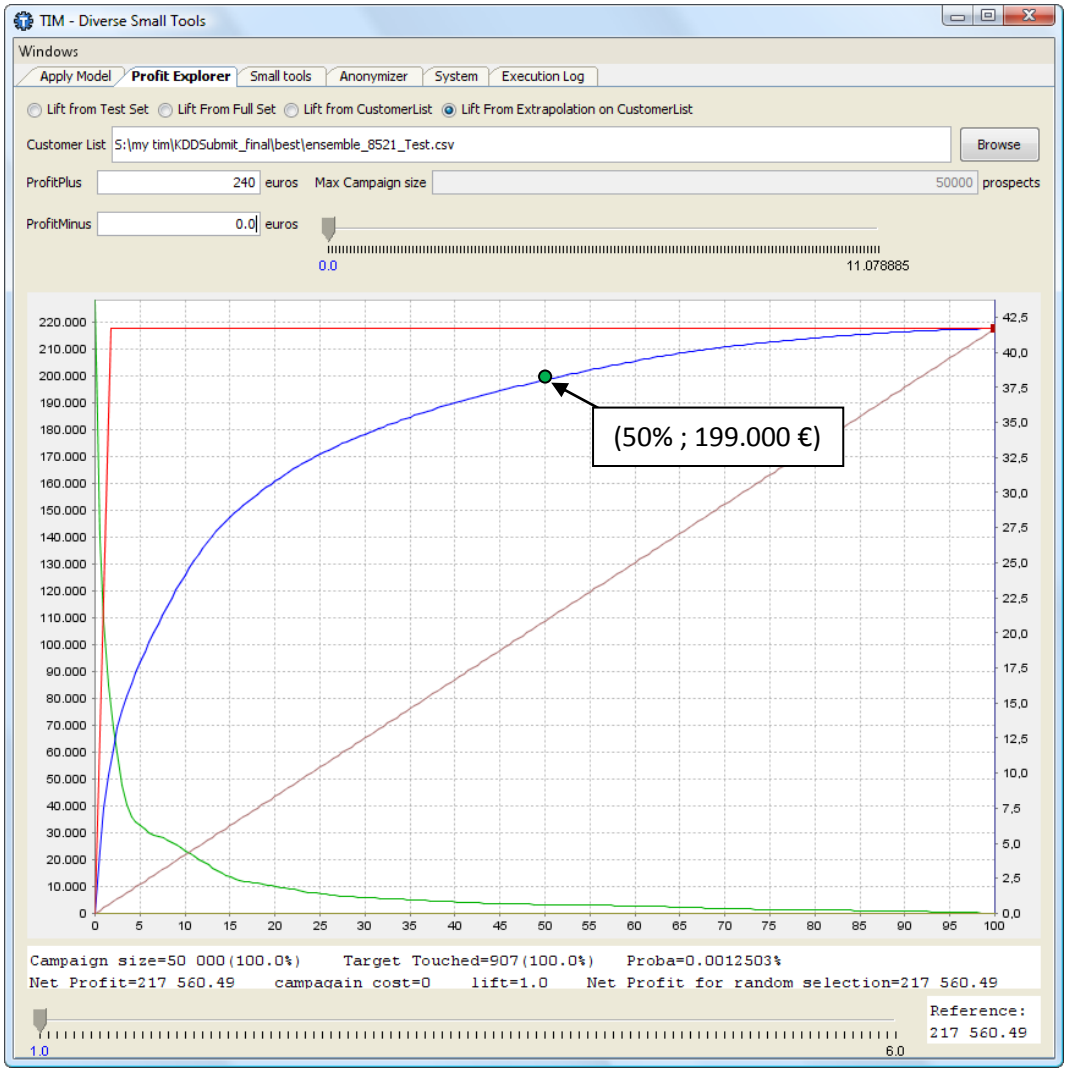
The “worst predictive model” is the “random ordering”: it is illustrated in pink: if you contact randomly 50% of your population, you will "touch" 50% of your targets.

The AUC of a predictive model is the “**A**rea **U**nder the (blue) **C**urve”. This area is normalized so that the AUC of the “perfect predictive model” (i.e. the area under the “red line”) is 100%. The AUC is a common measure of the quality of you predictive model. It is used for nearly all datamining completion (including the KDD2009 cup).

We will now make another assumption:

- The price of the “*internet connection subscription*” is 20 euros/month (or 240 euros/year).

