

# **DIRECT MARKETING CAMPAIGN! TARGETING! HOUSE HOLD INCOME >\$50K**



Xinyi Zheng, Enkelejda Gjergji, Yashagra Sharma

## **Executive Summary**

FOR OUR PROJECT, WE USED CENSUS DATA TO BUILD AND ANALYZE FOUR DIFFERENT MODELS TO PREDICT HOUSEHOLDS' INCOME GREATER THAN \$50K PER YEAR. THESE HOUSEHOLDS WILL BE TARGETED WITH A \$25 MILLION DIRECT MARKETING CAMPAIGN.

THE FOLLOWINGS ARE SOME OF THE PREDICTORS THAT CAN DETERMINE THE PROBABILITY OF HOUSEHOLD INCOME BEING GREATER THAN \$50K / YEAR!!

PREDICTORS : AGE, WORK CLASS,  
EDUCATION\_NUM, MARTIAL STATUS,  
OCCUPATION, RELATIONSHIP, RACE, SEX , NATIVE  
COUNTRY

!



# STATISTICAL DESCRIPTION OF SAMPLE

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         |

Table of statistics for numeric variables

|                | age       | fnlwgt      | capital_gain | capital_loss | hours_per_week |
|----------------|-----------|-------------|--------------|--------------|----------------|
| min            | 17.000000 | 14.000000   | 0.000000     | 0.000000     | 1.000000       |
| max            | 90.000000 | 1485.000000 | 99999.000000 | 4356.000000  | 99.000000      |
| median         | 36.000000 | 218.000000  | 0.000000     | 0.000000     | 40.000000      |
| mean           | 38.001809 | 248.516893  | 1072.274849  | 84.930678    | 40.311770      |
| std            | 13.450175 | 129.310778  | 7401.243733  | 396.647930   | 12.102392      |
| skew           | 0.593938  | 2.159925    | 11.957578    | 4.647479     | 0.199459       |
| kurt           | -0.117323 | 10.880462   | 154.558202   | 20.835491    | 2.959620       |
| missing values | 0.000000  | 0.000000    | 0.000000     | 0.000000     | 0.000000       |

## □ Two ways of imputing missing values

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

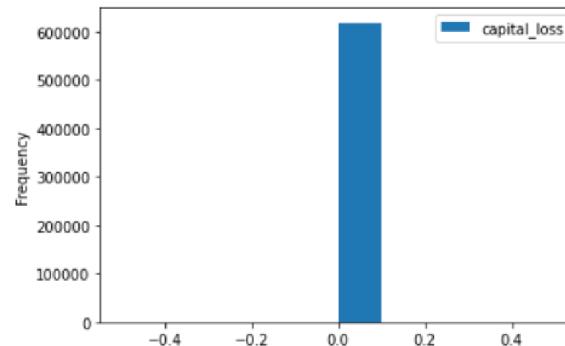
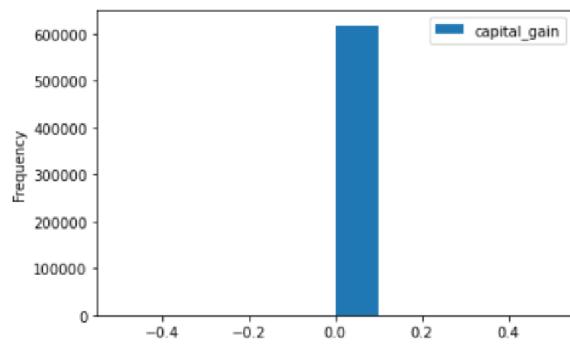
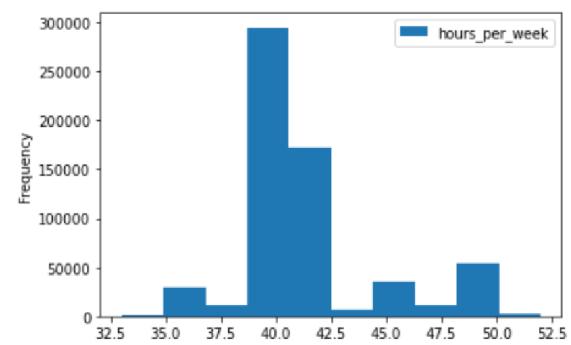
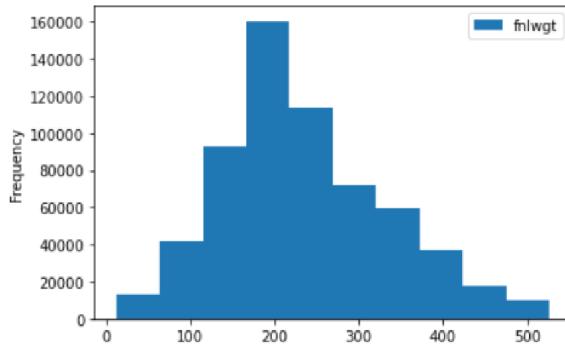
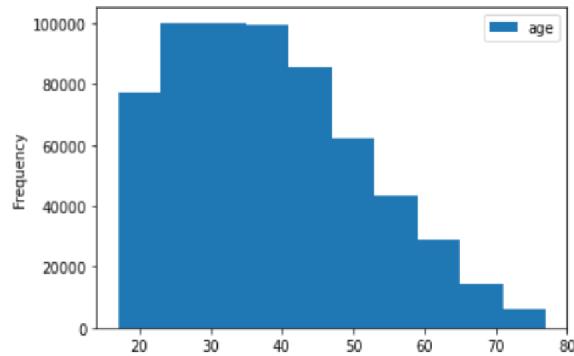
Hours-per-week, Fnlwgt -> **Rounded mean**

