Fraud prevention in electrical companies
UCM IV Modelling Week

Problem description

June 2010

neometrics

IV UCM Modelling Week

Contents

01 Problem description

- In some countries it is a common practice to **manipulate electrical meters** in order to reduce the electrical invoice in a fraudulent way.

- The **cost of visiting a given suspicious installation** which might turn out to be fraudulent is large. Thus, it's important to identify with as much prediction as possible, which installations should be checked. (**A checking is a 100% safe method to identify and cancel a fraud where it exists**)

- Having an **order of priority** about which installations should be checked and a probability of fraud associated to each installation will allow the company to **reduce the cost of the investigation substantially**

01 Problem description

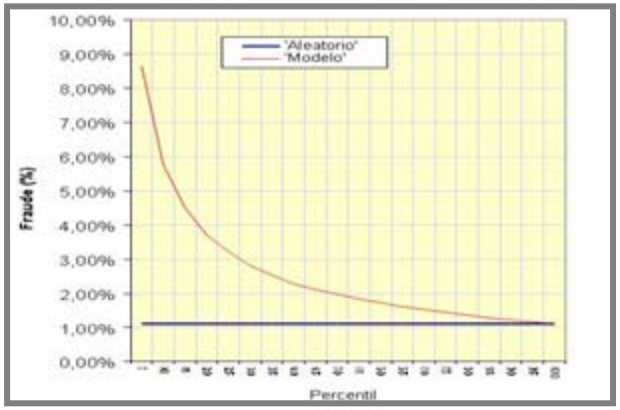
- This is a **widely spread problem** that is found in numerous energy production and distribution companies, especially in Latin America.

- We provide **cost and benefit data associated to the decision of wether to send or not the inspection,** so, as part of the solution **participants will be requested to identify the optimal cut probability which maximizes the ROI**. The results of every model will be analyzed and compared in terms of best economic ROI.

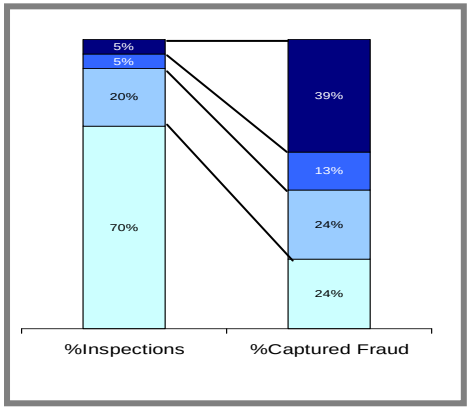
- This kind of problem gives participants the chance to **face a realistic problem** since it challenges them not only to generate a reliable predictive model but also deal with the need to **measure the economic impact of predictive models**, so important in real life business.

02 Mathematical approach (1/2)

- This is a typical predicting problem involving the creation of a **binary target model** (fraud/not fraud) .Of particular interest are the kind of event to model and the **complexity of the data** entry.
- The event to model is fraud. Modeling fraud in a binary target model usually involves a certain degree of complexity due to the fact that **fraud has different motivations** which are explained by different variables. This implies the use of **non linear components in its prediction**
- A successful predictive model will allow clear detection of those customers with a highest probability of having manipulated their electrical meters.



Cumulative response of the total population (example).

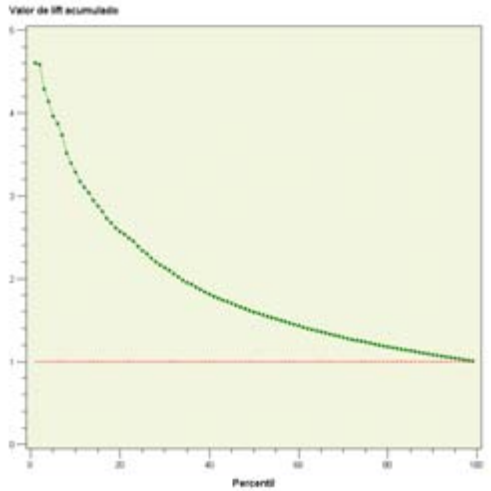


Captured Fraud by the model

02 Mathematical approach (2/2)

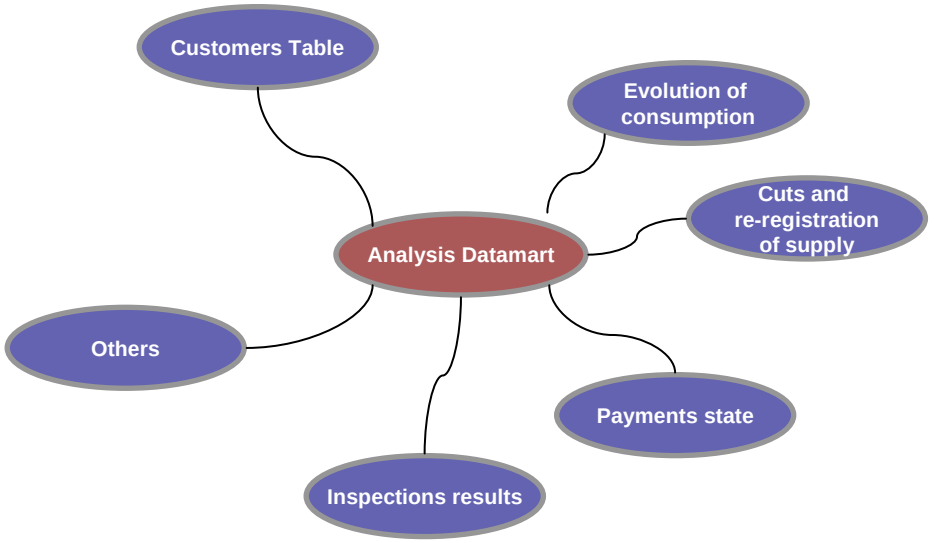
- **Any binary target model would be valid**. However, it is advisable to use techniques which can contain efficiently the **non-linear effects** above mentioned as well as the **possible interactions** between variables.
- It is suggested that participants **measure the performance of models using the lift chart**, one of the most common method used in data mining. **This chart should be incorporated into the ROI analysis as one of its components**.
- The **final model should be the one that maximizes the ROI**.

- Cumulative lift:
  - To construct this chart customers should be ordered on the X axis in descending order according to the fraud score, which the model has thrown out.
  - For each percentile, the Y axis represents the quotient between the **response rate obtained from the selection yielded by the model (the group of customers above the selected percentile) and the rate obtained through a random selection (without a model)**



03 Information to solve the problem

- Usual data in energy theft spotting



- The data provided for the resolution of the problem include the following:
  - A dataset with numerical and categorical **variables asociated to each customer and the target variable indicating fraud**.
  - An additional dataset with the **historical detail of monthly consumption** in the months immediately before the baseline date. These consumption data are specially relevant in fraud prediction and are **rich in information which can be treated either constructing calculated variables** which are later incorporated in a simple model or directly in more complex models.

