Competition II: Springleaf

Sha Li (Team leader)
Xiaoyan Chong, Minglu Ma, Yue Wang

CAMCOS Fall 2015
San Jose State University

Agenda

- Kaggle Competition: Springleaf dataset introduction
- Data Preprocessing
- Classification Methodologies & Results
  - Logistic Regression
  - Random Forest
  - XGBoost
  - Stacking
- Summary & Conclusion

Kaggle Competition: Springleaf

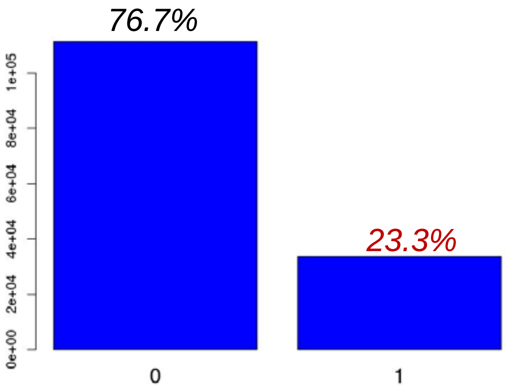
Objective: Predict whether customers will respond to a direct mail loan offer

- Customers: 145,231
- Independent variables: 1932
- “Anonymous” features
- Dependent variable:
  - target = 0: DID NOT RESPOND
  - target = 1: **RESPONDED**
- Training sets: 96,820 obs.
- Testing sets: 48,411 obs.

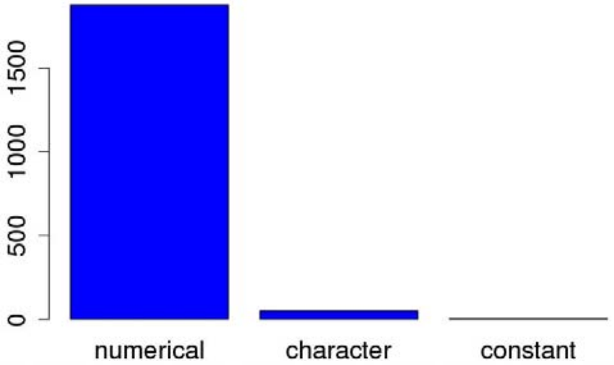
Dataset facts

- R package used to read file: *data.table::fread*

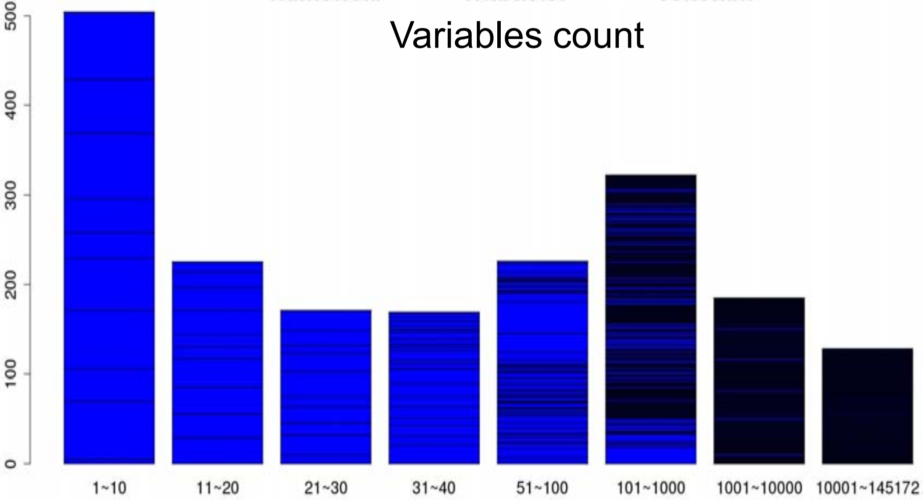
- Target=0 obs.: 111,458
- Target=1 obs.: 33,773
- Numerical variables: 1,876
- Character variables: 51
- Constant variables: 5
- Variable level counts:
  - 67.0% columns have levels <= 100



Class 0 and 1 count



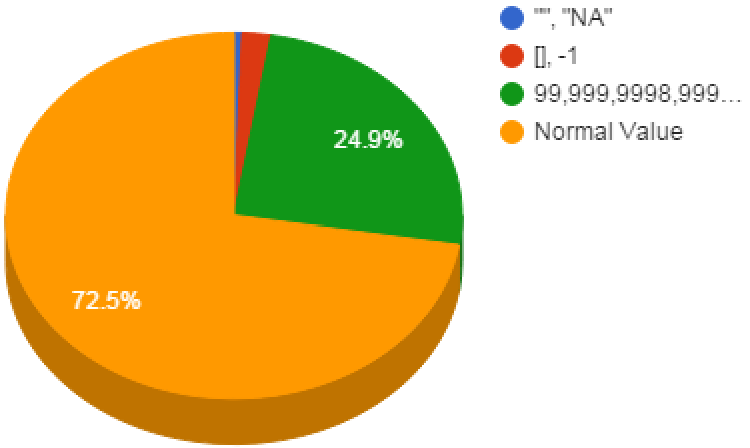
Variables count

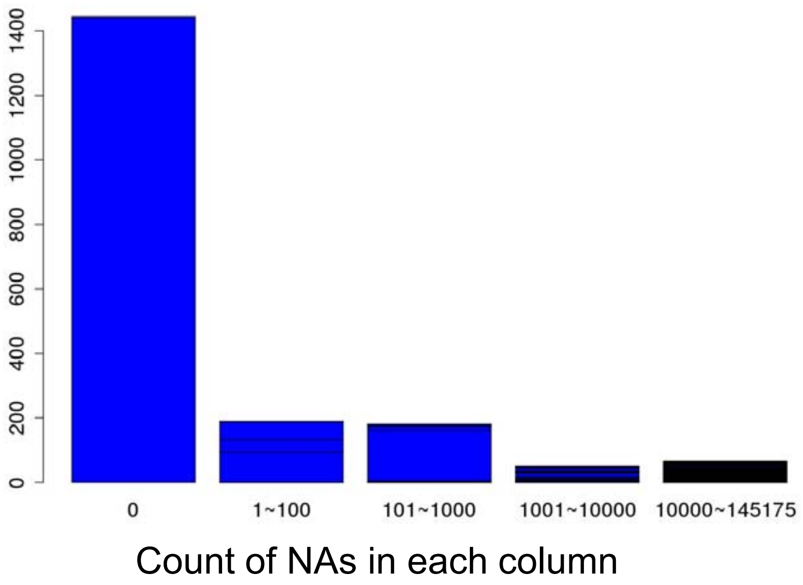


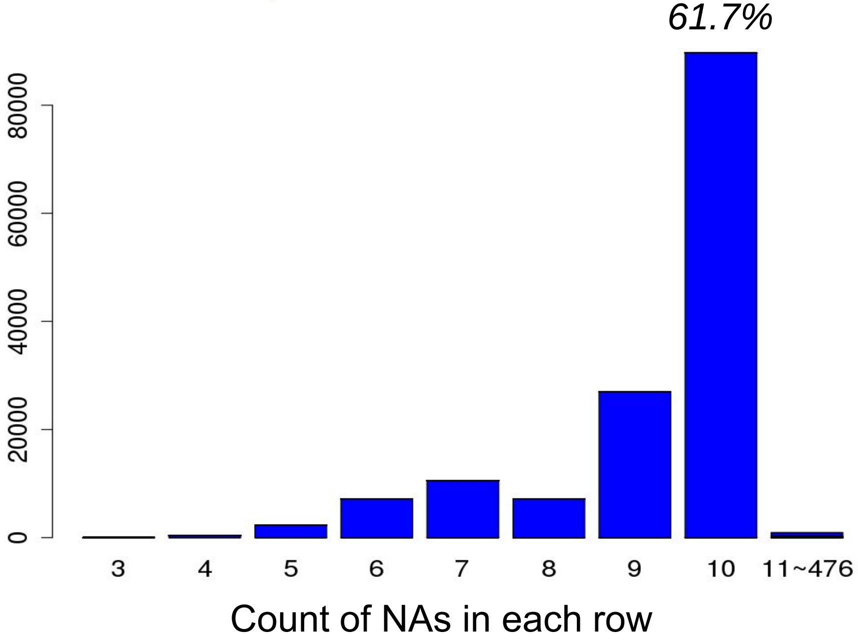
Count of levels for each column

Missing values

- “”, “NA”: 0.6%
- “[]”, -1: 2.0%
- -99999, 96, …, 999, …, 99999999: 24.9%
- 25.3% columns have missing values



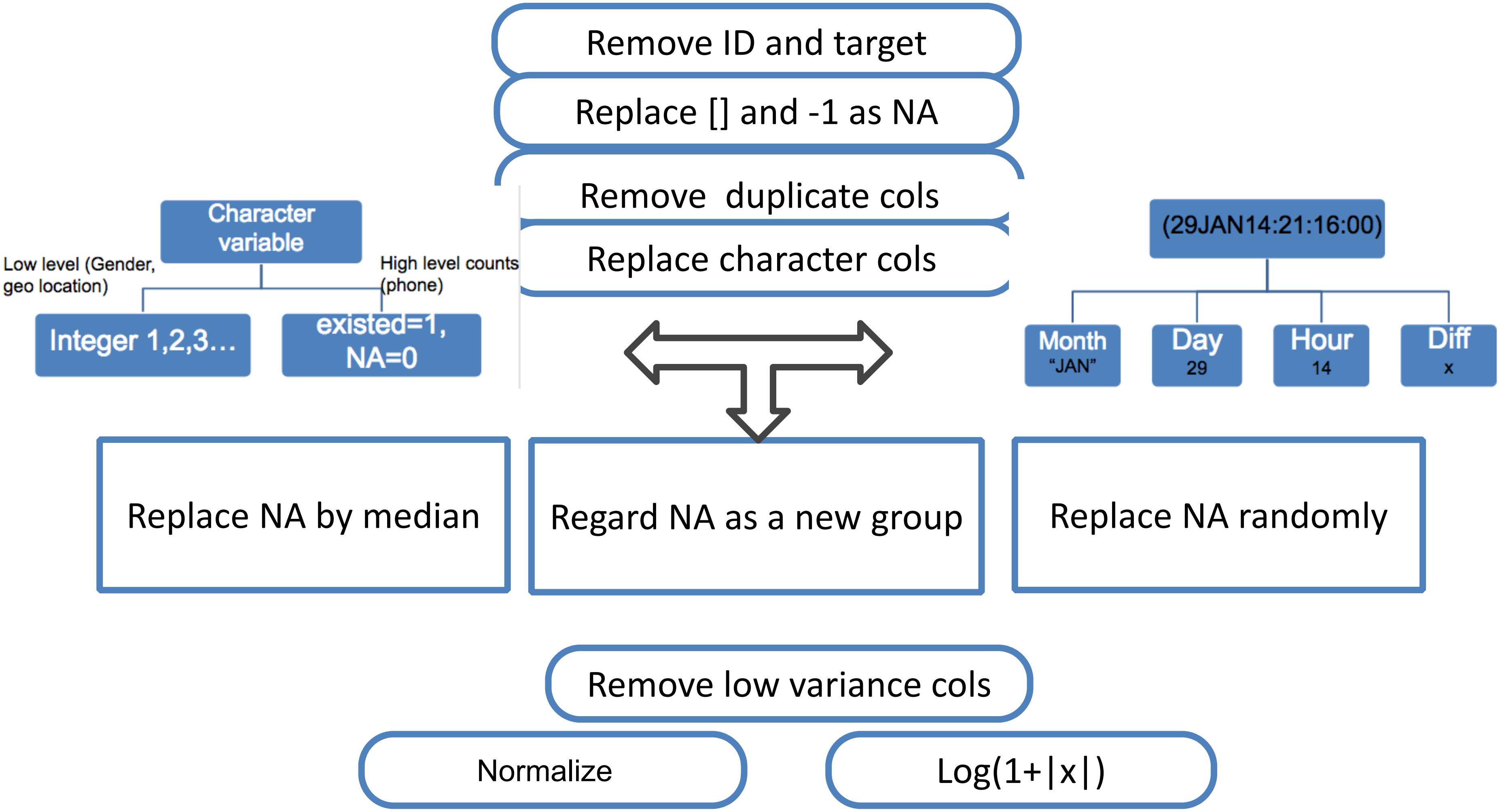




Challenges for classification

- Huge Dataset (145,231 X 1932)
- “Anonymous” features
- Uneven distribution of response variable
- 27.6% of missing values
- Deal with both numerical and categorical variables
- Undetermined portion of Categorical variables
- Data pre-processing complexity

Data preprocessing



Remove ID and target

Replace [] and -1 as NA

Remove duplicate cols

Replace character cols

Character variable

Low level (Gender, geo location)

High level counts (phone)

Integer 1,2,3…

existed=1, NA=0

(29JAN14:21:16:00)

Month “JAN”

Day 29

Hour 14

Diff x

Replace NA by median

Regard NA as a new group

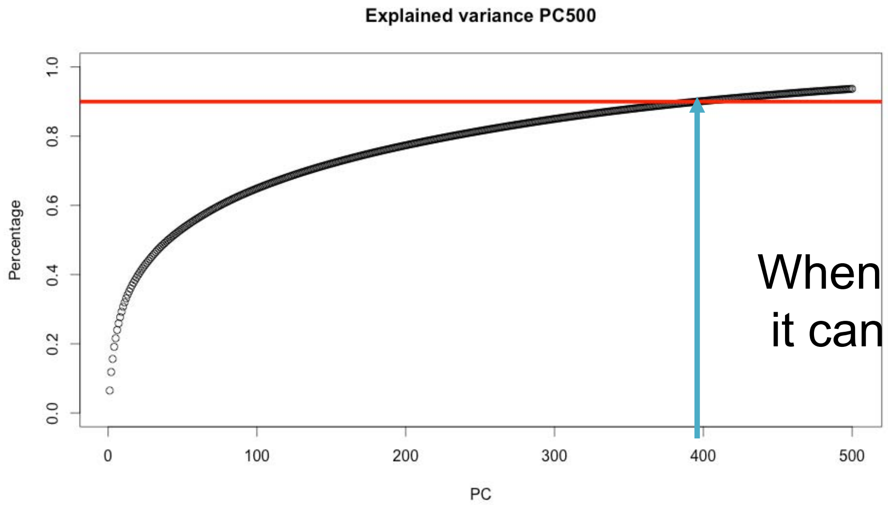
Replace NA randomly

Remove low variance cols

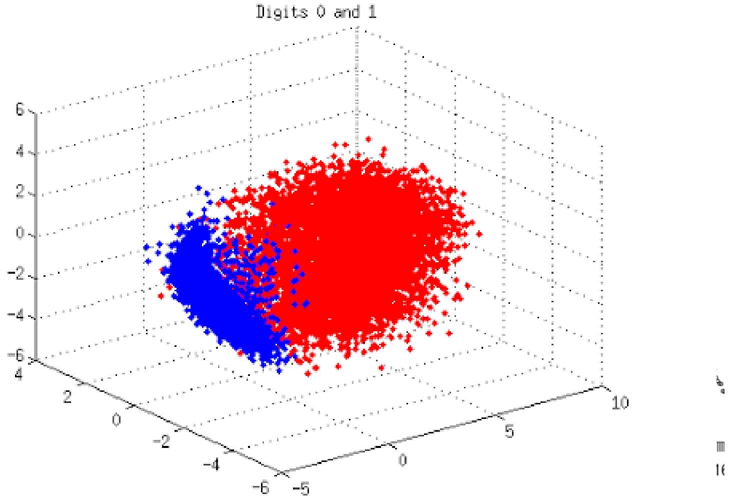
Normalize

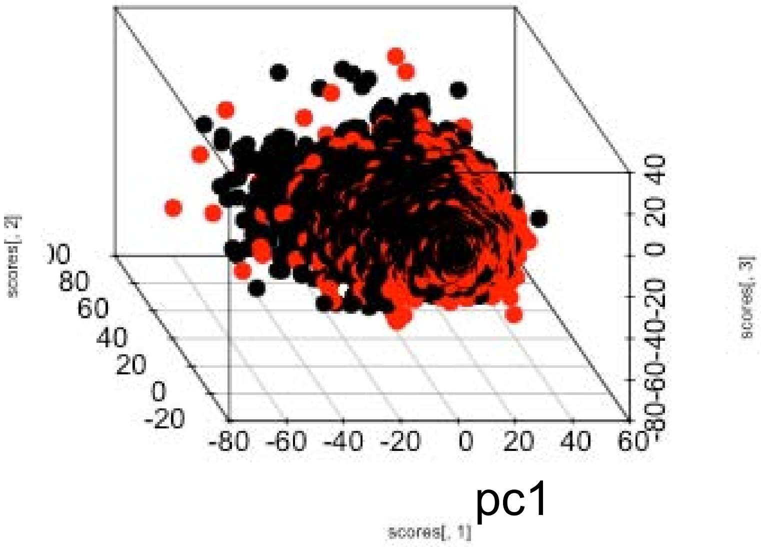
Log(1+|x|)

Principal Component Analysis



When PC is close to 400,
it can explain 90% variance.

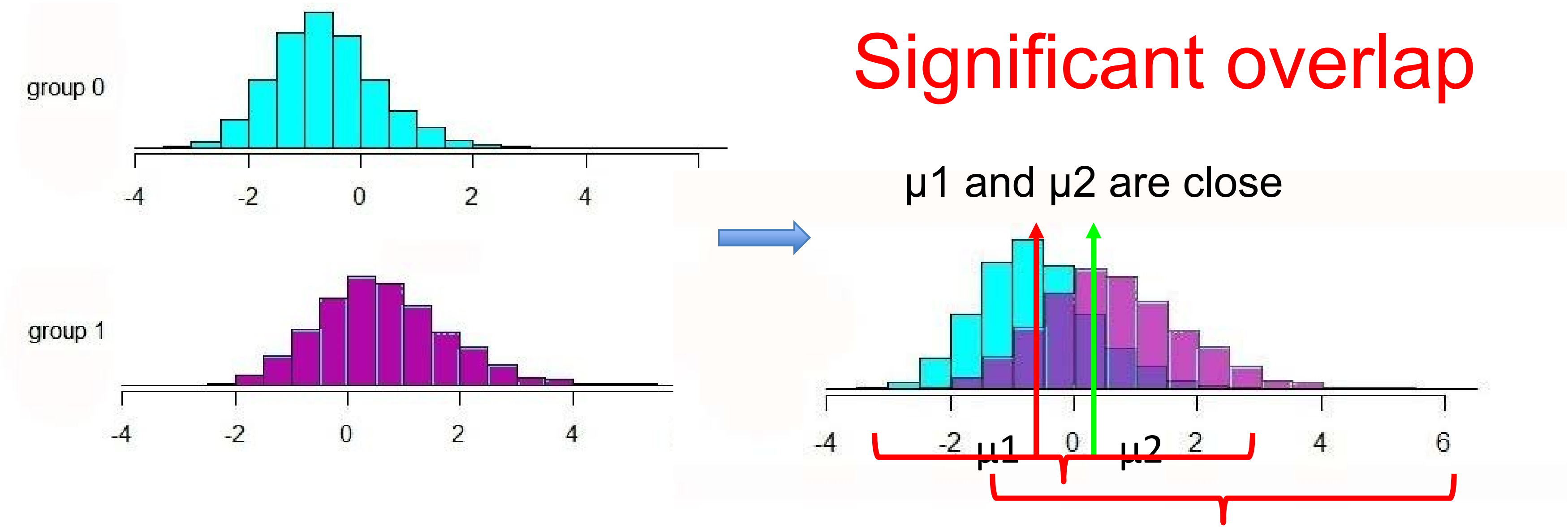




pc1

LDA: Linear discriminant analysis

- We are interested in the most discriminatory direction, not the maximum variance.
- Find the direction that best separates the two classes.



Significant overlap

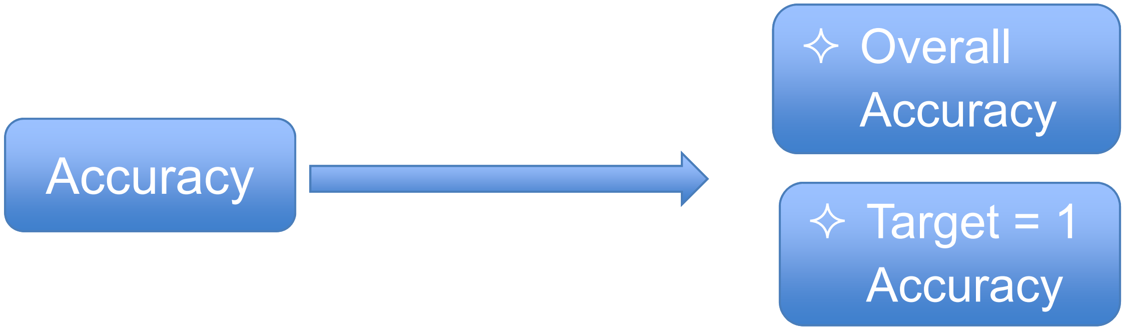
μ1 and μ2 are close

Var1 and Var2 are large

Methodology

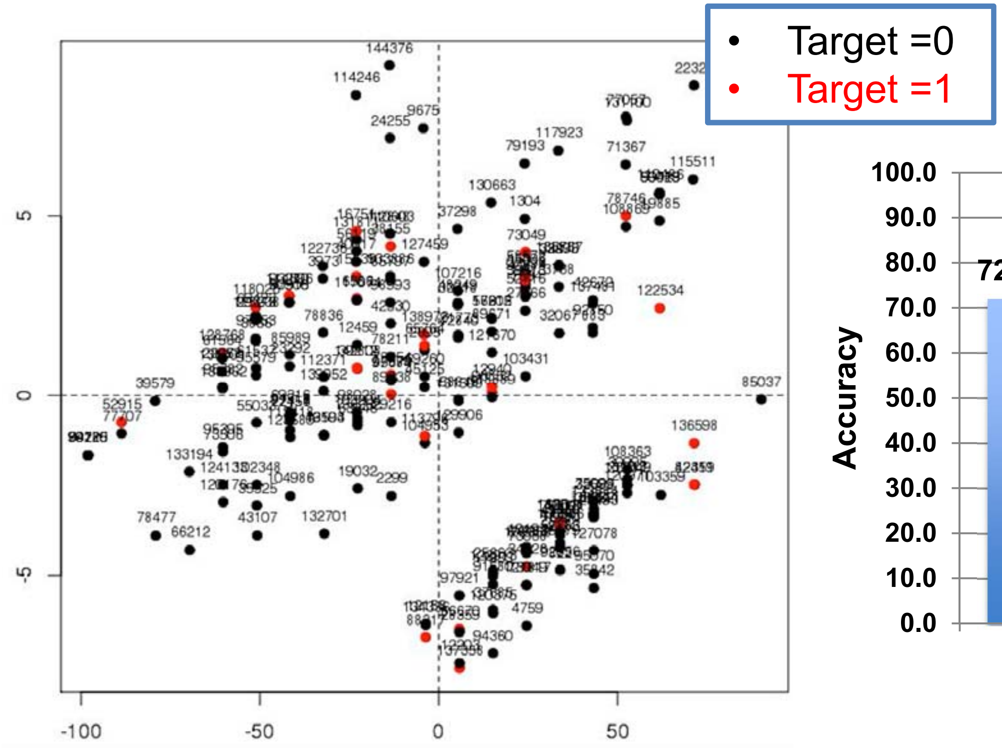
- K Nearest Neighbor (KNN)
- Support Vector Machine (SVM)
- Logistic Regression
- Random Forest
- XGBoost (eXtreme Gradient Boosting)
- Stacking

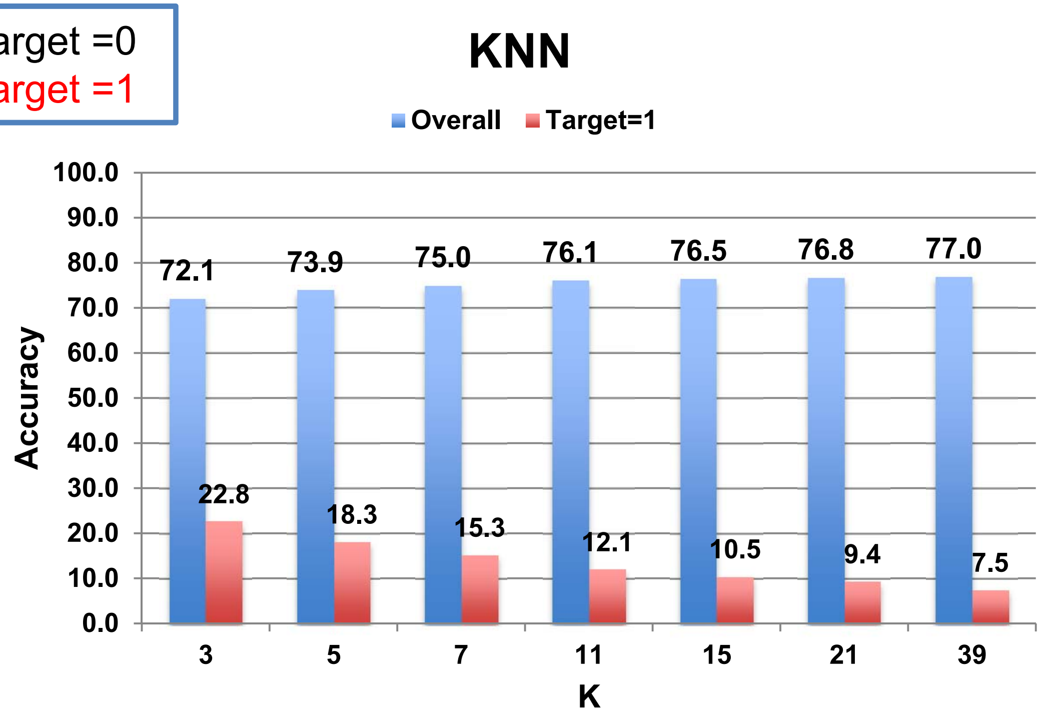
K Nearest Neighbor (KNN)



Accuracy

- Overall Accuracy
- Target = 1 Accuracy





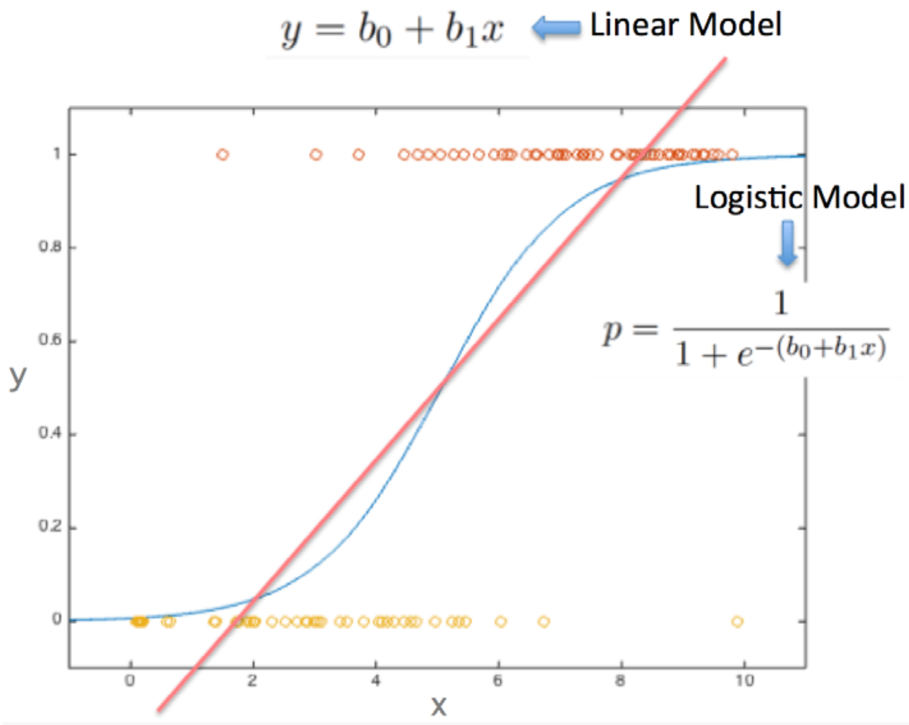
Support Vector Machine (SVM)

- Expensive; takes long time for each run
- Good results for numerical data

Confusion matrix		Prediction	
Truth		0	1
	0	19609	483
	1	5247	803

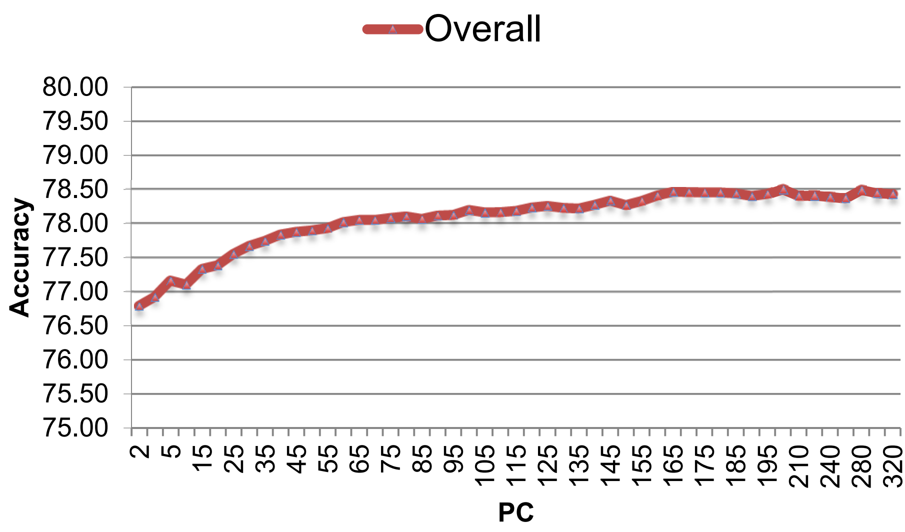
	Accuracy
Overall	78.1%
Target = 1	13.3%
Target = 0	97.6%

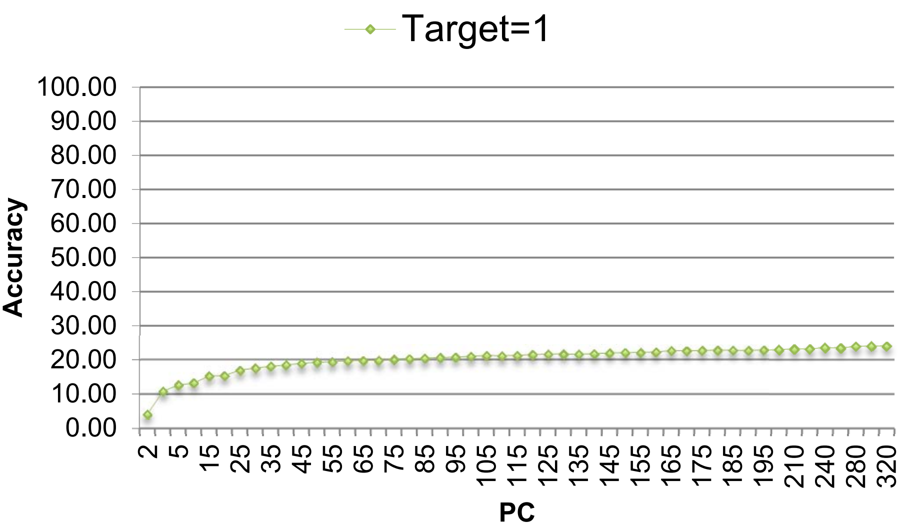
Logistic Regression



- Logistic regression is a regression model where the dependent variable is categorical.
- Measures the relationship between dependent variable and independent variables by estimating probabilities

Logistic Regression





Confusion matrix		Prediction	
Truth		0	1
	0	53921	3159
	1	12450	4853

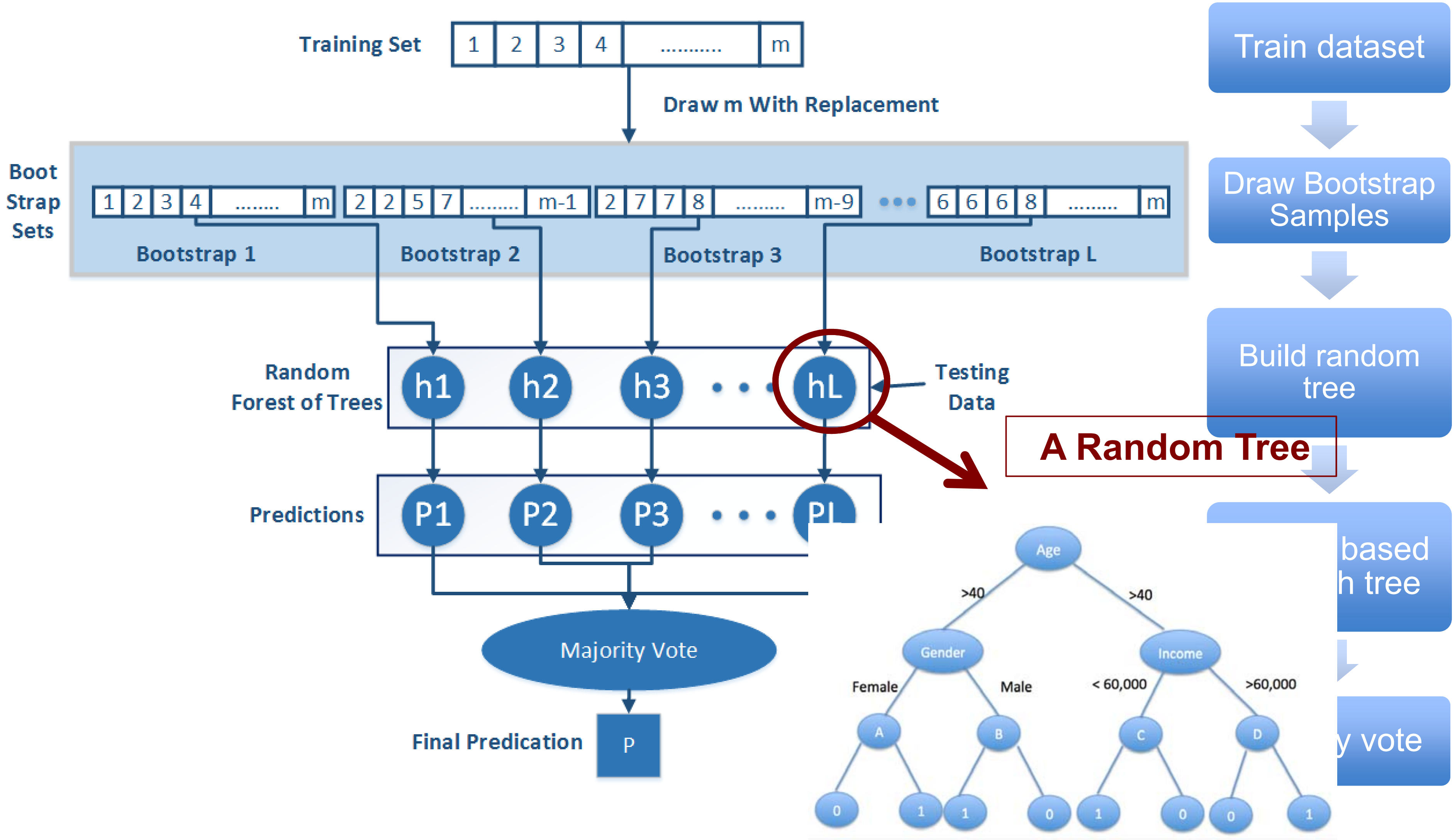
	Accuracy
Overall	79.2 %
Target = 1	28.1 %
Target = 0	94.5 %

Random Forest

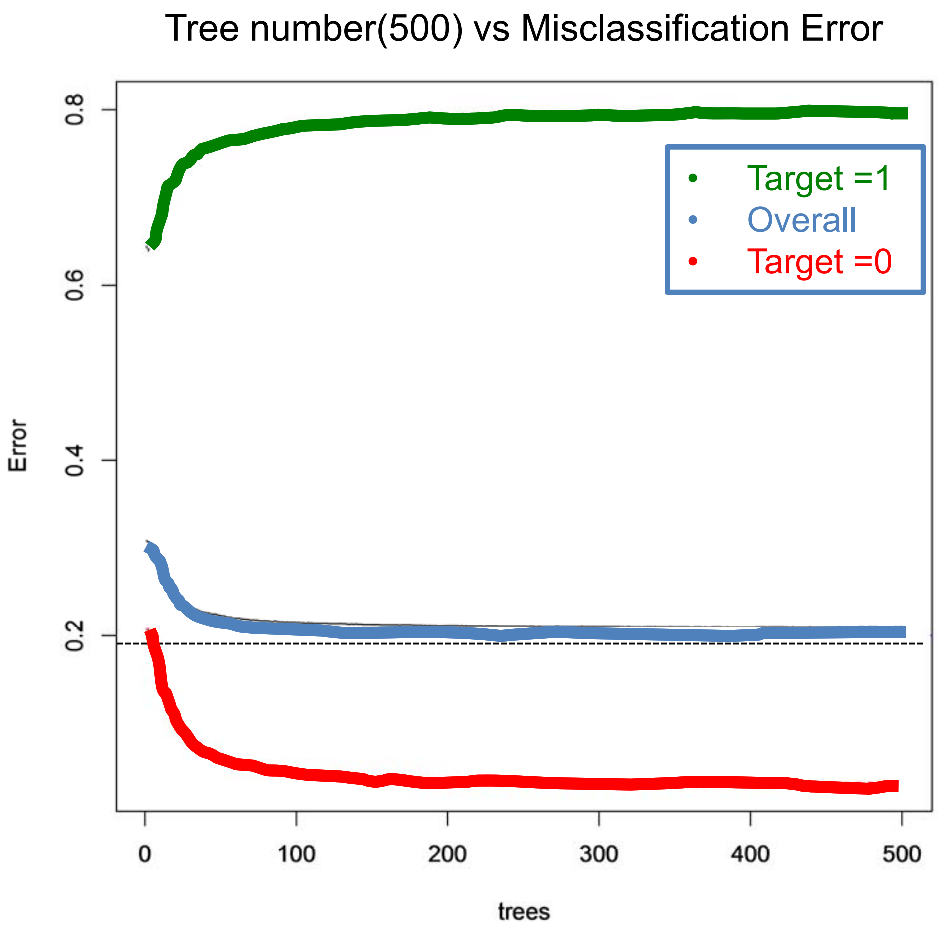
- Machine learning ensemble algorithm

  -- Combining multiple predictors
- Based on tree model
- For both regression and classification
- Automatic variable selection
- Handles missing values
- Robust, improving model stability and accuracy

Random Forest



Random Forest



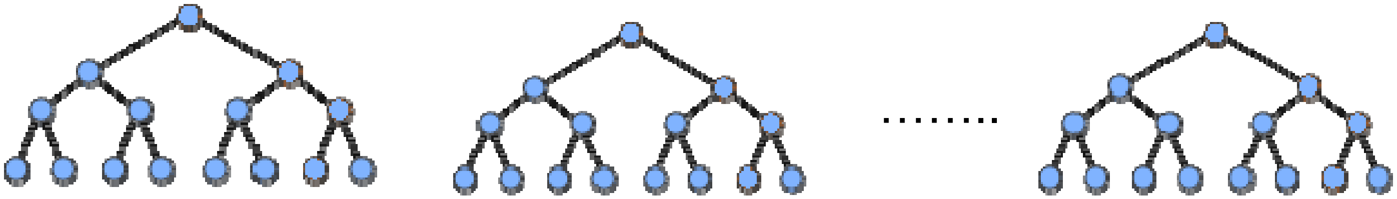
Confusion matrix		Prediction	
Truth		0	1
	0	36157	1181
	1	8850	2223

	Accuracy
Overall	79.3%
Target = 1	20.1%
Target = 0	96.8%

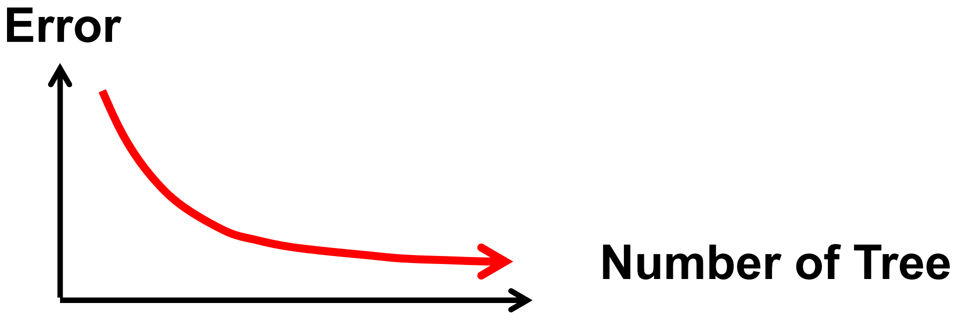
XGBoost

- Additive tree model: add new trees that complement the already-built ones
- Response is the optimal linear combination of all decision trees
- Popular in Kaggle competitions for efficiency and accuracy

Additive tree model

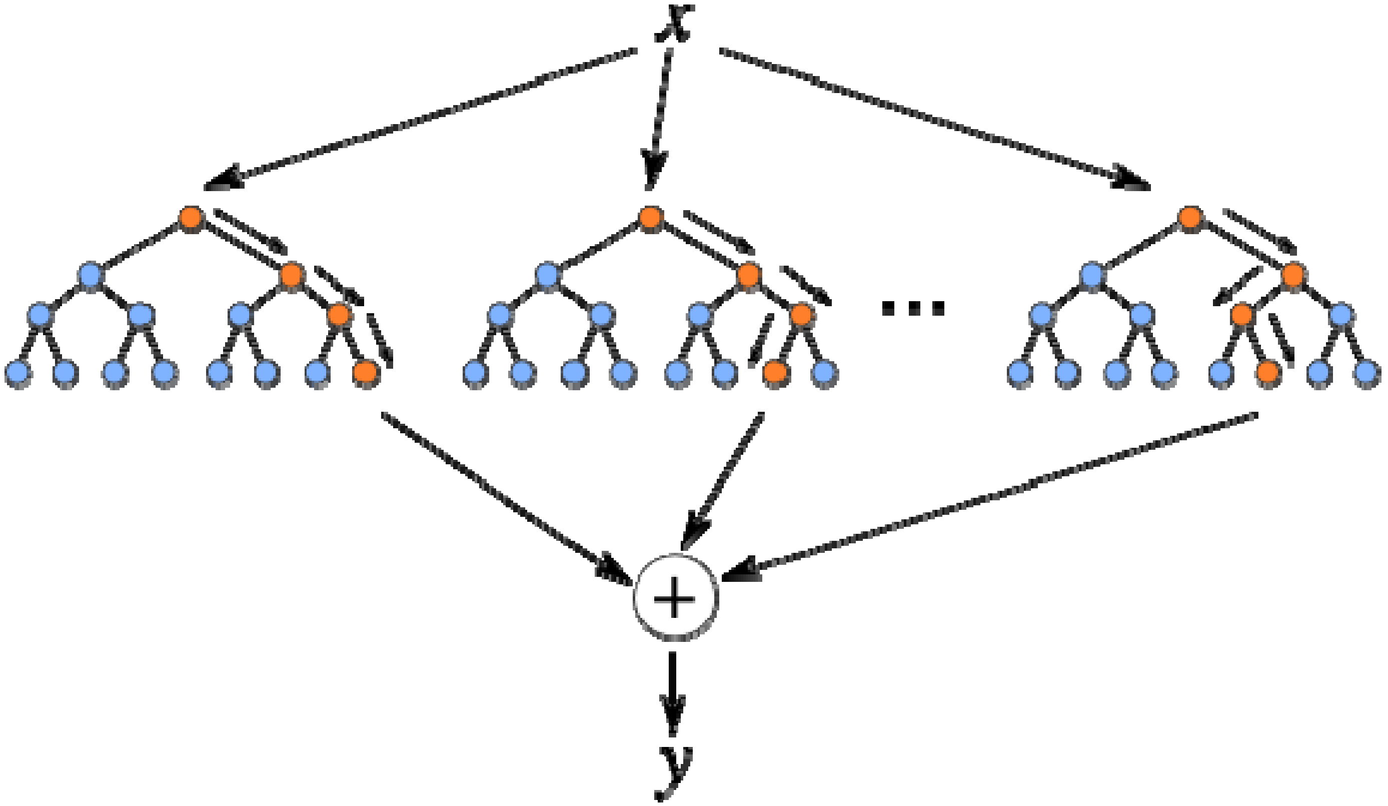


Greedy Algorithm

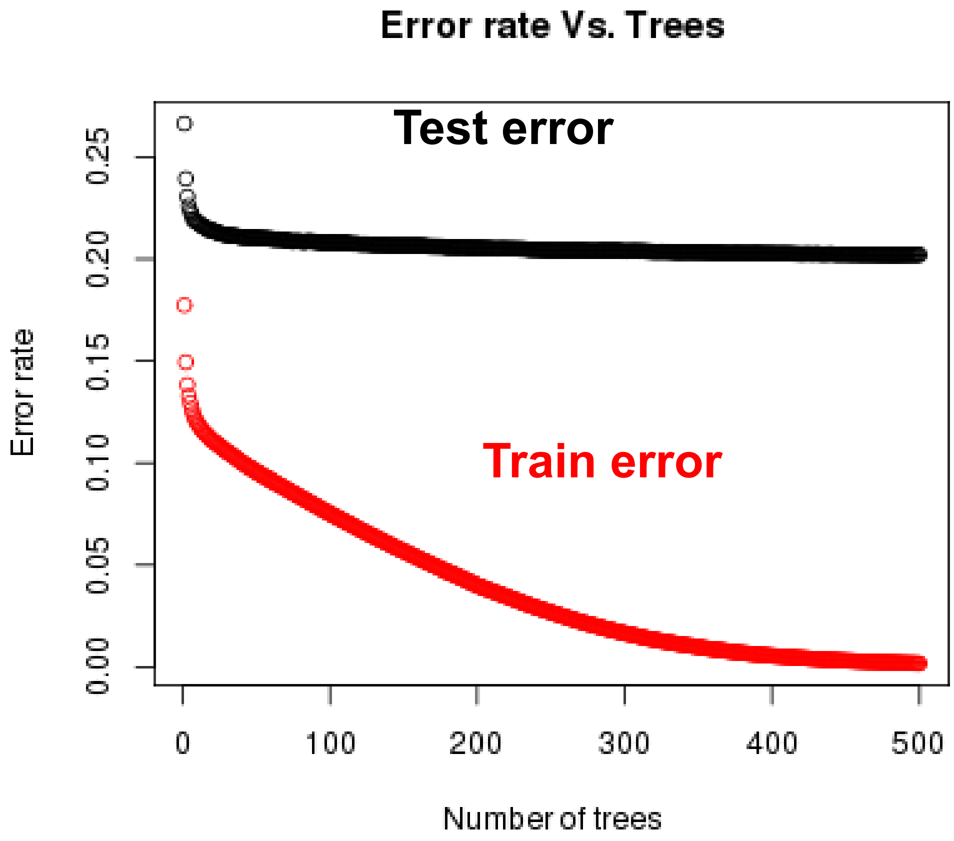


XGBoost

- Additive tree model: add new trees that complement the already-built ones
- Response is the optimal linear combination of all decision trees
- Popular in Kaggle competitions for efficiency and accuracy



