Contact Personalization using a Score Understanding Method

Vincent Lemaire, Raphael Féraud, Nicolas Voisine
Orange Labs, 2 avenue Pierre Marzin, 22307 Lannion Cedex - France
E-mail: vincent.lemaire@orange-ftgroup.com

***Abstract*— This paper presents a method to interpret the output of a classification (or regression) model. The interpretation is based on two concepts: the variable importance and the value importance of the variable. Unlike most of the state of art interpretation methods, our approach allows the interpretation of the model output for every instance. Understanding the score given by a model for one instance can for example lead to an immediate decision in a Customer Relational Management (CRM) system. Moreover the proposed method does not depend on a particular model and is therefore usable for any model or software used to produce the scores.**

I. Introduction

The most elaborate way, in a CRM system, to build knowledge on customer is to produce scores. Tools which produce scores allow to project, on a given population, quantifiable information. The score is an evaluation for all instances of a target variable to explain. The score (the output of a model) is computed using input variables which describe instances. Scores are then "injected" in the information system (IS), for example, to personalize the customer relationship.

Nevertheless, sometimes the scores are not directly usable. For example if a scoring model identifies a customer interested in churning, the score does not say anything on the action needed to avoid his cancellation. To prevent this intention to churn, the fragility of the customer and its causes have to be identified.

We propose to solve this problem by interpreting the classification produced by the model for every instance. To make possible the industrial implementation of this solution we propose a completely automatic method. The interpretation of the score is delivered for every instance to feed the information system. This knowledge could then be exploited to provide information personalized in the customer relationship management.

The proposed method is independent of the model used to build the scores. The most powerful model can be used without changing the difficulty of its interpretation. This interpretation method could thus remove one of the principal difficulty of the use of models like Support Vector Machines (SVM), Random Forest (RF) or artificial neural networks (ANN)in the marketing services.

II. Positioning and previous works

A. Variable importance

The field of machine learning abounds in techniques able to effectively solve problems of regression and/or classification. These techniques build a model from a training data base made up of a finite number of examples. The built model is used to associate an input vector to an output vector on a class label.

The large number of the models (linear regression, ANN, naive bayes, Random Forest (RF), Parzen window...) existing in the literature lead to a number of interpretation methods, generally specific to each model. The interpretation of the model is often based on: the parameters and the structure of the model [1], statistical tests on the coefficient's model [2], geometrical interpretations [3], rules [4] or fuzzy rules [5]. Resulting interpretations are often complex based on averages (for several individuals), for a given model (ANN, Decision Tree), or for a given task (regression OR classification).

Another approach consists in analysing the model as a black box with a sensibility analysis method. In these "What if?" analyses, the structure and the parameters of the model are only needed to compute the output of the model. This independence gives valid interpretation methods whatever the model.

To analyze in detail the state of the art approaches, notations which will be used below in this paper are introduced in table I.

V_j	:	an input variable j;
X	:	a vector of J dimension;
K	:	the number of training examples;
X_n	:	a example n;
X_{nj}	:	the component j of the vector X_n;
F	:	the predictive model;
p	:	the component p of the output vector;
$F^p(X)$	:	the output value of the component p of the output vector of the model;
and	:	$F_j^p(a;b) = F_j^P(a_1,\ldots,a_{j-1},b,a_{j+1},\ldots,a_J)$;

TABLE I
Notations

In this table $F_j^p(a;b)$ denotes the output p of the model when the component j, value a, is replaced by the value b. The proposed method analyses the outputs of the model one by one. Therefore the simplified notation F_j will be used (instead of F_j^p). All calculations presented in this paper are identical whatever the output p of the model.

Framling [6] introduces a variable importance measure, I, based on sensitivity analysis: $I(V_j|F,X_n,p = [F_j(X_n,\max(V_j)) - F_j(X_n,\min(V_j))]/[\max[F(X_n),\forall n] - \min[F(X_n),\forall n]]$; where $\max(V_j)$ and $\max(V_j)$ denotes

respectively the maximum and the minimum value of V_j.

