Correlation Explorations in a Classification Model

Vincent Lemaire
Orange Labs
2 avenue Pierre Marzin
22300 Lannion - France
+33 2 96 05 31 07
vincent.lemaire@orange-ftgroup.com

Carine Hue
GFI Informatique
11 rue Louis Broglie
22300 Lannion - France
chue@gfi.com

Olivier Bernier
Orange Labs
2 avenue Pierre Marzin
22300 Lannion - France
olivier.bernier@orange-ftgroup.com

ABSTRACT
This paper presents a new method to analyze the link between the probabilities produced by a classification model and the variation of its input values. The goal is to increase the predictive probability of a given class by exploring the possible values of the input variables taken independently. The proposed method is presented in a general framework, and then detailed for naive Bayesian classifiers. We also demonstrate the importance of "lever variables", variables which can conceivably be acted upon to obtain specific results as represented by class probabilities, and consequently can be the target of specific policies. The application of the proposed method to several data sets (data proposed in the PAKDD 2007 challenge and in the KDD Cup 2009) shows that such an approach can lead to useful indicators.

Categories and Subject Descriptors
G3 PROBABILITY AND STATISTICS [**Correlation and regression analysis**]: -; I.5 PATTERN RECOGNITION [**Design Methodology**]: Classifier design and evaluation

General Terms
Algorithms, Measurement, Economics, Experimentation.

Keywords
Exploration, Correlation, Classifier.

1. INTRODUCTION
Given a database, one common task in data analysis is to find the relationships or correlations between a set of input or explanatory variables and one target variable. This knowledge extraction often goes through the building of a model which represents these relationships (Han & Kamber, 2006). Faced with a classification problem, a probabilist model allows, for all the instances of the database and given the values of the explanatory variables, the estimation of the probabilities of occurrence of each class target.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Conference'04, Month 1–2, 2004, City, State, Country.
Copyright 2004 ACM 1-58113-000-0/00/0004…$5.00.

These probabilities, or scores, can be used to evaluate existing policies and practices in organizations. They are not always directly usable, however, as they do not give any indication of what action can be decided upon to change this evaluation.

Consequently, it seems useful to propose a methodology which would, for every instance in the database, (i) identify the importance of the explanatory variables; (ii) identify the position of the values of these explanatory variables; and (iii) propose an action in order to change the probability of the desired class. We propose to deal with the third point by exploring the model relationship between each explanatory variable independently from each other and the target variable. The proposed method presented in this paper is completely automatic.

This article is organized as follows: the second section gives the context of the proposed method within Orange. This method is implemented using: (i) a platform for customer analysis, (ii) a tool, named Khiops, to construct classification models and (iii) a tool, named Kawab, to examine the contribution of the input variables and which (iv) allows the exploration of the correlation for these models.

The third section positions the approach in relation to the state of the art in feature importance (or selection), value importance and correlation analysis. The fourth section describes the method at first from a generic point of view and then for the naive Bayes classifier.

Through three illustrative examples, the fifth section allows a discussion and a progressive interpretation of the obtained results. The purpose of the first use case (titanic) is to illustrate the importance of the so-called "lever variables". The aim of the second use case on the PAKDD challenge 2007 database is to show that our method can suggest useful actions in this case actions to increase the appetency to the product concerned by the challenge. The third use case is on Orange data and shows that with the platform CAP, the software Khiops and the functionalities of the Add-on for Khiops presented in this paper we have all the elements for a success story on "Data Mining Case Studies". The last section concludes this paper and gives some future trends.

2. CORRELATIONS EXPLORATIONS AS AN ELEMENT OF A COMPLETE DATA MINING PROCESS

This section describes the platform named CAP (Customer Analysis Platform) used to classify customers within the Orange information system. This platform (see section 2.1) implements a complete datamining process [10]. A data mining process is constituted of six main steps: business understanding, data understanding, data preparation, modeling, evaluation (interpretation) and deployment. The CAP platform implements in particular two important steps: the data preparation step and the deployment step. This platform also uses the Khiops software (see section 2.2) for the modeling step and uses an "add-on" for Khiops named Kawab for the interpretation or evaluation step (see sections 2.3 and 2.4).

2.1 The CAP platform

A heavy trend since the end of the last century is the exponential increase of the volume stored data. This increase does not automatically translate into richer information because the capacity to process data does not increase as quickly. With the current state of the technology, a difficult compromise must be reached between the implementation cost and the quality of the produced information. An industrial approach has been proposed in [11]ı allowing to increase considerably the capacity to transform data into useful information thanks to the automation of treatments and the focus on relevant data.

2.2 The Khiops Software

Khiops is a data preparation and modeling tool for supervised and unsupervised learning. It exploits non parametric models to evaluate the correlation between any type of variables in the unsupervised case and the predictive importance of input variables or pairs of input variables in the supervised case. These evaluations are performed by means of discretization models in the numerical case and value grouping models in the categorical case, which correspond to the search for an efficient data representation owing to variable recoding. The tool also produces a scoring model for supervised learning tasks, according to a naive Bayes approach, with variable selection and model averaging. The tool is designed for the management of large datasets, with hundreds of thousands of instances and tens of thousands of variables, and was successfully evaluated in international data mining challenges. This tool is used by more than 60 users in Orange. Example of published applications see [24, 20]. Khiops can be downloaded here: http://www.khiops.com/.

2.3 Variable Contribution

We proposed in [20] 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. This method has been sufficiently successful to be adopted by Orange business units which use commercial data-mining software like SAS™, Kxen™ or SPSS™.