## □ Numeric variable treated as categorical

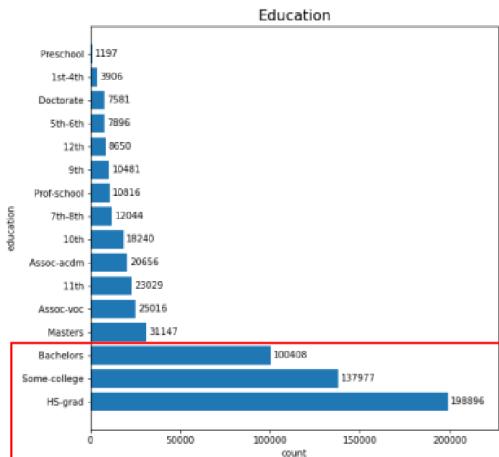
**Education - num**



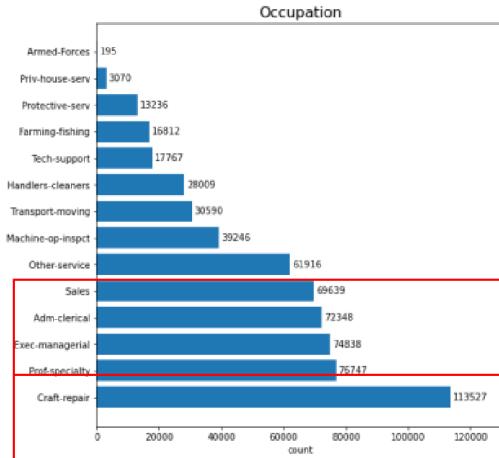
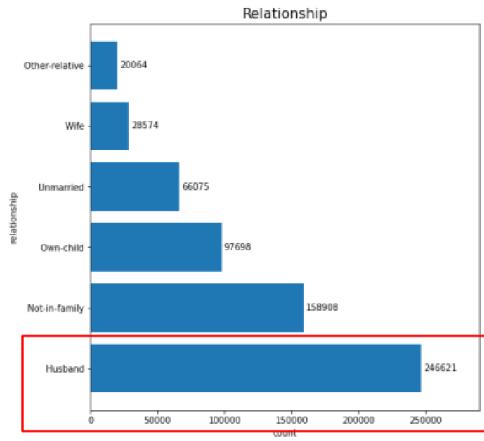
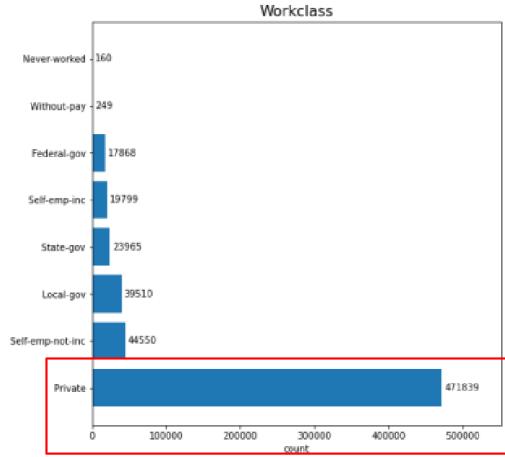
# HISTOGRAMS OF NUMERIC!VARIABLES



# EXPLORATORY DATA ANALYSIS



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

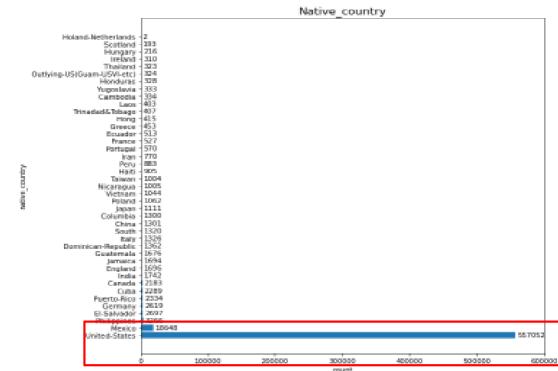
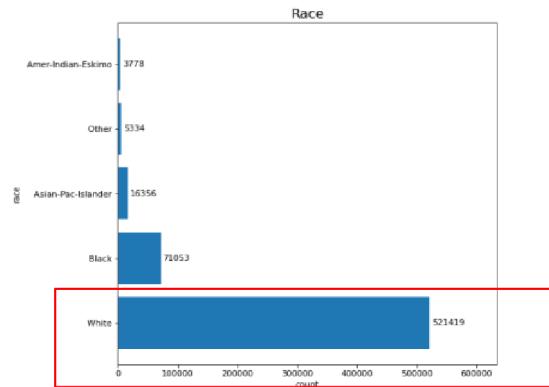
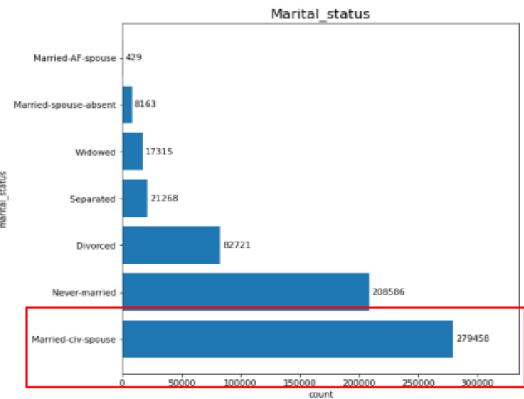


Top buckets: **Craft – repair**;  
Almost even split between **pro-specialty, exec-managerial, admin-clerical, sales and other services**

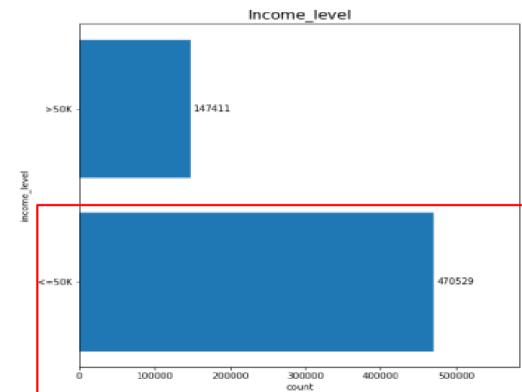
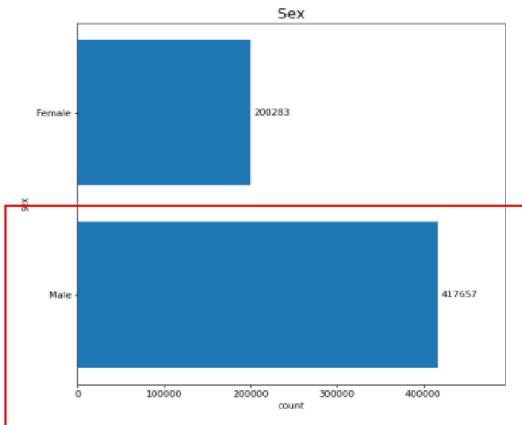
Top bucket: **Husband**



# EXPLORATORY DATA ANALYSIS CONTINUED



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

# MODEL 1

## - NATIVE

# BAYES

- 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

# MODEL 2

---

# LOGISTIC REGRES SION

- 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

# MODEL 3

---

# DECISION TREE

- 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

# MODEL 4

---

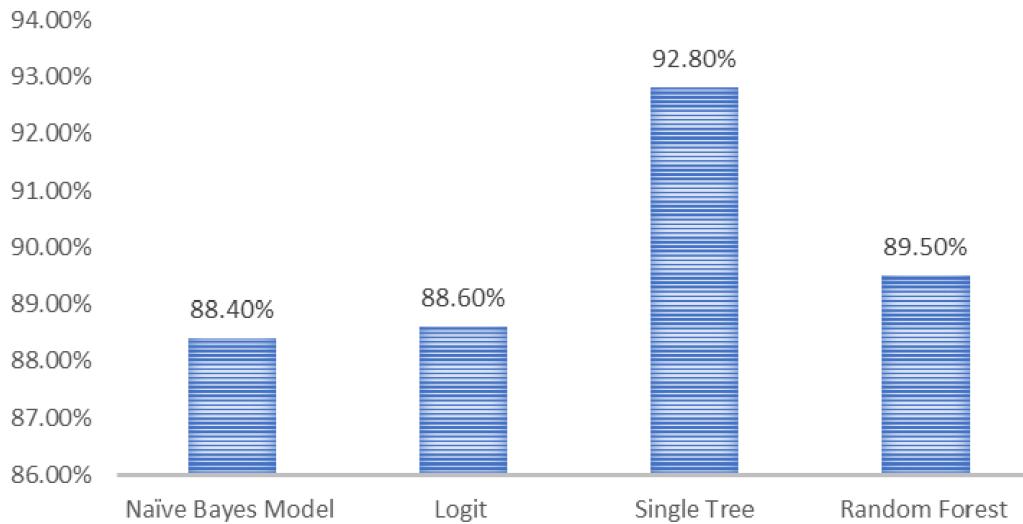
# RANDOM FOREST

# MODELS' COMPARISON AND RECOMMENDATIONS

Following Models are used in our Analysis in terms of prediction :-

- A) Naïve Bayes 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 Area Under the Curve percentage.

## MODEL COMPARISION



# CONCLU SION!

- 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**

# RECOMMENDATIONS!

- 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



# APPENDIX!



# Q1: HOW MANY OBSERVATIONS (ROWS) AND HOW MANY VARIABLES (COLUMNS) ARE THERE IN THE RAW DATA?

| Row ID         | Dimensions |
|----------------|------------|
| Number Rows    | 32561      |
| Number Columns | 15         |

Before reproduction

| Row ID         | Dimensions |
|----------------|------------|
| Number Rows    | 6179407    |
| Number Colu... | 15         |

After reproduction



## STATISTICAL DESCRIPTION OF DATA ( IN NUMERIC AND CATEGORICAL )!

### NUMERIC

Age  
Fnlwgt  
Capital-gain  
Capital-loss  
Hours-per-week

### CATEGORICAL

Workclass  
Education  
Education-num  
Marital-status  
Occupation  
Relationship  
Race  
Sex  
Native-country  
Income



## Q2: PRODUCE A TABLE OF VARIABLES SHOWING THEIR TYPES

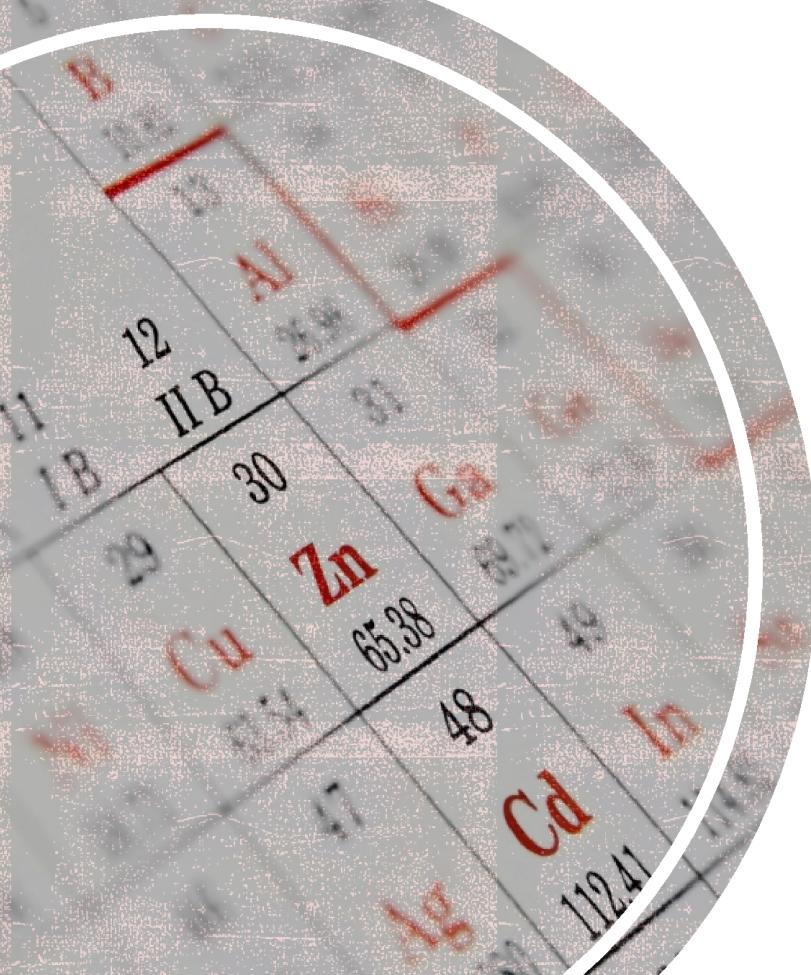
• We observed that there are :

6 variables as an int64

9 variables as an object

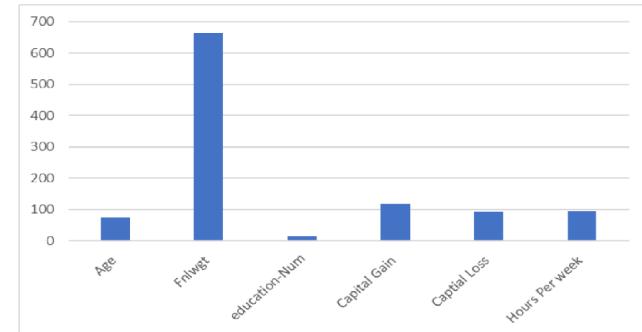
6 numeric variables and 9 categorical variables.

| Row ID | S Column Name  | S Column Type    |
|--------|----------------|------------------|
|        | age            | Number (integer) |
|        | workclass      | String           |
|        | fnlwgt         | Number (integer) |
|        | education      | String           |
|        | education-num  | Number (integer) |
|        | marital-status | String           |
|        | occupation     | String           |
|        | relationship   | String           |
|        | race           | String           |
|        | sex            | String           |
|        | capital-gain   | Number (integer) |
|        | capital-loss   | Number (integer) |
|        | hours-per-week | Number (integer) |
|        | native-country | String           |
|        | income         | String           |



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 education-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.



# Q4: FOR NUMERIC VARIABLES, PRODUCE A TABLE OF STATISTICS INCLUDING MISSING VALUES, MIN, MAX, MEDIAN, MEAN, STANDARD DEVIATION, SKEWNESS AND KURTOSIS.

| Row ID         | Column          | Min    | Max       | Mean      | Std. deviation | Variance   | Skewness | Kurtosis    | Overall sum | Median | 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      | 153,632,173 | 218    | 617940    |
| education-num  | education ... 1 | 16     | 18.015    | 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 |           |
| hours-per-week | hours-per... 1  | 99     | 40,312    | 12,099    | 146.389        | 0.214      | 3.022    | 24,910,684  | 40          | 617940 |           |



**Q5: HOW MANY OUTLIERS ARE PRESENT IN EACH NUMERIC VARIABLE? SHOW THE TALLIES IN A TABLE. SET THEM TO MISSING.**

| S | Outlier ...    | I       | Membe... | I | Outlier ... | D | Lower ... | D | Upper ... |
|---|----------------|---------|----------|---|-------------|---|-----------|---|-----------|
|   | age            | 6179407 | 26732    |   | -3          |   | 77        |   |           |
|   | fnlwgt         | 6179407 | 184651   |   | -45.5       |   | 526.5     |   |           |
|   | capital-gain   | 6148732 | 479390   |   | 0           |   | 0         |   |           |
|   | capital-loss   | 6179407 | 281762   |   | 0           |   | 0         |   |           |
|   | hours-per-w... | 6179407 | 1666477  |   | 32.5        |   | 52.5      |   |           |

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

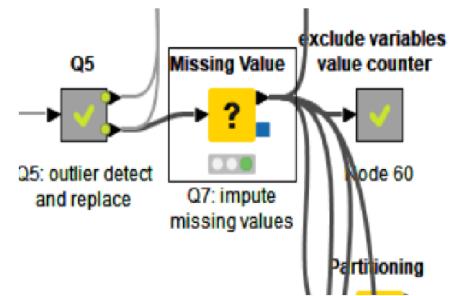


# Q6: COUNT THE UNIQUE VALUES OF EACH CATEGORICAL VARIABLE, INCLUDING MISSING VALUES. ARE THERE ANY UNUSUAL VALUES IN ANY OF THE CATEGORICAL VARIABLES?

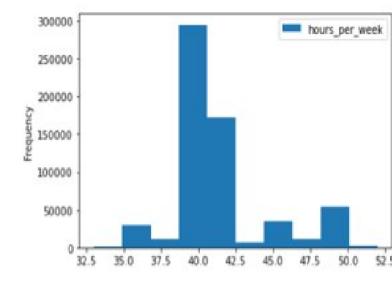
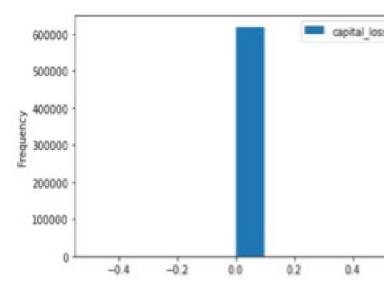
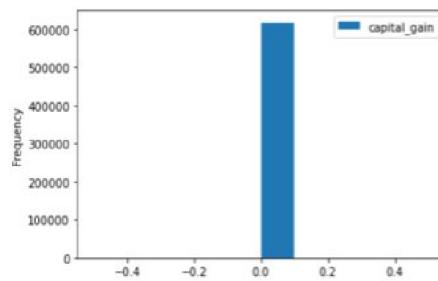
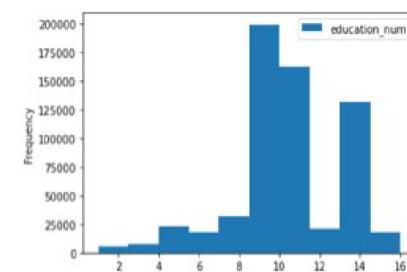
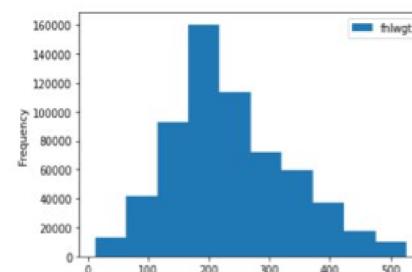
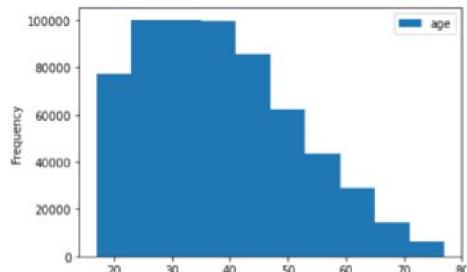
| Row ID  | S Row0  |
|---|---|
| workclass (Un...<br>education (Un...<br>education-nu...<br>marital-status...<br>occupation (U...<br>relationship (...<br>race (Unique ...<br>sex (Unique c...<br>native-countr...<br>income (Uniqu... | State-gov(239009), Self-emp-not-inc(446198), Private(4375003), Federal-gov(177820), Local-gov(394822), ?(346139), Self-emp-inc(196394), Without-pay(2440), Never-worked(1582)<br>Bachelor(1007005), HS-grad(1990368), 11th(229045), Masters(309876), 9th(104065), Some-college(1376186), Assoc-acdm(206375), Assoc-voc(251446), 7th-8th(121493), Doctora...<br>13(1007005), 9(1990368), 7(229045), 14(309876), 5(104065), 10(1376186), 12(206375), 11(251446), 4(121493), 16(77106), 15(106932), 3(77413), 6(183643), 2(40204), 1(12032...<br>Never-married(2087316), Married-civ-spouse(2796937), Divorced(825539), Married-spouse-absent(80737), Separated(211614), Married-AF-spouse(4327), Widowed(172937)<br>Adm-clerical(724014), Exec-managerial(749790), Handlers-cleaners(280034), Prof-specialty(767104), Other-service(621478), Sales(596737), Craft-repair(787586), Transport-moving(...<br>Not-in-family(1587372), Husband(2468038), Wife(285148), Own-child(979045), Unmarried(658632), Other-relative(201172)<br>White(5209899), Black(712338), Asian-Pac-Islander(166166), Amer-Indian-Eskimo(37577), Other(53427)<br>Male(4178696), Female(2000711)<br>United-States(5456824), Cuba(23395), Jamaica(17186), India(16851), ?(112699), Mexico(186308), South(13083), Puerto-Rico(23085), Honduras(3324), England(16410), Canada(21...<br><=50K(4705296), >50K(1474111) |

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

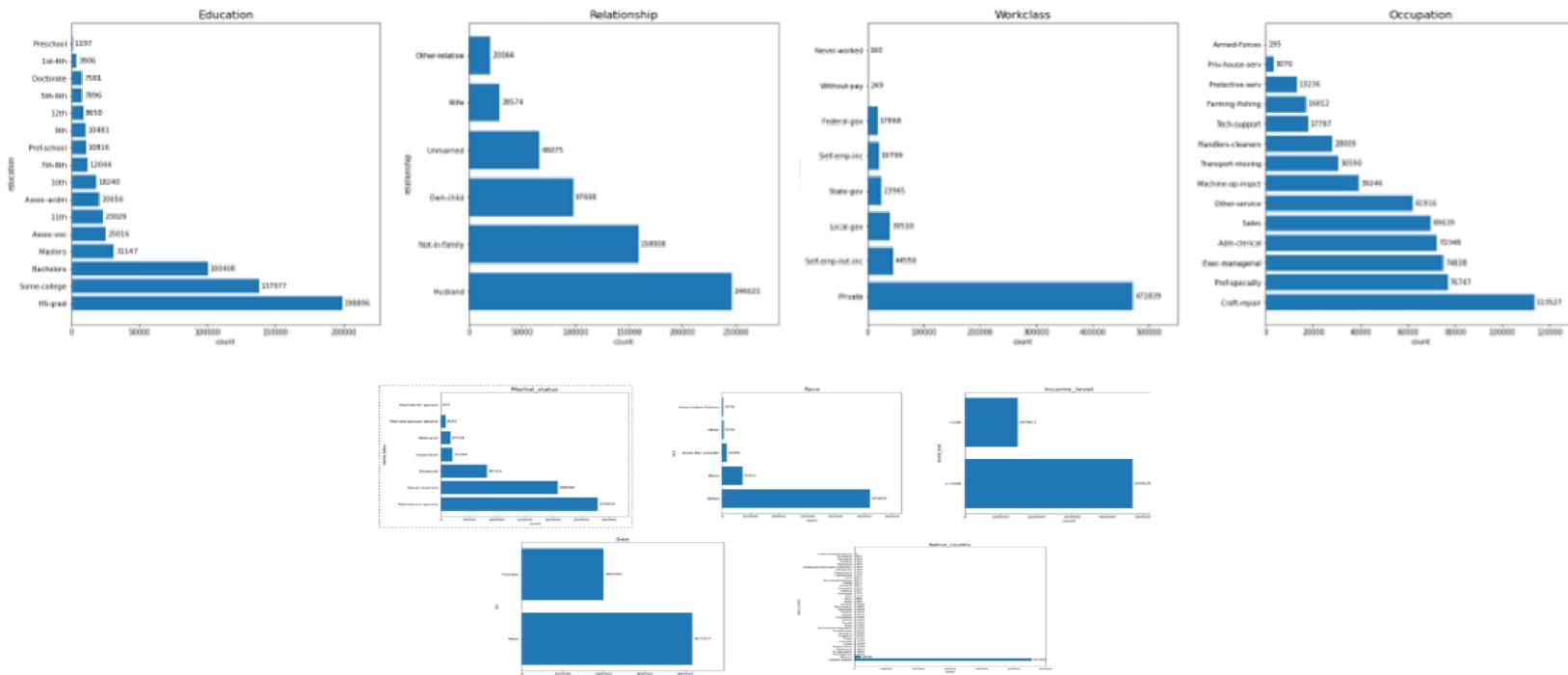
**Q7: IMPUTE THE MISSING VALUES. BE SURE TO EXPLAIN HOW YOU DID THAT IN YOUR PRESENTATION.**



## Q8: PRODUCE A HISTOGRAM OR BOXPLOT FOR EACH OF THE NUMERIC VARIABLES.



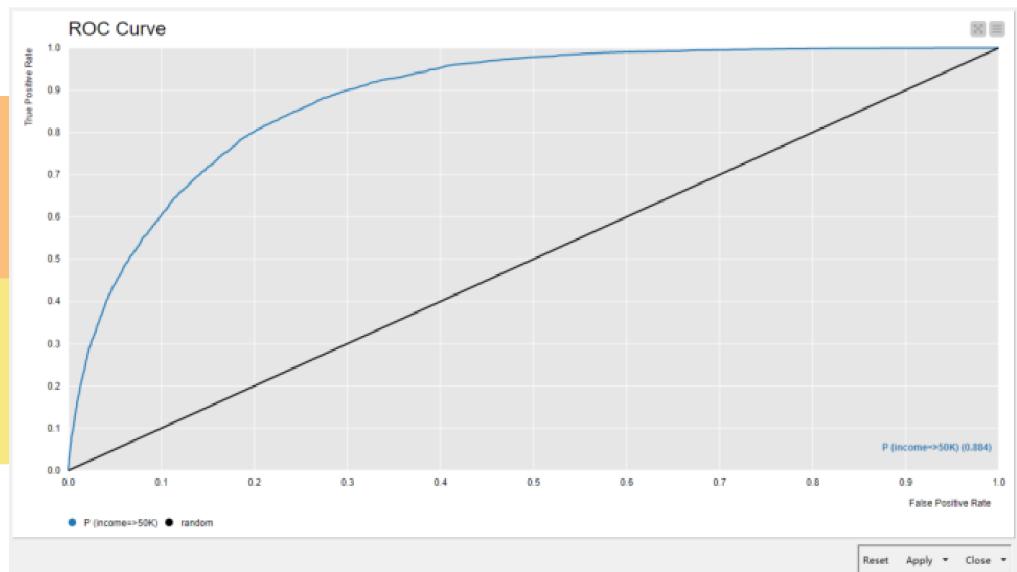
# Q9: PRODUCE A BAR CHART FOR EACH OF THE CATEGORICAL VARIABLES SHOWING THE COUNTS FOR EACH UNIQUE VALUE



