

KDD Cup 2009 presentation: University of Melbourne

Hugh Miller, Sandy Clarke, Stephen Lane, Andrew Lonie, David Lazaridis, Slave Petrovski and Owen Jones

June 2009

Competition results



Competition results with predictions



Challenge Introduction

Some of the main features of the data that needed consideration:

- Large number of observations
- Large number of predictors, many of limited value
- Missing values
- Continuous and categorical values
- Categorical variables with a large number of levels
- Continuous distributions with peculiar distributions
- Variable interactions
- Class imbalances

Feature selection

Our approach required a substantial reduction in dimensionality. We ranked the predictors for each of the three responses according to the following procedure:

- The training data was split into two halves. Half was used to generate predictions.
- The AUC of these predictions on the second half was calculated.
- For categorical predictors, the prediction was the mean of the responses for each level.
- For continuous predictors, the variable was split into bins based on 1% quantiles and took the mean of response.
- The procedure was repeated and results averaged to increase reliability.

Feature selection

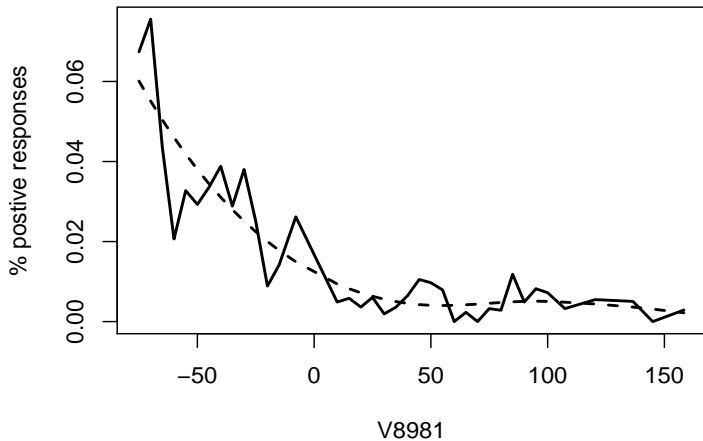
The approach was very simple, but had the following advantages:

- Speed
- Stability with respect to scale
- Comparability between categorical and continuous variables
- Ability to detect nonlinear relationships

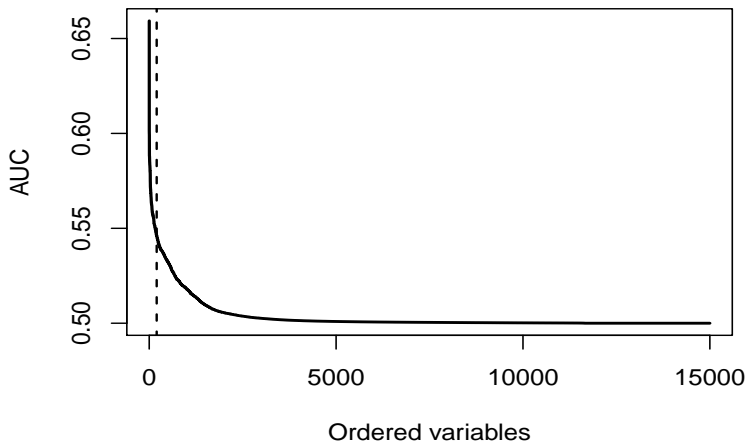
Appetency and Churn feature selections produced good results.
Upselling was more variable.

Feature selection

Figure shows the quantile fit for the most important variable in the Churn model.



AUC scores for predictors in Churn model



Treatment of categorical variables

Having categorical variables with a large number of levels proved highly undesirable. We attempted to aggregate the levels. If a categorical variable had > 25 levels, we replaced it:

- kept any levels with exposure > 1000 observations worth of exposure.
- aggregating any levels with exposure between 500-999 together
- aggregating any levels with exposure between 250-499 together
- aggregating any levels with exposure between 1-250 together.

Producing final models

For each response we built a gradient boosting machine:

- Shallow decision trees were used as a base learner.
- Incorporated shrinkage to increase robustness.
- Used Bernoulli loss.
- Response classes were weighted to better balance trees.
- Aimed to model around 200 predictors.
- Models were tuned using cross-validation and the 10% feedback on the test dataset.

Producing final models

Decision trees have a number of features that make them suitable for this year's competition

- Handling of missing variables.
- Robustness against extreme values.
- Handling categorical and continuous variables.
- Models interactions between predictors.
- Can model nonlinear dependencies.

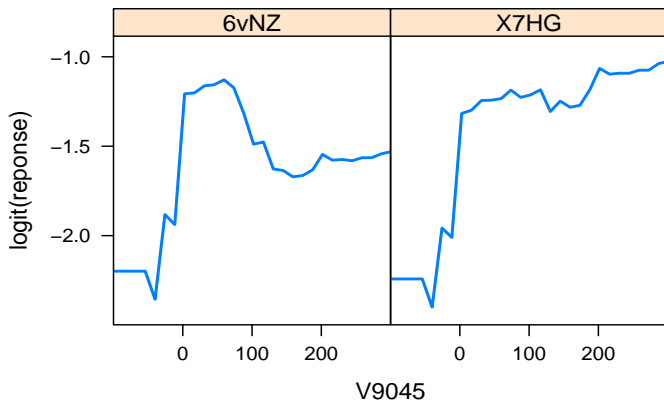
Some model parameters

	Model		
	Churn	Appetency	Upselling
Number of variables	198	196	201
Class weight	12	20	12
Shrinkage parameter	0.01	0.01	0.01
Number of trees	1300	1300	3000
Tree depth	5	3	5

Table: Model parameters for boosted tree models

Interaction example

Partial expectation plot of the logit transform of the probability of a positive response, by V9045 across two levels of V14990.



Relative variable importance

Rank	Churn		Appetency		Upselling	
	Name	Rel. Inf.	Name	Rel. Inf.	Name	Rel. Inf.
1	V8981	20.13	V9045	23.78	V9045	45.52
2	V14990	10.25	V8032	13.56	V14990	7.86
3	V10533	4.65	V14995	10.79	V8981	5.32
4	V14970	4.60	V14990	6.07	V12507	4.96
5	V5331	2.36	V5826	3.72	V6808	4.65
6	V14995	2.19	V8981	3.23	V1194	2.58
7	V14822	2.10	V10256	3.03	V14970	2.16
8	V9045	2.00	V12641	2.72	V14871	1.33
9	V2570	2.00	V14772	1.72	V1782	1.15
10	V14923	1.88	V14939	1.68	V10256	1.05
11	V14765	1.19	V14867	1.62	V5026	0.96
12	V14904	1.14	V14970	1.42	V8032	0.91
13	V5702	1.13	V11781	1.14	V14786	0.81
14	V11047	1.12	V14871	0.89	V7476	0.62
15	V14778	0.97	V14788	0.86	V11781	0.59

Computational details

- Almost all work done in R, with a small amount in SAS.
- The gbm package was used for the boosted models.
- Computer that fit the most models had 2.66Ghz Core 2 Duo processor, 2Gb Ram.