Kxen at CoIL Challenge 2000

Author: Michel Bera/Bertrand Lamy

Issue date: May 4, 2000

1. Modeling phase description

The model has been built done using KXEN product: KXEN Components. The free text describing the data has been converted in the product’s own description format. We then used KXEN Robust Regression (K2R)engine, which builds a polynomial model for the scoring task.

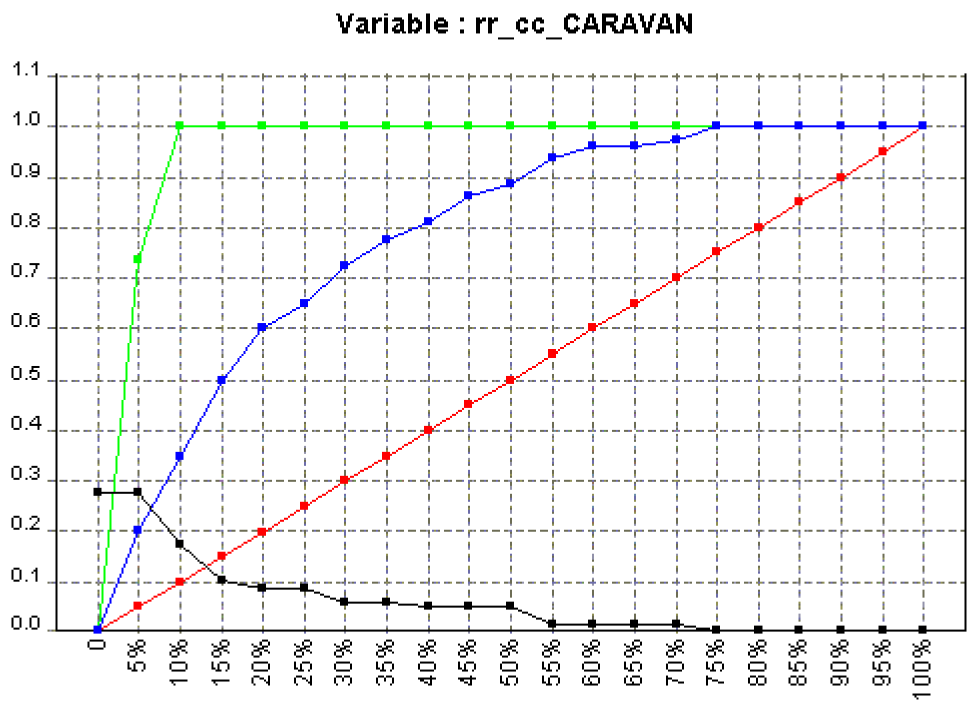
We used a built-in Estimation/Validation split strategy on the randomized data set.

The overall modeling process took about 50 seconds on a regular laptop. Applying the model to the evaluation data set took 6 seconds.

Our goal is to build as fast as possible a "robust" (or consistent) model, in other words, having good generalization expectations.

The computed lift curve (see next graphic) gives an overview of model performance.

Results show that at 20% of the population, the model can detect 60% of the relevant targeted population. So we can expect to select about 138 caravan insurance policies in the 800 customers selected list for evaluation (as there are about 5.5% insurance holders in the data, and the whole test contains about 4000 customers).



Lift curve for the model (blue), as compared to "perfect imaginary" model (green) and random selection (red). The black is the corresponding probability of having an insurance policy (for the lift curves : horizontal axis, percentage of population; vertical axis, percentage of target business obtained).