Let’s enter the number “240” inside the “profit plus” parameter box. We obtain:

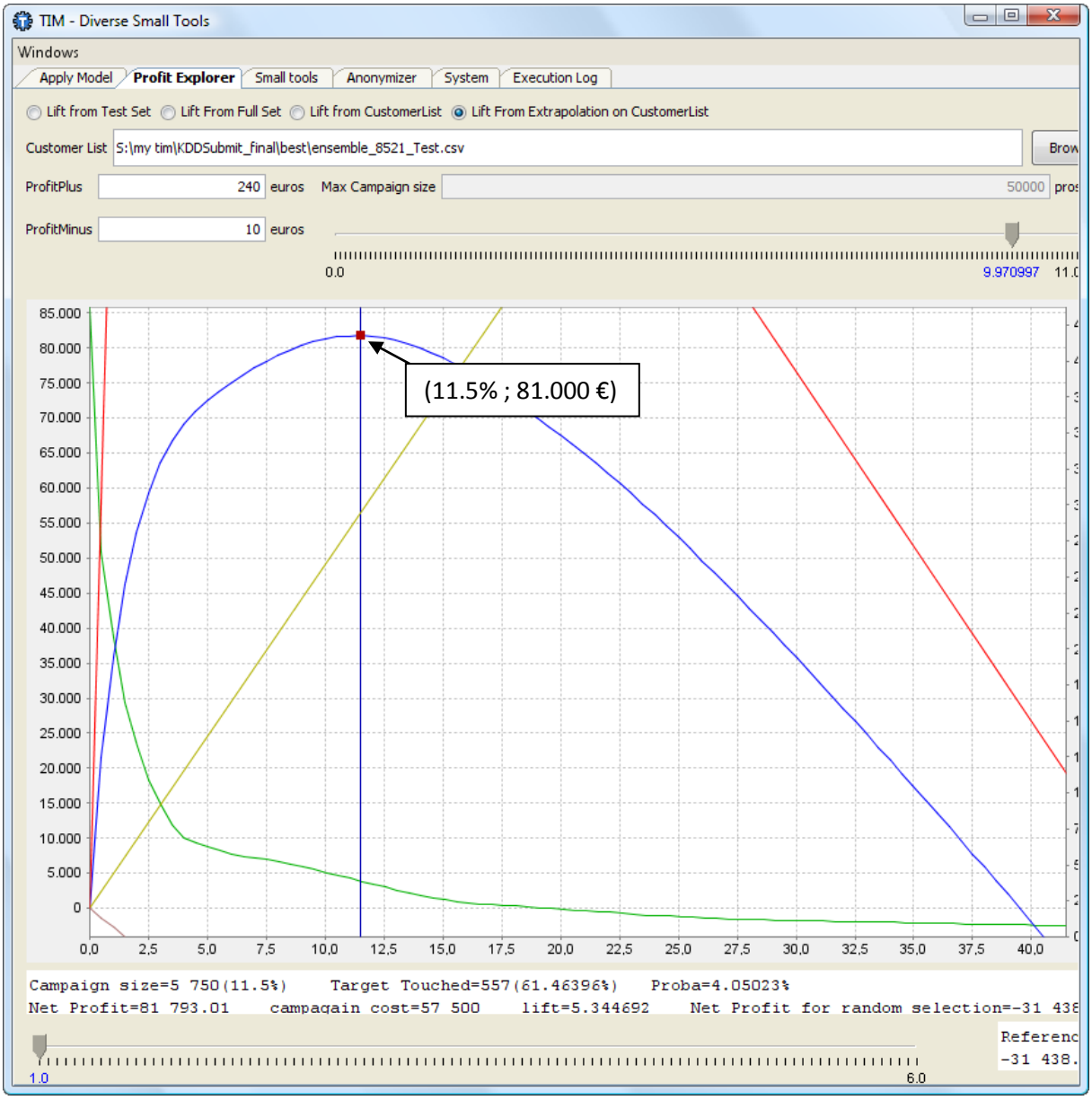


Please note that the graphic did NOT change (compared to the graphic on page 2)! Only the meaning of the vertical axis changed: Now, the vertical axis represents the net return of the marketing campaign in euro. The point (50% ; 199.000 €) means that, if we use the predictive model to select the "right" 50% of the total population, the marketing campaign will bring us 199.000 €. With these set of parameters (ProfitPlus=240 € and ProfitMinus = 0€), the "best" marketing campaign is a campaign where you "contact" everybody (i.e. the point (100% ; 217.000€) is the marketing campaign with the highest ROI). This is completely un-realistic.