# QUES 10 : MODEL 1 - NATIVE BAYES

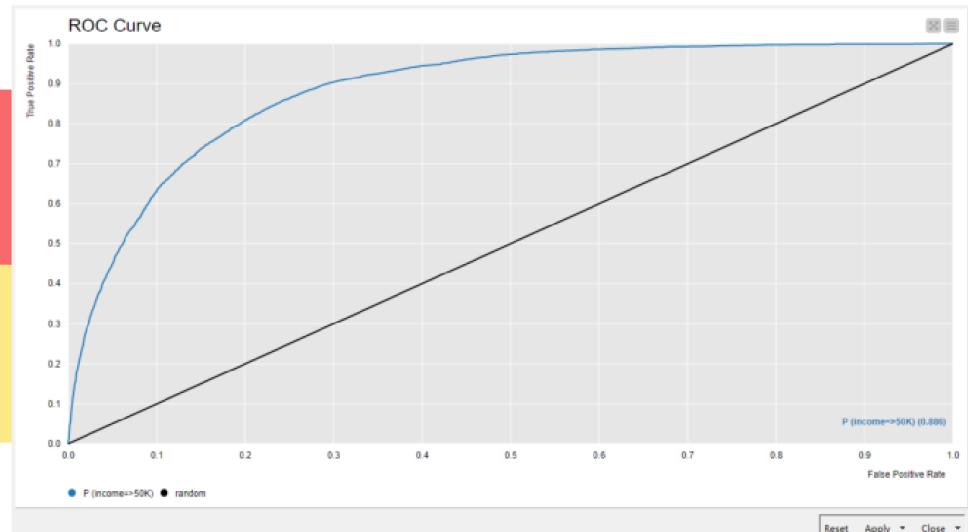
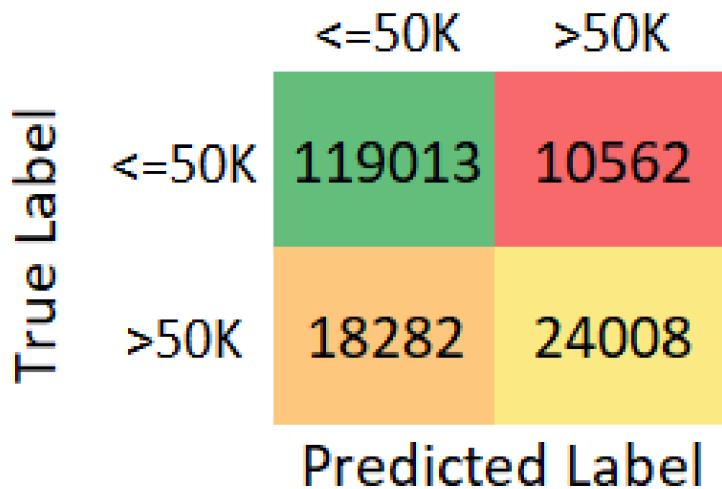
## CONFUSION TABLE / ROC CURVE! / METRICS

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



| RowID   | TruePositives | FalsePositives | TrueNegatives | FalseNegatives | Recall | Precision | Sensitivity | Specificity | F-measure | Accuracy | Cohen's kappa |
|---------|---------------|----------------|---------------|----------------|--------|-----------|-------------|-------------|-----------|----------|---------------|
| >50K    | 31298         | 20909          | 108666        | 10992          | 0.74   | 0.60      | 0.74        | 0.84        | 0.66      |          |               |
| <=50K   | 108666        | 10992          | 31298         | 20909          | 0.84   | 0.91      | 0.84        | 0.74        | 0.87      |          |               |
| Overall |               |                |               |                |        |           |             |             |           | 0.81     | 0.54          |

# REGRESSION CONFUSION TABLE / ROC CURVE! / METRICS



| RowID   | TruePositives | FalsePositives | TrueNegatives | FalseNegatives | Recall | Precision | Sensitivity | Specificity | F-measure | Accuracy | Cohen's kappa |
|---------|---------------|----------------|---------------|----------------|--------|-----------|-------------|-------------|-----------|----------|---------------|
| >50K    | 24008         | 10562          | 119013        | 18282          | 0.57   | 0.69      | 0.57        | 0.92        | 0.62      |          |               |
| <=50K   | 119013        | 18282          | 24008         | 10562          | 0.92   | 0.87      | 0.92        | 0.57        | 0.89      |          |               |
| Overall |               |                |               |                |        |           |             |             |           | 0.83     | 0.52          |