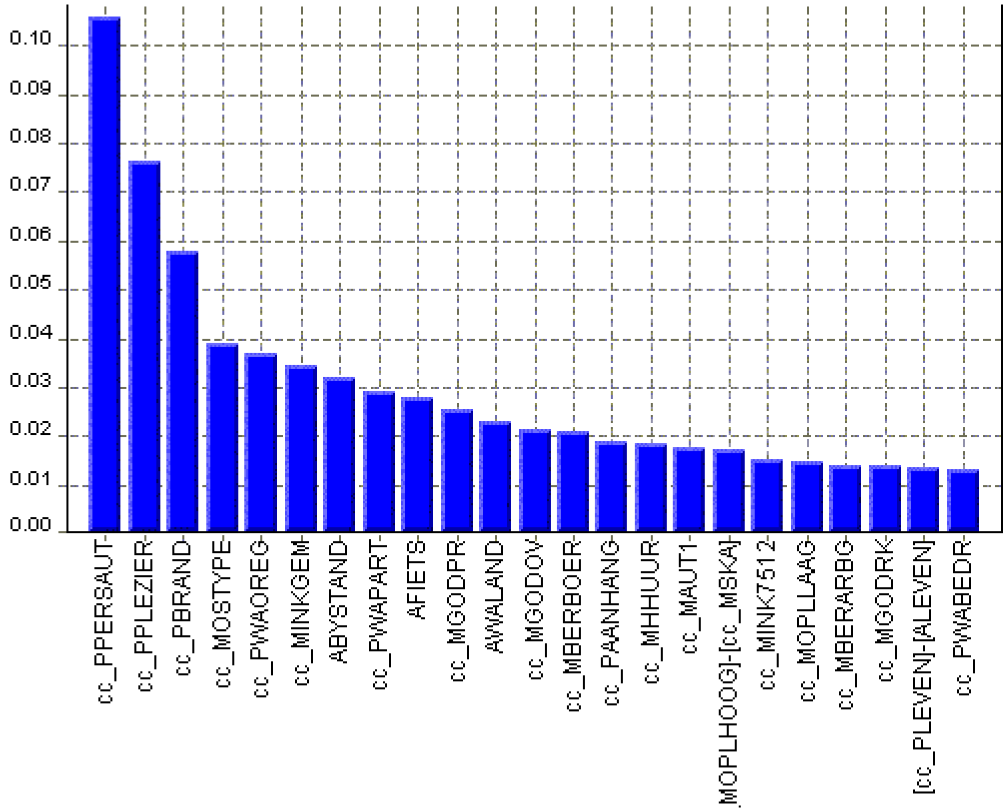
2. Description Task

The description task is prepared by the KXEN Components, as the model computes on the fly a number of indicators and charts, during the modeling phase. The following charts are direct screen shots of the software.

To describe how the model computes and builds its score function, we use the notion of "Variables Contributions". It expresses the relative contribution of each input variable in the resulting score, thus the propensity of having a caravan insurance policy.

Some contributions are only expressed for one variable alone(thus expressing the contribution in the model of this variable), while some contribution may be related to 2 or more variables. In this last case, the contribution expresses the fact that the important information is brought by the "additional knowledge" brought by the second variable when the first one is already known.

The contribution chart gives an overall picture of the important variables:



We can then have a look at the most important variables in depth, to see how the different categories influence.

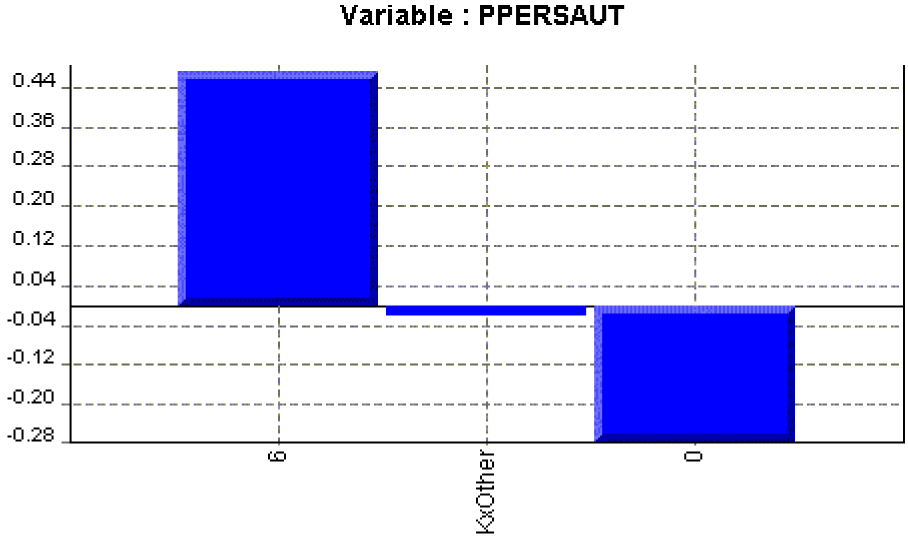
The following graph should be read with the following conventions;