We will now make another assumption, to correct for this un-realistic situation:

- Each time, we contact one individual to propose him an "*internet connection subscription*", we will lose 10 euros (price of the stamp, plus price of the small "welcome gift" or whatever).

Let's enter the number "10" inside the "*profit minus*" parameter box. We obtain:



Now, the best marketing campaign Is represented by the point (11.5% ; 81.000 €): , if we use the TIMi predictive model to select the "right" 11.5% of the total population, the marketing campaign will bring us 81.000 €.

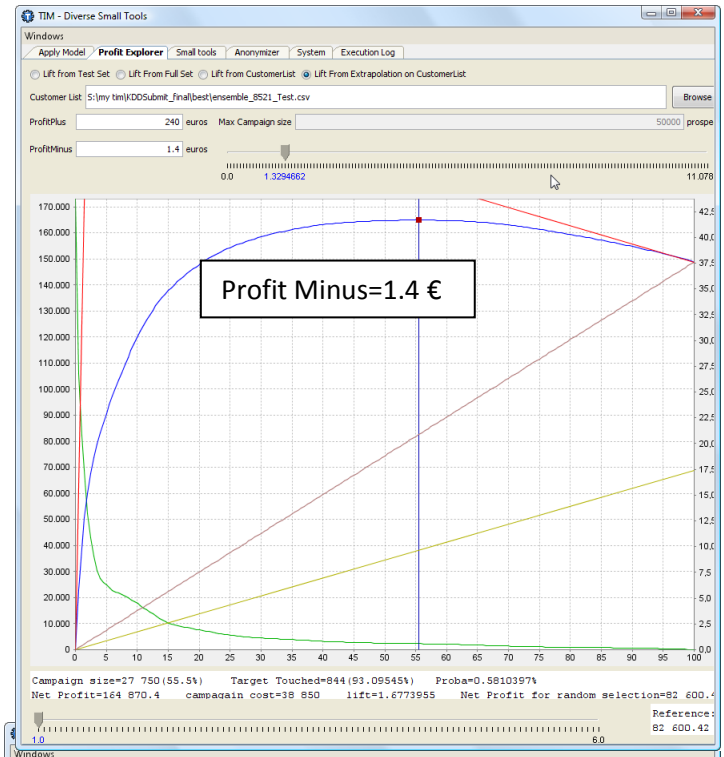
This value of 81.000€ is computed based on a total population of 50.000 individuals. In reality, the "Orange" company has a lot more "*mobile phone subscriptions*". Let's assume that the real number of "*mobile phone subscriptions*" is 5.000.000 (and not 50.000). This number is 100 times bigger that the population that has been analyzed up to now (in the chart above). Thus, in the real-world (when the real "scale" of the population is used), the marketing campaign will bring us 81.000 € x 100 = 8.100.000 €.

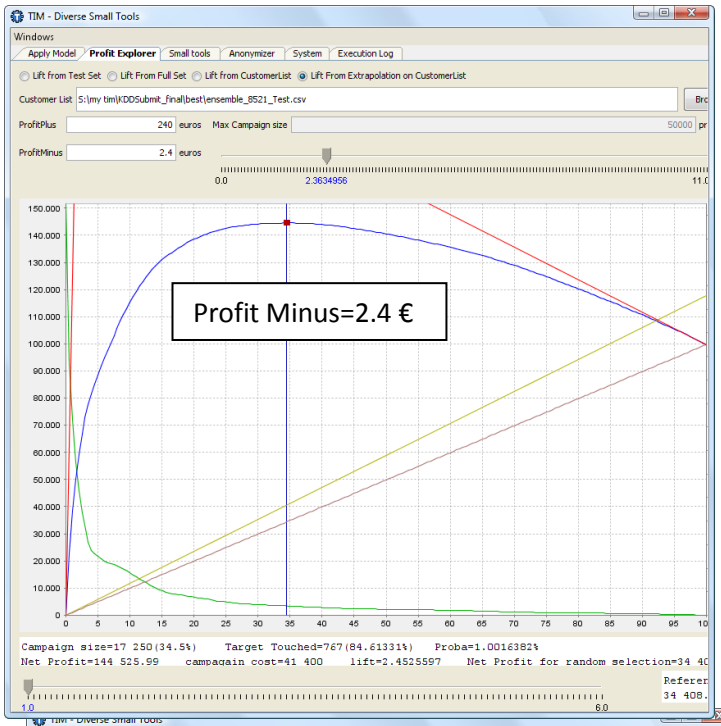
If you play a little with the “Profit Minus parameter”, you will notice that:

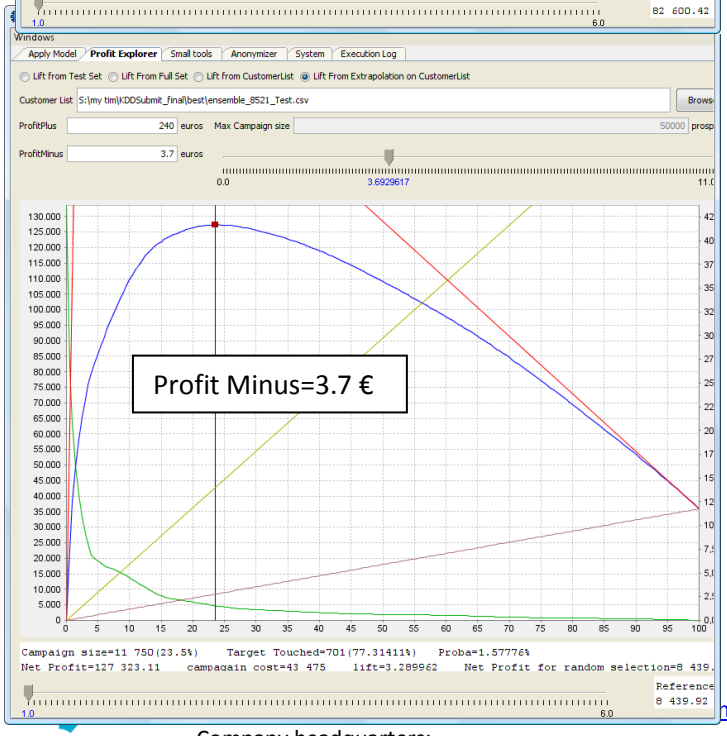
> *The higher the “profit minus” parameter, the smaller the size of the marketing campaign.*

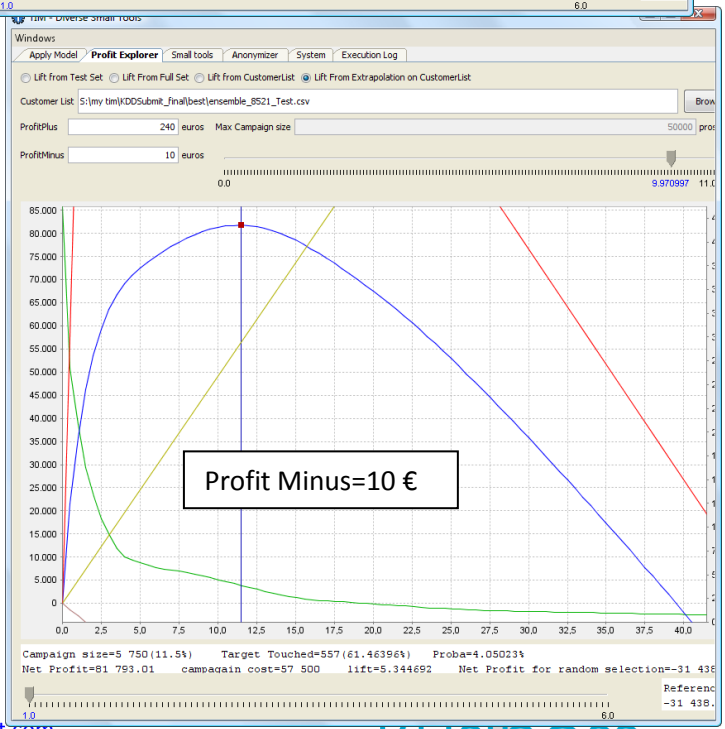
In other words, **if** it’s very expensive to “contact” people, **then** you must be very careful about who you are contacting: the size of the marketing campaign reduces more and more to “focus” only on the most “likely-to-buy” clients (at the start of the lift, on the left).

