Business-Insight Top winner at the PAKDD 2010 cup

The objective of the 14th Pacific-Asia Knowledge Discovery and Data Mining conference (PAKDD 2010) is Re-Calibration of a Credit Risk Assessment System Based on Biased Data.

There were 3 datasets used for the Challenge. They were collected during period from 2006 to 2009, and came from a private label **credit card** operation of a **Brazilian credit company** and its partner shops.

The prediction targets to detect are the "bad" clients. A client is labeled as "bad" (target variable=1) if he/she made 60 days delay in any payment of the bills contracted along the first year after the credit has been granted. In short, the clients that do not pay their debt are labeled as "bad".

The datasets that are available to the participants were:

1. **modelling** (50,000 samples). This is the only dataset where we have the "target" information (we know the "bad" clients). There are 26.08% of "bad" clients
2. **leaderboard** (20,000 samples). This dataset was used on the internet web-site of the competition to give instantaneous feedback about the accuracy of the different models developed by the different teams. The real-time LeaderBoard stimulates the competitors' daily participation because everybody can see how the other teams are performing.
3. **prediction** (20,000 samples). This is the only dataset that was used to obtain the final ranking of the competition.

The datasets consist of 52 explanatory variables of several types.

The important aspect to emphasize is that the Modelling and Leaderboard datasets include only "approved customers". This "approval" was computed using an old predictive model that is already in use in the bank. As a consequence, only a part of the "market" is monitored.

However, for the purpose of monitoring the decision support system's performance and collecting data for future model re-calibration, some clients have received the credit they had applied for, even if the current decision system classified them as "bad" clients. The Prediction dataset is the only one that contains "approved" and "non-approved" customers. The percentage of targets in the Prediction dataset is thus expected to be higher than inside the Modelling and Leaderboard datasets. There is thus a large sampling bias between the data sets.

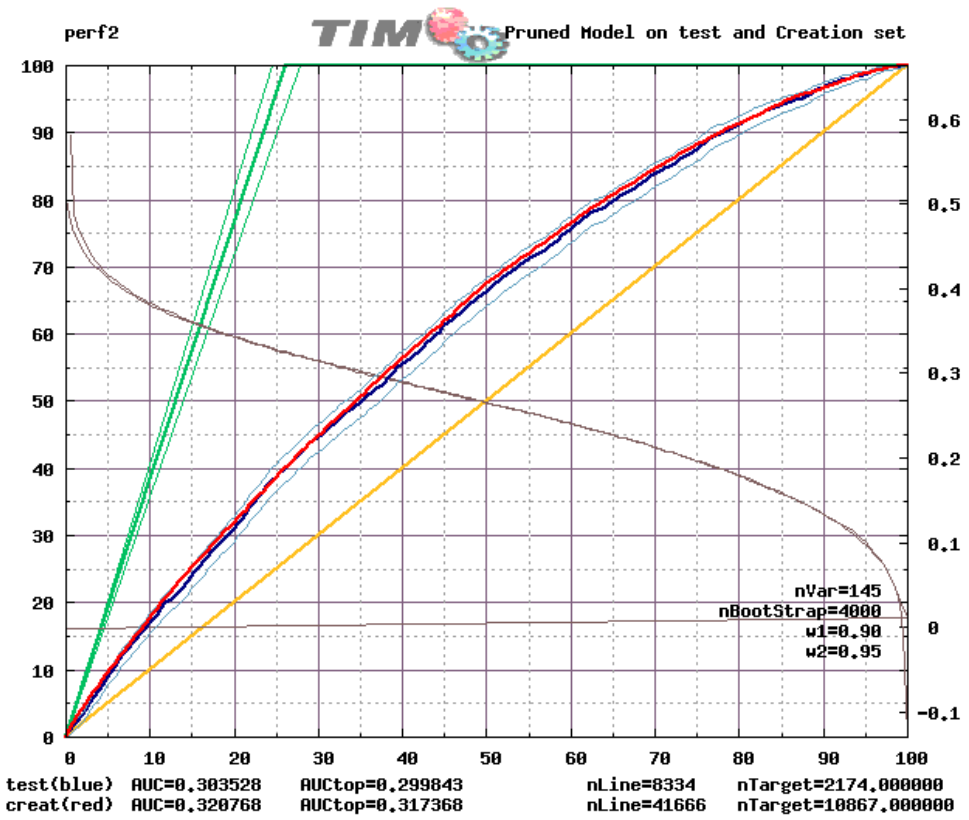
This competition thus focuses on the credit scoring model's generalization capacity from partial biased data sets available for modeling.

The performance of the predictive model (that is used to compute the final ranking of the teams) is measured using area under the receiver operating characteristic curve (AUC) on the Prediction data set of 20,000 further accounts.

The final ranking is:

PAKKD2010 Results			
Ranking	**Team Name**	**Institution**	**AUC_ROC**
1	GH (Grzegorz Haranczyk)	GH	0.645
2	Winners(Navin Loganathan,Ranjani Subramaniam, Shobha Prabhakar, Sankar Deivanayagam)	LatentView Analytics	0.641
2	Latentview (Priya Balakrishnan, Kiran.PV, Syluvai Anthony, Mahadevan Balakrishnan)	LatentView Analytics	0.641
4	uq (Vladimir Nikulin)	University of Queensland	0.638
5	Tabnak (Yasser Tabandeh)	Shiraz University	0.637
6	Wiggle Puppy (Daniel Felix)	Independent	0.634
7	**Kranf (Frank Vanden Berghen)**	**Business-Insight**	**0.633**
7	TZTeam (Didier Baclin) **(This team was also using TIM!)**	None	0.633
7	Mjahrer (Michael Jahrer)	Commendo Consulting	0.633
10	Abhyuday (Abhyuday Desai)	Kiran Analytics,Inc.	0.629
10	iDO95 (Max Wang, Amy Yu)	Alliance Data	0.629

This is the lift that we obtained using TIM:



Some remarks:

1. Business-Insight obtained the 7th place amongst a total of 94 teams that participated to the PAKDD 2010 world-competition.

2. There are 2 users of the "TIM software suite" inside the TOP 10 winners of the competition.

3. Only one person (Frank Vanden Berghen) did work on the PAKDD competition and only for 1 day. Frank was using TIM.

4. Given more time, we could have "cleaned" the dataset better.
   In particular, it seems that one of the most important variables for prediction is the "location". Many columns encode the "location" inside the competition datasets: "professional_borough", "professionnal_city", "residential_borough", "residential_city". Unfortunately, these variables are extremely noisy: for example here are all the different sppling of the same city: ACHOEIRO DE ITAPEMIRIM, CACHOEIRO DE ITAPEMIRIM, CACHOEIRO D EITAPEMIRIM, CACHOEIRO D ITAPEMIRIM, CACHOEIRO DE ITAPEMIRIM, CACHOEIRO DE ITAMPEMIRIM, CACHOEIRO DE ITAPEMIRIM, CACHOEIRO DE ITPEMIRIM, CACHOEIRO ITAPEMIRIM, CACHOEIROIRA ITAPEMIRIM. We had no time to manually correct all the spelling mistakes inside these 4 variables. We have now added a new automated text-mining spelling-correction operator inside our ETL tool (Anatella) to easily cope (in a matter of minute) with such situation for the next competition.

5. The competition-winning-team managed to obtain a slightly higher score than TIM because:

   a. They did a better job cleaning the data (it takes time!).

   b. They did work on the competition for more than 2 months.

   c. More importantly: they managed to find extra-variables linked to the "location": area, population in 2005, density in 2005, GDP, GDP per capita PPP, Human Development Index (HDI), literacy rate, infant mortality, life expectancy. These extra-variables are important because, as we have noticed ourselves the "location" concept is a very important one.

6. Another Team than Business-Insight was also using TIM as their predictive datamining tool: The "TZTeam". The "TZTeam" is composed of only one individual: Didier Baclin. He did a very fine job on the competition despite the fact that he worked on it for less than 2 hours (from a personal communication)! The comments from Didier Baclin about the TIM software are:

    a. …a great piece of data analysis software…

    b. TIM is a very fast and easy to use software which can be used on massive datasets to model binary or continuous outcomes just like in this PAKDD competition.

    c. TIM … is fast and easy to use which makes it a great tool for exploring modeling possibilities.

    d. TIM allows the use for several variable selection techniques…

    e. It was easy to score the …. dataset thanks to TIM's built in scoring module.

This competition once again demonstrates the superior accuracy of the "TIM software suite" for datamining. TIM is a world-level datamining software that delivers the most accurate and robust models.

Business-Insight Top winner at the AUSDM 2009 cup

by **Frank** on Thu Dec 03, 2009 12:11 pm

The goal of the AusDM 2009 Analytic Challenge was to encourage the discovery of new algorithms for ensembling or 'blending' sets of expert predictions. Ensembling is the process of combining multiple sets of expert predictions so as to result in a single prediction of higher accuracy than those of any of the individual experts.

From previous data mining competitions such as the Netflix Prize, it has become apparent that for many predictive analytics problems, the best approach for maximizing prediction accuracy is to generate a large number of individual predictions using different algorithms and/or data, and ensembling these sub-results for a final prediction.

The AusDM 2009 challenge organizers provided sets of predictions obtained from the two leading teams in the Netflix Prize competition; Belkor's Pragmatic Chaos, and The Ensemble. The objective of the Netflix competition is to guess which rating a user will give to a specific movie. The ratings are in a "star" scale: from "1 star" to "5 stars". The final objective for the Netflix company is to use these "guesses" (or predictions) inside a recommander system to "suggest" movies to potential netflix customers.

There were three datasets provided:

- a small dataset with 30,000 sets of predictions from 200 experts (different algorithms or variations)
- a medium dataset with 40,000 sets of predictions from 250 experts
- a large dataset with 100,000 sets predictions from 1151 experts

Only the medium and large datasets were used inside the competition. Each of the three datasets were evenly divided into a "Test" subset containing only the individual expert predictions and a "training" subset containing both the expert predictions and the actual values for training. The training values were obtained from the Netflix Prize dataset by the organizers of the AusDM Challenge.

There were 4 different tasks in the AUSDM 2009 competition:

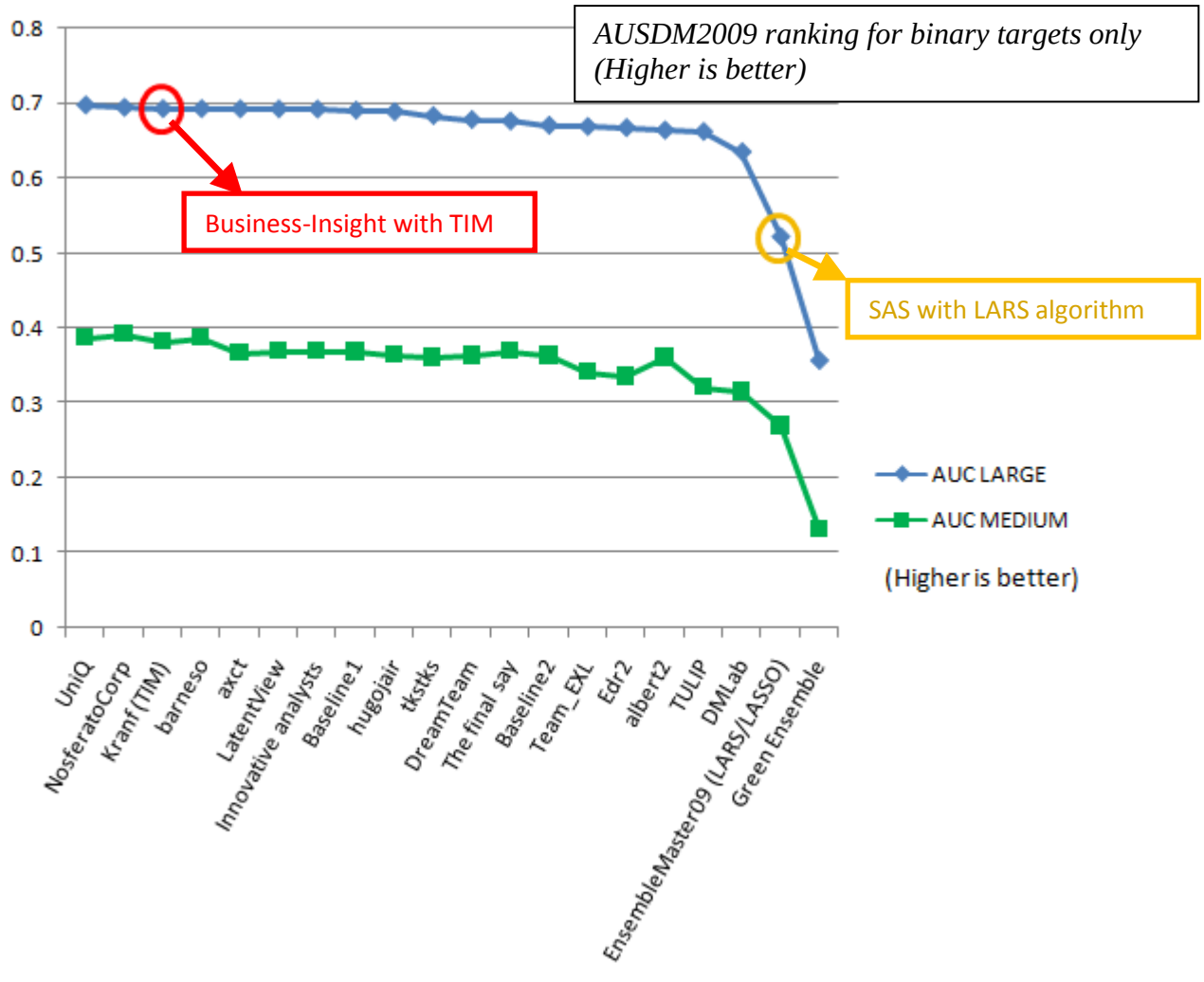
- *RMSE Large* (training dataset of 1,151 vars and 50000 rows) and *RMSE medium* (training dataset of 250 vars and 50000 rows): These are two continuous prediction problems: The objective is to predict exactly the rating (the number of "star" given to a movie multiplied by 1000: this gives 1000,2000,3000,4000 or 5000) of the movie.