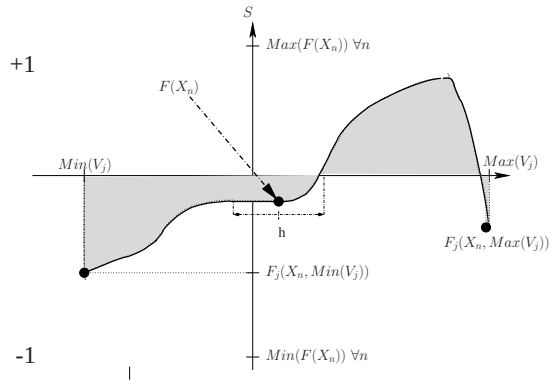

Fig. 1. What if simulation: Output values of the model vs. values of V_j.

This measurement is interesting but can be misleading when F is not monotonous (see Figure 1). In this illustrative example, the variable V_j is important for the model F: according to the values of this variable an example can be classified in class $+1$ or -1. However $F(X_n, \max(V_j))$ and $F(X_n, \min(V_j))$ are close, which leads to underestimate the importance of the variable V_j. Moreover, this method is based on extremums variable and thus very sensitive to noise.

Another approach is based on the variation of the model output for a variation h of the variable V_j and an example X_n (see Fig. 1). When h tends towards zero, this measurement corresponds to the partial derivative of the model compared to the variable V_j. In this case, measurement is local and can give an erroneous importance measurement: the partial derivative at the point $F(X_n)$ is null for this example whereas the variable V_j is important. When h is larger, as in the previous case, this measurement can be misleading when F is not monotonous. The problem is the same when these measurements are averaged on all examples.

Feraud et al. [7] proposes a method based on the integral of the variations of the outputs model. This measurement is well adapted to non monotonous functions. On the illustrative example (see Fig. 1), this measurement is related to the surface under the curve. As this surface is important, the variable V_j is important. The principal drawback of this method is that it does not take into account the distribution of the examples to define the interval of integration.

We propose a method of variable importance measurement based on the integral of the output variations of the model using the probability distributions of the examples. This measurement was tested successfully for classification problems in [8] and of regression in [9]. This method will be used in this paper as the "variable importance" definition.

B. Variable influence

For a given problem, a subset of relevant variables can be chosen using the variable importance measurement. This variable selection increases the model robustness and facilitates the model interpretation. However, the notion of variable importance, for an instance X_n, is not sufficient to interpret its classification.

One way to complete the interpretation is to analyse the importance of the value of the considered variable V_j on the output value of the model. In Figure 1 the example X_n belongs to the class -1. What indicates the value of the variable V_j for this example? Is it possible to change its class by modifying the V_j value? We propose to answer questions such as these ones using a measurement of the value of a given variable V_j for an example X_n. The importance of the value of a variable will be called its "influence".

To produce an interpretation of the model Féraud et al. [7] propose to segment examples and then characterize each cluster using the variables importance and influences inside every cluster. In this paper the objective is to propose a method which produces, automatically (without human assistance), an interpretation of the score for each example (instead for each cluster).

Therefore an "influence measurement" relative to every example will be proposed in the next section. Among existing methods the method proposed in [6] by Framling is the closest. But Framling uses extremums and an assumption of monotonous variations of the output model versus the variations of the input variable. The proposed "influence" measure is based on the distribution of the examples and is therefore more robust to outliers.

III. Method description

A. Importance of an input variable for an example

Considering[1] the model F, the example X_n, the input variable V_j and the variable to be explained p, the sensitivity of the model $S(V_j/F, X_n, p)$ is defined as the sum of the variations observed on the output p when perturbing the example X_n using the probability distribution of the input variable V_j.

The perturbed output of the model F, for an example X_n is the model output for this example but having replaced the value of the variable V_j with the value for an example k. The measured variation, for the example X_n, is then the difference between the "true output" $F_j(X_n)$ and the "perturbed output" $F_j(X_n, X_k)$ of the model.

The sensitivity of the model is then the mean value of $||F_j(X_n) - F_j(X_n, X_k)||^2$ for the probability distribution of the variable V_j. Approximating the variable probability distribution by the empirical distribution of the examples:

$$S(V_j|F, X_n, p) = \sum_{k=1}^{K} ||F_j(X_n) - F_j(X_n; X_k)||^2 \quad (1)$$

A sensitivity distribution is available by carrying out this sensitivity measurement on the output p and whatever is the input variable[2] V_j. The importance of the variable V_j to the

[1]Definitions I and I_v are presented here for one variable V_j, of the input vector of the model, and one output p, of the output vector. These definitions are the same whatever the considered variables j and p.