For Orange business unit which uses our in house software, Khiops, we have developed an “add-on” to compute contribution indicators especially for the naive Bayes classifier. This add-on implements five importance or contribution indicators for the naive Bayes classifier: two indicators that we proposed (Minimum of variable probabilities difference and Modality Probability) and three other indicators generally found in the state of the art [26]. All these indicators are based on the comparison between the probability of the reference class knowing the value of all the explanatory variables and the probability of the reference class knowing the value of all except on explanatory variable.

2.4 Correlation Exploration

The purpose of this paper is to describe a new method capable of analyzing the correlations in the constructed classification model to propose an action in order to change the customer response. This method is implemented as an add-on for Khiops named Kawab, but as described in this paper, can be used whatever the modeling software used during the datamining process. This method will be downloadable at www.khiops.com in June 2009.

3. BACKGROUND

Machine learning abounds with methods for supervised analysis in regression and/ or classification. Generally these methods propose algorithms to build a model from a training database made up of a finite number of examples. The output vector gives the predicted probability of the occurrence of each class label. In general, however, this probability of occurrence is not sufficient and an interpretation and analysis of the result in terms of correlations or relationships between input and output variables is needed. The interpretation of the model is often based on the parameters and the structure of the model. One can cite, for example: geometrical interpretations [6], interpretations based on rules [30] or fuzzy rules [1], statistical tests on the coefficient's model [23]. Such interpretations are often based on averages for several instances, for a given model, or for a given task (regression or classification).

Another approach, called sensitivity analysis, consists in analyzing the model as a black box by varying its input variables. In such "what if" simulations, the structure and the parameters of the model are important only as far as they allow accurate computations of dependant variables using explanatory variables. Such an approach works whatever the model. A large survey of "what if" methods, often used for artificial neural network, is available in [21, 19].

3.1 Variable importance

Whatever the method and the model, the goal is often to analyze the behavior of the model in the absence of one input variable, or a set of input variables, and to deduce the importance of the input variables, for all examples. The reader can find a large survey in [14]. The measure of the importance of the input variables allows the selection of a subset of relevant variables for a given problem. This selection increases the robustness of models and simplifies the understanding of the results delivered by the model. The variety of supervised learning methods, coming from the statistical or artificial intelligence communities often implies importance indicators specific to each model (linear regression, artificial neural network ...).

Another possibility is to try to study the importance of a variable for a given example and not in average for all the examples. Given a variable and an example, the purpose is to obtain the variable importance only for this example: for additive classifiers see [25], for Probabilistic RBF Classification Network see [27], and for a general methodology see [20]. If the model is restricted to a naive Bayes Classifier, a state of art is presented in [22, 26]. This importance gives a specific piece of information linked to one example instead of an aggregate piece of information for all examples.

3.2 Importance of the value of an input variable

To complete the importance of a variable, the analysis of the value of the considered variable, for a given example, is interesting. For example Féraud et al. [12] propose to cluster examples and then to characterize each cluster using the variables importance and importance of the values inside every cluster. Framling [13] uses a "what if" simulation to place the value of the variable and the associated output of the model among all the potential values of the model outputs. This method which uses extremums and an assumption of monotonous variations of the output model versus the variations of the input variable has been improved in [20].

3.3 Instance correlation between an explanatory variable and the target class

This paper proposes to complete the two aspects presented above, namely the importance of a variable and the importance of the value of a variable. We propose to study the correlation, for one instance and one variable, between the input and the output of the model.

For a given instance, the distinct values of a given input variable can pull up (higher value) or pull down (lower value) the model output. The proposed idea is to analyze the relationship between the values of an input variable and the probability of occurrence of a given target class. The goal is to increase (or decrease) the model output, the target class probability, by exploring the different values taken by the input variable. For instance for medical data one tries to decrease the probability of a disease; in case of cross-selling one tries to increase the appetency to a product; and in government data cases one tries to define a policy to reach specific goals in terms of specific indicators (for example decrease the unemployment rate).

This method does not explore causalities, only correlations, and can be viewed as a method between:

- selective sampling [28] or adaptive sampling [29]: the model observes a restricted part of the universe materialized by examples but can "ask" to explore the variation space of the descriptors one by one separately, to find interesting zones.

- and causality exploration [18, 15]: as example D. Choudat [7] propose the imputability approach to specify the probability of the professional origin of a disease. The causality probability is, for an individual, the probability that his disease arose from exposures to professional elements. The increase of the risk has to be computed versus the respective role of each possible type exposures. In medical applications, the models used are often additive models or multiplicative models.

3.4 Lever variables

In this paper we also advocate the definition of a subset of the explanatory variables, the "lever variables". These lever variables are defined as the explanatory variables for which it is conceivable to change their value. In most cases, changing the values of some explanatory variables (such a sex, age...) is indeed impossible. The exploration of instance correlation between the target class and the explanatory variables can be limited in practice to variables which can effectively be changed.

The definition of these lever variables will allow a faster exploration by reducing the number of variable to explore, and will give more intelligible and relevant results. Lever variables are the natural target for policies and actions designed to induce changes of occurrence of the desired class in the real world.