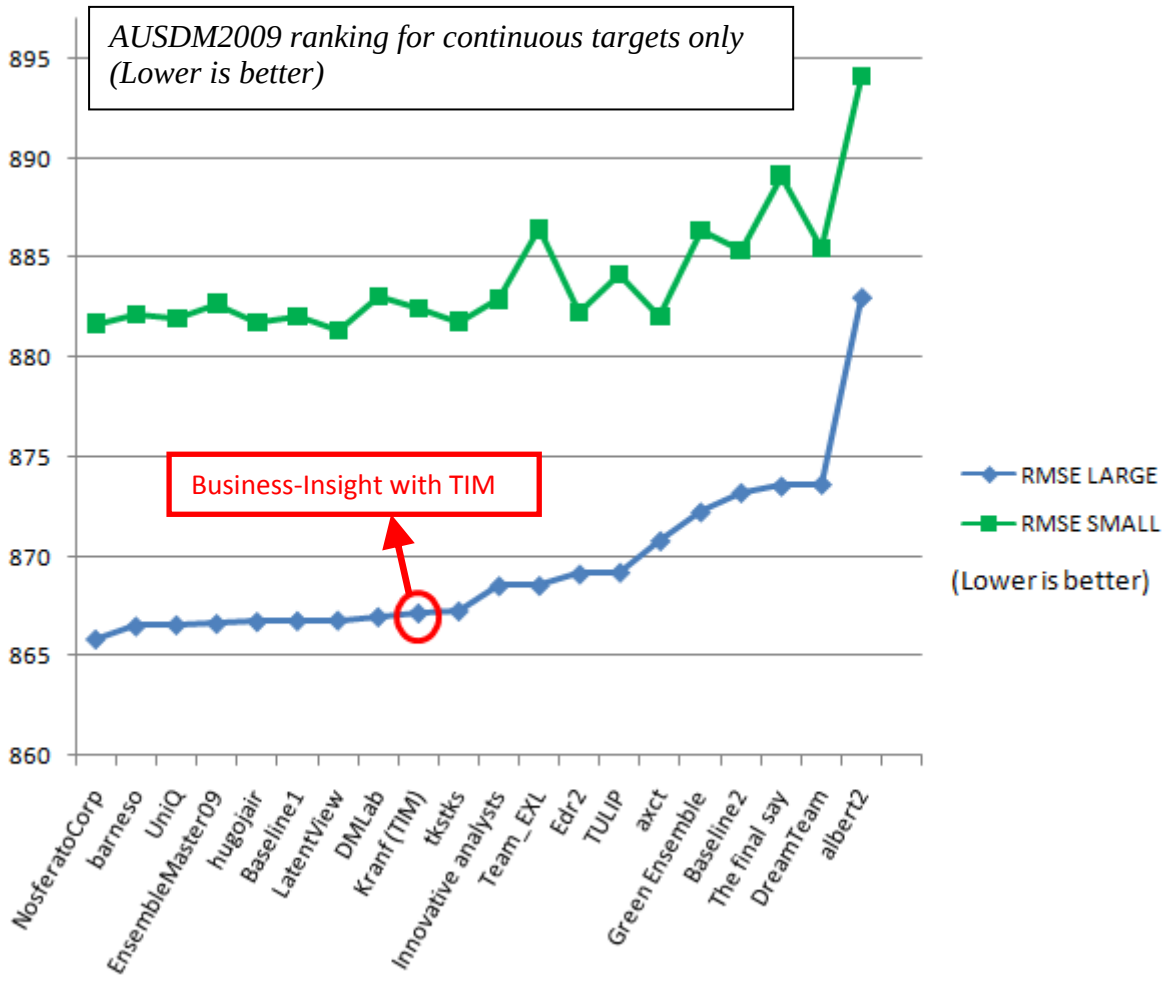
- *AUC Large* and *AUC Medium*: These are two binary prediction problems: The objective is to predict exactly if the rating is "5 stars" or "1 star".

We used TIM without any special treatments on the dataset obtained from the AusDM 2009

website. Our final rank is shown here (see "Kranf" entry) (Inside the different tables, the organization committee is using the terms "Gini" and "AUC" indifferently):

Champions League - Final Standings									
	Medium				Large				
Team Name	RMSE	Rank	Gini	Rank	RMSE	Rank	Gini	Rank	Points
NosferatoCorp	881.650	2	0.392	1	865.861	1	0.6941	2	9
UniQ	881.884	5	0.3879	2	866.582	3	0.6972	1	15
barneso	882.125	8	0.387	3	866.514	2	0.6923	4	23
LatentView	881.322	1	0.3693	6	866.809	7	0.692	6	33
Kranf	882.383	10	0.3819	4	867.172	9	0.6924	3	38
hugojair	881.722	3	0.3643	10	866.738	5	0.6884	9	41
Baseline1	881.98	6	0.3687	8	866.785	6	0.6903	8	42
Innovative analysts	882.879	12	0.3691	7	868.541	11	0.6918	7	55
axct	881.985	7	0.3674	9	870.801	15	0.6921	5	56
tkstks	881.755	4	0.3613	13	867.273	10	0.6831	10	57
EnsembleMaster09	882.656	11	0.2703	19	866.643	4	0.522	19	76
Edr2	882.211	9	0.3353	16	869.127	13	0.6671	15	81
DMLab	883.038	13	0.3153	18	866.985	8	0.635	18	83
The final say	889.101	19	0.3694	5	873.55	18	0.6759	12	84
Team_EXL	886.394	18	0.3421	15	868.555	12	0.6686	14	85
Baseline2	885.299	15	0.3635	12	873.219	17	0.6702	13	87
DreamTeam	885.412	16	0.3635	11	873.638	19	0.6778	11	87
TULIP	884.128	14	0.3217	17	869.216	14	0.662	17	93
albert2	894.071	20	0.3606	14	883.029	20	0.6638	16	106
Green Ensemble	886.367	17	0.1327	20	872.25	16	0.3572	20	109
TUB09	7858.22	21	0.0315	21	8120.22	21	0.1234	21	126





AUDSM2009 ranking for continuous target

We did perform best on the binary prediction problems: on these problems our ranking at the competition is 3 (on the large dataset) and 4 (on the small dataset). Our global ranking is 5.

A complete report that gives more explanations about all the techniques used by all the competitors is available here:
http://www.business-insight.com/downloads/AusDM09EnsemblingChallenge.pdf

Now, some comments about the competition:

1. The organization committee was very efficient. Many thanks!

2. The old engine of TIM (and RANK) was previously based only on the LASSO algorithm. On this dataset, the LARS/LASSO algorithm as implemented inside the SAS software behaved very poorly. The low-quality SAS implementation might be the cause of this low ranking. This poor behavior is illustrated by the "EnsembleMaster09" team that used the LASSO algorithm on the "large AUC" task and obtained a very poor accuracy: they were "the last but one" with an AUC of 52% (we obtained an AUC of 69% with TIM) (see the attachment "lasso.png" for the exact rankings on the LARGE AUC task).

3. For the RMSE challenge, we used only one simply, straight-forward "continuous" predictive model. Instead of one continuous predictive model, we could have used 5 "binary" predictive models. In this case, the final prediction is simply the results of a continuous model applied on a dataset containing only 5 columns that are the predictions obtained from the 5 "binary" predictive models. This should have given better results than one "big" continuous predictive model. We didn't try this approach by lack of time and because we had no idea that our final ranking at the competition would be so high. If we had known in advance our final ranking, we would have invested a little bit more time in the competition on these continuous predictive models.

4. All the competitors that obtained a better ranking than us were using "ensembling" techniques: their "final" predictive model is indeed a mixture (an ensemble) of many different predictive models. This technique is:

    a. prohibitively slow because it requires to build thousands of different models.

    b. not applicable in a real-world industrial context because of the complexity of the deployment of these "meta-models".

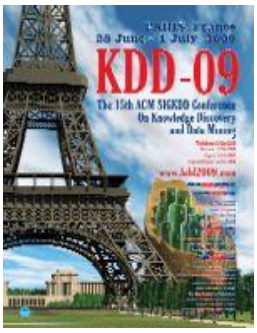
    c. somewhat disturbing because the columns of the competition dataset are already the output of many different predictive models built using "ensemble techniques". If we start using "ensemble techniques" to combine the output of predictive models built using "ensemble techniques", we can push this logic even further and do the following:

        - build many predictive models using ensemble technique (in the same way that the AUSDM2009 competition dataset was generated) (iteration 1)
        - build many predictive models using ensemble technique to accurately combine all the models build in iteration 1 (as the first competitors of the AUSDM2009 competition did) (iteration 2)
        - build many predictive models using ensemble technique to accurately combine all the models build in iteration 2 (iteration 3)
        - build many predictive models using ensemble technique to accurately combine all the models build in iteration 3 (iteration 4)
        - ...

