Chapter 20

PAKDD-2007: A Near-Linear Model for the Cross-Selling Problem

Thierry Van de Merckt
VADIS Consulting, Belgium

Jean-François Chevalier
VADIS Consulting, Belgium

ABSTRACT

This chapter presents VADIS Consulting’s solution for the cross-selling problem of the PAKDD-2007 competition. For this competition, the authors have used their in-house developed tool RANK, which automates a lot of important tasks that must be done to provide a good solution for predictive modelling projects. It was for them a way of benchmarking their 3 years of investment effort against other tools and techniques. RANK encodes some important steps of the CRISP-DM methodology: Data Quality Audit, Data Transformation, Modelling, and Evaluation. The authors have used RANK as they would do in a normal project, however with much less access to the business information, and hence the task was quite elementary: they have audited the data quality and found some problems that were further corrected, they have then let RANK build a model by applying its standard recoding, and then applied automatic statistical evaluation for variable selection and pruning. The result was not extremely good in terms of prediction, but the model was extremely stable, which is what the authors were looking for.

The challenge proposed this year by the PAKDD is a good example of a usual predictive task in analytical CRM projects. It concerns a financial company having customers of credit card as well as of home loan (mortgage) products. Both of these products have been on the market for many years, although for some reason 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 – although this challenge is reduced when the right methods are used for building a model.

A modeling dataset of 40,700 customers with 40 modeling variables, plus a target variable, is...
provided. 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 if otherwise (40,000 random samples). A prediction dataset (8,000 sampled cases) is provided with similar variables but withholding the target variable, in order to judge the real performance of the model by the PAKDD contest committee. 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.

Our goal in competing on this contest was to apply our standard processes and methods and to submit our solution without extensive search and tuning. We decided that the time spent should not be more than 1 man day in total, which was the case. On the technical side, our challenge – as usual – was to make sure that our model would be robust in order to avoid any bad surprise when applied on the prediction dataset.

APPLIED METHODOLOGY

Our methodology for building analytical solutions is based on CRISP-DM. In order to support our consultants in applying this methodology in a rigorous and consistent way, we have developed a platform called RANK that automates some of the major steps of the process, as shown in figure 1.

Since PAKDD07 contest provides the data sets, the target, and the data dictionary, the first three steps are not applicable. Hence, the process is the following:

- **Audit** – Evaluation of the data quality, its consistency, etc.
- **Transformation** – Preparation of the data for modeling: defining types, binning, re-coding, deriving new variables, linearization of the vector space, normalization, etc.
- **Modeling** – building the model itself, by choosing the best technique, the set of relevant variables, etc.
- **Evaluation** – asserting the model stability, its statistical relevance, etc. And reviewing the business relevance (this last important step is not applicable to the contest).

The last two steps (Learning and Deployment) are not applicable to the PAKDD07 contest.

RANK provides a great help for all these steps to the analyst.

Audit

The audit allows analyzing the distribution of the variables and to spot anomalies. An example is given in the next figure.

This variable indicates the Number of Bureau Enquiries in the last 6 months for Mortgages. Maximum actual value is 97. Special values are: