COMPARATIVE APPROACH ON ADULT DATA SETS

- Ben Mathew John(2012EET3000)
- Rahul Saluja(2012EET2984)

- Objective
- Data Set
- Understanding the dataset
- Feature Reduction
- Classification Results

Objective

- Given a set of features, whether we can predict the income of a person.
- Role of different features and intuitively interpret their importance's.
- How the various classification algorithms performs.

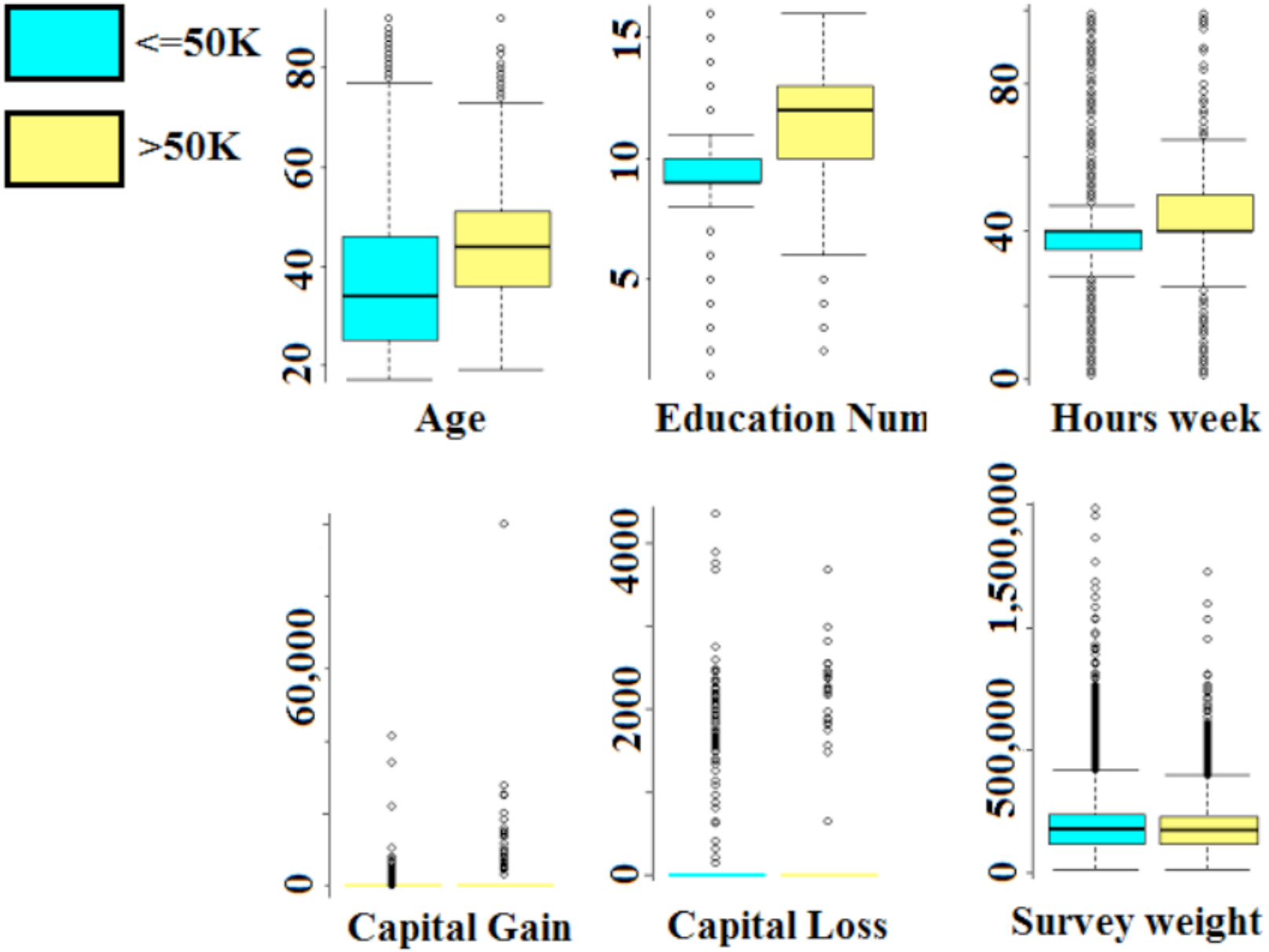
DATA SET

Attribute	Values	Missing
Employment Class	Private (68%), Self employed 1 (8%), Local Gov(6%), State Gov(4%), Unknown (5%), Self employed 2 (3%), Federal Gov(3%), No Pay(1.5%), Never Worked (0.5%)	1836
Education Level	High School (32%), Some college (22%), Bachelors (16%), Masters (5%), Vocational (4%), 11th (4%), Assoc Academic (3%), 10th (3%), 7-8th (2%), Professional School (2%), 9th (2%), 12th (2%), Doctorate (1%), 5-6th (1%), 1-4th (1%), Preschool (1%)	0