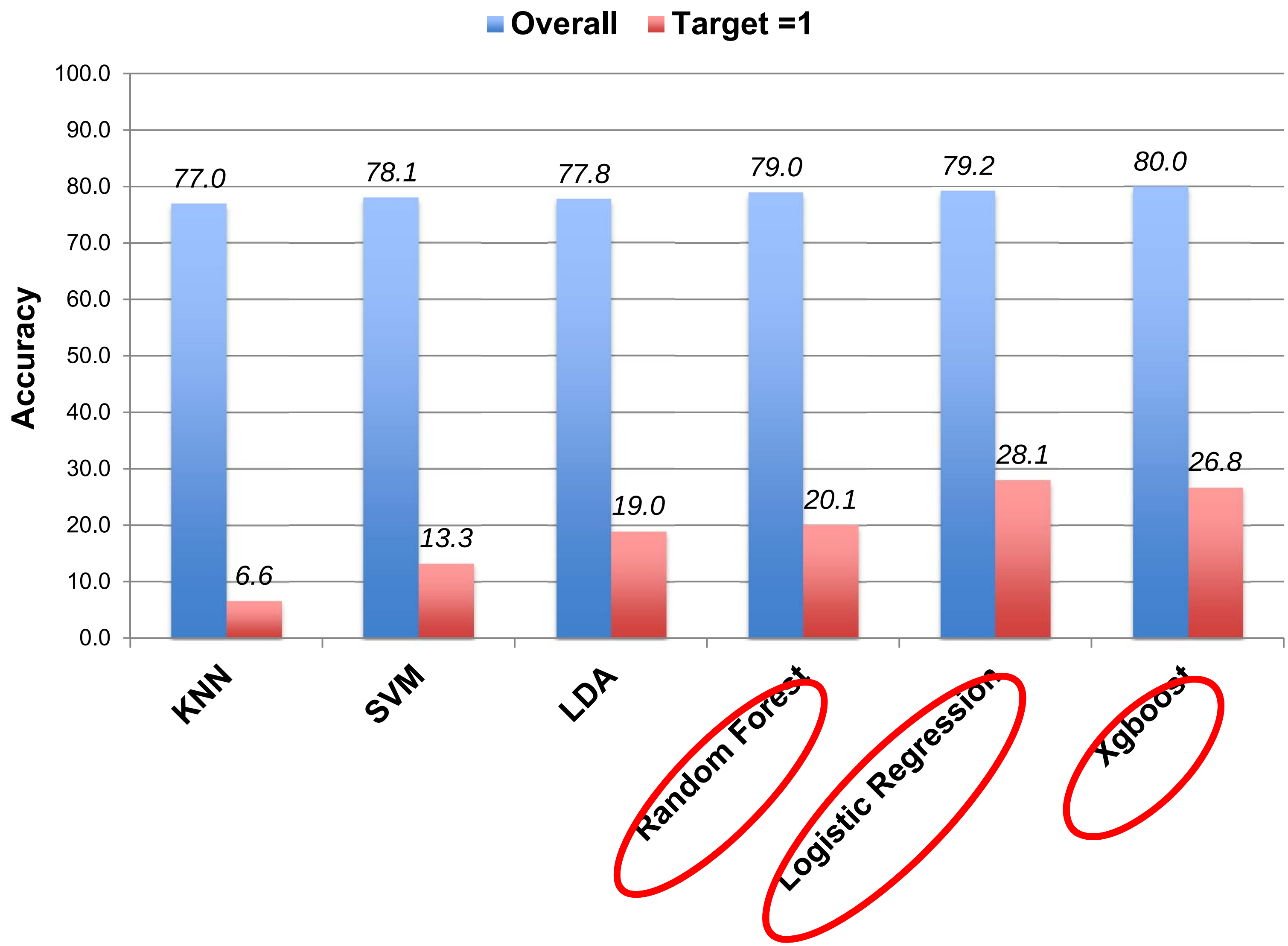
XGBoost



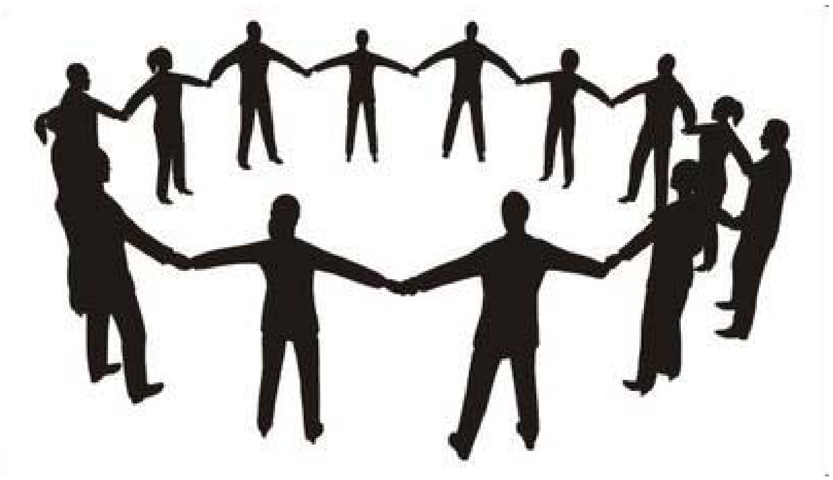
Confusion matrix		Prediction	
		0	1
Truth	0	35744	1467
	1	8201	2999

	Accuracy
Overall	80.0%
Target = 1	26.8%
Target = 0	96.1%

Methods Comparison

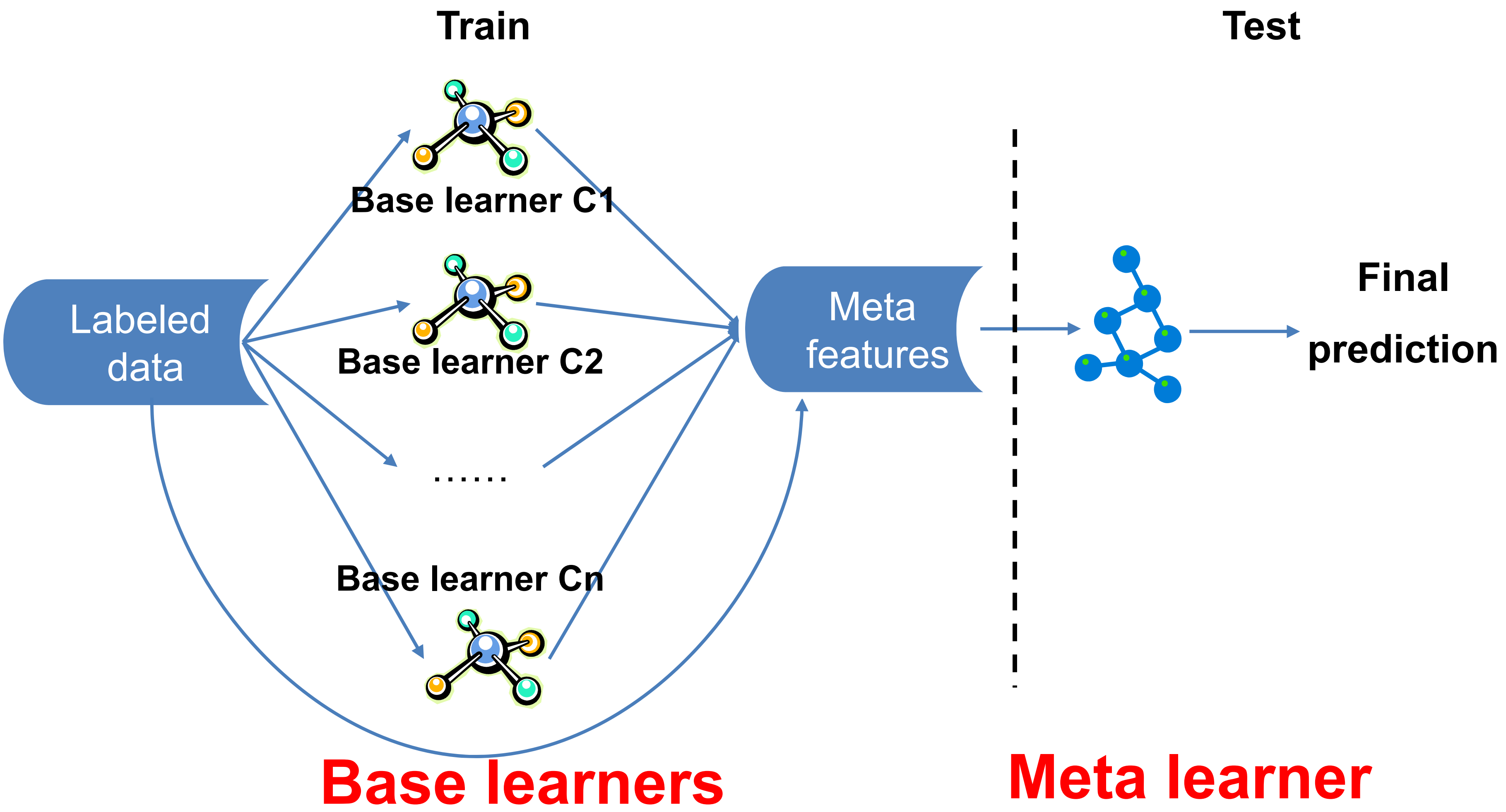


Winner or Combination ?



Stacking

- Main Idea: Learn and combine multiple classifiers



Train

Test

Labeled data

Base learner C1

Base learner C2

……

Base learner Cn

Meta features

Final prediction

Base learners

Meta learner

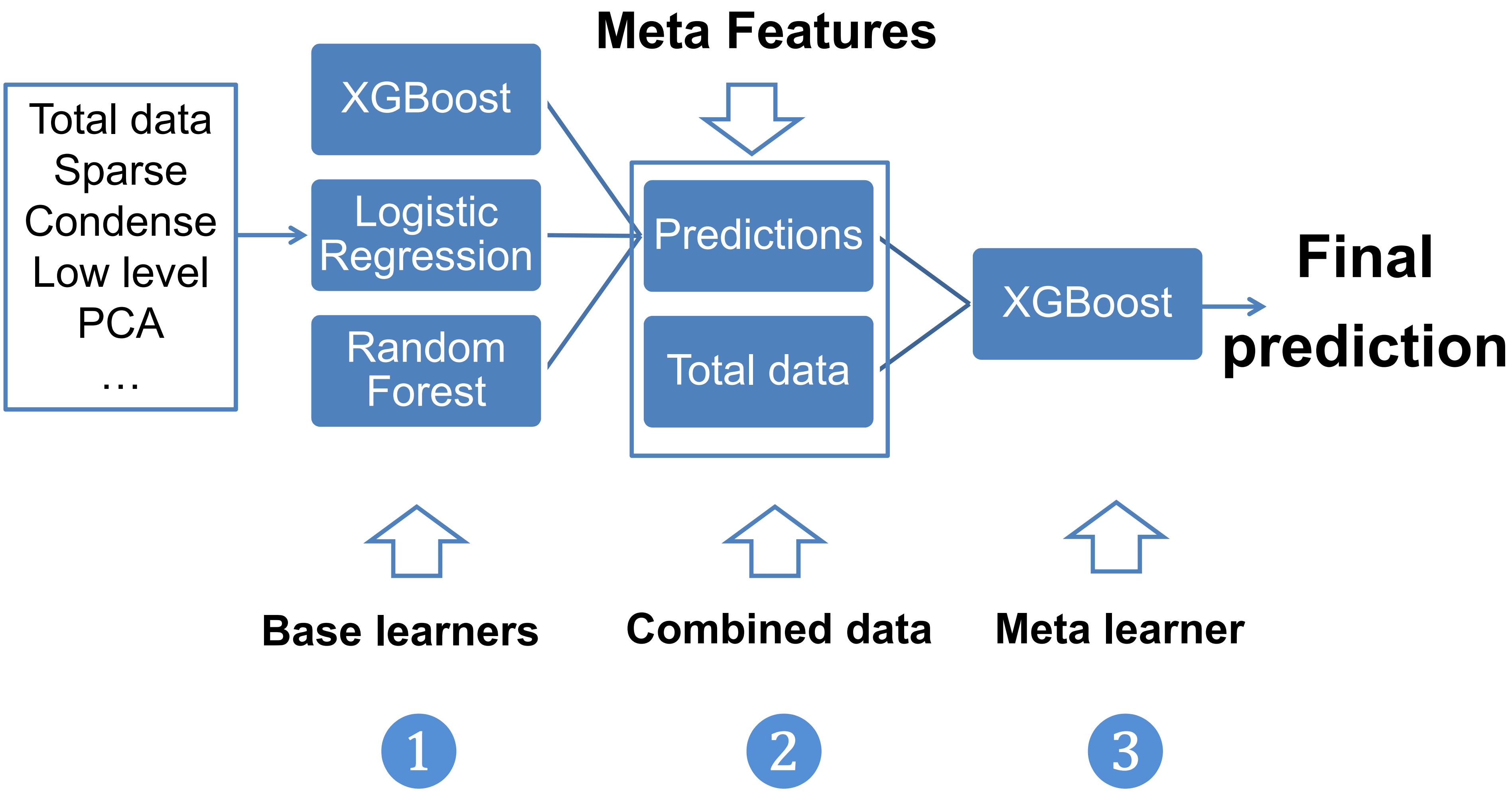
Generating Base and Meta Learners

- **Base model—efficiency, accuracy and diversity**
  - Sampling training examples
  - Sampling features
  - Using different learning models