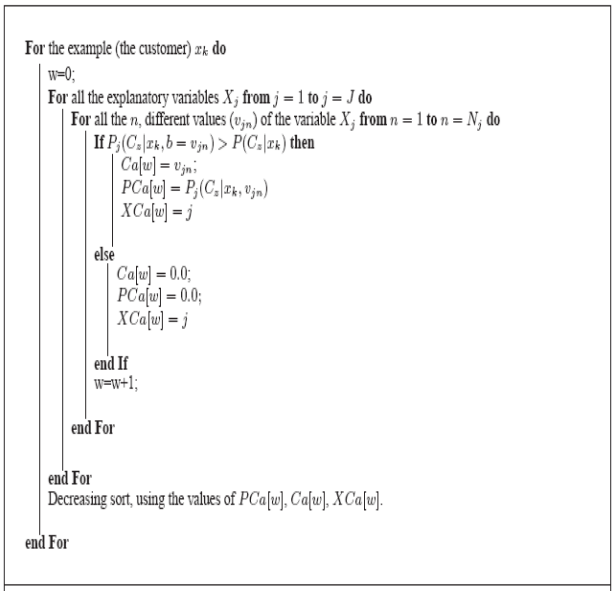
4. CORRELATION EXPLORATION – METHOD DESCRIPTION

In this section, the proposed method is first described in the general case, for any type of predictive model, and then tested on naive Bayes classifiers.

4.1 General case

Let C_z be the target class among T target classes. Let f_z be the function which models the predicted probability of the target class $f_z(X=x) = P(C_z \mid X=x)$, given the equality of the vector X of the J explanatory variables to a given vector x of J values. Let v_{jn} be all the n different possible values of the variable X_j.

The Algorithm 1 describes the proposed method. This algorithm tries to increase the value of $P(C_z \mid X = x_k)$ successively for each of the K examples of the considered sample set using the set of values of all the explanatory variables or lever variables. This method is halfway between selective sampling [28] and adaptive sampling [29]. The model observes a restricted part of the universe materialized by examples but can "ask" to explore the variation space of the descriptors one by one separately, to find interesting zones. The next subsections describe the algorithm in more details.



Algorithm 1: Exploration and ranking of the score improvements

4.1.1 Exploration of input values

For the instance x_k, $P(C_z \mid x_k)$ is the "natural" value of the model output. We propose to modify the values of the explanatory variables or lever variables in order to study the variation of the model output for this example. In practice, we propose to explore the values independently for each explanatory variable. Let $P_j(C_z \mid x_k, b)$ be the output model f_z given the example x_k but for which the value of its j^{th} component has been replaced with the value b. For example, the third explanatory variable is modified among five variables: $P_3(C_z \mid x_k, b) = f_z(x_k^1, x_k^2, b, x_k^4, x_k^5)$. By scanning all the variables and for each of them all the set of their possible values, an exploration of "potential" values of the model output is computed for the example x_k.

4.1.2 Domain of exploration

The advantage of choosing the empirical probability distribution of the data as domain of exploration has been showed experimentally in [5, 19, 20]. A theoretical proof is also available for linear regression in [8] and for naive Bayes classifiers in [26]. Consequently the values used for the J explanatory variables will be the values of the K examples available in the training database. This set can also be reduced using only the distinct values: let N_j be the number of distinct values of the variable X_j.

4.1.3 Results ranking

The exploration of the explanatory variables or of the lever variables is done by scanning all the possible values taken by the examples in the training set. When the modification of the value of the variable leads to an improvement of the probability predicted by the model, three pieces of data are kept (i) the value which leads to this improvement (Ca); (ii) the associated improved probability (PCa); and (iii) the variable associated to this improvement (XCa). These triplets are then sorted according to the improvement obtained on the predicted probability. Note: if no improvement is found, the tables CA and PCa only contain null values.

It should also be possible (i) to explore jointly two or more explanatory variables; (ii) or to use the value (Ca[0]) which best improves the output of the model ($P(C_z \mid X = x)$) (this value Ca[0] is available at the end of the Algorithm) and then to repeat again the exploration on the example x_k on its others explanatory variables. These other versions are not presented in this paper but will be the focus of future works.

Algorithm 1: Exploration and ranking of the score improvements

4.1.4 Cases with class changes

When using Algorithm 1, the predicted class can change. Indeed it is customary to use the following formulation to designate the predicted class of the example x_k:

$$\arg\max_{z} P(C_z \mid x_k)$$

Using Algorithm 1 for x_k belonging to the class t ($t \neq z$) could produce $P(C_z \mid x_k, b) > P(C_t \mid x_k)$. In this case the corresponding value (Ca) carries important information which can be exploited.

The use of Algorithm 1 can exhibit three types of values (Ca):

- values which do not increase the target class probability;
- values which increase the target class probability but without class change (the probability increase is not sufficient);
- values which increase the target class probability with class change (the probability increase is sufficient).

The examples whose predicted class changes from another class to the target class are the primary target for specific actions or policies designed to increase the occurrence of this class in the real world.

4.2 Case of a naive Bayesian classifier

A naive Bayes classifier assumes that all the explanatory variables are independent knowing the target class. This assumption drastically reduces the necessary computations. Using the Bayes theorem, the expression of the obtained estimator for the conditional probability of a class C_z is:

$$P(C_z \Big| x_k) = \frac{P(C_z)\prod_{j=1}^{J} P(X_j = v_{jk} \big| C_z)}{\sum_{t=1}^{T} P(C_t)\prod_{j=1}^{J} P(X_j = v_{jk} \big| C_t)} \quad (1)$$

The predicted class is the one which maximizes the conditional probabilities. Despite the independence assumption, this kind of classifier generally shows satisfactory results [17]. Moreover, its formulation allows an exploration of the values of the variables one by one independently.

The probabilities $P(X_j = v_{jk} \mid C_z)$ ($\forall$ *j, k, z*) are estimated using counts after discretization for numerical variables or grouping for categorical variables [3]. The denominator of the equation above normalizes the result so that $\sum_z P(C_z \mid x_k)=1$.

