

Faculté des arts et des sciences - Secteur des sciences

Département d'informatique et de recherche
opérationnelle

Data Mining in Banking Applications

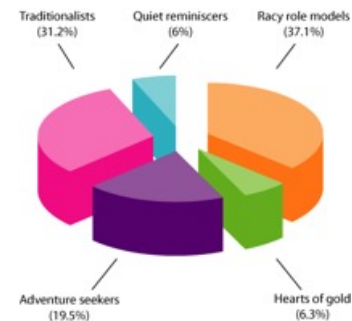
Predicting Business Revenue

Yoshua Bengio and al., Université de Montréal

What is Data Mining?

Data mining is one of the many commercial applications of machine learning

Data Mining in Bank Applications, e.g.



Why is Data Mining so Interesting Here?

- Because it is essential for efficient marketing, which relies on the size and potential of the customers
- Because it allows to tackle accurately your customer needs
- Because the commercial banking market is lucrative and offers challenges that can be solved with data mining

What Are the Main Challenges?

Our algorithms are specially designed to face the main difficulties encountered with banking data:

- ▣ Very large quantities of data
- ▣ Missing values
- ▣ Large number of inputs potentially relevant for prediction
- ▣ Non-linear dependencies
- ▣ High variability in the data

How Does it Work?

Comparative study of the following algorithms:

- Standard statistical methods
(logistic and linear regression)
- Well-known methods in machine learning
(Neural networks, decision trees, Adaboost)
- Custom methods, based on machine learning and tailored to the specific challenges of this banking data

Prediction of Revenue

- From business profiles containing 669 inputs:
 - Balance of accounts,
 - transactions statistics,
 - bank transactions
 - credit card data, etc
 - The objective is to classify into 3 business categories:
 - Very small < 1M\$,
 - Small 1 - 10M\$
 - Medium 10 - 100M\$.
-

Training of a Predictor

- We train a model from the cases (training samples) for which the revenue is known (*labeled cases*):

Sample 1: Inputs for business 1, revenue of business 1

Sample 2: Inputs for business 2, revenue of business 2

Etc. for all the businesses for which revenue is known.

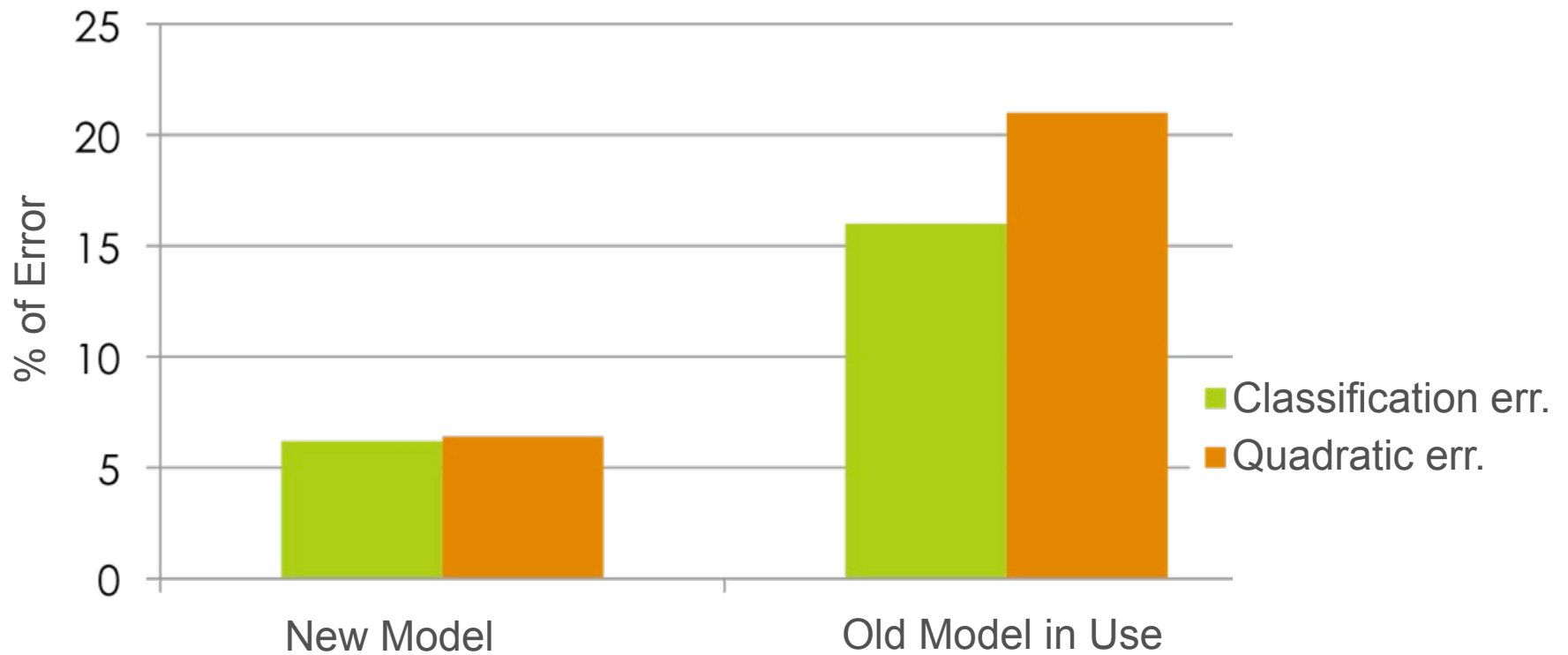


- The predictor can now be applied in *unlabeled cases*, where the revenue is not known, i.e., the majority of businesses.

Bank Data

Number of samples	2008/03	2007/09	2006/09
Learning	77370	70522	47057
Testing	37003	36710	11765
Labeled	114373	107232	58822
Unlabeled	273220	278432	328465

Performance of our Model



Achievements

- Faster Training (from 1 month to 1 day)
 - Faster Predictions
 - Improved user's confidence with the estimation of the importance of each input variable
 - Saves automatically the model and the processing
 - Stabilization of the classification year after year
-

Conclusion

It has been shown that our statistical learning model proposed for data mining is **3 times better** than the currently used model.

Perspectives

A good market development strategy depends strongly on the performance of tools like ours.

Value Added

Our predictions are a good asset for:

- Business development
 - Tailored marketing
- Performance Management
 - Dashboards to monitor branches

Yoshua Bengio, PhD



- Full Professor at U. Montreal
- 17 years of industrial R&D
- More than 200 publications and 4300 citations
- NSERC Research Chair and Canada Research Chair
- Winner of Urgel Archambault's prize 2009
- Fellow of CIRANO and of CIFAR
- Founder of LISA laboratory (1993)
- <http://www.iro.umontreal.ca/~bengioy>
- V.P. R&D, ApSTAT Technologies