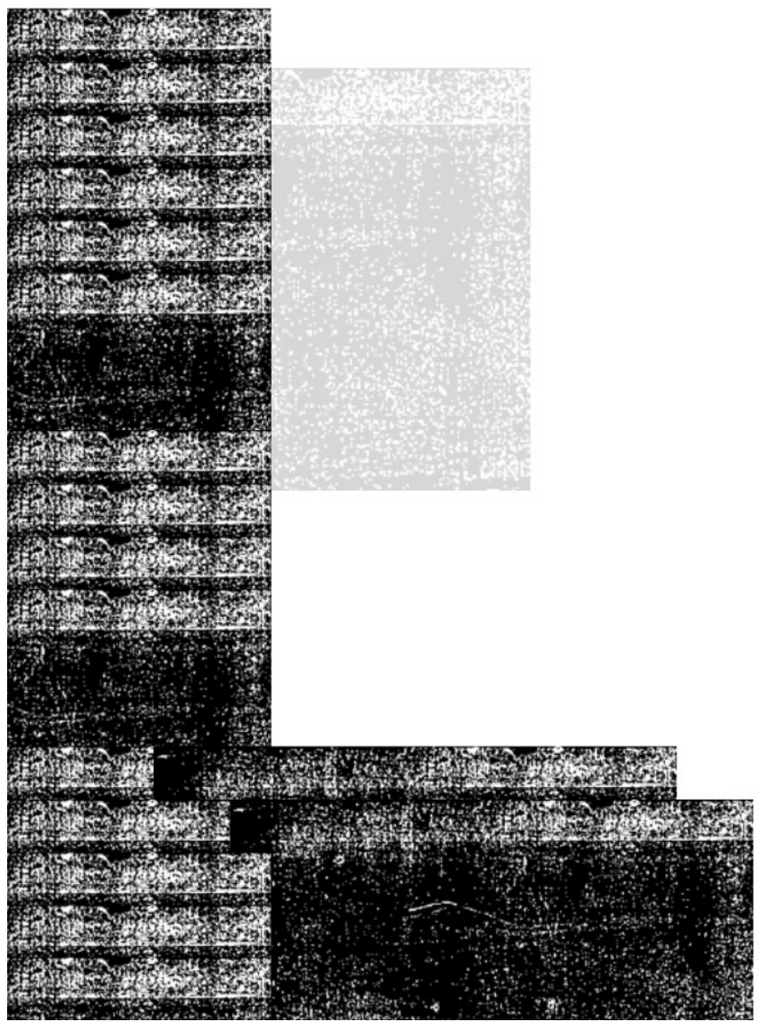
Xinyi Zheng, Enkelejda Gjergji, Yashagra Sharma

Executive Summary



US Census dataset contains **6,179,373,392** records. **Central Limit Theorem** and **Stratified Sampling** methods are employed to conduct the analysis, producing a **sample of 617,940 observations and 15 variables.**

❑ Outliers in numeric variables

Outlier column	Outlier count
hours-per-week	166,184
capital-loss	27,953
capital-gain	48,220
fnlwgt	18,689
age	2,608

❑ Two ways of imputing missing values

Age, capital-loss, capital-gain -> **Most frequent value**

Hours-per-week, Fnlwgt -> **Rounded mean**

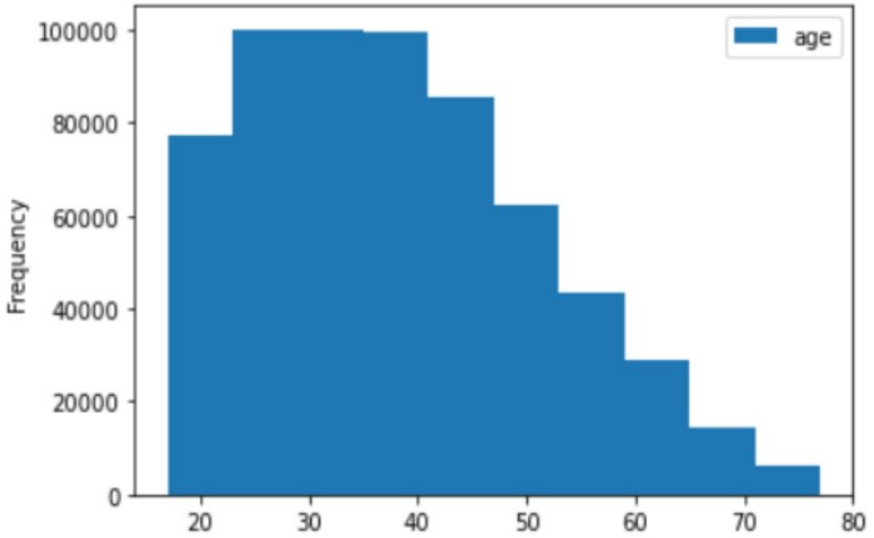
❑ Numeric variable treated as categorical

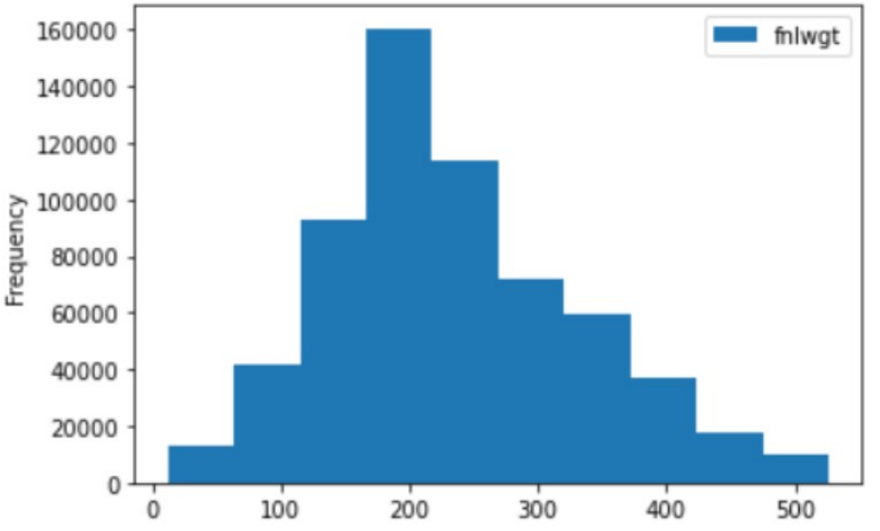
Education - num

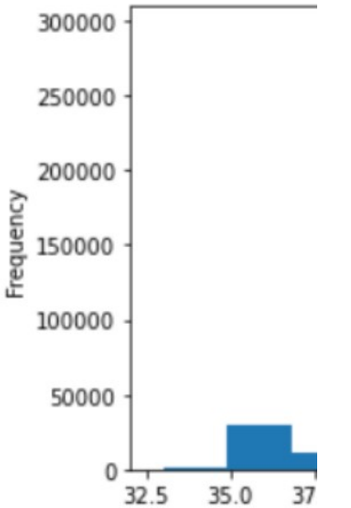
Table of statistics for numeric variables

	age	fnlwgt	capital_gain
min	17.000000	14.000000	0.000000
max	90.000000	1485.000000	99999.000000
median	36.000000	218.000000	0.000000
mean	38.001809	248.516893	1072.274849
std	13.450175	129.310778	7401.243733
skew	0.593938	2.159925	11.957578

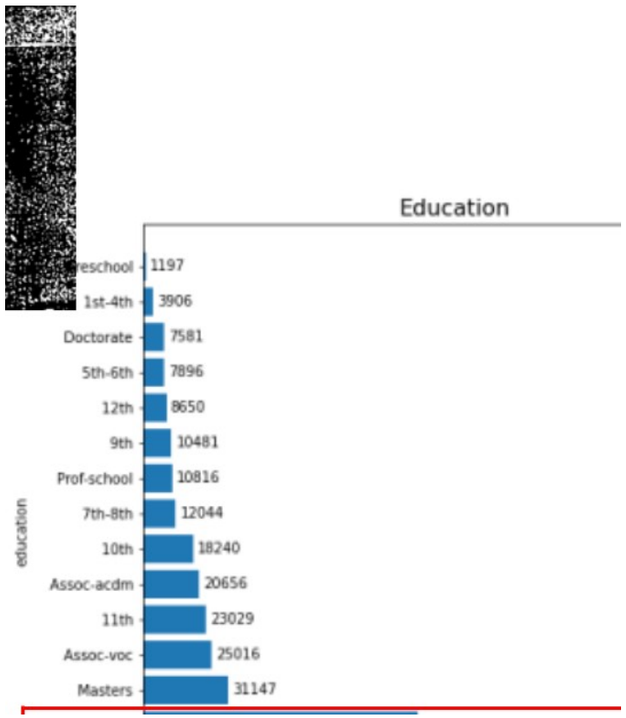
[redacted]