Here are some simulations, for different values of the “profit minus” parameter (the TIM software is producing these simulations in real-time when you move the slider-bar) (don’t forget to multiply by 100 to have an idea of the ROI of the real marketing campaign, as explained on the previous page):



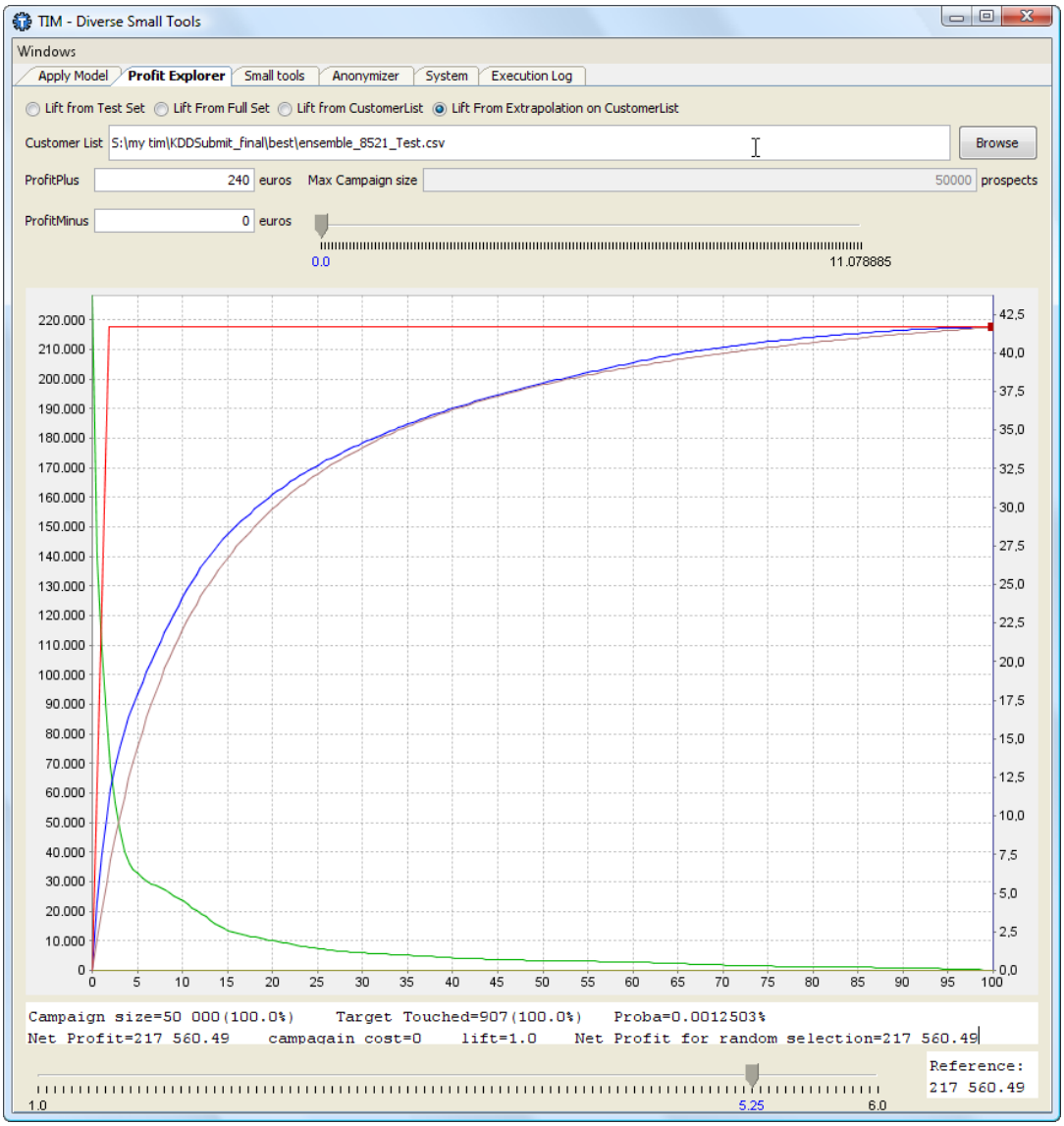






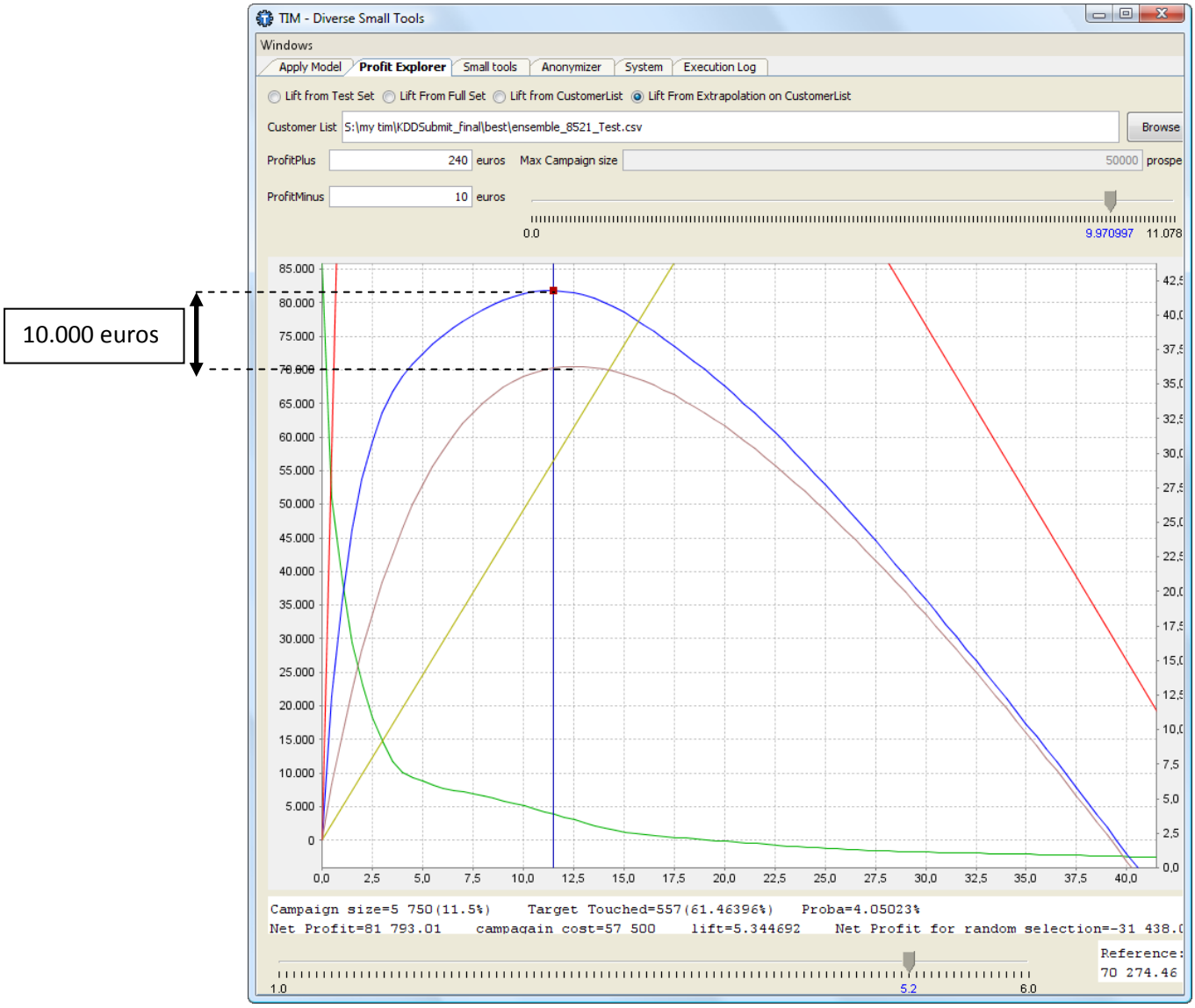
Let's now go back to the main question: "*What's the difference in ROI between a lift of 85% and a lift of 82%?*". We just illustrated the ROI of a predictive model with a lift of 85% (see the graphics above) how does this compare to the ROI of a predictive model with a lift of 82%?

Let's go back to the first graph. We will adjust the "pink reference line" so that the pink curve matches the performance of a predictive model with a lift of 82% (we play with the slider on the bottom of the window).