- **Meta learner**
  - Majority voting
  - Weighted averaging
  - Kmeans

    Unsupervised
  - Higher level classifier — Supervised(XGBoost)

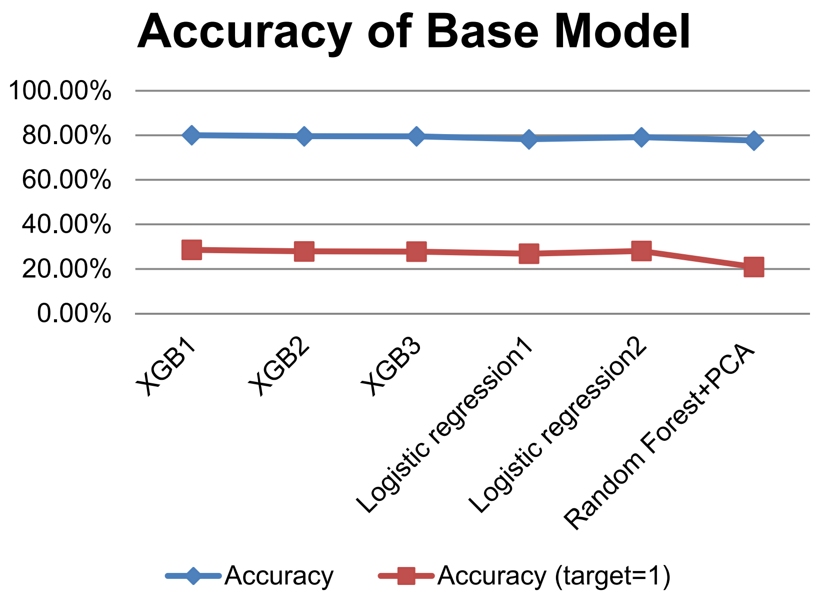
Stacking model

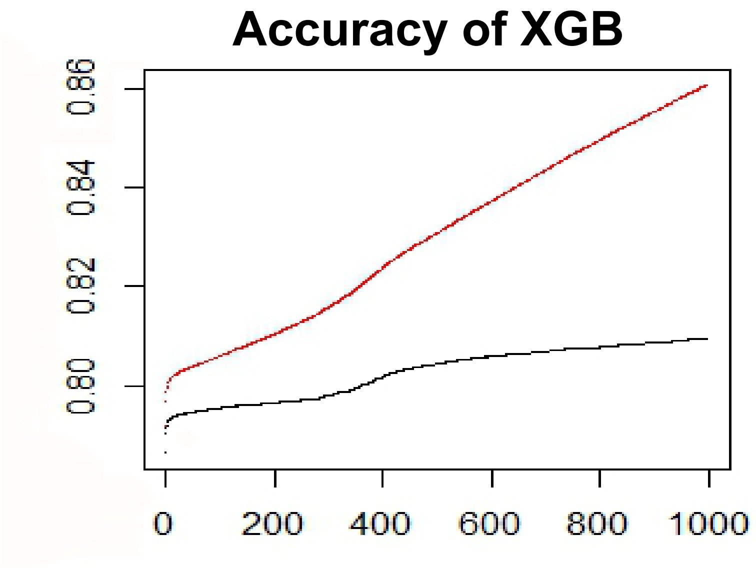


Stacking Results

Base Model	Accuracy	Accuracy (target=1)
XGB + total data	80.0%	28.5%
XGB + condense data	79.5%	27.9%
XGB + Low level data	79.5%	27.7%
Logistic regression+ sparse data	78.2%	26.8 %
Logistic regression+ condense data	79.1%	28.1%
Random forest + PCA	77.6%	20.9%

Meta Model	Accuracy	Accuracy (target=1)
XGB	81.11%	29.21%


Accuracy of Base Model

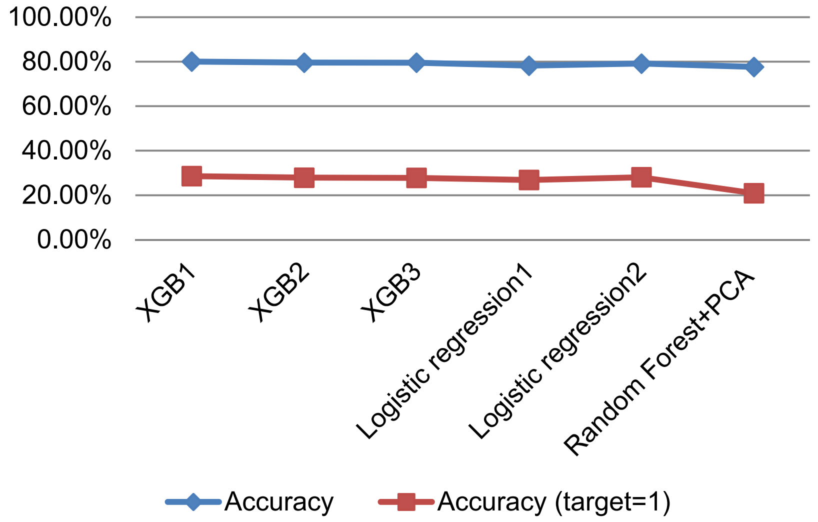

Accuracy of XGB

Stacking Results

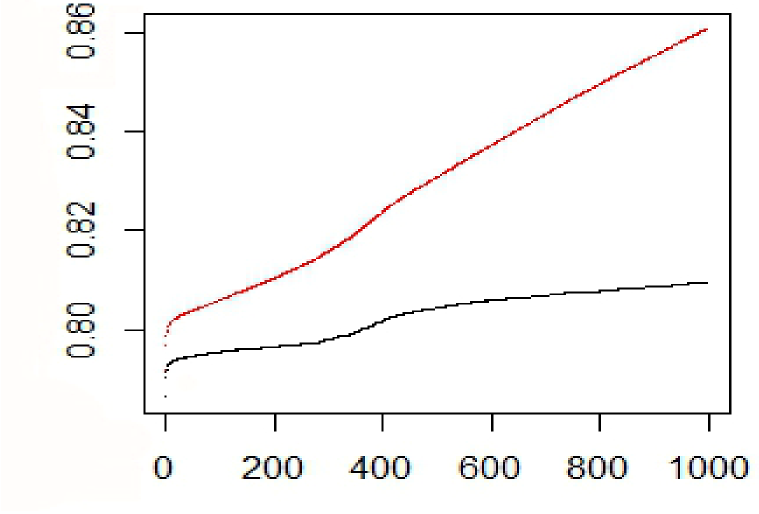
Base Model	Accuracy	Accuracy (target=1)
XGB + total data	80.0%	28.5%
XGB + condense data	79.5%	27.9%
XGB + Low level data	79.5%	27.7%
Logistic regression+ sparse data	78.2%	26.8 %
Logistic regression+ condense data	79.1%	28.1%
Random forest + PCA	77.6%	20.9%

Meta Model	Accuracy	Accuracy (target=1)
XGB	81.11%	29.21%
Averaging	79.44%	27.31%
Kmeans	77.45%	23.91%

Accuracy of Base Model



Accuracy of XGB



Summary and Conclusion

- Data mining project in the real world
  - Huge and noisy data
- Data preprocessing
  - Feature encoding
  - Different missing value process:

    New level, Median / Mean, or Random assignment
- Classification techniques
  - Classifiers based on distance are not suitable
  - Classifiers handling mixed type of variables are preferred
  - Categorical variables are dominant
  - Stacking makes further promotion
- Biggest improvement came from model selection, parameter tuning, stacking
- Result comparison： Winner result: 80.4%

  Our result: 79.5%

Acknowledgements

We would like to express our deep gratitude to the following people / organization:

- Profs. Bremer and Simic for their proposal that made this project possible
- Woodward Foundation for funding
- Profs. Simic and CAMCOS for all the support
- Prof. Chen for his guidance, valuable comments and suggestions

QUESTIONS

?