The use of the Algorithm 1 requires to compute $P(C_z \mid X = x_k)$, and $P_j(C_z \mid X = x, b)$ which can be written in the form of Equations 2 and 3:

$$P(C_z \Big| x_k) = \frac{\overbrace{P(C_z)\prod_{j=1}^{J} P(X_j = v_{jk} \big| C_z)}^{e^{L_z}}}{\sum_{t=1}^{T} P(C_t)\prod_{j=1}^{J} P(X_j = v_{jk} \big| C_t)} \quad (2)$$

$$P_j(C_z \Big| x_k, b) = \frac{\overbrace{P(C_z)\prod_{j=1, j\neq q}^{J} P(X_j = v_{jk} \big| C_z) P(X_q = b \big| C_z)}^{e^{L_{z'}}}}{\sum_{t=1}^{T} \left[P(C_t)\prod_{j=1}^{J} P(X_j = v_{jk} \big| C_t)\right] P(X_q = b \big| C_t)} \quad (3)$$

In Equations 2 and 3 numerators can be written as e^{Lz} and $e^{Lz'}$ with:

$$L_z = \log(P(C_z)) + \sum_{j=1}^{J} \log(P(X_j = v_{jk} \mid C_z))$$

and

$$L_{z'} = \log(P(C_z)) + \sum_{j=1, j\neq q}^{J} \left[\log(P(X_j = v_{jk} \mid C_{z'})) + \log(P(X_q = b \mid C_{z'}))\right]$$

This formulation will be used below.

4.2.1 Implementation details on very large databases

To measure the reliability of our approach, we tested it on marketing campaigns of France Telecom (results not allowed for publication until now). Tests have been performed using the PAC platform [11] on different databases coming from decision-making applications. The databases used for testing had more than 1 million of customers, each one represented by a vector including several thousands of explanatory variables. These tests raise several implementation points enumerated below:

- To avoid numerical problems when comparing the "true" output model $P(C_z \mid x_k)$ and the "explored" output $P_j(C_z \mid x_k, b)$, $P(C_x \mid x_k)$ is computed as:

$$P(C_x \Big| x_k) = \frac{1}{\sum_{t=1}^{T} e^{L_t - L_x}}$$

  where

$$L_t = \log(P(C_t)) + \sum_{j=1}^{J} \log(P(X_j = v_{jk} \big| C_t))$$

- To reduce the computation time: the modified output of the classifier can be computed using only several additions or subtractions since the difference between L_z (used in Equation 2) and $L_{z'}$ (used in Equation 3) is:

  $L_{z'} = L_z - \log(P(x_q=v_{jk} \mid C_z)) + \log(P(X_q=b \mid C_z))$

- Complexity: For a given example x_k, the computation of tables presented in Algorithm 1 is of complexity $O(\sum_{j=1}^{d} N_j)$.

This implementation is "real-time" and can be used by an operator who asks the application what actions to do, for example to keep a customer.

5. EXPERIMENTATIONS

In this section we describe the application of our proposed method to three illustrative examples. This first example, the Titanic database, illustrates the importance of lever variables. The second example illustrates the results of our method on the dataset used for the PAKDD 2007 challenge. Finally, we present the results obtained by our method on the KDD Cup 2009.

5.1 The Titanic database

5.1.1 Data and experimental conditions

In this first experiment the Titanic (www.ics.uci.edu/~mlearn/) database is used. This database consists of four explanatory variables on 2201 instances (passengers and crew members). The first attribute represents the class trip (status) of the passenger or if he was a crew member, with values: 1st, 2nd, 3rd, crew. The second (age) gives an age indication: adult, child. The third (sex) indicates the sex of the passenger or crew: female or male. The last attribute (survived) is the target class attribute with values: no or yes. Readers can find for each instance the variable importance and the value importance for a naive Bayes classifier in [26].

Among the 2201 examples in this database, a training set of 1100 examples randomly chosen has been extracted to train a naive Bayes classifier using the method presented in [3]. The remaining examples constitute a test set. As the interpretation of a model with low performance would not be consistent, a prerequisite is to check if this naive Bayes classifier is correct. The model used here [16] gives satisfactory results:

- Accuracy on Classification (ACC) on the train set: 77.0%; on the test set: 75.0%;
- Area under the ROC curve (AUC) (Fawcett, 2003) on the train set: 73.0%; on the test set: 72.0%.

The purpose here is to the see another side of the knowledge produced by the classifier: we want to find the characteristics of the instances (people) which would have allowed them to survive.

5.1.2 Input values exploration

Algorithm 1 has been applied on the test set to reinforce the probability to survive. Table 1 shows an abstract of the results: (i) it is not possible to increase the probability for only one passenger or crew; (ii) the last column indicates that, for persons predicted as surviving by the model (343 people), the first explanatory variable (status) is the most important to reinforce the probability to survive for 118 cases; then the second explanatory variable (age) for 125 cases; and at last the third one (sex) for 100 cases. (iii) For people predicted as dead by the model (758) the third explanatory variable (sex) is always the variable which is the most important to reinforce the probability to survive.

Table 1: Ranking of explanatory variables

	Size	Status / Age / Sex
Predicted 'yes'	343	118 / 125 / 100
Predicted 'no'	758	0 / 0 / 758

These 758 cases predicted as dead are men and if they were women their probability to survive would increase sufficiently to survive (in the sense that their probability to survive would be greater than their probability to die). Let us examine then, for these cases, additional results obtained by exploring the others variables using Algorithm 1:

- the second best variable to reinforce the probability to survive is (and in this case they survive):
  - for 82 of them (adult + men + 2nd class) the second explanatory variable (age);
  - for 676 of them (adult + men + (crew or 3rd class)) the first explanatory variable (status);

- the third best variable to reinforce the probability to survive is (and in this case nevertheless they are dead):
  - for 82 of them (adult + men + 2nd class) the first explanatory variable (status);
  - for 676 of them (adult + men + (crew or 3rd class)) the second explanatory variable (age).

