A Combination of Boosting and Bagging for KDD Cup 2009 – Fast Scoring on a Large Database

Jianjun Xie

Viktoria Rojkova

Siddharth Pal

Stephen Coggeshall

ID Analytics, Inc.

San Diego, California, USA

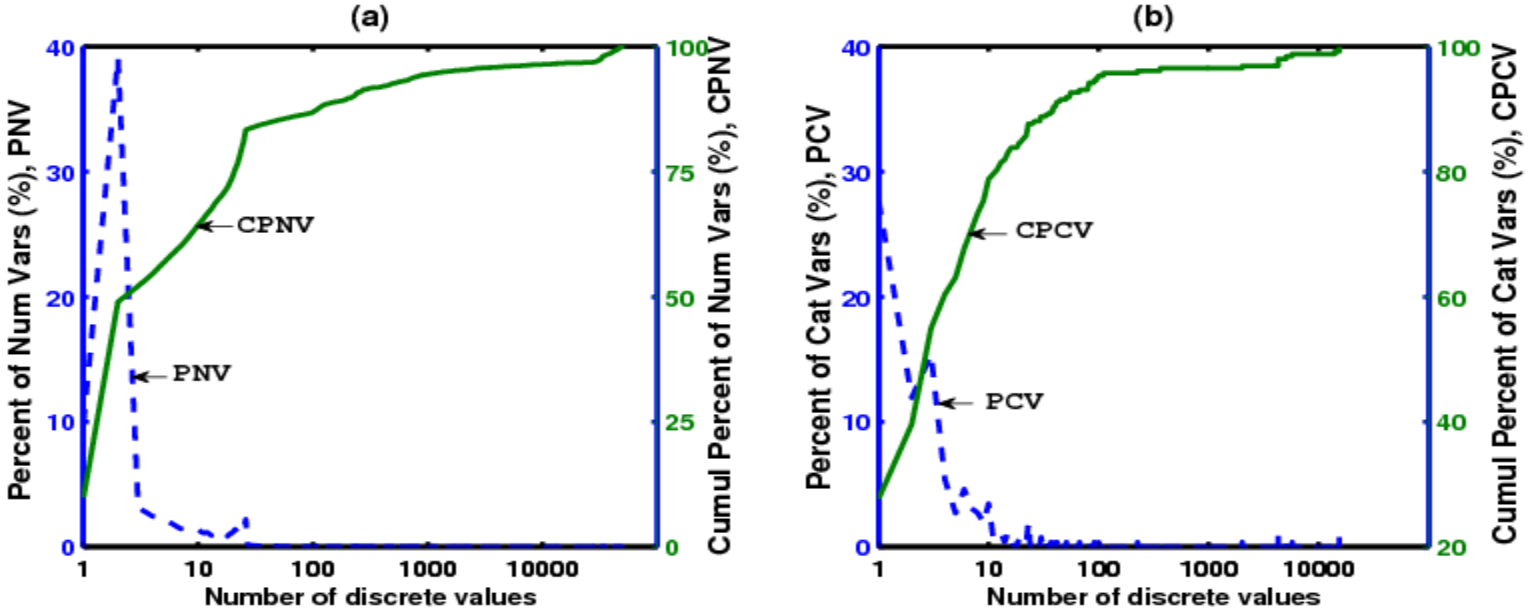
June 28, 2009

ID Analytics provides on-demand identity intelligence solutions:

- ID Score®
- ID Score®-Action
- ID Analytics® Credit Optics™
- ID Analytics for Authentication™
- ID Analytics for Compliance™
- ID Analytics for Data Defense™
- ID Network® Attributes
- My ID Monitoring and Alerts™