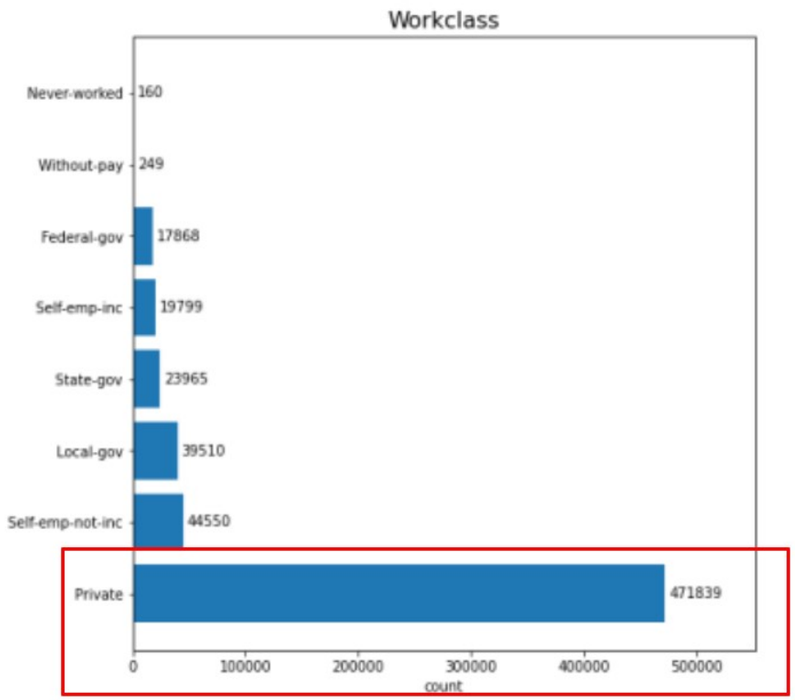


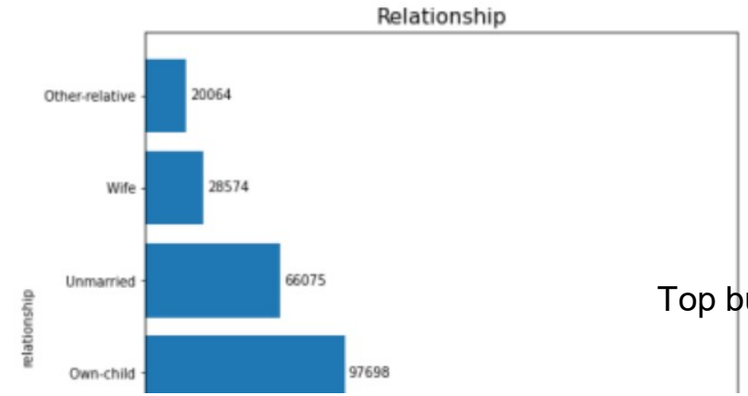


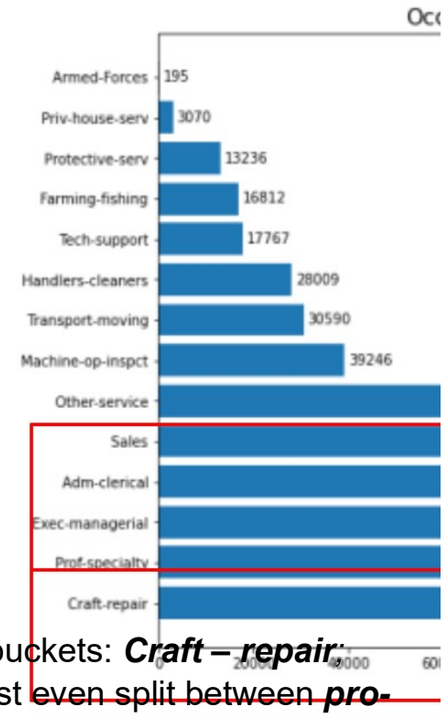
[redacted]



Top buckets: ***Bachelor holders***, those with ***some college*** *and* ***HS-grad***

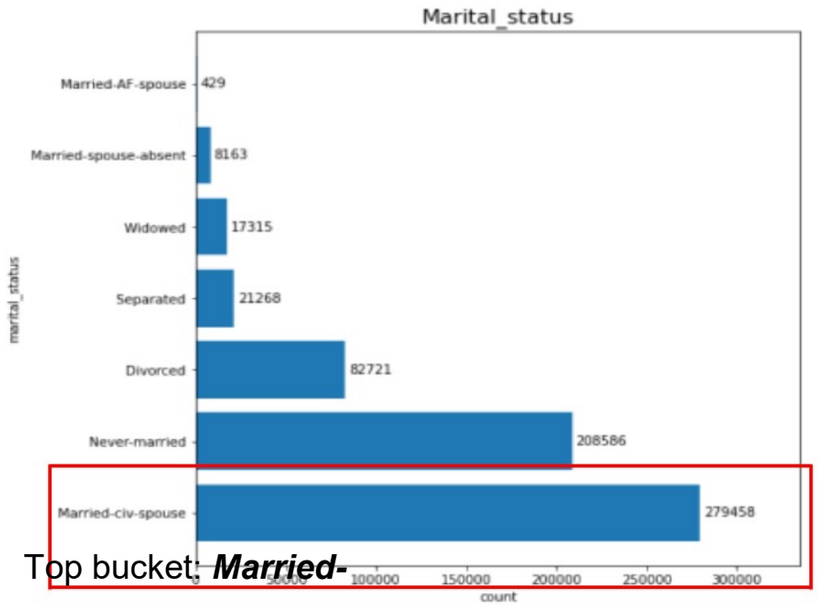




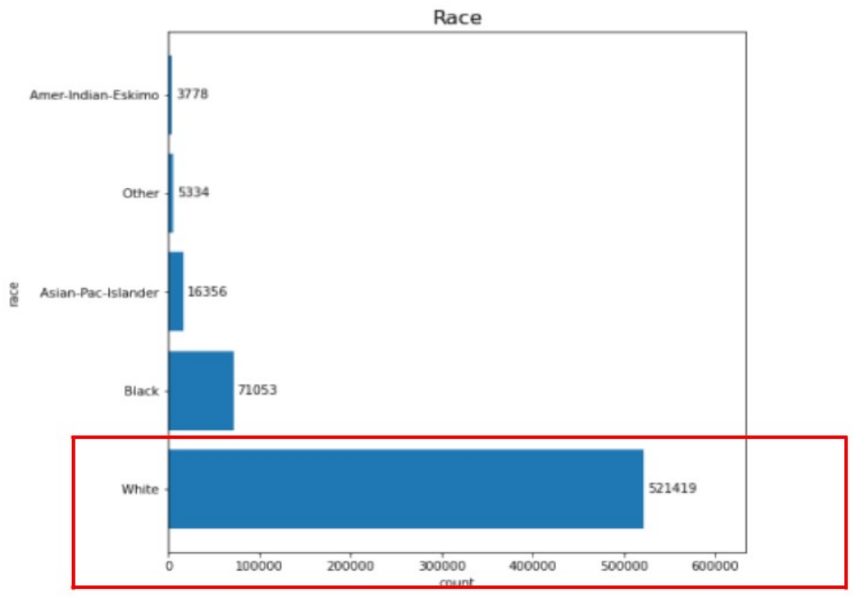


Top buckets: ***Craft – repair***; Almost even split between ***pro-specialty, exec-managerial, admin-clerical, sales and other services***