Relationship	Husband (41%), Not-in-family (26%), Own child (16%), Unmarried (11%), Wife (4%), Other relative (2%)	0
Race	White (85%), Black (10%), Asian / Pacific Islander (3%), American Indian / Eskimo (1%), Other (1%)	0
Marital Status	Married-civ-spouse (46%), Never-married (33%), Divorced (14%) Separated (3%), Widowed (2%), Married-AF-spouse (1%), Married-spouse-absent (1%)	0
Occupation	15 categories	1843
Country	42 categories: USA (90%)	583
Salary [Label]	<=$50K (76%), >$50K (24%)	0
Gender	Male (67%), Female (33%)	0

Numerical attribute

	Min	Max	Mean	Std deviation	Unique
Age	17	90	38	13.6	73
Survey Wgt.	12285	1484705	189778.367	105549.978	21648
Capital gain	0	99999	1077.649	7385.292	119
Capital loss	0	4356	87.3	402.9	92
Hours per week	1	99	40.43	12.34	94

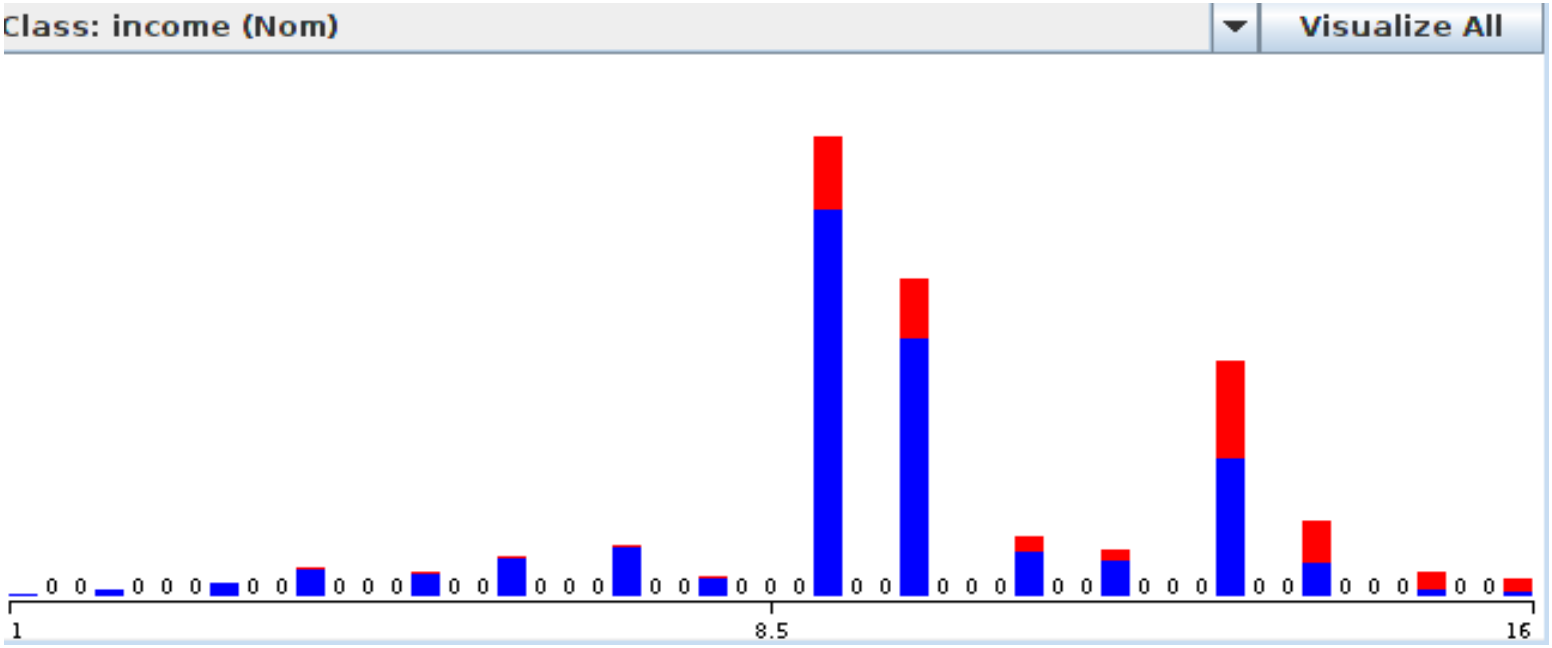
	<=50k	>50k
Capital Gain	96%	79%
Capital Loss	98%	90%