[2]The importance is not intrinsic to one input variable but to all variables. The distribution is established for all the input variables and using all the examples

example X_n, $I(V_j|F, X_n, p)$, is then defined as the rank o of the model sensitivity, $S(V_j|F, X_n, p)$, in the sensitivity distribution $S(V_j|F, X_i, p)\ \forall i, j$:

$$I(V_j|F, X_n, p) = \\ P[(S(V_j|F, X_i, p)\forall i, \forall j) \leq S(V_j|F, X_n, p)] \geq o \quad (2)$$

This measurement provides the variable importance of an input variable to an example relatively to all others examples and all others input variables. This relative measurement gives relevant information to every instance.

B. Influence on an example of an input variable value

An input variable can "pull up" (high value) or "pull down" (low value) the model output. For the example X_n the "natural" value of the output model p is by definition $F(X_n)$ (which can also be denoted by $F_j(X_n, X_n)$). The perturbed value considering the input variable V_j is $F_j(X_n, X_k)$.

The distribution of $F_j(X_n, X_k)$ represents the "potential" values for the example X_n if its variable V_j was different. The position of the natural value of X_n ($F(X_n)$) within this distribution gives information on the value of V_j (X_{nj}). The influence of the variable V_j on an example X_n is then defined, $I_v(V_j|F, X_n, p)$, as the rank r of the "natural" output model within the "potential values":

$$I_v(V_j|F, X_n, p) = P[(F_j(X_n, X_k)\forall k) \leq F(X_n)] \geq r. \quad (3)$$

For example, for a two classes classification problem (output -1 or $+1$), a high value of the rank of I_v shows a positive influence on the class $+1$ and a negative one on the class -1. Reciprocally a low value of the rank of I_v shows a positive influence on the class -1 and negative one on the class $+1$.

C. Automation of the interpretation: discussion

In business applications of CRM, scores identify customers most interested to react positively to a marketing campaign. For example, rather than to send a mail to all its customers to offer a product, a company will prefer to target the subset of its customers having the most "appetency" for the product. The marketing campaign will be less expensive, and the customers who are not interested by the product will have a lower probability to receive the publicity's product in their post-box (or mailbox).

The score interpretation brings additional information to improve the effectiveness of marketing campaigns. The score understanding provides means to support and personalize commercial action. For example if a customer is identified as fragile because he wishes to renew his mobile phone, the telecommunication company will be able to react by proposing a subscription with a reduction on the purchase price of a mobile phone. If the fragility of another customer corresponds to an under use of its "pay monthly plan", the company will be able to propose a better adapted plan.

In our system (see Figure 2), scores and score interpretations are evaluated in the deployment phase. Customer identifiers having the highest scores and the corresponding interpretation are send to the CRM system. This system uses the score understanding to personalize customer relationships.

The proposed method in this paper analyses the sensitivity of the model output p considering each input variable independently.

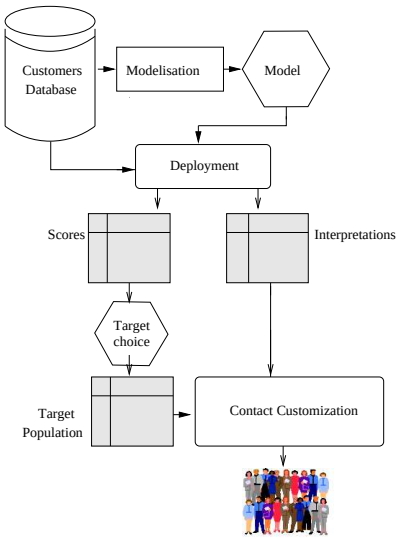

Fig. 2. Application architecture

The different steps needed to obtain the score understanding can require a long computation time. To speed up this computation two solutions are possible. The first solution extracts "an abstract" of each input variable using for example the method presented in [10] or centile information for continuous value and the method presented in [11] for categorical variables. The second one consists in memorising the $S(.)$ distribution.

IV. Illustration on a toy example

A. Toy example

A toy example has been constructed to test and observe the model interpretation method proposed in this paper. This toy example is presented in Figure 3. In this figure the class -1 is in black and the class $+1$ is in gray. The Figure 4 illustrates "a priori" influence zones of the two dimensions: (1) areas of points A and C: examples where both V_1 and V_2 influence the class, (2) area of point B: examples where only V_1 influences the class, (3) areas of point D and F: examples where only V_2 influences the class and (4) area of point E: examples where any dimension influences the class.

Data: 1000 examples for the training set and 1000 for the test set, were randomly drawn ($V_1 \in [0:2], V_2 \in [0:2]$).

Models - Two types of model were tested on this toy example: (1) a Neural Network [12] (NN) using one hidden layer, a sigmoid activation function, the standard back propagation algorithm (stochastic version) and the squared error for cost function. Using a cross validation procedure the number of hidden units has been fixed to 4; (2) a Parzen Window [13] (PW) using an Gaussian Kernel and the L2 norm

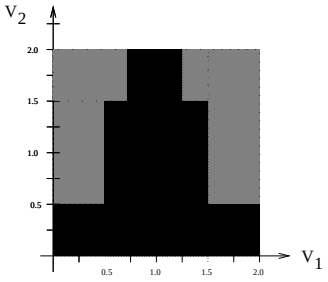

Fig. 3. Toy example: two classes

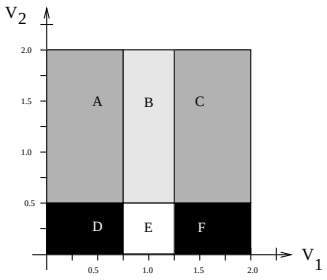

Fig. 4. Influence zones

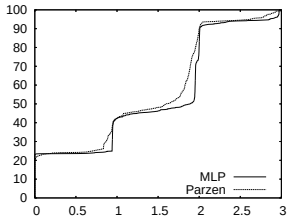

Fig. 5. Ordered sensibility distribution for V_1.

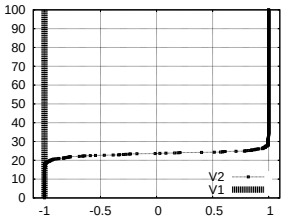

Fig. 6. Ordered "potential" output for the test point 'F' and V_1,V_2 using the MLP.

$(P(y_i|X_n) = \left(\sum_{n\ y=y_i} K(X_n, X_k)/\sum_n K(X_n, X_k)\right)$ where $K(X_n, X_k) = exp(||X_n - X_k||^2)/(2\sigma^2))$. The parameter σ was fixed to 0.1 using a cross validation procedure. Whatever the model the data were standardized before training.

B. Construction of the elements of the interpretation

Among the 1000 test examples, 6 representative examples of influence zones of variables V_1 and V_2 were selected to illustrate the method. Their location is indicated in the figure 4 and they are named from A to F: A(0.25,1.50), B(1.00,1.50), C(1.75,1.50), D(0.25,0.25), E(1.00,0.25), F(1.75,0.25).

The interpretation as of the these 6 examples requires the following steps (for $n \in \{A, B, C, D, E, F\}$):

- for $I(V_j/F, X_n, p)$:
  (1.1) calculation of $S(V_j/F, X_i, p)\ \forall j, \forall i$
  (1.2) sorting $S(.)$
  (1.3) determination of the rank of $S(V_j/F, X_n, p)$;
- for $I_v(V_j/F, X_n, p)$:
  (2.1) calculation of the $F(X_n, X_k)\ \forall k$;
  (2.2) sorting $F(.)$
  (2.3) determination of the rank of $F(X_n)$;

C. Results and discussion

The Figure 6 shows the sensibility distribution ($S(.)$, equation 1) obtained for V_1 using the NN and the PW on the training set. The x-coordinate represents a sensitivity value and the y-coordinate its corresponding rank in the distribution. The sensibility ranks progresses by stages (the result, not presented here, is the same for $V2$). Sensibility distributions are constituted of some important modalities relatively to the considered classification problem and the models used. These distributions concatenate the effect of individual sensibilities and influence zones: zones where the input variables have no interest, zones where they have high interest and transitory zones.

Figure 5 presents the distributions of "potential" output for the test point F and both the input variables V_1, V_2 using the NN. The obtained distribution using the input variable V_1 has an only one modality: $F(X_n, X_k) = -1.0\ \forall k$. This result is consistent since this variable has no influence for this example F. The obtained distribution using the input variable V_2 has 3 modes : $F(X_n, X_k) = -1$, $-1 \leq F(X_n, X_k) \leq +1$, $F(X_n, X_k) = +1$.