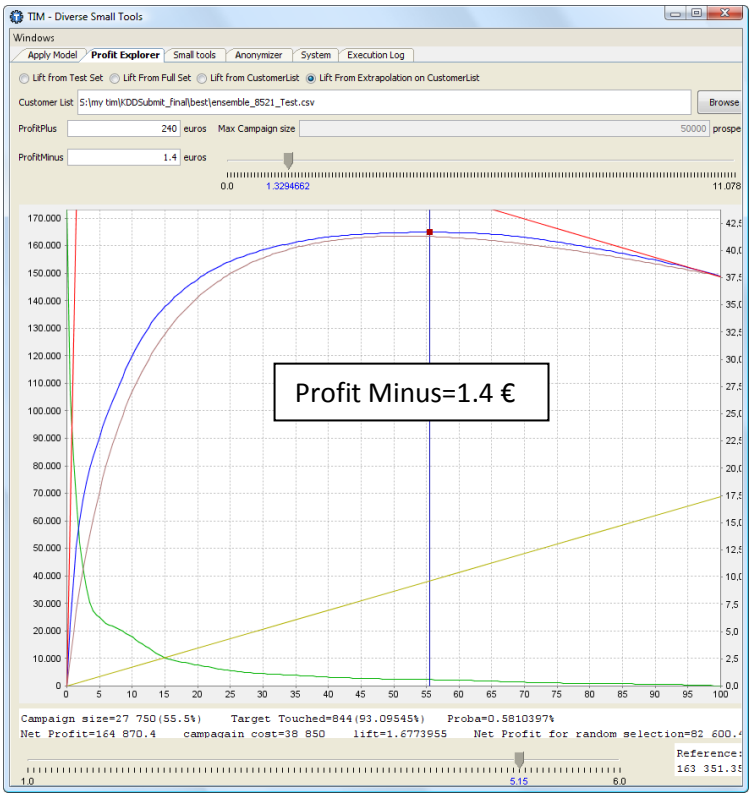
Please note that the pink curve (that represents the model with AUC=82%) and the blue curve (that represents the model with AUC=85%) are nearly the same (the difference between the two curves is only 3% AUC).

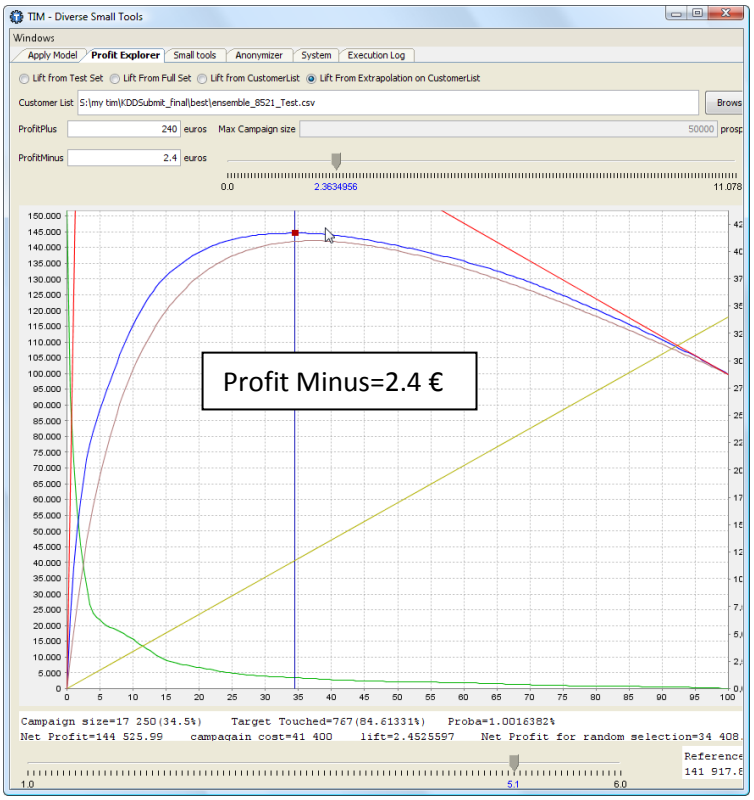
Let's enter, once again, the number "10" inside the "*profit minus*" parameter box. We obtain:

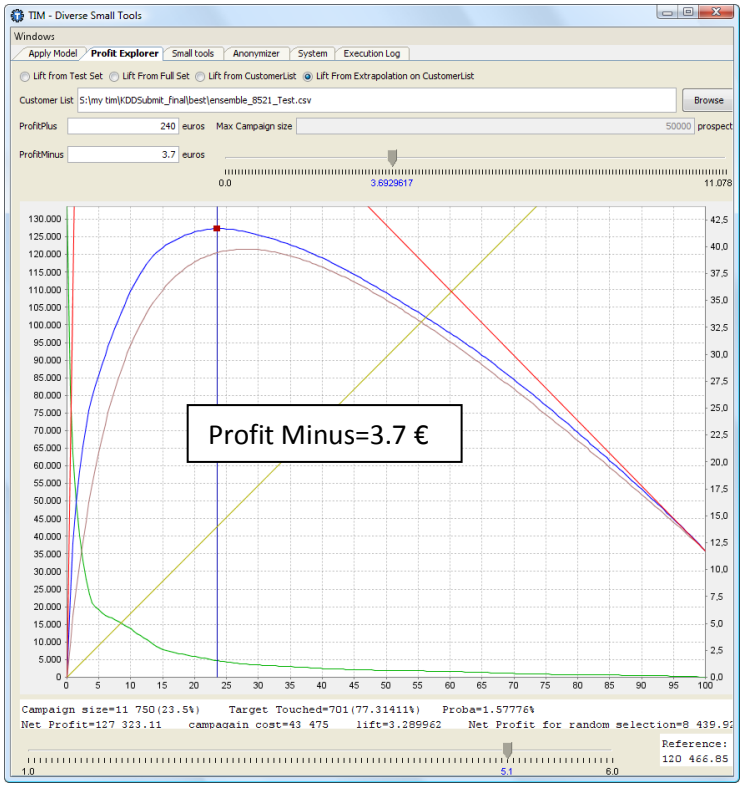


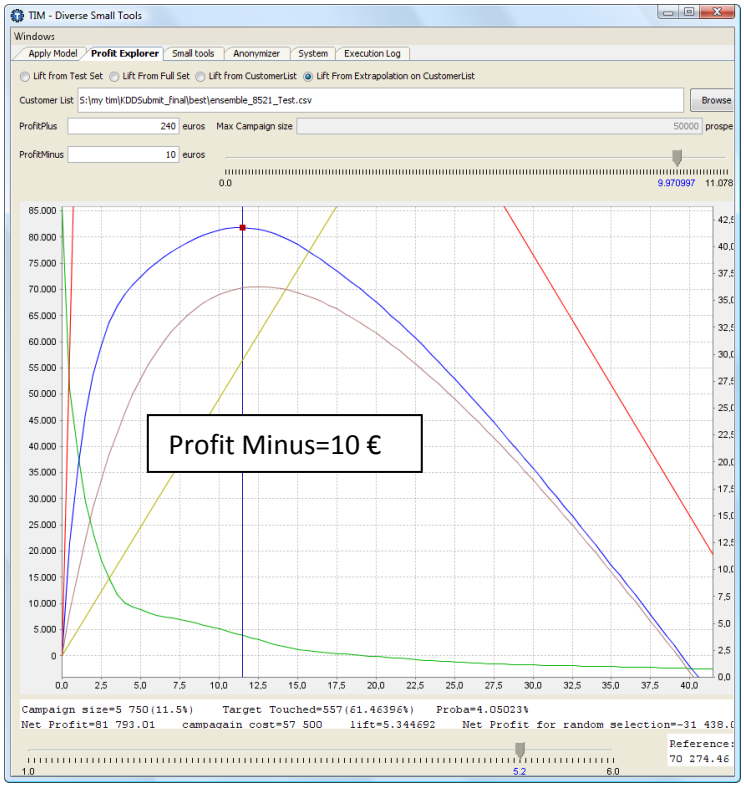
The difference in ROI between these 2 predictive models (the first model with AUC=85% and the second model with AUC=83%), for this set of parameters is 10.000 euros. We still have to multiply this number by 100 to have an idea of the ROI of the real-size marketing campaign, as explained on page 5. Thus, the difference of ROI between the 2 predictive models for the real-size marketing campaign is 1.000.000 euros.

Here are some simulations, for different values of the “profit minus” parameter (don’t forget to multiply by 100 to have an idea of the ROI of the real-size marketing campaign, as explained on the page 5):









Business-Insight SPRL E-mail: sales@business-insight.com
Company headquarters:
Address: Chemin des 2 Villers, 11 - 7812 Ath (V.N.D.) - Belgium
Phone (global): +32 479 99 27 68

A final word of caution: don't forget to multiply all the ROI's given inside this document by a security factor, as explained on page 1 (point 3).

This is a really crude analysis but it still gives an idea about the "order of magnitude" of the difference in ROI between predictive models that appears really close. A very small difference of a few percents in AUC can directly means hundred thousands of euros of difference in ROI.

NOTE: the analytical engine of TIM has been updated in September 2009 (4 months after the KDDCup 2009) and now delivers predictive models that are even more accurate. The accuracy improvement is usually around 2% AUC.

NOTE: SAS, KXEN, SPSS did participate to the competition because they all were "gold sponsors" but they did not reveal the accuracy of their models (they were too frightened to do that! ;-)).

NOTE: Rankings obtained using segmentation techniques have usually a lift between 20 and 30%. Don't use segmentation technique anymore: you are losing money when you use segmentation technique!