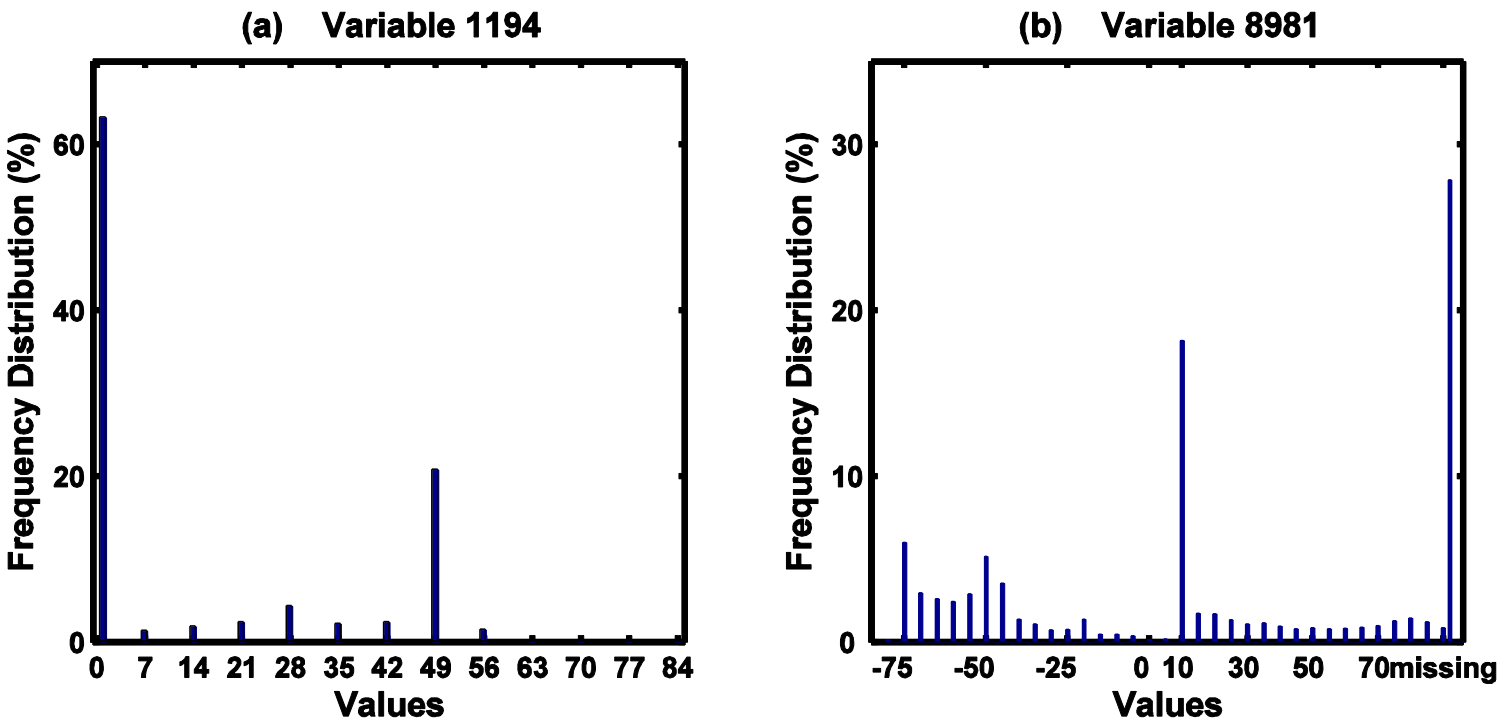
Data Analysis

- 50% of 14740 numerical variables have 1 or 2 discrete values
- 80% of 260 categorical variables have categories < 10
- 79% of 14740 numerical variables have > 98% population filled by 0



Histogram Patterns

- Check any sampling bias between training and testing
- Discover variable scramble logic



Improve Label Balance

Churn	Appetency	Up-selling	Frequency	Percentage
-1	-1	-1	41756	83.51%
-1	-1	1	3682	7.36%
-1	1	-1	890	1.78%
1	-1	-1	3672	7.34%

Down-sampling rate for each modeling task.

Label	Down-sampling on negative examples	Positive rate after sampling
Churn	70%	10.17%
Appetency	20%	8.31%
Up-selling	90%	8.12%

Stochastic Gradient Boosting Tree (TreeNet)

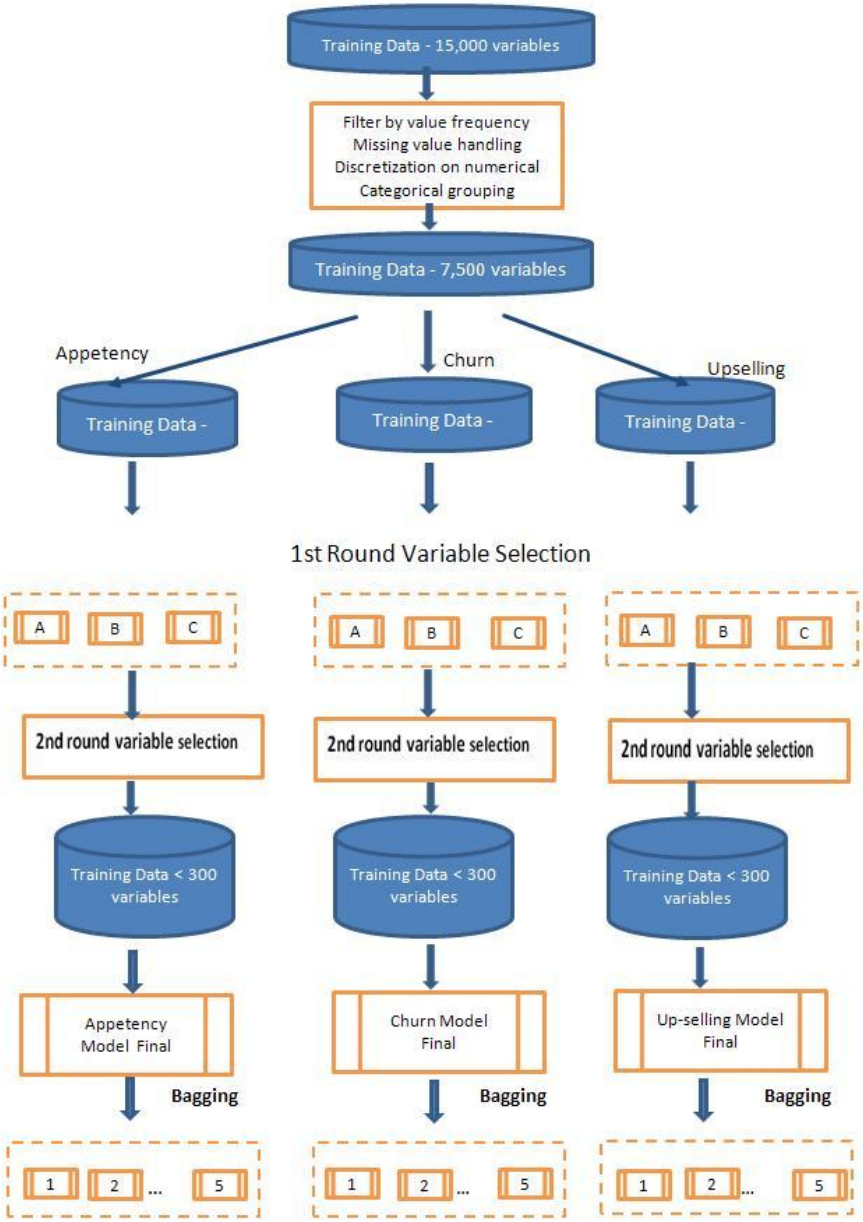
- Fits many small trees
- Uses a function of the errors as weights
- Stochastic gradient boosting
  - Randomly selects the sample for each iteration of the boosting
  - Tests the model on held out data to obtain best parameters

Combine Boosting and Bagging

- A single TreeNet model may not see complete picture of all data
  - Inside TreeNet: split training and testing, sampling at each tree split
  - Outside TreeNet: down-sampling on negative population to improve label balance
- Bagging Boosted tree models together get improved performance
  - Iterations of bootstrap sampling (typically 5)
  - Performance (AUC) improve 0.3-0.5% on most out-of-bagging validations over a single model

Modeling Workflow

1. Data preprocessing
2. Variable reduction
3. Variable selection
4. Single model building
5. Bagging



Modeling on Slow Track

- Model performance of small dataset is lower than that of large dataset
- Competition rule requires small model compete with large model
- Our strategy
  - Unscramble the small dataset
  - Map variables between small and large
  - Focus model on large dataset
  - Use small as Reference (additional 10% validation feedback)

Unscramble the Small Dataset

- Unscramble the variable mapping
  - Compare the frequency distribution of each variable (histogram)
  - Able to map 194 out of 230 variables by histogram alone.
- Unscramble the example order
  - Construct a key using mapped variables (key = $Var_i Var_j \ldots Var_n$)
  - Cut key out in original order: sequence ID, key
  - Sort key files by key
  - Paste sequence ID files together

Final Results

AUC results of our final models on test dataset.

Dataset	Churn		Appetency		Up-selling		Scores	
	10%	100%	10%	100%	10%	100%	10%	100%
Large (fast)	0.7333	0.7565	0.8705	0.8724	0.9308	0.9025	0.8354	0.8448
Large (slow)	0.7390	0.7614	0.8714	0.8761	0.9023	0.9061	0.8376	0.8479
Small (slow)	0.7612	0.7611	0.8544	0.8765	0.9155	0.9057	0.8437	0.8478

Key Factors to Achieve the Results

- Combination of boosting and bagging
- Variable preprocessing and selection
- Proper imbalanced data handling

id:analytics™