Figures 6 and 5 show that it could be interesting to use a rank range instead of a single rank. Quintiles, Q_1, Q_2, Q_3, Q_4 and Q_5, will be now used with the respective labels: "Very weak", "Weak", "Average", "Strong", "Very Strong". Each rank belongs to one of these quintiles (value of Q in the Table II) and has therefore the corresponding label. The joint observation of Table II), Figure II and Figure 5 shows a total coherence in the obtained results.

The influence of an input variable (I_v) has to be evaluated also in conjunction with the variable importance (I). If $I = 0$ the corresponding I_v is unimportant. Variables with a small I should not be used in the interpretation. In this case the interpretation has to be based only on the important variables (in these cases the value I_v is not presented in the Table II).

Interpretation using the MLP				
V_j, X_n	S	I	$F(X_n)$	I_v
V_1, X_A	1.24	Q_4 (o=63)	+1.00	Q_5 (r=99)
V_2, X_A	0.96	Q_3 (o=49)	+1.00	Q_5 (r=99)
V_1, X_B	2.70	Q_5 (o=89)	-1.00	Q_1 (r=14)
V_2, X_B	0.00	-	-	-
V_1, X_C	1.24	Q_4 (o=63)	+1.00	Q_5 (r=99)
V_2, X_C	0.93	Q_2 (o=31)	+1.00	Q_5 (r=99)
V_1, X_D	0.00	-	-	-
V_2, X_D	3.03	Q_5 (o=95)	-1.00	Q_2 (r=22)
V_1, X_E	0.00	-	-	-
V_2, X_E	0.00	-	-	-
V_1, X_F	0.00	-	-	-
V_2, X_F	3.05	Q_5 (o=98)	-1.00	Q_2 (r=21)
Interpretation using the Parzen window				
V_j, X_n	S	I	$F(X_n)$	I_v
V_1, X_A	1.16	Q_4 (o=63)	+0.99	Q_4 (r=74)
V_2, X_A	0.97	Q_3 (o=53)	+0.99	Q_4 (r=74)
V_1, X_B	2.28	Q_5 (o=89)	-0.99	Q_2 (r=25)
V_2, X_B	0.00	-	-	-
V_1, X_C	1.16	Q_4 (o=63)	+0.99	Q_4 (r=75)
V_2, X_C	0.90	Q_2 (o=35)	+0.99	Q_4 (r=67)
V_1, X_D	0.00	-	-	-
V_2, X_D	2.96	Q_5 (o=96)	-0.99	Q_1 (r=12)
V_1, X_E	0.00	-	-	-
V_2, X_E	0.00	-	-	-
V_1, X_F	0.00	-	-	-
V_2, X_F	3.02	Q_5 (o=90)	-0.99	Q_1 (r=12)

TABLE II
INTERPRETATION OF THE 6 TEST POINTS

D. Two examples of obtained interpretations

Two interpretations using Table II are presented here. The first interpretation is for the test point A using the Parzen

Window. The interpretation contains 3 elements: (1) the point belongs to the class $+1$ with a probability (the CRM score) of 0.99 (the value of $F(X_A)$) because:

* (2): V_1 which is very important indicates that it belongs strongly to the class $+1$
* (3): V_2 which is moderately important indicates that it belongs strongly to the class $+1$

The second interpretation is for the test point D [3] using the MLP. The interpretation contains 2 elements: (1) the point belongs to the class -1 with a probability (the CRM score) of 1.00 (the value of $F(X_D)$) because:

* (2): V_2 which is very important indicates that it belongs strongly to the class -1

The inspection of obtained interpretations, Table II, on all points of the figure 3 shows that interpretations are consistent whatever the tested model; thus is an important advantage of the proposed method. The interpretation method is also usable for other applications: the importance (I) and the influence (I_v) (of an input variable) being known, the class of an example (a customer in our application of this method) could be changed or reinforced.

V. Transposition to a real application

A. Introduction to the "Why" and "How" notions

The aim of the transposition detailed in this section is a proof of concept, intended for a Orange™ Business Unit, of the interpretation method presented in this paper. The purpose is to show that the interpretation method can be used in the context of CRM.