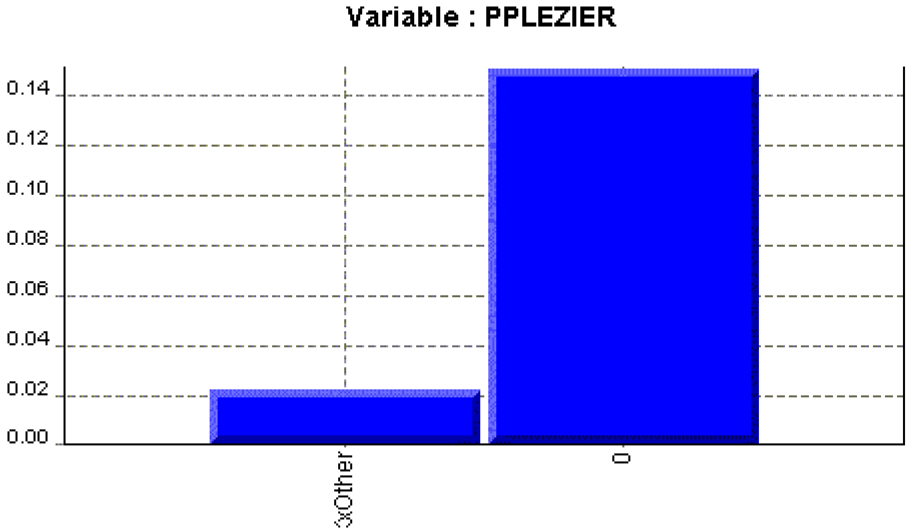
- Positive/negative bars reflects a positive/negative impact on having a caravan insurance policy.
- Categories are ordered from left to right in decreasing order of insurance policy propensity (categories on the left have more chance to have a policy than categories on the right).

- The height of the bar is the combination of the propensity of having a policy with the number of customer in this categories (thus a bar is high if it has a strong different propensity, or if it represents a large number of customers).
- In some variables, non-significant categories, or non-consistent categories (categories upon one should not rely for the problem), are grouped in a "trash" category named 'KxOther'.

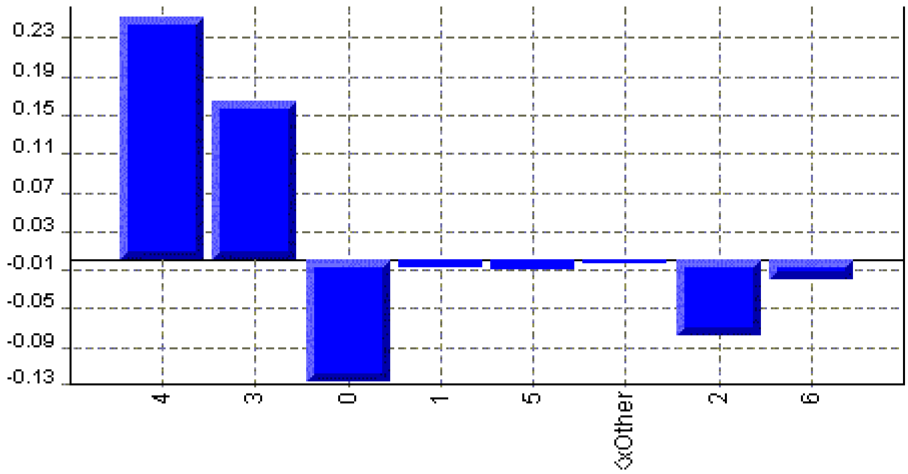
Categories importance clearly show you where you should focus to select possible policy owner.

Following the above described reading conventions, we can now browse the most important variables found by the model.