Of course, in this case, most explanatory variables are not in fact lever variables, as they cannot be changed (age or sex). The only variable that can be changed is status, and even in this case, only for passengers, not for crew members. The change of status for passengers means in fact buying a first class ticket, which would have allowed them a better chance to survive. The other explanatory variables enable us to interpret the obtained survival probability in terms of priority given to women and first class passengers during the evacuation.

5.2 Application to sale: results on the PAKDD 2007 challenge

5.2.1 Data and experimental conditions

The data of the PAKDD 2007 challenge are used (http://lamda.nju.edu.cn/conf/pakdd07/dmc07/): The data are not on-line any more but data descriptions and analysis results are still available. Thanks to Mingjun Wei (participant referenced P049) for the data (version 3).

The company, which gave the database, has currently a customer base of credit card customers as well as a customer base of home loan (mortgage) customers. Both of these products have been on the market for many years, although for some reasons the overlap between these two customer bases is currently very small. The company would like to make use of this opportunity to cross-sell home loans to its credit card customers, but the small size of the overlap presents a challenge when trying to develop an effective scoring model to predict potential cross-sell take-ups.

A modeling dataset of 40,700 customers with 40 explanatory variables, plus a target variable, had been provided to the participants (the list of the 40 explanatory variables is available at http://perso.rd.francetelecom.fr/lemaire/data_pakdd.zip). This is a sample of customers who opened a new credit card with the company within a specific 2-year period and who did not have an existing home loan with the company. The target categorical variable "Target_Flag" has a value of 1 if the customer then opened a home loan with the company within 12 months after

opening the credit card (700 random samples), and has a value of 0 otherwise (40,000 random samples).

A prediction dataset (8,000 sampled cases) has also been provided to the participants with similar variables but withholding the target variable. The data mining task is to produce a score for each customer in the prediction dataset, indicating a credit card customer's propensity to take up a home loan with the company (the higher the score, the higher the propensity).

The challenge being ended it was not possible to evaluate our classifier on the prediction dataset (the submission site is closed). Therefore we decide to elaborate a model using the 40 000 samples in a 5-fold cross validation process. In this case each 'test' fold contains approximately the same number of samples as the initial prediction dataset. The model used is again a naive Bayes classifier (Boullé, 2008; Guyon, Saffari, et al., 2007). The results obtained on the test sets are:

- Accuracy on Classification (ACC): 98.29% ± 0.01% on the train sets and 98.20% ± 0.06% on the test sets.
- Area under the ROC curve (AUC): 67.98% ± 0.74% on the train sets and 67.79% ± 2.18% on the test sets.
- Best results obtained on one of the folds: Train set AUC=68.82%, Test set AUC=70.11%.

Table 2: PAKDD 2007 challenge: the first three best results

id participant	AUC for test set	Rank	Modeling Technique
P049	70.01%	1	TreeNet + Logistic Regression
P085	69.99%	2	Probit Regression
P212	69.62%	3	MLP + n-Tuple Classifier

Table 2 shows the first three best results and corresponding method of winners of the challenge. Results obtained here by our model are coherent with those of the participants of the challenge.

5.2.2 Input values exploration

The best classifier obtained on the test sets in the previous section is used. This naive Bayes classifier (Boullé, 2007) uses 8 variables out of 40 (the naïve Bayes classifier takes into account only input variables which have been discretized (or grouped) in more than one interval (or group) see (Boullé, 2006)). These 8 variables and their intervals of discretization (or groups) are presented in Table 3. All variable are numerical except for the variable "RENT_BUY_CODE" which is symbolic with possible values of 'O' (Owner), 'P' (Parents), 'M' (Mortgage), 'R' (Rent), 'B' (Board), 'X' (Other).

The lever variables were chosen using their specification (see http://lamda.nju.edu.cn/conf/pakdd07/dmc07/ or the appendix A). These lever variables are those for which a commercial offer to a customer can change the value. We define another type of variable which we will explore using our algorithm, the observable variables. These variables are susceptible to change during a life of a customer and this change may augment the probability of the target class, the propensity to take up a home loan. In this case, the customers for which this variable has changed can be the target of a specific campaign. For example the variable "RENT_BUY_CODE" can not be changed by any offer but is still observable. The customer can move from the group of values [O,P] ('O' Owner, 'P' Parents) to [M,R,B,X] ('M' Mortgage, 'R' Rent, 'B' Board, 'X' Other). Among the eight variables (see Table 3) chosen by the training method of the naive Bayes classifier, two are not considered as 'lever' variables or observable variables ("AGE_AT_APPLICATION" and "PREV_RES_MTHS") and will not be explored.

Table 3: Selected explanatory variables (there is no reason in [2] to have two intervals for each variable, it is here blind chance).