04 ROI Analysis

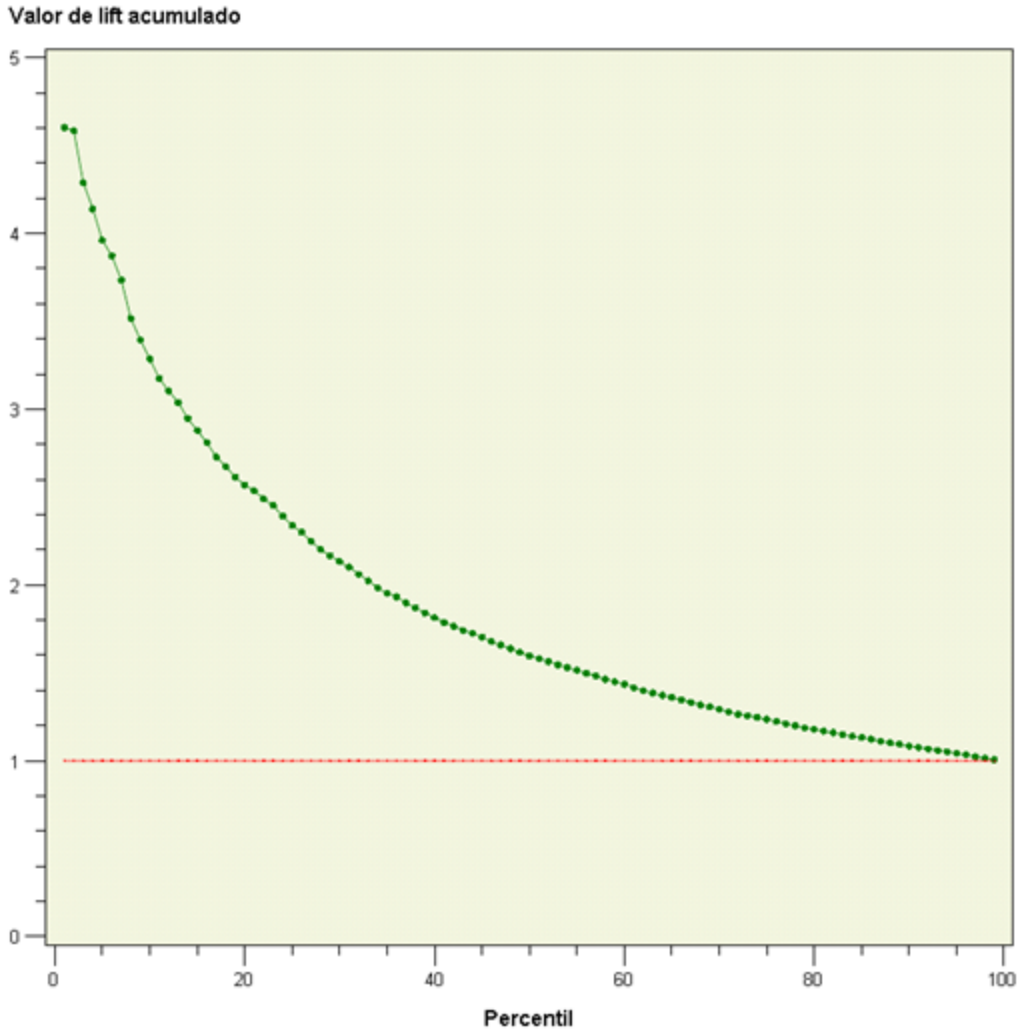
- The proposal is to make the ROI analysis for the constructed model finding the **optimal number of inspections to be carried out in order to maximize ROI** following the score results from the model and taking into account these aspects:

  - The **number of detected frauds predicted by the model will be given by the lift curve**, which is usually used in data mining to measure the reliability of a predictive model.
  - The hypothesis are :
    - **One year period**
    - **Average consumption for detected fraud** (If fraud is detected, the **gross saving would be the mean income obtained by the company per customer per year**)
  - For each **inspection there will be a fixed cost** 15.000 MU, (to simplify the same cost is hypothesized for all inspections).
  - A **fixed cost is required for the maintenance of the inspection crew** regardless of their number.

- Finally, participants are required to extend their analysis by **incorporating a developing cost of the model**. The cost always exists and depends on the number of hours the analyst of the company has spent or can be a fixed cost for external consultancy. The participants can analyze assuming costs in function of different hypothesis of investment amortization (1 year, 2 years, 3 years…).

05 Reference model

- As a reference, we provide the following results for the problem yielded by a Neo Metrics model:
  - Expected profitability : 185.830.000 MU
  - Real profitability : 182.930.000 MU
- This is **not the final model that NM implemented in its client**.



06 Notes

- In consumption table there are more customers than in train table, which could be ignored.
- The average invoice per customer per year can be calculated by mean_pago variable (mensual payment average) multiplying by 12.
- The mean cost per inspection is 15.000 MU (Monetary Unit)
- The anual cost of the inspection crew is 100.000.000 MU (conversion rate into Euros x 0.0015)
- The falso_target variable divides data into two sets: In the first set the variable has 0/1 values indicating no fraud / fraud . Participants should use this part of data to train the model. The second set of this variable is not informed and it should be used to validate and calculate the profitability of the model.
- As a summary, to obtain the profitability, the participants should construct following elements:
  - A predictive model which assigns a fraud probability to each inspection.
  - The optimal selection of inspections should the one that maximizes the expected profitability.
- The expected and real profitability of the selection above are required and participants should compare the optimal selection with the actual of the company (The selection of every inspections)

neometrics

Value on every interaction