The way to improve customer's relationship is described in the following example. A campaign is designed to reduce customers' churn. The score (probability that a customer, X_n, churns) interpretation has to explain (i) "Why" the trained model indicates that the customer has this score and (ii) "How" it is possible to decrease this score.

The "Why" and "How" information are not useful for all customers. Marketers need this information only for customers on which the campaign will be applied. These customers are selected using their churn probability (high scores). These customers are named "the target".

Using the "Why" and "How" information, marketers will write a more personalized script to retain customers. The commercial script can be personalized for each customer relationship. In the discussion between the teleoperator and the customer is rarely possible to influence more than one aspect of this customer (one input variable of the classification model which produces scores). Therefore an only one variable will be kept in the Why and How interpretations as described in the next section.

[3]For the point D which belongs to the class -1, and reciprocally for the point A of the class $+1$, a low rank of I_v indicates a positive influence on the class -1 and negative one on the class $+1$, see section III-B

B. Implementation

The Why notion uses the definition of I presented in section III-A. This definition is used, here, only for the most important variable. This variable describes a "profile" on the customer X_n and we define a Why notion by:

$$Why(X_n|F,p) = \underset{V_j}{\operatorname{argmax}} \left[I(V_j|F,X_n,p) \right] \tag{4}$$

The computation time of $Why(X_n)$ is in $O(KJ)$. This computation can be simplified only if the V_{dj}, the number d of different values of the variable V_j are considered. In this case the computation time of $Why(X_n)$ is in $O\left(\left(\sum_{j=1}^{j} V_{dj}\right) J\right)$. Computation time can exceed a day (since more than one million of customers are concerned) and become useless in the CRM-Analytics loop (see Figure 2). To reduce this computation time, variables which have more than 100 different values are discretized using centiles. Therefore a variable has now a maximum of T modalities ($T \leq 100, \forall j$). The why notion uses then for $S(.)$ the computation:

$$S(V_j|F,X_n,p) = \sum_{t=1}^{T} ||F_j\left(X_n\right)) - F_j\left(X_n; V_{tj}\right)||^2 \; P(V_{tj}) \tag{5}$$

where $P(V_{tj})$ is the probability of V_{tj}.

The "How" interpretation looks for values of variables that positively change the score of a customer ("pull down" value for churn or vice versa "pull up" value for "appetency"). This interpretation is tied to I_v (see equation 3). Here for the Orange Business Unit application, the "How" is limited to the more positive variable, such as ($F_j(.,.) \in$ [0:1]):

$$How(X_n|F,p) = \underset{V_j}{\operatorname{argmin}} \left[\underset{t}{\operatorname{argmin}} \left[F_j(X_n, V_{tj}) \right] \right] \tag{6}$$

Here the problem is to prevent churn and to find the "worst" variable. Furthermore, variables that cannot be changed, such as sex, birthday or address, are not tested.

C. Experiments on Orange scores

Orange scores are calculated with the SAS™, Kxen™ or Khiops™ software (depending on the Business Unit and the country). Results presented here have been obtained using the Kxen software using a model close to a ridge regression. However the structure of the model is not used as detailed above in this paper.

For confidentiality reasons results of the "why" and "how" approaches on recent Orange scores are not presented. Only the "Why" information is illustrated on an older model of churn. This model is computed on a table of 100000 customers. The target is composed of 10 % of customers.

Input variables are defined as follow: indicators of telephone use; flags on the possession of service or product; indicators on customer (sex, senior (yes/no), ...); indicators of customer environment; indicators of customer purchasing behaviour; ...

Why	% of the Target	Usage	Product 1	Product 2	Service 1	Customer Indication	Customer Environment	Customer Behavior	...
Usage	58%	0.19	0.00	1.04	0.68	0.99	0.99	0.07	...
Product 1	17%	2.10	6.77	1.20	1.03	1.23	0.95	3.05	...
Product 2	15%	1.85	0.00	0.49	1.15	0.79	1.01	1.06	...
Product 3	6%	1.97	0.08	1.16	3.74	0.66	0.99	1.40	...
...	...	...	...	...	...	...	...	...	...

TABLE III
"WHY" RESULTS

Table III shows on the first column the name of the most important variable using the definition equation of 4. The second column indicates the percentage of customers for which this variable is the most important. From the third to the last, columns gives ratios. For example the cell at the intersection of the "Usage" column and the "Product 1" line gives the ratio between the mean value of the input variable "Usage" and the mean variable of customer for which the "Product 1" input variable is the most important (in the "Why" sense). This cell indicates customers who have a mean greater than the mean population.