        There are no limits to the number of iterations that you can do, using ensemble technique. The question to ask is: "*Does this really worth it? Does the additional accuracy in AUC or RMSE justify using such cumbersome technique?*". Indeed, if you are using TIM to create your predictive models, the prediction accuracy of the "simple" predictive models delivered by TIM is already so high that I personally think that it does not worth the trouble of using "ensemble techniques". This is not true if you are using another datamining tool. Anyway, the TIM software directly offers you "out of the box" all the required tool to do "ensemble technique", if you really want to go in this direction.

We spent less than one-half-man-hour (and around 6 computing hours) on these 4 tasks and, thanks to TIM, we are now in the "top winners" of the competition. We are very pleased by the efficiency of TIM, both in terms of computing speed and accuracy.

Frank



Business-Insight Top winner at the KDD2009 cup

by **Frank** on Mon Oct 19, 2009 8:45 am

Business-Insight took part to the world-famous datamining competition: the "KDD2009 cup".

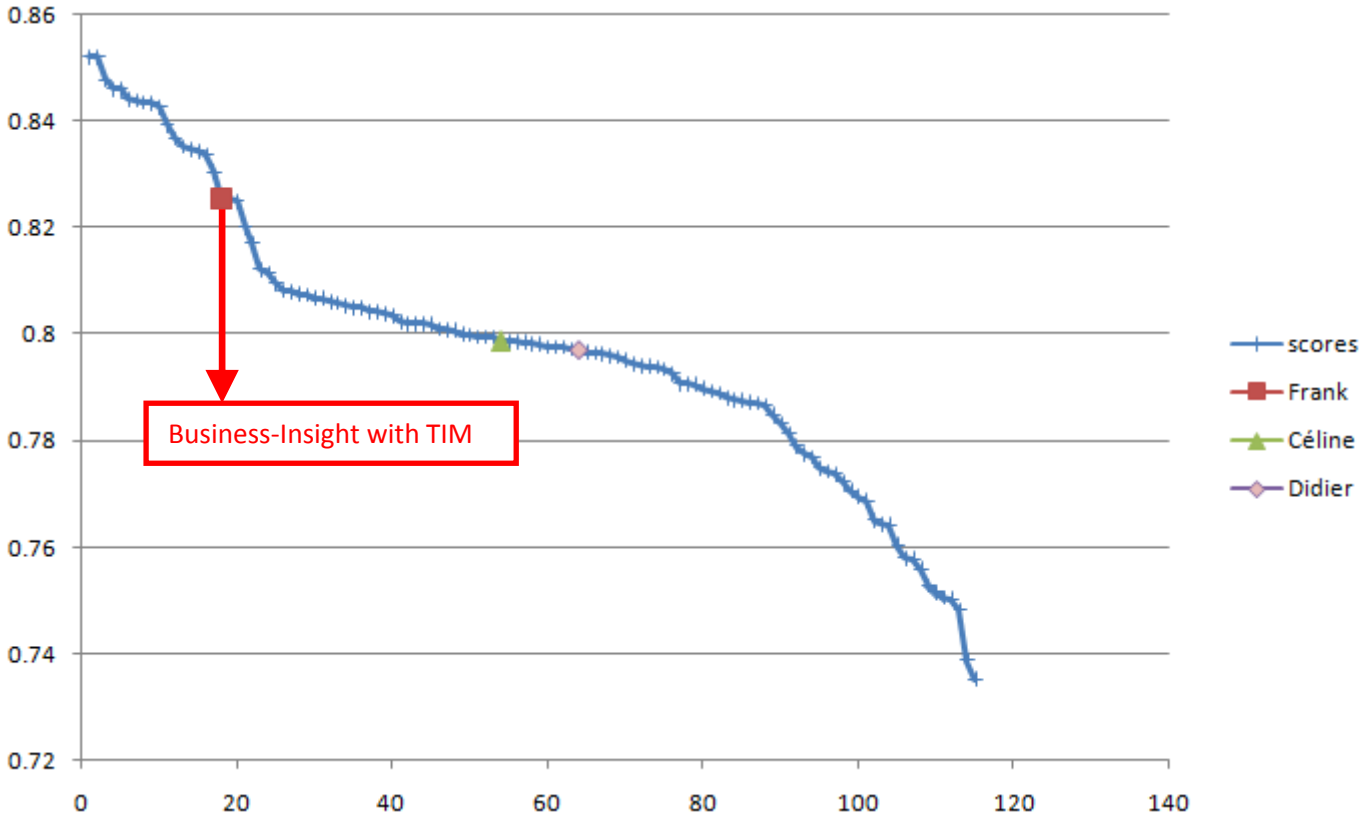
The rules of such competitions are always more or less the same and are very simple: for example: you receive some data collected in 2006 and 2007 and you must predict what will happen in 2008. The organizers of the competition are comparing your predictions with the real events of 2008 (they are the only one to know what happened in 2008). The team with the less prediction errors gets the first place.