Explanatory Variables	Interval 1 or Group 1	Interval 2 or Group 2
RENT_BUY_CODE	M,R,B,X	O,P
PREV_RES_MTHS	]-∞,3.5[	[3.5,+∞ [
CURR_RES_MTHS	]-∞,40.5[	[40.5,+∞ [
B_ENQ_L6M_GR3	]-∞,0.5[	[0.5,+∞ [
B_ENQ_L3M	]-∞,0.5[	[3.5,+∞ [
B_ENQ_L12M_GR3	]-∞,1.5[	[1.5,+∞ [
B_ENQ_L12M_GR2	]-∞,0.5[	[0.5,+∞ [
AGE_AT_APPLICATION	]-∞,45.5[	[45.5,+∞ [

Algorithm 1 has been applied on the 40700 instances in the modeling data set. The 'yes' class of the target variable is chosen as target class (C_z = 'yes'). This class is very weakly represented (700 positive instances out of 40700). The AUC values presented in Table 2 or on the challenge website does not show if customers are classified as 'yes' by the classifier. Exploration of lever variables does not allow in this case a modification of the predicted class. Nevertheless Table 4 and Figure 1 show that a large improvement of the 'yes' probability (the probability of cross-selling) is possible.

In Table 4 the second column (C2) presents the best $P_f(C_z \mid x_k, b)$ obtained, the third column (C3) the initial corresponding $P(C_z \mid x_k, b)$, the fourth column (C4) the initial interval used in the naive Bayes formulation (used to compute $P(C_z \mid x_k, b)$) and the last column (C5) the interval which gives the best improvement (used to compute $P_f(C_z \mid x_k, b)$). This table shows that:

- for all lever or observable variables, there exists a value change that increases the posterior probability of occurrences of the target class;
- the variable that leads to the greatest probability improvement is B_ENQ_L3M (The number of Bureau Enquiries in the last 3 months), for a value in [1.5,+∞[rather than in]- ∞,1.5[; This variable is an observable variable, not a lever variable, and means that a marketing campaign should be focused on customers who contacted the bureau more than once in the last three months;
- nevertheless, none of those changes leads to a class change as the obtained probability ($P_f(C_z \mid x_k, b)$) stays smaller than $P(C_z \mid x_k)$.

Table 4: Best $P(C_z)$='yes' obtained

C1: explored variable	C2	C3	C4	C5
RENT_BUY_CODE	0.6	0.26	[O,P]	[M,R,B,X]
CURR_RES_MTHS	0.36	0.21	[40.5,+ ∞ [	]-∞,40.5[
B_ENQ_L6M_GR3	0.25	0.10	]-∞,0.5[	[0.5,+∞[
B_ENQ_L3M	0.12	0.12	]-∞,1.5[	[1.5,+ ∞[
B_ENQ_L12M_GR3	0.36	0.16	]-∞,0.5[	[1.5,+∞[
B_ENQ_L12M_GR2	0.36	0.24	[0.5,+∞[	]- ∞,0.5[

In Figure 1 the six dotted vertical axis represent the six lever or observable variables as indicated on top or bottom axis. On the left hand size of each vertical axis, the distribution of $P(C_z \mid x_k)$ is plotted (□) and on the right hand size the distribution of $P_j(C_z \mid x_k, b)$ is plotted (■). Probability values are indicated on the y-axis. In this Figure only the best $P_j(C_z \mid x_k, b)$ (PCa[0] in Algorithm 1) is plotted. This figure illustrates in more details the same conclusions as given above.

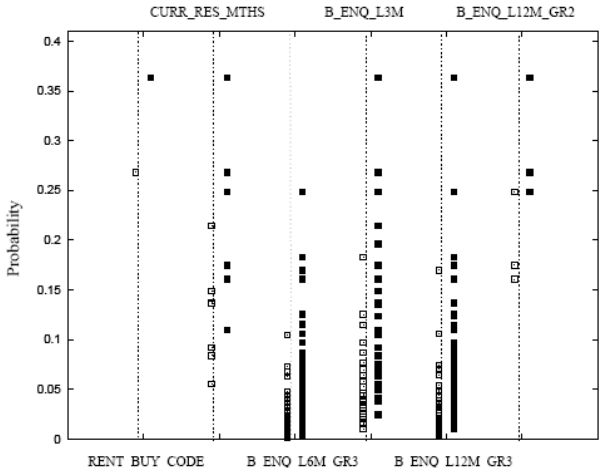


Fig 1: Obtained results on $P_j(C_z \mid x_k, b)$.

5.3 Test on the KDD Cup 2009

5.3.1 Task description

The KDD Cup 2009 offers the opportunity to work on large marketing databases from the French Telecom company Orange. The goal is to predict the propensity of customers to switch provider (churn), buy new products or services (appetency), or buy upgrades or add-ons proposed to them to make the sale more profitable (up-selling).

In this section we consider only the problem of churn. The churn rate is also sometimes called the attrition rate. In its broadest sense, the churn rate is a measure of the number of individuals or items moving into or out of a collection over a specific period of time. The term is used in many contexts, but is most widely applied in business. For instance, it is an important factor for mobile telephone networks and pay TV operators.

In this study the moment of churn is the moment when the client cancels ("closes") his Orange product or service. A churner is a client having a product or service at time t_n and having no product at time t_{n+1} For more details see the presentation at: http://perso.rd.francetelecom.fr/lemaire/kddcup/ChallengePresentation_03192009.pdf. Churn has high cost as to conquer a customer is more expensive than to keep a customer.

5.3.2 Data and experimental conditions

In this paper we consider the small dataset available at the end of the fast challenge. Both training and test sets contain 50,000 examples. This real life dataset has numerical and categorical variables: the first 190 variables are numerical and the last 40 are categorical. These 230 variables are currently used by the marketing teams.

We used the Khiops software to elaborate a naive Bayes classifier. The performance of this classifier, on the small dataset, can be found on the challenge website with the name "reference". The AUC obtained on the training set is 0.6791 and 0.6827 on 10% of the test set. Results on 100% of the test set will be available only at the end of the challenge and can not be divulged for the time being.

5.3.3 Input values exploration

Algorithm 1 has been applied on the training set to reinforce the probability to stay loyal (the probability of not churning, the reference class C_z). Only the variable which reinforces the most the probability of not churning is kept.

Table 5 presents the list of variables which can reinforce the probability of not churning (note that sometimes it is not possible to reinforce this probability, therefore the sum of the second column is not equal to 50000).

In this table the first column gives the identifier of the input variable, the second column (C2) gives the number of client who see their probability of not churning reinforced using the variable indicates in column one, the third column (C3) indicates for the corresponding line the number of customers for whom the reinforcement leads to a change of class for the reference class, the fourth column (C4) the number of customers for whom the initially predicted class is the reference class and the last column (C5) the number of customers for whom the reinforcement does not change the predicted class for the reference class.

For the challenge, the meaning of the variables is not revealed so it is not possible to see if these variables are lever variables. But the results of the table 5 indicate a high potential of the proposed method for Orange. Even using an "action" which does not lead to a class change, as for example when using the input variable 189, clients are pushed far from the churn "boundary" (see Figure 2).

Two types of action are possible:

- Preventive (or Push) Action: to prevent a customer from churning an operator can propose an offer, a service, to a customer and in this case his corresponding attribute changes to decrease (or to increase) a probability output of the model;
- Reactive Action: a modification of an attribute of customer is observed and detected since with this value this customer goes near the churn boundary.

Table 5: List of best variables

Variable	C2	C3	C4	C5
6	16	0	16	0
7	4	0	4	0
73	3885	371	3465	49
74	775	0	775	0
81	122	47	68	7
113	884	52	828	4
126	42148	5627	34949	1572
189	446	51	390	5
193	61	0	61	0
205	96	0	96	0
206	372	0	372	0
210	2	0	2	0
212	168	0	168	0
213	8	0	8	0
218	377	62	275	40
227	254	0	254	0
228	382	0	382	0

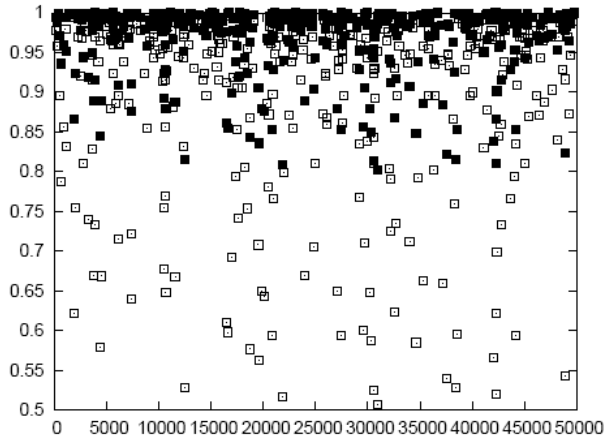


Figure 2: Obtained results on $P_j(C_z|x_k$, b) using the input variable 189. The horizontal axis indicates the number of the client in the training dataset. On the vertical axis: the distribution of $P(C_z|x_k)$ is plotted using a □ and the distribution $P_j(C_z|x_k$, b) is plotted using a ■. In this Figure only the best $P_j(C_z|x_k$, b) (PCa[0] in Algorithm 1) is plotted and only for clients for who a reinforcement is possible but who were already classified as "loyal" (column C4 in Table 5).

6. CONCLUSION

In this paper we proposed a method to study the influence of the input values on the output scores of a probabilistic model. This method is a part of a complete data mining process adopted by several Orange business units.

The method has first been defined in a general case valid for any model, and then been detailed for naive Bayes classifier. We also demonstrate the importance of "lever variables", explanatory variables which can conceivably be changed. Our method has first been illustrated on the simple Titanic database in order to show the need to define lever variables. Then, on the PAKDD 2007 challenge databases, a difficult problem of cross-selling, the results obtained show that it is possible to create efficient indicators that could increase sells. Finally we demonstrated the applicability of our method to the KDD Cup 2009.

The case study presented on the Titanic dataset illustrates the point of applying the proposed method to accident research. It could be used for example to analyze road accidents or air accidents. In the case of the air accidents any new plane crash is thoroughly analyzed to improve the security of air flights. Despite the increasing number of plane crashes, the relative frequency of those in relation to the volume of traffic is decreasing and air security is globally improving. Analyzing the correlations between the occurrence of a crash and several explanatory variables could lead to a new approach to the prevention of plane crashes.

This type of relationship analysis method has also great potential for medicine applications, in particular to analyze the link between vaccination and mortality. The estimated 50% reduced overall mortality currently associated with influenza vaccination among the elderly is based on studies neither fully taking into account systematic differences between individuals who accept or decline vaccination nor encompassing the entire general population. The proposed method in this paper could find interesting data for infectious diseases research units. Another potential area of application is the analysis of the factors causing a disease, by investigating the link between the occurrence of the disease and the potential factors.

Three main future works are also under consideration:

- the study of the temporal evolution of predicted scores when the values of the explicative variables are likely to change;
- the possibility to learn iteratively a new predictive model after having modified the data according to the best action found after correlation exploration;
- performing a controlled test on the 'lever variables' to see if the action of moving from one set of values to another affects churn/response/survival etc - i.e. if there is in fact an underlying causality in the lever variable (this causality can not be concluded from correlations).

The proposed method is very simple but efficient. It is now implemented in an add-on of the Khiops software—(see http://www.khiops.com), and its user guide (including how to obtain the software) is available at: http://perso.rd.francetelecom.fr/lemaire/understanding/Guide.pdf

This tool could be useful for companies or research centers who want to analyze classification results with input values exploration.