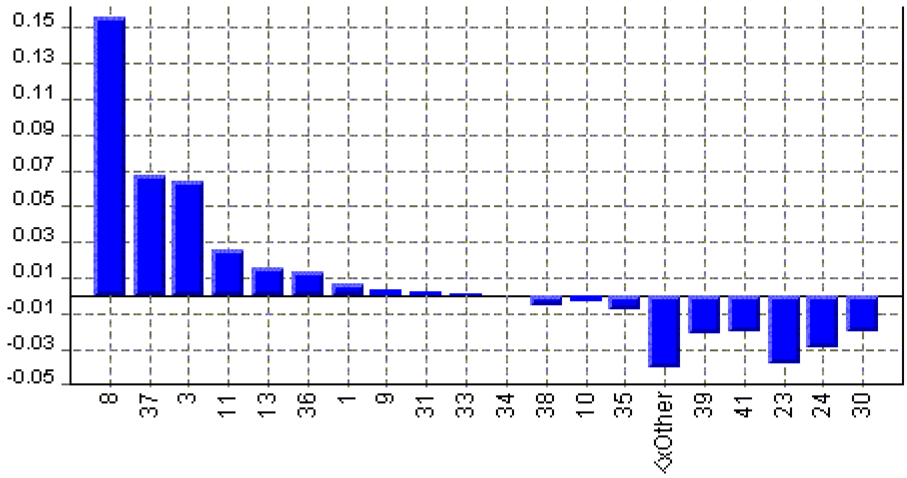




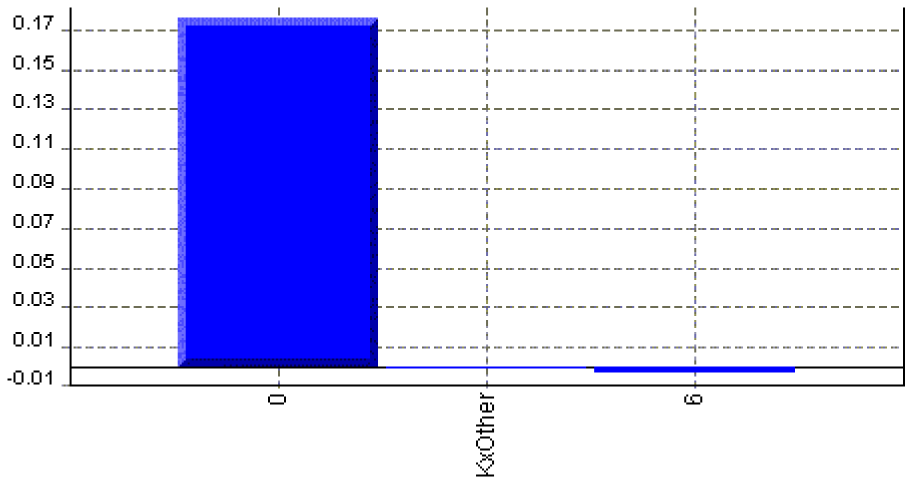
Variable : PBRAND

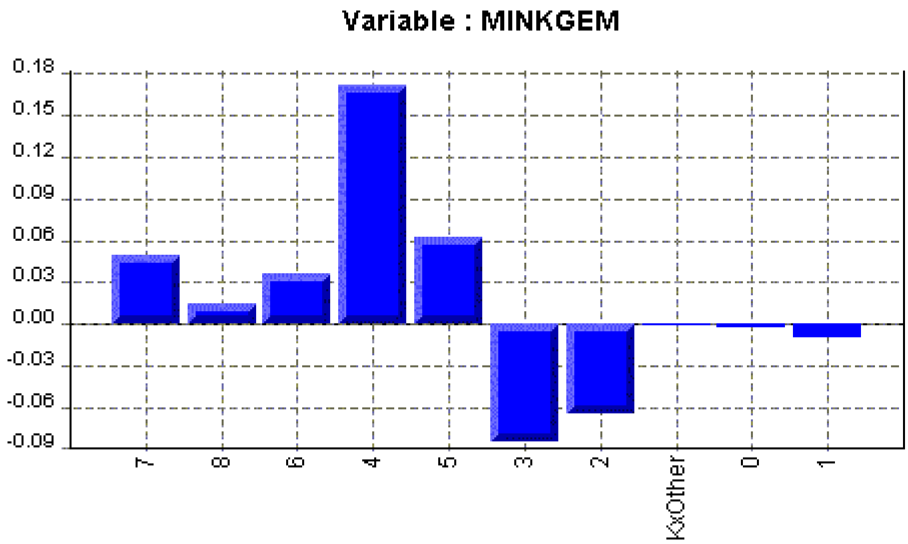


Variable : MOSTYPE



Variable : PWAOREG





3. Additional comments

Kxen engines are built on Vladimir Vapnik theory.

http://www.research.att.com/info/vlad

Our purpose at Kxen is not to be the best, but to be among the bests, in a very short time for the complete process (preprocessing and processing).

Our goal at Kxen is to build in a short period of time (here on the Coil2000 competition 50 seconds for modelization and 6 seconds for application) a "robust" (or consistent) model, in other words, having good generalization expectations.

However, in this example, we would have loved to have the raw data too, to compute our own preprocessing automatically through the KXEN Components.

Thanks a lot to all the Coil2000 organization for giving us the opportunity to work on a new difficult problem, and to share our passion for data with all our friends here!

Michel Bera and Bertrand Lamy