Data mining in the real world: What do we need and what do we have?

Françoise Soulié Fogelman
KXEN
25 Quai Gallieni - 92 158 Suresnes cedex - France
Francoise.SoulieFogelman@kxen.com
http://www. kxen.com

Historically, data mining has been in the hands of small teams of expert statisticians who produce a few models per year. However, recently companies invested heavily in building huge data warehouses (from a few terabytes to peta-bytes) that contain millions of records and thousands of variables; for example, 5,000 variables on 150 million customers and prospects. That has changed the economics of data mining. Now businesses want a return on that investment and are looking well beyond reporting and basic statistics [1]. When they review their business activities, they see the need for 100s or 1000s of predictive models per year. Of course, very few companies can produce that many today, due to a lack of expert staff and appropriate tools. Some actually do generate that many models and we will provide examples, including:

- A broadband communications company moved from 5 cross-sell models per year to 1600
- A wireless communications company that produces 700 CRM models per year
- A national retailers cut time-to-model by 90% and scores 75M households in 30 minutes
- A marketing research firm built 370 propensity-to-buy models on a PC in an afternoon

We will analyze model production from an industrial viewpoint and review constraints such as large data volumes (records and variables), ability to produce robust models with little intervention, and fast algorithms automatically parameterized. Typical model training would be 300,000 records with 600 variables, trained in less than one hour total on a personal computer, including data coding, variable selection, and modeling; and application of the model on a few million records in an additional hour. We will review how we achieve this level of productivity through the extensive application of Vapnik's SRM framework [2].

Lastly, we will discuss future challenges including the automatic coding of multi-media data (text, images, audio, and movies), integration of model training and application into vertical software packages, combining analytic data sets, and combining models. We feel that the data mining role is businesses is on a fast track today and Machine Learning practitioners will play a major role, provided we take into account key industrial constraints.

References

1. Davenport, Thomas (2006) "Competing on analytics," *Harvard Business Review*, January
2. Vapnik, V.N. (1995) *The Nature of Statistical Learning Theory*, Springer