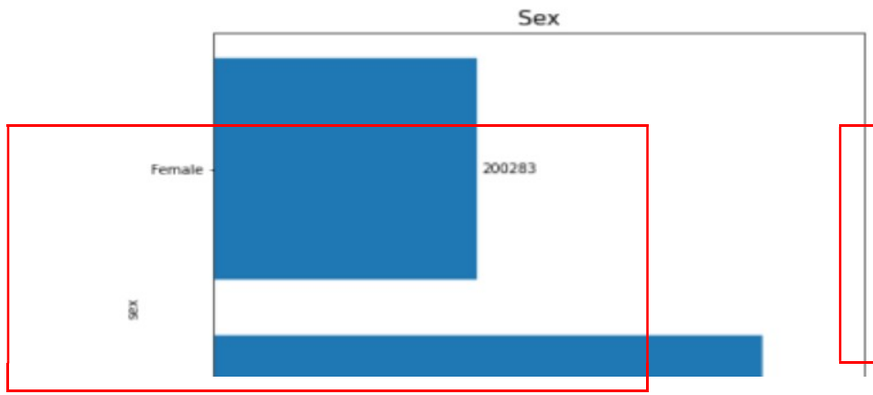
Top bucket: ***Husband***

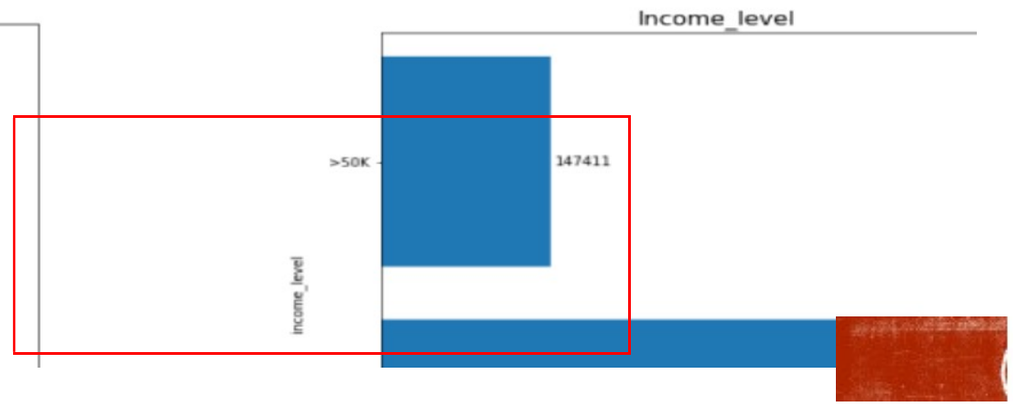


Top bucket: ***Married-civilian-spouse***



Top buckets: ***White Male, US Natives***

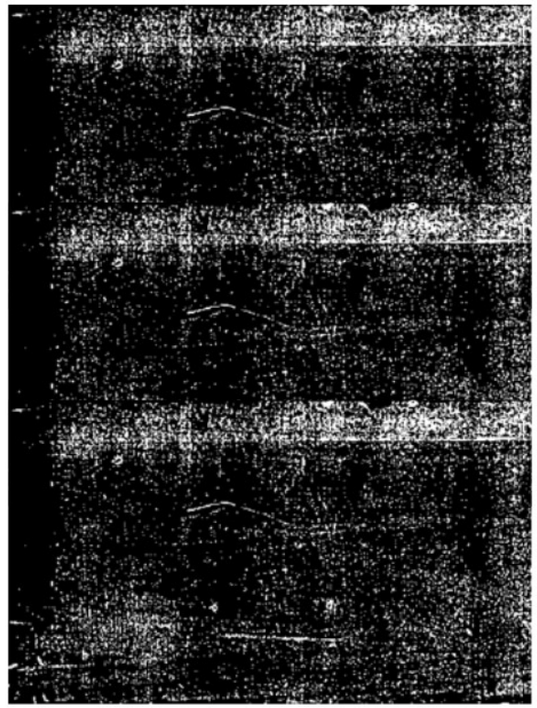




- Naïve Bayes model 's Area Under Curve is **88.4%** and True Positive rate is **74%**.

- 14 predictors are used to predict Income: Age, Fnlwgt, Capital-gain, Capital-loss, Hours-per-week, work class , Education, Education – num, Marital Status, Occupation, Relationship, Race , Sex , Native Country.

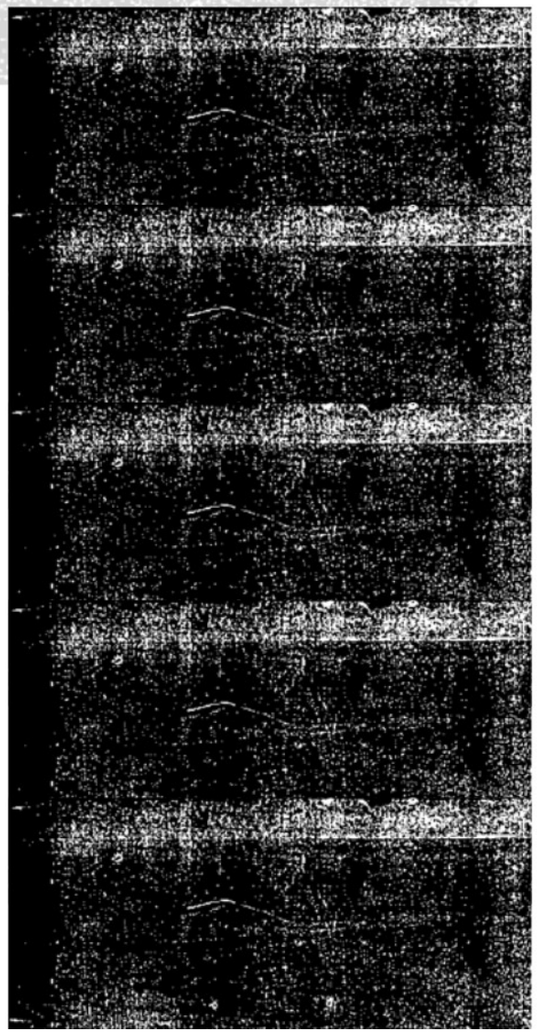
- The finding is that Naïve Bayes model can use US Census data to predict whether income exceeds $50K/yr. well



- Logistic Regression's Area Under Curve is **88.6%** and True Positive rate is **57%**

- The same 14 predictors are tested to predict Income, only 4 predictors have a significant relationship with Income: Age, Hours-per-week, Marital-status, Race

- The finding is that Logistic Regression model can use US Census data to predict whether income exceeds $50K/yr well and people who are older, used to get married and whose race is white or Asian-Pac-Islander are more likely to earn over $50K/yr

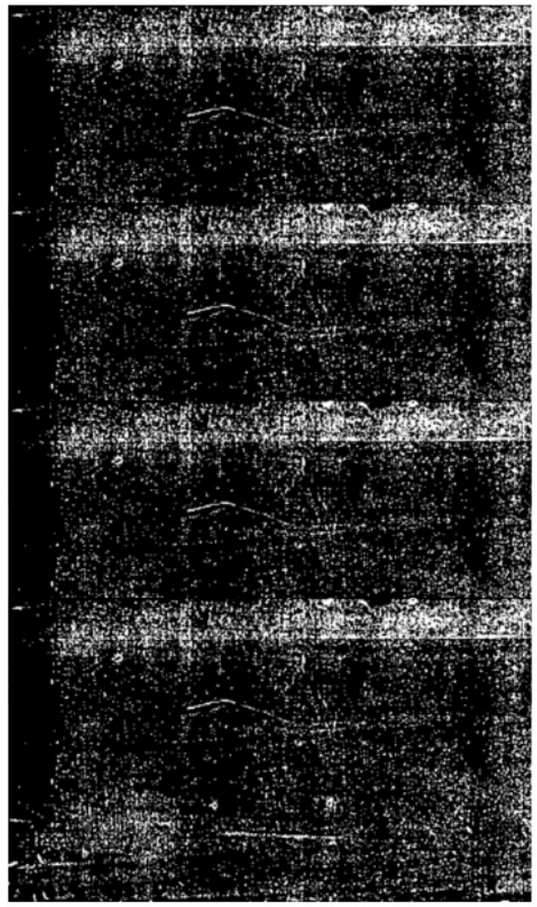


- Decision Tree's Area Under Curve is **92.8%** and True Positive rate is **93%**

- Relationship, Education-num, Marital-status are top 3 most useful variables for decision rules to predict Income

- The finding is that Decision Tree model can use US Census data to predict whether income exceeds $50K/yr well

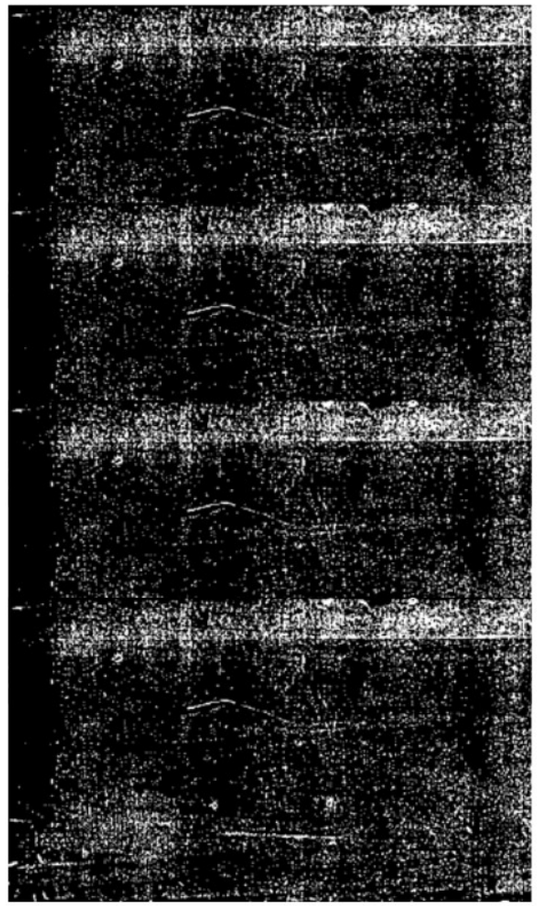
- The person whose occupation is "Prof-specialty", native-country is "United-States", age is between 30.5 and 61.5, education-year is "15" and relationship is "Husband", is more likely to have an income >50k

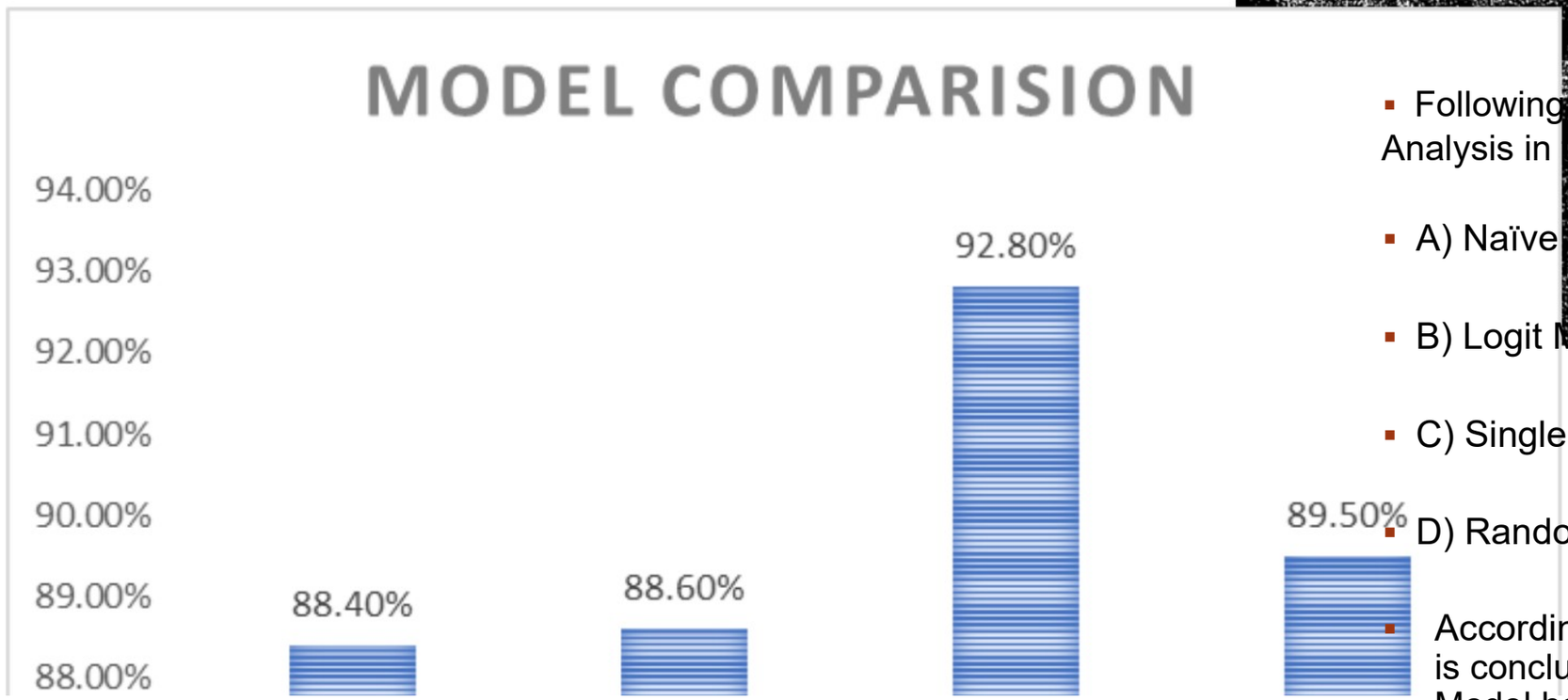


- Random forest's Area Under Curve is **89.5%** and True Positive rate is **94%**

- Marital-status, Relationship, Education, Education-num are top 4 most useful variables for decision rules to predict Income

- The finding is that Random Forest model can use US Census data to predict whether income exceeds $50K/yr well and Marital-status is identified as an important variable in Logit Regression, Single Tree and Random Forest



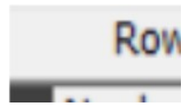


- Following Models are used in our Analysis in [redacted] rediction :-
- A) Naïve [redacted] Model – 88.4%
- B) Logit Model – 88.60%
- C) Single Tree Model – 92.80%
- D) Random Forest – 89.5%
- According to the analysis, it is concluded that the Single Tree Model has the highest Arear Under the Curve percentage.

- Age & Native Country - People who are **35.5** and **61.5** years and older
- Education_num - **15 years of education**
- Relationship - **Husband**
- Occupation - **Prof.-Specialty**

- Target the Age Group Between 30 years old and 60
- Target people who marked their relationship as “Husband” and marital status “Married”
- People who have achieved higher education – Bachelors and more

Before reproduction



After reproduction

NUMERIC	CATEGORICAL
Age	Workclass
Fnlwgt	Education
Capital-gain	Education-num
Capital-loss	Marital-status
Hours-per-week	Occupation
	Relationship
	Race
	Sex
	Native-country
	Income

Row ID	S Column Name	S Column Type
age	age	Number (integer)
workclass	workclass	String
fnlwgt	fnlwgt	Number (integer)
education	education	String
education-num	education-num	Number (integer)
marital-status	marital-status	String
occupation	occupation	String
relationship	relationship	String
race	race	String
sex	sex	String
capital-gain	capital-gain	Number (integer)
capital-loss	capital-loss	Number (integer)

- We obse[illegible]here are :

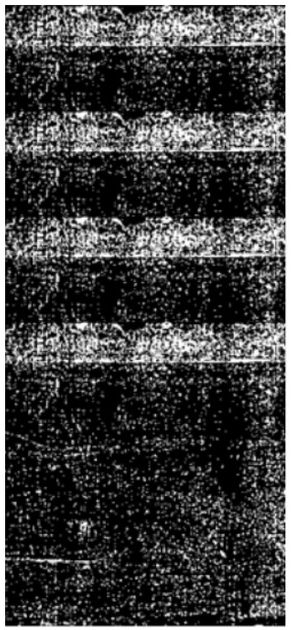
6 variables [illegible]64

9 variables [illegible]ject

6 numeric [illegible]s and 9 categorica[illegible]s

Q3: SOME OF THE VARIABLES APPEAR TO BE NUMERIC BUT SHOULD BE TREATED AS CATEGORICAL. YOUR BEST CLUE IS WHETHER A VARIABLE HAS ONLY A FEW DISCRETE VALUES. WHICH NUMERIC VARIABLES SHOULD BE TREATED AS CATEGORICAL?

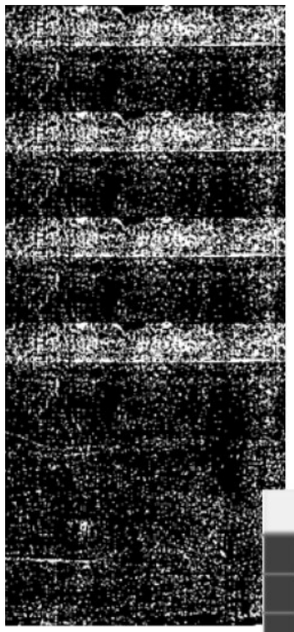
- Since educucation-num only has a few distinct values, it should be treated as a categorical. After this change, there are now 5 numeric variables and 10 categorical variables.



Row ID	S Column	D Min	D Max	D Mean	D Std. deviation	D Variance	D Skewness	D Kurtosis	D Overall sum	D Median	I Row count
age	age	17	90	37.991	13.452	180.967	0.594	-0.127	23,476,338	36	617940
fnlwgt	fnlwgt	12	1,485	248.458	129.37	16,736.542	2.133	10.594	153,532,173	218	617940
education-num	education...	1	16	10.015	2.618	6.852	-0.36	0.653	6,188,596	10	617940
capital-gain	capital-gain	0	99,999	1,083.049	7,421.357	55,076,53...	11.904	153.363	669,259,192	0	617940
capital-loss	capital-loss	0	4,356	84.406	395.608	156,505.675	4.664	20.986	52,157,680	0	617940

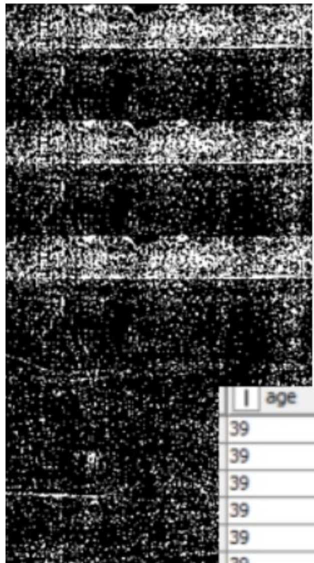
S Outlier ...	I Membe...	I Outlier ...	D Lo
age	6179407	26732	-3
fnlwgt	6179407	184651	-45.5
capital-gain	6148732	479390	0

There are outliers present in all 5 numeric variables which is now set to missing.

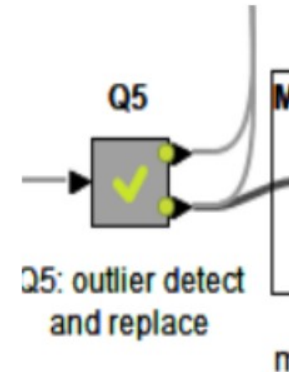


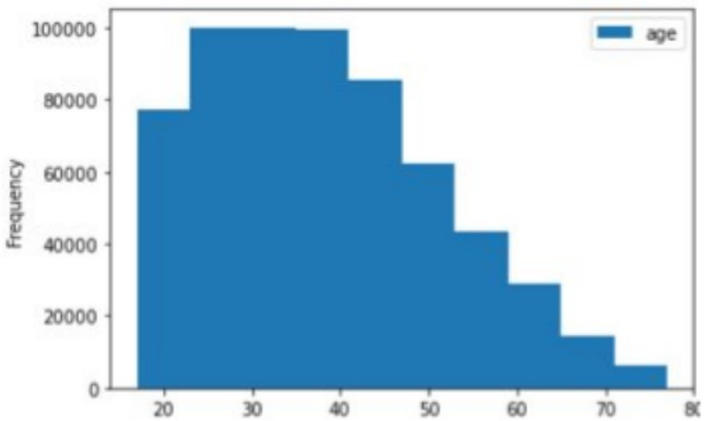
Row ID	S Row0
workclass (Un...	State-gov(239009), Self-emp-not-inc(446198), Private(4375003), Federal-gov(177820), Local-gov(394822), ?(346139), Self-emp-inc(19639
education (Un...	Bachelors(1007005), HS-grad(1990368), 11th(229045), Masters(309876), 9th(104065), Some-college(1376186), Assoc-acdm(206375), Ass
education-nu...	13(1007005), 9(1990368), 7(229045), 14(309876), 5(104065), 10(1376186), 12(206375), 11(251446), 4(121493), 16(77106), 15(106932)
marital-status...	Never-married(2087316), Married-civ-spouse(2796937), Divorced(825539), Married-spouse-absent(80737), Separated(211614), Married-AF
occupation (U...	Adm-clerical(724014), Exec-managerial(749790), Handlers-cleaners(280034), Prof-specialty(767104), Other-service(621478), Sales(696737

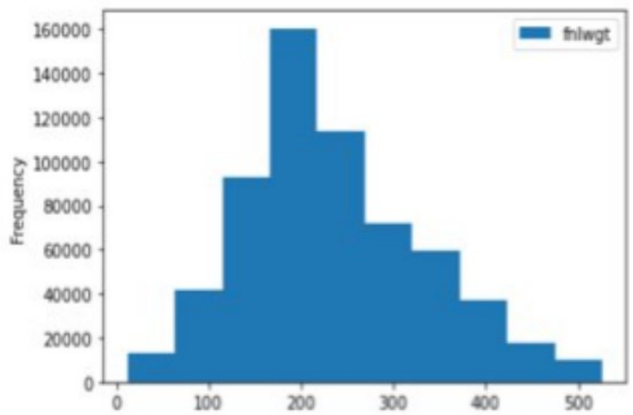
We found some missing value as "?" And we replace it by using NAN

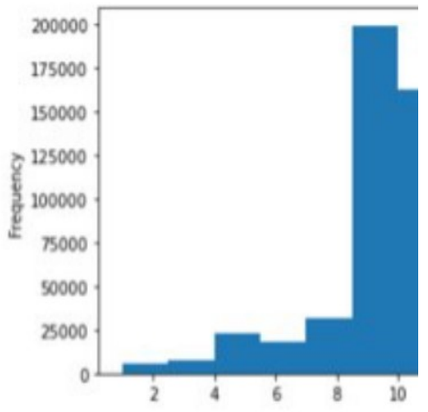


I age	S workclass	I fnlwgt	S education	S educati...	S marital-...	S occupa...	S relation...	S race	S sex	I capital-...	I capital-l...	I hours-p...	S native-...	S income
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	State-gov	78	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K



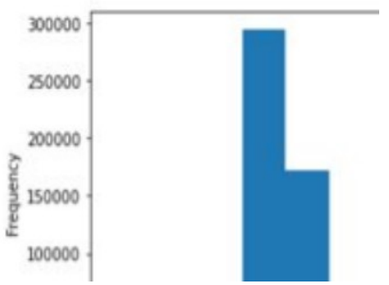

age
Frequency
100000
80000
60000
40000
20000
0
20
30
40
50
60
70

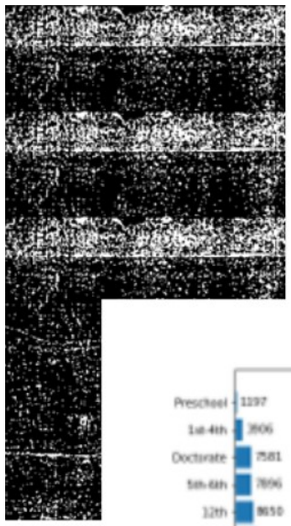

fnlwgt
Frequency
160000
140000
120000
100000
80000
60000
40000
20000
0
0
100
200
300
400
500

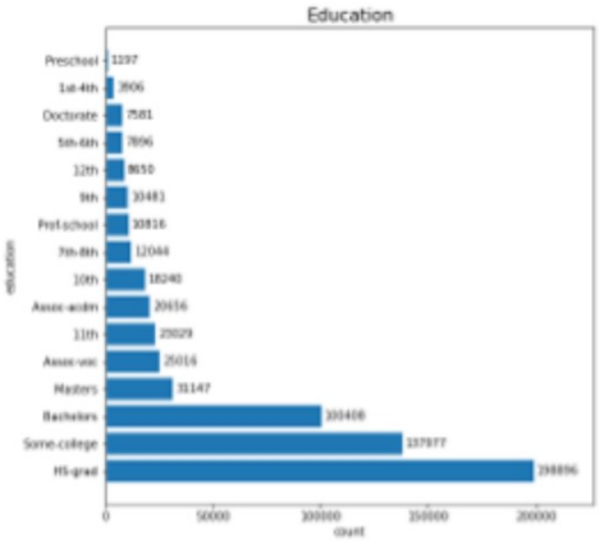

Frequency
200000
175000
150000
125000
100000
75000
50000
25000
0
2
4
6
8
10

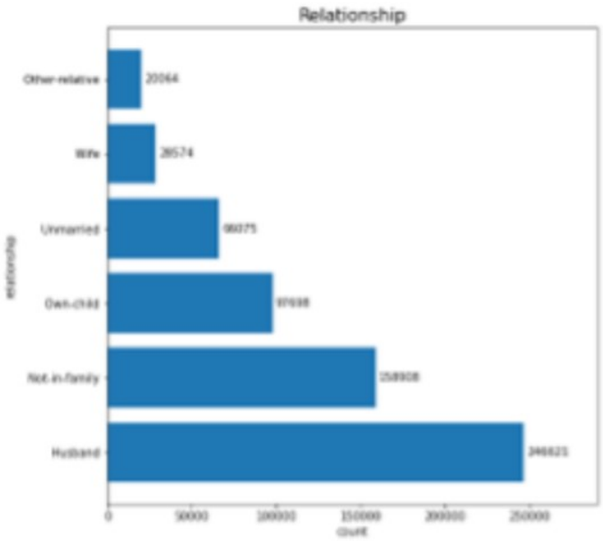

capital_gain
Frequency
600000
500000
400000
300000
200000

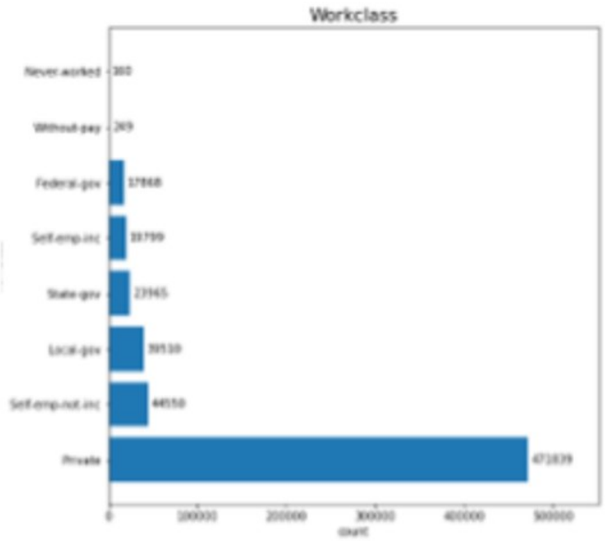

capital_loss
Frequency
600000
500000
400000
300000
200000

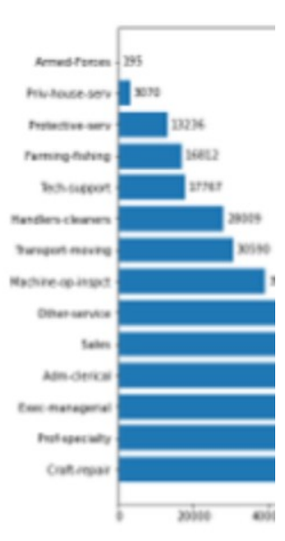

Frequency
300000
250000
200000
150000
100000

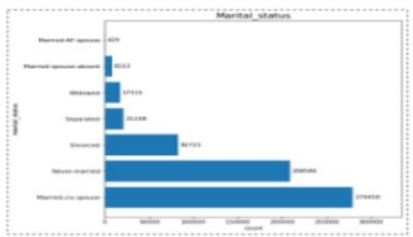


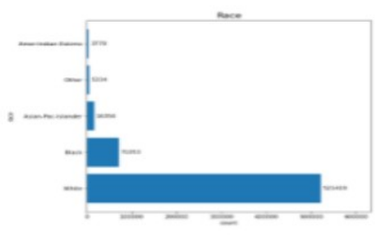

Education
Preschool 1197
1st-4th 3906
Doctorate 7581
5th-6th 7896
12th 9650
9th 10481
Prof-school 10816
7th-8th 12044
10th 18040
Assoc-acdm 20656
11th 25029
Assoc-voc 25014
Masters 31147
Bachelors 101408
Some-college 137977
HS-grad 298896
education
0
50000
100000
150000
200000
count