This year, there were 2 different challenges for the KDD cup, each using a different datasets. The first dataset/challenge had no or very little interest outside the academic and university world because this dataset contains 15000 columns (do you know many companies that possess such a large dataset? I know none).

The second dataset was a "standard dataset" usually found in enterprise (with around 300 columns). This second dataset was only used inside the KDD competition for the “small challenge” (which is named this way because this last dataset is smaller than the first one) also called the “second challenge”. The second dataset was used to create predictions for the Orange Telecom operator (“Orange” is the number 1 of the French Telecom). The final ranking is based on the average quality (AUC) of 3 predictive models. The 3 predictive models to built are one Churn model, one Upselling Model and one "Propensity-to-buy" (appetency) model.

The KDD cup has draw a lot of attention this year because the task to fulfill (for the second challenge) is a very common task inside the telecom industry and represents accurately the kind of tasks that are encountered in "real life" (in opposition to the purely "abstracts" tasks that are commonly proposed in the university world). The competition was a real struggle because everybody wanted to demonstrate his superiority on "real world tasks".

The final results of the competition are here: http://www.kddcup-orange.com/winners.php?page=slow. You will notice that, on this page, the results obtained with the "small dataset" are mixed with the results of the "large dataset". I extracted the result obtained on the "small dataset" only and put them inside an excel file "KDD_results_small.xls" in attachement. Here is an graphical illustration of the scores for the best performers on the "small dataset" only:



scores of top performers

Using the TIM datamining software, Business-Insight obtained the 18th rank (on a total of over 1200 companies that took part in the competition) with a best score of 0.825 (see the chart above). Two other teams (Céline&Didier) participated to the competition using a very old beta-version of TIM and obtained the rank 54 and 64.

To summarize: for the "small dataset" challenge (higher means better):

Position	TEAM Name	Churn	Appetency	Upselling	Global
1	IBM research	76.51	88.19	90.92	85.21
18	Frank (final)	73.97	84.34	89.63	82.55
54	Céline (test 7)	72.30	81.47	85.84	79.87
64	Didier (grouped)	72.53	81.14	85.44	79.70

Datamining softwares are all about speed: if your software is faster, you can make more computations, use better parameter settings and, at the end, obtain better predictions. This is why the winner of the KDD this year is IBM (they were not using SPSS!): they don't have a good datamining software but they were using a large number of PC's and a very large crew: more than 15 people working full-time during 1 month, exclusively on this.

The companies that are inside the TOP20 of the KDD2009 ranking are all using (except Business-Insight) techniques that are extremely costly in terms of computing time and in terms of "man power". This is unrealistic. In real situations, in banks, in Telco, in insurance, you don't have such computing power or so much "man power". In opposition, we only used our own personal laptop to obtain all the results (...and I only worked on the competition

during the evenings). At the end, the results obtained are quite spectacular, especially when you take into account the very small computing power that was used.

This rank places the Business-Insight company as the best datamining company in Europe. 

Frank

Additionnal notes:

1. The small dataset is the most interesting dataset for this challenge because it allows to get a score of 85.21 while the large dataset only gives you a score of 84.93. ...and, of course, this is on the small dataset that the Business-Insight obtained the best results! The small dataset is also the dataset that resemble the most to "real-life" dataset usually available in enterprises.

2. The results obtained on the large dataset were all obtained in 5 days. Given more time (and more CPU power!), we could have obtained higher scores.

3. You can notice that there is a strong difference of performance between the TOP20 and the rest of the competitors: you can easily see that in the excel file in attachment.

Business-Insight top winner at the PAKDD2007

The PAKDD2007 is the 11th Pacific-Asia Knowledge Discovery and Data Mining conference. These results were published the 1st may 2007.

Problem description:

Cross-Selling: Credit Card 2 Mortgage. We try to sell a “mortage” to customers that already have a “credit card”.