7. REFERENCES

[1] J. M. Benitez, J. L. Castro, and I. Requena. Are artificial neural networks black boxes. IEEE Transactions on Neural Networks, 8(5):1156–1164, 1997. Septembre.

[2] M. Boullé. Compression-based averaging of selective naive Bayes classifiers. Journal of Machine Learning Research, 8:1659–1685, 2007.

[3] M. Boullé. Khiops: outil de préparation et modélisation des données pour la fouille des grandes bases de données. In Extraction et gestion des connaissances (EGC'2008), pages 229–230, 2008.

[4] M. Boullé. Modl: a bayes optimal discretization method for continuous attributes. Machine Learning, 65(1):131–165, 2006.

[5] L. Breiman. Random forest. Machine Learning, 45, 2001.stat-www.berkeley.edu/users/breiman/Breiman.

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

[7] D. Choudat. Risque, fraction étiologique et probabilité de causalité en cas d'expositions multiples, i : l'approche théorique. Archives des Maladies Professionnelles et de l'Environnement, 64(3):129–140, 2003.

[8] G. Diagne. Sélection de variables et méthodes d'interprétation des résultats obtenus par un modèle boite noire. Master's thesis, UVSQ-TRIED, 2006.

[9] T. Fawcett. Roc graphs: Notes and practical considerations for data mining researchers. Technical Report HPL-2003-4, HP Labs, 2003., 2003.

[10] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. A*dvances in Knowledge Discovery and Data Mining*, chapter From data mining to knowledge discovery : An overview. AAAI/MIT Press, 1996.

[11] R. Féraud, M. Boullé, F. Clérot, and F. Fessant. Vers l'exploitation de grandes masses de données. In *Extraction et Gestion des Connaissances (EGC)*, pages 241–252, 2008.

[12] R. Féraud and F. Clérot. A methodology to explain neural network classification. *Neural Networks*, 15(2):237–246, 2002.

[13] K. Främling. *Modélisation et apprentissage des preferences par réseaux de neurones pour l'aide à la decision multicritère*. PhD thesis, Institut National des Sciences Appliquées de Lyon, 1996.

[14] I. Guyon. *Feature extraction, foundations and applications*. Elsevier, 2005.

[15] I. Guyon, C. Constantin Aliferis, and A. Elisseeff. *Computational Methods of Feature Selection*, chapter Causal Feature Selection, pages 63–86. Chapman and Hall/CRC Data Mining and Knowledge Discovery Ser., 2007.

[16] I. Guyon, A. Saffari, G. Dror, and J. Bumann. Report on preliminary experiments with data grid models in the agnostic learning vs. prior knowledge challenge. In *IJCNN: International Joint Conference on Neural Networks*, 2007.

[17] D. Hand and K. Yu. Idiot's Bayes - not so stupid after all? *International Statistical Review*, 69(3):385–399, 2001.

[18] M. S. Kramer, J. M. Leventhal, T. A. Hutchinson, and A. R. Feinstein. An algorithm for the operational assessment of adverse drug reactions. i. background, description, and instructions for use. *Journal of the American Medical Association*, 242(7):623–632, 1979.

[19] V. Lemaire and R. Féraud. Driven forward features selection: a comparative study on neural networks. In *International Conference on Neural Information Processing*, 2006.

[20] V. Lemaire, R. Féraud, and N. Voisine. Contact personalization using a score understanding method. In *International Joint Conference on Neural Network*, Hong-Kong, October 2008.

[21] P. Leray and P. Gallinari. Variable selection. Technical Report ENV4-CT96-0314, University Paris 6, 1998.

[22] M. Možina, J. Demšar, M. Kattan, and B. Zupan. Nomograms for visualization of naive Bayesian classifier. In *PKDD '04: Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 337–348, New York, USA, 2004. Springer-Verlag New York, Inc.

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

[24] P. Poirier, C. Bothorel, E. Guimier De Neef, and M. Boullé. Automating opinion analysis in film reviews : the case of statistic versus linguistic approach. In *LREC 2008 Workshop on Sentiment Analysis: Emotion, Metaphor, Ontology and Terminology*, pages 94–101, 2008.

[25] B. Poulin, R. Eisner, D. Szafron, P. Lu, R. Greiner, D. S. Wishart, A. Fyshe, O. Pearcy, C. Macdonell, and J. Anvik. Visual explanation of evidence with additive classifiers. In *IAAI*, 2006.

[26] M. Robnik-Sikonja and I. Kononenko. Explaining classifications for individual instances. *IEEE TKDE*, 20(5):589–600, 2008.

[27] M. Robnik-Sikonja, A. Likas, C. Constantinopoulos, and I. Kononenko. An efficient method for explaining the decisions of the probabilistic rbf classification network. currently under review, partialy available as TR, `http://lkm.fri.uni-lj.si/rmarko`, 2009.

[28] N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proc. 18th International Conf. on Machine Learning*, pages 441–448.Morgan Kaufmann, San Francisco, CA, 2001.

[29] A. Singh, R. Nowak, and P. Ramanathan. Active learning for adaptive mobile sensing networks. In *IPSN '06: Proceedings of the fifth international conference on Information processing in sensor networks*, pages 60–68, New York, NY, USA, 2006. ACM Press.

[30] S. Thrun. Extracting rules from artifcial neural networks with distributed representations. In M. Press, editor, *Advances in Neural Information Processing Systems*, volume 7, Cambridge, MA, 1995. G. Tesauro, D. Touretzky, T. Leen