Benchmarking Tree Methods

Arie Slobbe and Anna Winkler

Data

Goal: Classify Individuals' Income as "≦50K" or ">50K"

Age	Race
Employer Type	Sex
Education Level	Capital Gain
Education Number	Capital Loss
Marital	Hours Per Week
Occupation	Relationship

Loss Function:

|false positive | = |false negative|

Data Source:

Barry Becher, 1994 Census Data

Incomplete Data:

Included in analysis

	age	type_employer	education_lev	education_num	marital	occupation	relationship	race	sex	capgain	caploss	hr_per_week	country
1	39	State-gov	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States
2	50	Self-emp-not-inc	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States
3	38	Private	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States
4	53	Private	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States
5	28	Private	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba
6	37	Private	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States
7	49	Private	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica
8	52	Self-emp-not-inc	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States
9	31	Private	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States
10	42	Private	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States
11	37	Private	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States
12	30	State-gov	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India
13	23	Private	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States
14	32	Private	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States
15	40	Private	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?
16	34	Private	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	Mexico
17	25	Self-emp-not-inc	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States
18	32	Private	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States
19	38	Private	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States
20	43	Self-emp-not-inc	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States
21	40	Private	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States
22	54	Private	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States
23	35	Federal-gov	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States
24	43	Private	11th	7	Married-civ-spouse	Transport-moving	Husband	White	Male	0	2042	40	United-States
25	59	Private	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States
26	56	Local-gov	Bachelors	13	Married-civ-spouse	Tech-support	Husband	White	Male	0	0	40	United-States
27	19	Private	HS-grad	9	Never-married	Craft-repair	Own-child	White	Male	0	0	40	United-States

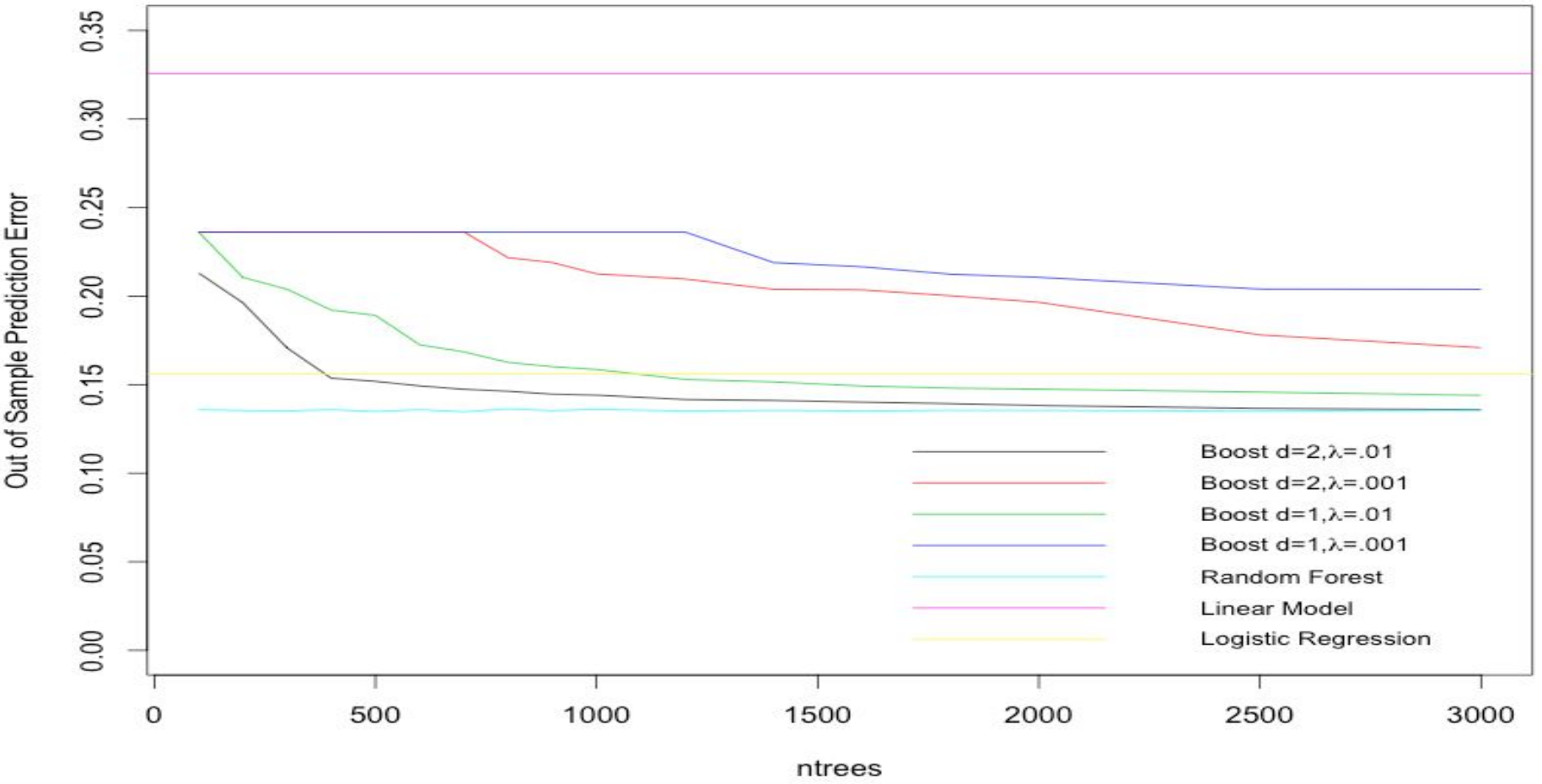
Problem

Task: Benchmark Tree Methods

Previous Methods:

Naive-Bayes	Ensemble Selection	Coupled Learning	MAPTAN + BMA	LPBoost
83.8%	91.3%	74.4%	82.1%	85.0%

Benchmark Results



Comparison

Logistic Regression	Boost d = 2, λ = .01	Random Forest
84.4%	86.4%	86.5%

Naive-Bayes	Ensemble Selection	Coupled Learning	MAPTAN + BMA	LPBoost
83.8%	91.3%	74.4%	82.1%	85.0%

Further Research

Random Forests:

- minimum size of terminal nodes
- vary sample size for bootstrap

Boosting

- cross validating depth and shrinkage interaction

Explore ways to deal with missing data

Sources

https://archive.ics.uci.edu/ml/datasets/Adult

An Empirical Evaluation of Supervised Learning for ROC Area. ROCAI. 2004 -- Ensemble Selection

ftp.esat.kuleuven.be/pub/SISTA/hamers/BH_clm.pdf - coupled kernels

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.518.4804&rep=rep1&type=pdf - LP Boost

http://digital.csic.es/bitstream/10261/3149/1/DSMAPTAN.pdf - MAPTAN