Difficulty: Target is small: 1.71%

Data:

40,700 Card customers with the company within a specific 2-year

700 Targets (1.71%): opened a home loan with the company within 12 months after opening the credit card

40 modelling variables

Evaluation:

Prediction dataset without target (8000 customers)

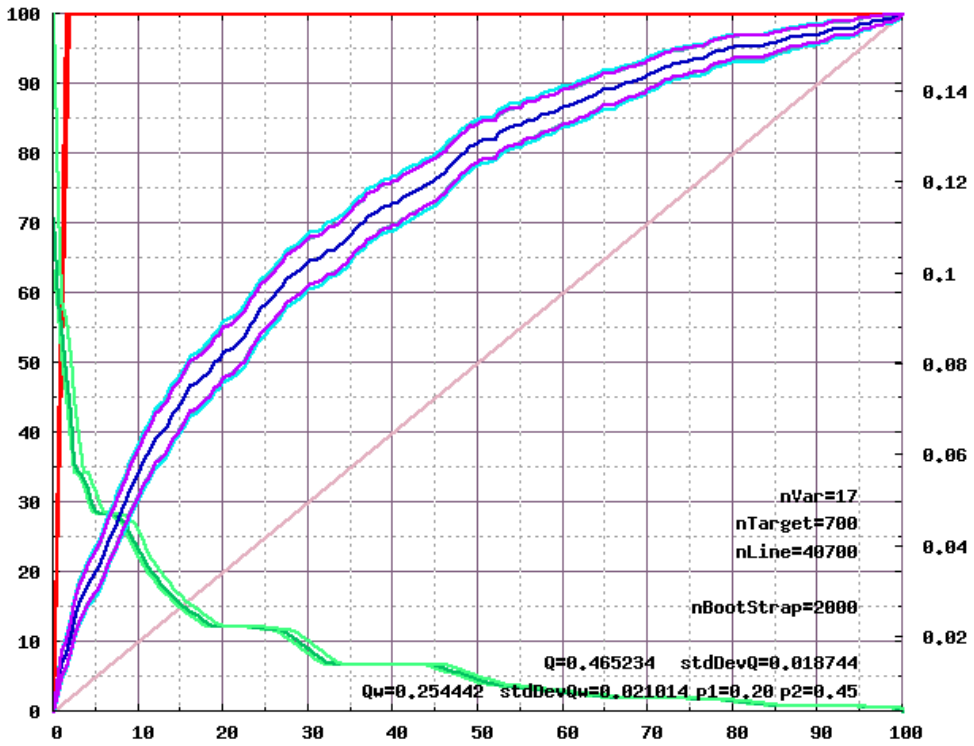
Criterion: Area under the lift curve (AUC)

TIM Model:

17 variables.

Quality on build (AUC) ~= 73.26 % ± 1.8

The lift of the TIM model:



The ancestor of TIM obtained the 6th RANK of the PAKDD2007 competition (amongst 47 competitors).

ID & Link to Report	Prediction AUC *	Rank	Modeling Technique	Remark
P049	70.01%	1	TreeNet + Logistic Regression	**Grand Champion (Tie)**
P085	69.99%	2	Probit Regression	**Grand Champion (Tie)**
P212	69.62%	3	MLP + n-Tuple Classifier	**First Runner-Up (Tie)**
P054	69.61%	4	TreeNet	**First Runner-Up (Tie)**
P088	69.42%	5	TreeNet	**In Top 10**
P248	69.28%	6	Ridge Regression	**In Top 10**
P134	69.14%	7	2-Layer Linear Regression	**In Top 10**
P126	69.10%	8	Logistic Regression + Decision Stump + AdaBoost + VFI	**In Top 10**
P227	68.85%	9	Logistic Average of Single Decision Functions	**In Top 10**
P178	68.69%	10	Logistic Regression	**In Top 10**
P249	68.58%	11	Unspecified Ensemble	**In Top 20**
P056	68.54%	12	Decision Tree + Neural Network + Logistic Regression	**In Top 20**
P041	68.28%	13	Scorecard Linear Additive Model	**In Top 20**
P021	68.04%	14	Random Forest	**In Top 20**
P148	68.02%	15	Expanding Regression Tree + RankBoost + Bagging	**In Top 20**
P116	67.58%	16	Logistic Regression	**In Top 20**
P149	67.56%	17	J48 + BayesNet	**In Top 20**
P083	67.54%	18	Neural Network + General Additive	**In Top 20**
P172	67.50%	19	Decision Tree + Neural Network	**In Top 20**
P078	66.71%	20	Decision Tree + Neural Network + Logistic Regression	**In Top 20**