Table III shows a main profile, which is pointed by the "Usage" variable, that contains 58 % of the "target population". The analysis of the first line of this table indicates (1) for the first column: customers with weak usage of some services (5 times smaller than the mean population); (2) for the second column: customers with no services or product of type "Product 1"; and so on. Therefore a possible marketing campaign can be build to push service usage or to suggest adequate services for their consumption. Others lines and cell of the table III can be analysed using the same process.

15 models have been tested (for this churn problem) with different numbers of input variables. All tests demonstrate that the approach is useful. The "Why" approach allows to detect profiles in high scores and to provide relevant interpretation. The "How" approach seeks the best value that will allow to reinforce (or change) a score.

D. Discussions

The Orange case shows the usefulness of the approach to detect high scores profiles. The profiles interpretation is easy since it contains only the most important variable which characterizes the profile itself.

However profile built using only the most important variable is not always the best choice. If all high scores have the same most important variable the second most sensitive variable has to be considered and so on. When the model has a lot of input variables the profile could be difficult to analyse. This is another obstacle for marketing use of the interpretation method.

VI. CONCLUSION

A method to interpret results of a predictive model has been presented. Experimental results on a toy problem using two different models and experimental results using another model (from a commercial software) were performed. Results show a very nice behavior of the method. At the moment this method is being industrialized in Orange CRM applications.

Even if the method was elaborated for black box models there are still ways to improve the approaches to speed up computing of sensitivity. The sensitivity analysis of specific model (i.e. logistic regression) could be accelerated by finding an analytic sensitivity function for the model. For example the method is exact for naive bayes model which is used in the Khiops software[4]. The proposed method will be added to the Khiops software next year. Future work concerns the extension of the method to obtain an instance selection method.

ACKNOWLEDGMENTS

Authors would like to thank Claude Riwan and the Score Team of Orange France for their contribution to the experimentation of the method presented in this paper.

REFERENCES

[1] Y. Arcadius, J. Akossou, and R. Palm, "Consequences of variable selection on the interpretation of the results in multiple linear regression," in *Biotechnol. Agron. Soc. Environ.*, vol. 9, 2005, pp. 11–18.

[2] J. Nakache and J. Confais, *Statistique explicative appliquée*. TECHNIP, 2003.

[3] J. J. Brennan and L. M. Seiford, "Linear programming and l1 regression: A geometric interpretation," *Computational Statistics & Data Analysis*, 1987.

[4] S. Thrun, "Extracting rules from artifcial neural networks with distributed representations," in *Advances in Neural Information Processing Systems*, M. Press, Ed., vol. 7. Cambridge, MA: G. Tesauro, D. Touretzky, T. Leen, 1995.

[5] J. M. Benitez, J. L. Castro, and I. Requena, "Are artificial neural networks black boxes," *IEEE Transactions on Neural Networks*, vol. 8, no. 5, pp. 1156–1164, 1997, septembre.

[6] K. Främling, "Explaining results of neural networks by contextual importance and utility," in *AISB*, 1996.

[7] R. Féraud and F. Clérot, "A methodology to explain neural network classification," *Neural Networks*, vol. 15, no. 2, pp. 237–246, 2002.

[8] V. Lemaire and C. Clérot, "An input variable importance definition based on empirical data probability and its use in variable selection," in *International Joint Conference on Neural Networks IJCNN*, 2004.

[9] V. Lemaire and R. Féraud, "Driven forward features selection: a comparative study on neural networks," in *International Conference on Neural Information Processing*, Hong-Kong, October 2006.

[10] M. Boulllé, "Khiops: a statistical discretization method of continuous attributes," *Machine Learning (ML)*, vol. 55, no. 1, pp. 53–69, 2004.

[11] M. Boullé, "A bayes optimal approach for partitioning the values of categorical attributes," *Journal of Machine Learning Research*, 2005.

[12] J. A. Anderson, *An introduction to neural network*. MIT Press, 1995.

[13] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Stat.*, pp. 1065–1076, 1962.

[4] `http://www.francetelecom.com/en/group/rd/offer/software/applications/providers/khiops.html`