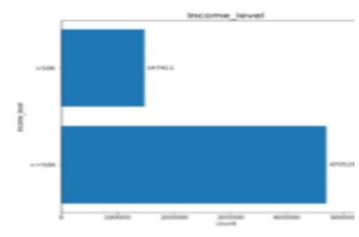

Relationship
Other-relative 20064
Wife 28574
Unmarried 98075
Own-child 97698
Not-in-family 158908
Husband 246621
relationship
0
50000
100000
150000
200000
250000
count


Workclass
Never-worked 100
Without-pay 269
Federal-gov 17868
Self-emp-inc 19799
State-gov 23965
Local-gov 39510
Self-emp-not-inc 44550
Private 471839
0
100000
200000
300000
400000
500000
count

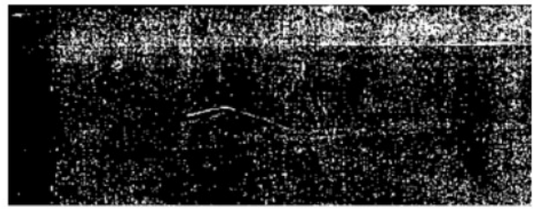

Armed-Forces 295
Priv-house-serv 3070
Protective-serv 13236
Farming-fishing 16812
Tech-support 17767
Handlers-cleaners 28019
Transport-moving 30590
Machine-op-inspct
Other-service
Sales
Adm-clerical
Exec-managerial
Prof-specialty
Craft-repair
0
20000


Marital_status

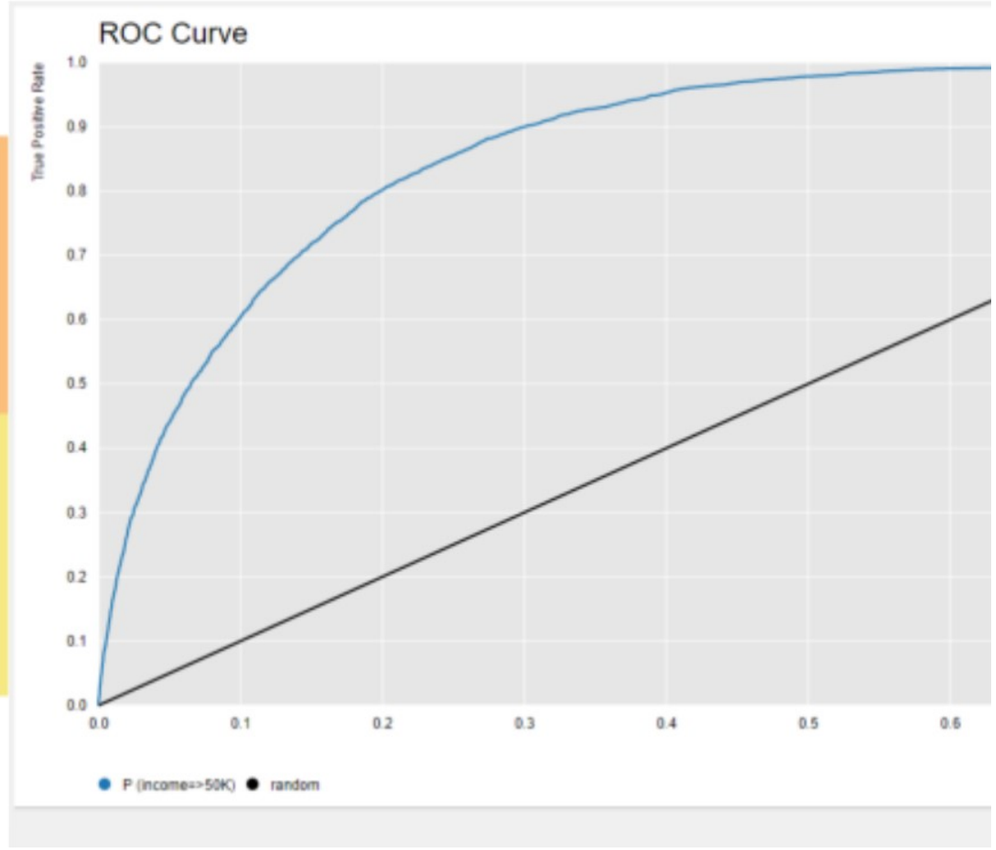

Race

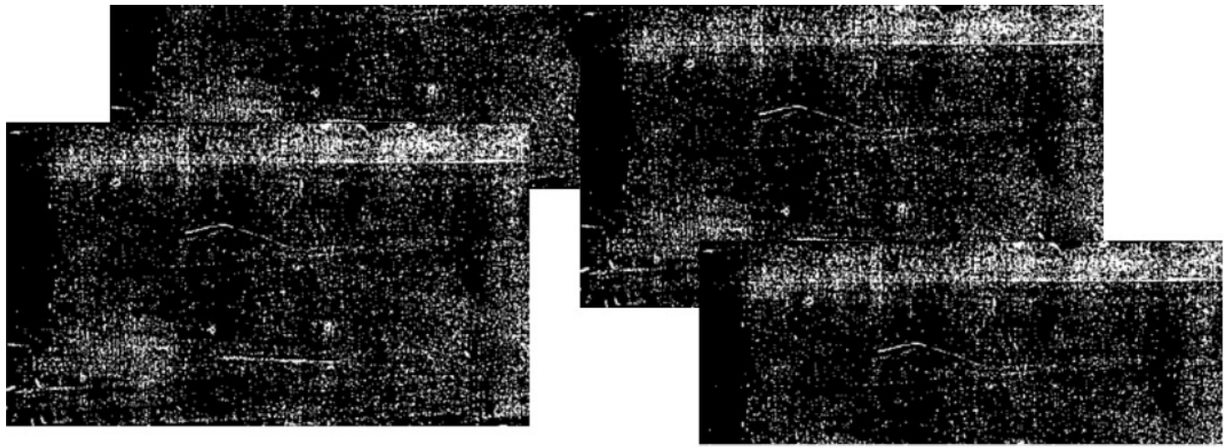

Income_level

Sex

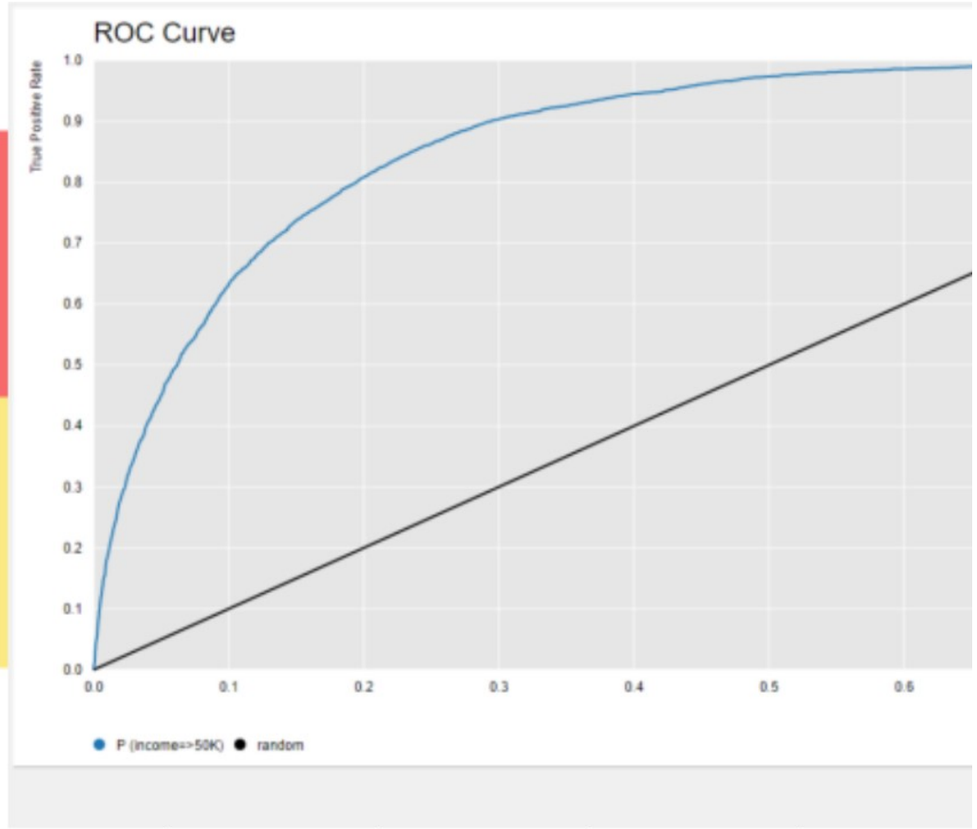


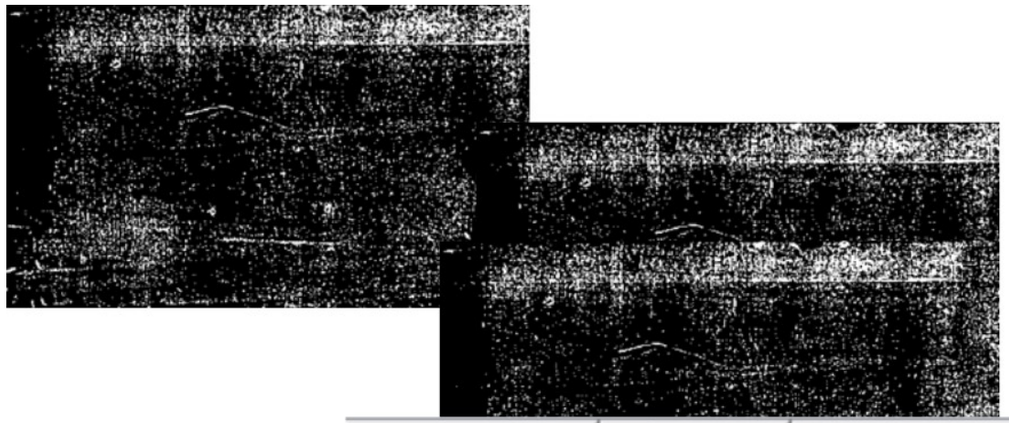
True Label \ Predicted Label	<=50K	>50K
<=50K	108666	20909
>50K	10992	31298



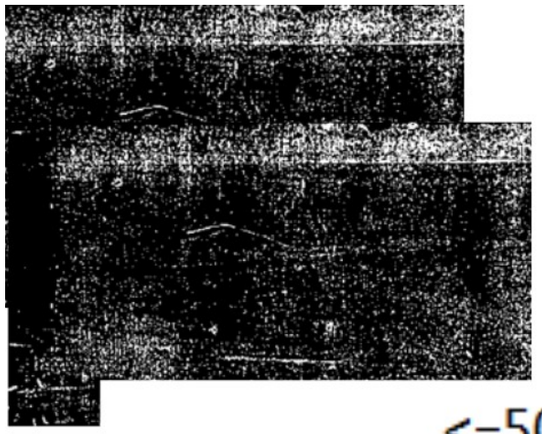


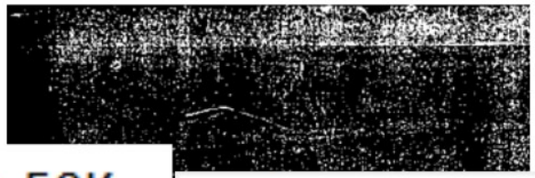
True Label \ Predicted Label	<=50K	>50K
<=50K	119013	10562
>50K	18282	24008

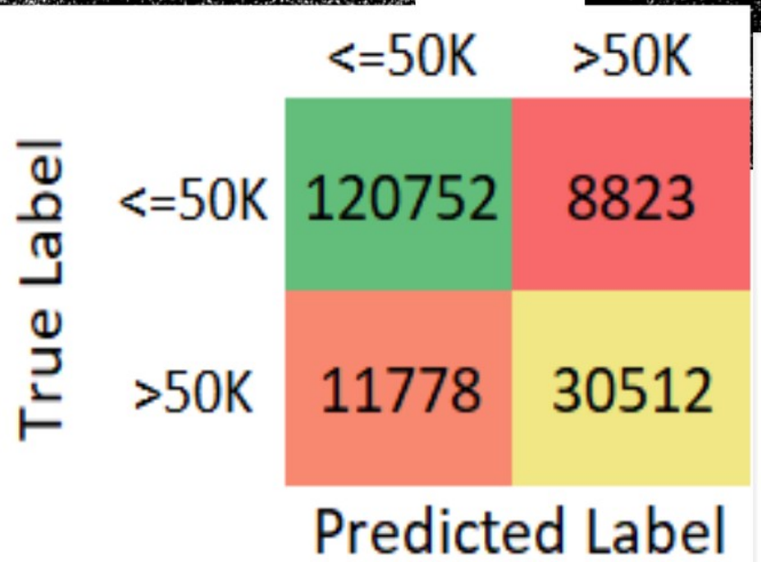


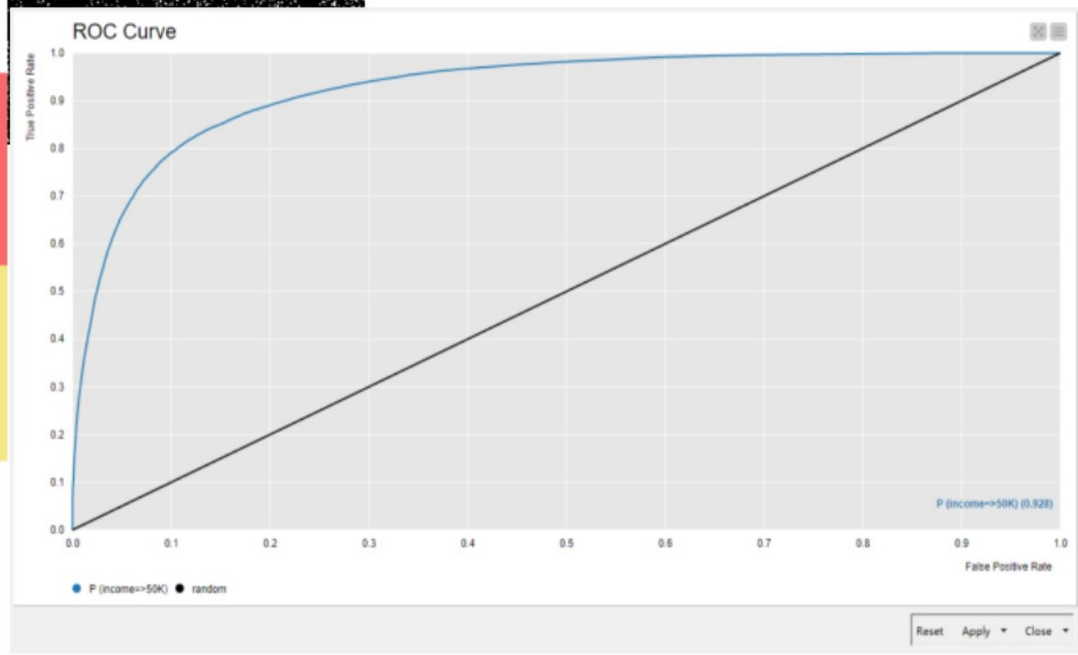


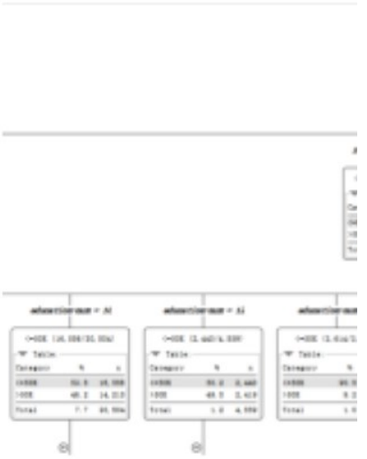
Row ID	S Logit	S Variable	D Coeff.	D Std. Err.	D z-score
Row1	<=50K	age	-0.028	0	-61.598
Row2	<=50K	fnlwgt	-0	0	-8.39
Row3	<=50K	hours-per-week	-0.067	0.001	-51.421
Row27	<=50K	Divorced_marital-status	-0.362	0.022	-16.695
Row29	<=50K	Separated_marital-status	-0.357	0.04	-9.013
Row52	<=50K	White_race	-0.105	0.018	-5.742
Row28	<=50K	Married-spouse-absent_marital-status	-0.34	0.062	-5.47
Row54	<=50K	Amer-Indian-Eskimo_race	0.358	0.073	4.934
Row30	<=50K	Married-AF-spouse_marital-status	-0.467	0.141	-3.319
Row53	<=50K	Asian-Pac-Islander_race	-0.149	0.056	-2.68
Row31	<=50K	Widowed_marital-status	-0.111	0.042	-2.631
Row55	<=50K	Other_race	0.164	0.075	2.184
Row87	<=50K	Outlying-US(Guam-USVI-etc)_native...	13.83	8,701.044	0.002
Row25	<=50K	Preschool_education	13.077	9,072.319	0.001
Row97	<=50K	Constant	7.37	6,842.552	0.001
Row96	<=50K	Holand-Netherlands_native-country	6.423	8,701.052	0.001
Row10	<=50K	Without-pay_workclass	13.553	20,580.814	0.001
Row72	<=50K	Cambodia_native-country	-2.61	8,702.801	-0
Row20	<=50K	Doctorate_education	-2.64	9,075.185	-0

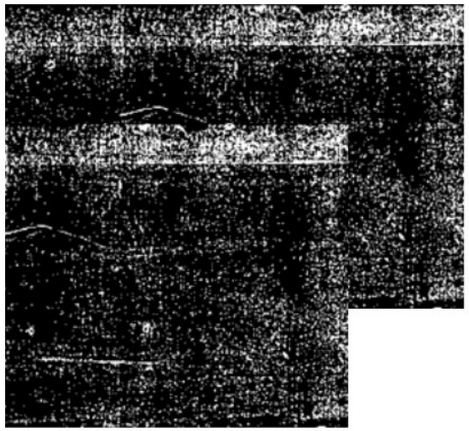





<=50K
>50K
True Label
<=50K
120752
8823
>50K
11778
30512
Predicted Label

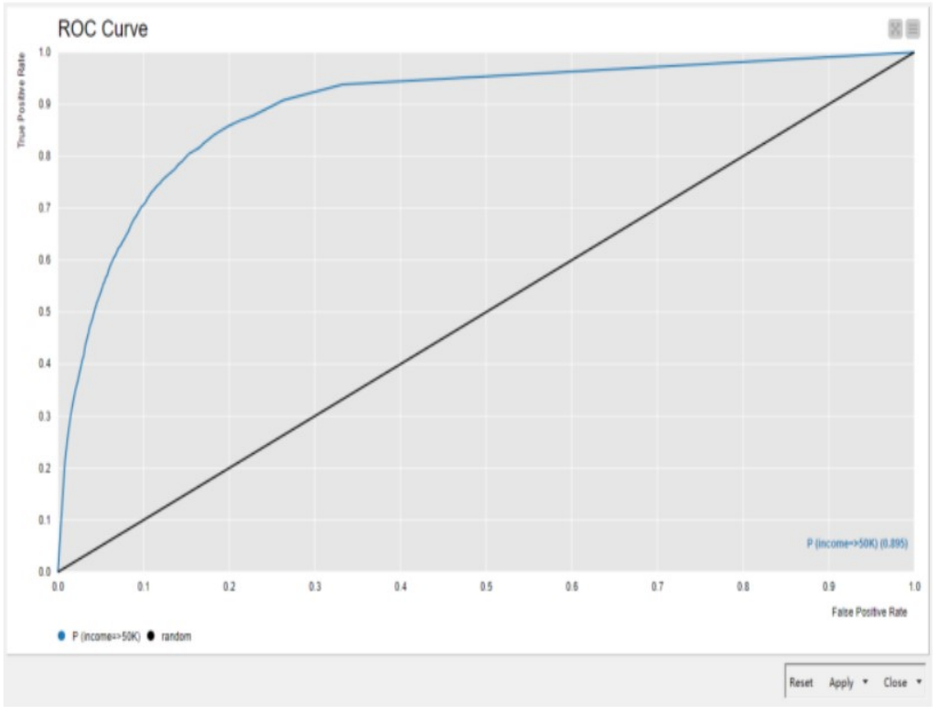

ROC Curve
True Positive Rate
False Positive Rate
P (income=>50K) (0.928)
P (income=>50K)
random
Reset
Apply
Close

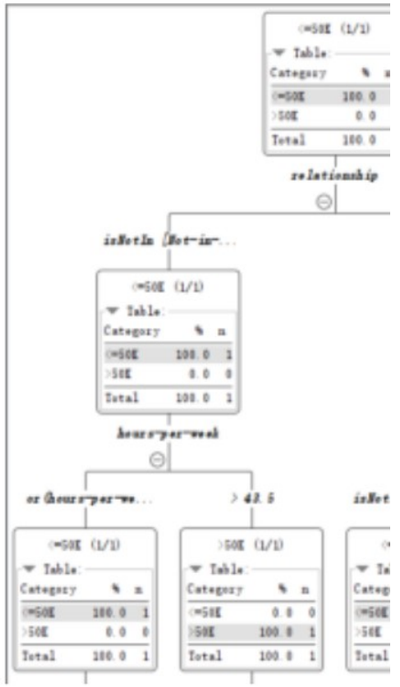




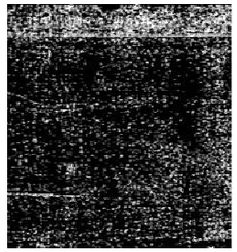
Row ID	Condition	Outcome	Record count	Number of corre
Row1286	$marital-status$ = "Never-married" AND $relationship$ = "Own-child"	<=50K	53,170	52,780
Row1360	$native-country$ = "United-States" AND $education-num$ = "9" AND $rela...	<=50K	14,450	14,037
Row1	$hours-per-week$ <= 44.5 AND age <= 29.5 AND $education-num$ = "13" ...	<=50K	6,693	6,407
Row93	$occupation$ = "Adm-clerical" AND $education-num$ = "9" AND $relations...	<=50K	5,620	5,488
Row109	$occupation$ = "Other-service" AND $education-num$ = "9" AND $relation...	<=50K	4,902	4,839
Row204	$occupation$ = "Adm-clerical" AND $education-num$ = "10" AND $relation...	<=50K	4,813	4,693
Row648	$workclass$ = "Private" AND age <= 34.5 AND $occupation$ = "Craft-re...	<=50K	4,122	3,450
Row1458	$education-num$ = "9" AND $native-country$ = "United-States" AND $rela...	<=50K	4,106	4,037
Row1013	$education-num$ = "4" AND $relationship$ = "Husband"	<=50K	3,980	3,614
Row110	$occupation$ = "Sales" AND $education-num$ = "9" AND $relationship$ = ...	<=50K	3,668	3,496
Row1411	$occupation$ = "Adm-clerical" AND $education-num$ = "10" AND $relation...	<=50K	3,610	3,555
Row1086	$occupation$ = "Prof-specialty" AND $native-country$ = "United-States"...	>50K	3,063	2,842
Row113	$occupation$ = "Machine-op-inspct" AND $education-num$ = "9" AND $rela...	<=50K	3,044	3,011
Row220	$occupation$ = "Other-service" AND $education-num$ = "10" AND $relatio...	<=50K	2,847	2,802
Row751	$workclass$ = "Private" AND $occupation$ = "Exec-managerial" AND $educ...	>50K	2,788	2,563
Row240	$marital-status$ = "Never-married" AND $education-num$ = "12" AND $rel...	<=50K	2,638	2,411
Row305	$education-num$ = "6" AND $relationship$ = "Not-in-family"	<=50K	2,566	2,548
Row118	age <= 36.5 AND $workclass$ = "Private" AND $occupation$ = "Craft-re...	<=50K	2,397	2,359
Row341	age > 43.5 AND $race$ = "White" AND age <= 66.5 AND age > 28.5 A...	>50K	2,333	2,141
Row233	age <= 50.5 AND $occupation$ = "Craft-repair" AND $education-num$ = ...	<=50K	2,199	2,106
Row111	$occupation$ = "Transport-moving" AND $education-num$ = "9" AND $relat...	<=50K	2,153	2,034
Row1252	$occupation$ = "Other-service" AND $relationship$ = "Wife"	<=50K	2,116	1,732
Row221	age <= 40.5 AND $occupation$ = "Sales" AND $education-num$ = "10" AN...	<=50K	2,045	2,024
Row1485	$native-country$ = "Mexico" AND $relationship$ = "Other-relative"	<=50K	1,855	1,855
Row107	$occupation$ = "Handlers-cleaners" AND $education-num$ = "9" AND $rela...	<=50K	1,848	1,841
Row1472	$education-num$ = "10" AND $native-country$ = "United-States" AND $rel...	<=50K	1,801	1,739
Row1016	$native-country$ = "United-States" AND $occupation$ = "Prof-specialty"...	>50K	1,791	1,550

True Label \ Predicted Label	<=50K	>50K
<=50K	121289	8286
>50K	16944	25346





Row ID	#splits (le...	#splits (level 1)	#splits (level 2)	#candidates (level 0)	#candidates (le
marital-status	25	25	40	30	47
relationship	20	30	43	20	39
education	14	23	55	19	36
education-num	11	23	65	21	37
age	7	20	52	22	37
workclass	5	10	19	29	44
occupation	5	21	52	16	29
sex	5	11	12	18	38
hours-per-week	5	22	31	24	61
native-country	2	8	17	24	40



	Naive Bayes(>50K)	Logistic Regression(>50K)	Single Tree(>50K)	Random Forest(>50K)
Accuracy	0.814383382	0.832170599	0.880132662	0.853198732
Missclassification Rate	0.185616618	0.167829401	0.119867338	0.146801268
True Positve Ratio	0.740080397	0.56769922	0.721494443	0.599337905
False Positive Ratio	0.174739675	0.081512637	0.068091839	0.063947521
Specificity	0.838633996	0.866841473	0.931908161	0.936052479
Precision	0.599498152	0.694474978	0.775695945	0.753627498
Prevalence	0.303767492	0.201146249	0.228871498	0.195688476