<=50K
>50K
20
40
60
80
Age
5
10
15
Education Num
0
40
80
Hours week
0
60,000
Capital Gain
0
2000
4000
Capital Loss
0
500,000
1,500,000
Survey weight

Mapping between education number and education level attributes

Educ_num	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Education	Preschool	1st-4th	5th-6th	7th-8th	9th	10th	11th	12th	HS-grad	Some-college

Educ_num	*11*	*12*	*13*	*14*	*15*	*16*
Education	Assoc-voc	Assoc-acdm	Bachelors	Masters	Prof-school	Doctorate



Comparison of occupation attribute values within the salary classes

	Exec	Prof	Cleaners	Other	Missing	Farming	Machine
<=50K	9%	9%	5%	13%	7%	4%	7%
>50K	25%	24%	0%	2%	3%	2%	3%
% change	194%	158%	-100%	-87%	-64%	-58%	-55%

	Adm	Tech	Protect	Transport	Sales	Craft	Army	House
<=50K	13%	2%	2%	5%	11%	13%	0%	0%
>50K	6%	4%	3%	4%	12%	12%	0%	0%
% change	-51%	39%	35%	-21%	16%	-7%	0%	0%

Rules discovered for >50K salary class

> 50K Rules	
marital_stat = Married-civ-spouse	+ cap_gain > 5095.5 + cap_loss > 1782.5 + hrs_per_week > 41.5 + occup = Prof-specialty + age > 34.5 + fnlwgt ≤ 110267 + education = Bachelors
occup = Exec-managerial	+ age > 37.5 + fnlwgt ≤ 160045
occup = Sales	+ age > 42.5 + relationship = Husband
occup = Tech-support	+ fnlwgt > 112507
occup = Protective-serv	+ age > 48
cap_gain > 7073.5	
fnlwgt > 211972 and ≤ 372272.5 and ≤ 221579 + age ≤ 37.5	

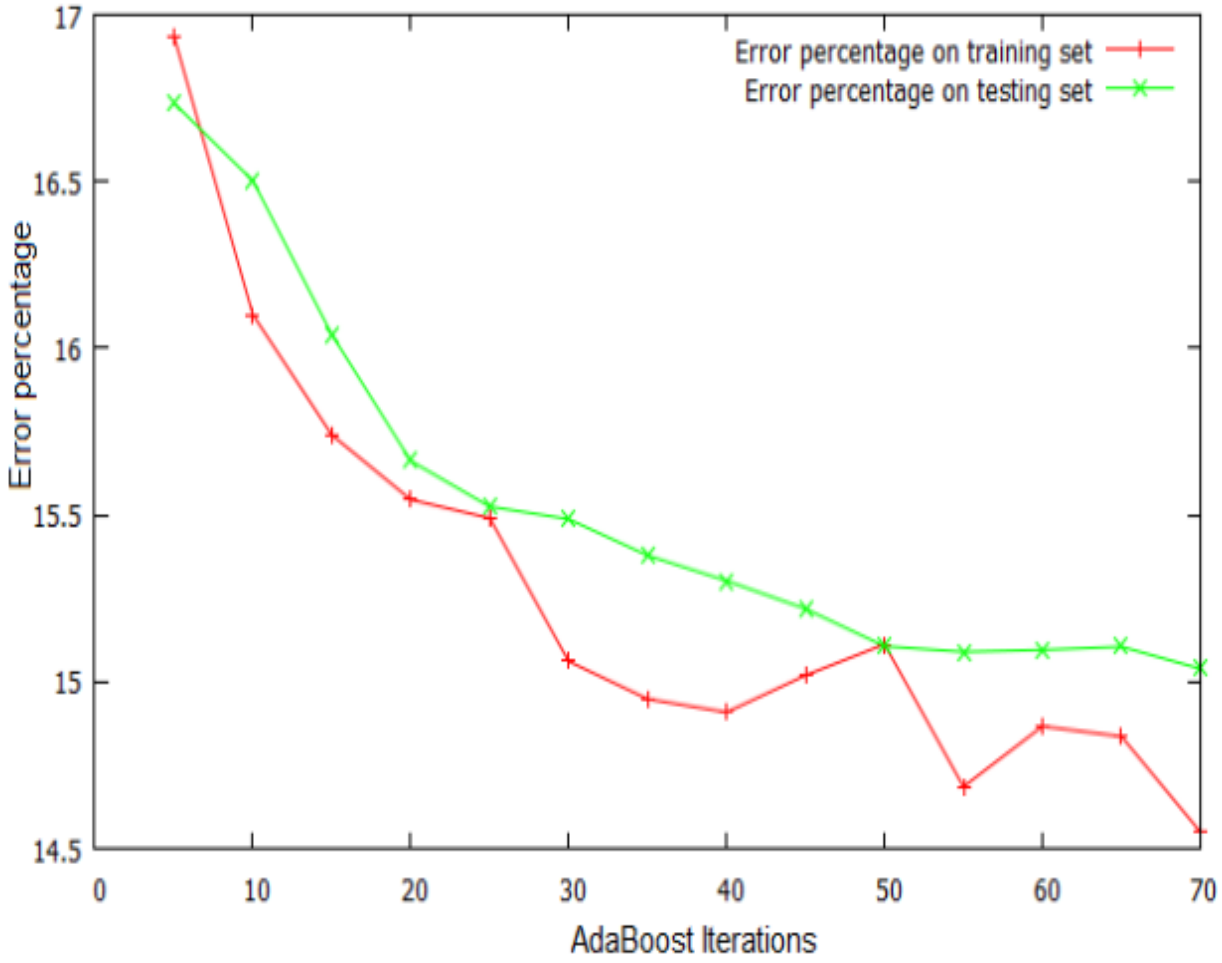
Naïve Bayes	Accuracy	Precision >50K	ROC
Unmodified	83.47	0.715	0.892
Forward Selection	83.41	0.715	0.892
Forward Selection with Binning	85.26	0.694	0.912

Numerical attribute binning results

hrs_per_week		age		capital loss		capital gain	
Bin Qty	Accuracy	Bin Qty	Accuracy	Bin Qty	Accuracy	Bin Qty	Accuracy
2	83.1	2	83.14	2	84.21	2	81.75
3	83.37	3	83.1	5	84.46	10	82.41
4	83.22	4	83.3	15	84.55	25	83.15
5	83.33	5	83.43	25	84.63	50	83.3
6	83.31	10	83.47	50	84.68	100	83.38
7	83.38	11	83.54	100	84.78	150	83.43
8	83.34	12	83.53	125	84.83	250	83.45
10	83.4	13	83.49	150	84.81	500	83.71
20	83.43	14	83.48	200	84.87	1000	84.02
30	83.44	15	83.54	250	85.03	1250	84.15
40	83.44	20	83.53	500	85.05	1500	84.17
50	83.45	25	83.51	750	85.06	2500	84.18
75	83.44	35	83.51	1000	85.06	5000	84.18

RESULTS

Classifier	Accuracy	Memory
SMO polykernel	85.26	424 M
Multilayer Perceptron	83.36	
Bonzaiboost-n200-d2	86.40	62 M
Bonzaiboost-n1000-d1	86.9	62 M
Boostexter	87.4	45 M


Error percentage on training set
Error percentage on testing set
Error percentage
AdaBoost Iterations
17
16.5
16
15.5
15
14.5
0
10
20
30
40
50
60
70

THANK YOU