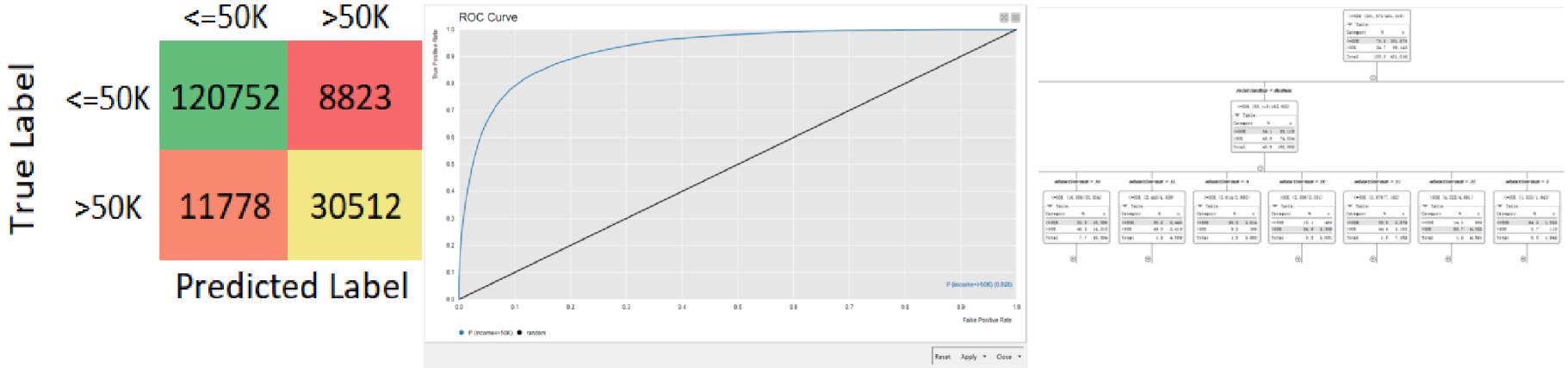


# QUES 11 : MODEL 2 – LOGISTIC REGRESSION COEFFICIENTS

| Row ID | Logit | Variable                                  | Coeff. | Std. Err.  | z-score | P> z  |
|--------|-------|---|--------|------------|---------|-------|
| Row1   | <=50K | age                                       | -0.028 | 0          | -61.598 | 0     |
| Row2   | <=50K | fnlwgt                                    | -0     | 0          | -8.39   | 0     |
| Row3   | <=50K | hours-per-week                            | -0.067 | 0.001      | -51.421 | 0     |
| Row27  | <=50K | Divorced_marital-status                   | -0.362 | 0.022      | -16.695 | 0     |
| Row29  | <=50K | Separated_marital-status                  | -0.357 | 0.04       | -9.013  | 0     |
| Row52  | <=50K | White_race                                | -0.105 | 0.018      | -5.742  | 0     |
| Row28  | <=50K | Married-spouse-absent_marital-status      | -0.34  | 0.062      | -5.47   | 0     |
| Row54  | <=50K | Amer-Indian-Eskimo_race                   | 0.358  | 0.073      | 4.934   | 0     |
| Row30  | <=50K | Married-AF-spouse_marital-status          | -0.467 | 0.141      | -3.319  | 0.001 |
| Row53  | <=50K | Asian-Pac-Islander_race                   | -0.149 | 0.056      | -2.68   | 0.007 |
| Row31  | <=50K | Widowed_marital-status                    | -0.111 | 0.042      | -2.631  | 0.009 |
| Row55  | <=50K | Other_race                                | 0.164  | 0.075      | 2.184   | 0.029 |
| Row87  | <=50K | Outlying-US(Guam-USVI-etc)_native-country | 13.83  | 8,701.044  | 0.002   | 0.999 |
| Row25  | <=50K | Preschool_education                       | 13.077 | 9,072.319  | 0.001   | 0.999 |
| Row97  | <=50K | Constant                                  | 7.37   | 6,842.552  | 0.001   | 0.999 |
| Row96  | <=50K | Holand-Netherlands_native-country         | 6.423  | 8,701.052  | 0.001   | 0.999 |
| Row10  | <=50K | Without-pay_workclass                     | 13.553 | 20,580.814 | 0.001   | 0.999 |
| Row72  | <=50K | Cambodia_native-country                   | -2.61  | 8,702.801  | -0      | 1     |
| Row20  | <=50K | Doctorate_education                       | -2.64  | 9,075.185  | -0      | 1     |
| Row21  | <=50K | Prof-school_education                     | -2.522 | 9,066.361  | -0      | 1     |
| Row14  | <=50K | Masters_education                         | -1.845 | 9,061.882  | -0      | 1     |
| Row71  | <=50K | Columbia_native-country                   | 1.581  | 8,702.801  | 0       | 1     |
| Row11  | <=50K | Bachelors_education                       | -1.432 | 9,066.548  | -0      | 1     |
| Row22  | <=50K | 5th-6th_education                         | 1.196  | 9,064.162  | 0       | 1     |
| Row68  | <=50K | Philippines_native-country                | -1.134 | 8,702.801  | -0      | 1     |
| Row65  | <=50K | Canada_native-country                     | -1.118 | 8,702.801  | -0      | 1     |



# QUES 12 : MODEL 3 – DECISION TREE CONFUSION TABLE / ROC CURVE! / METRICS



| RowID   | TruePositives | FalsePositives | TrueNegatives | FalseNegatives | Recall | Precision | Sensitivity | Specificity | F-measure | Accuracy | Cohen's kappa |
|---------|---------------|----------------|---------------|----------------|--------|-----------|-------------|-------------|-----------|----------|---------------|
| >50K    | 120752        | 11778          | 30512         | 8823           | 0.93   | 0.91      | 0.93        | 0.72        | 0.92      |          |               |
| <=50K   | 30512         | 8823           | 120752        | 11778          | 0.72   | 0.78      | 0.72        | 0.93        | 0.75      |          |               |
| Overall |               |                |               |                |        |           |             |             |           | 0.88     | 0.67          |



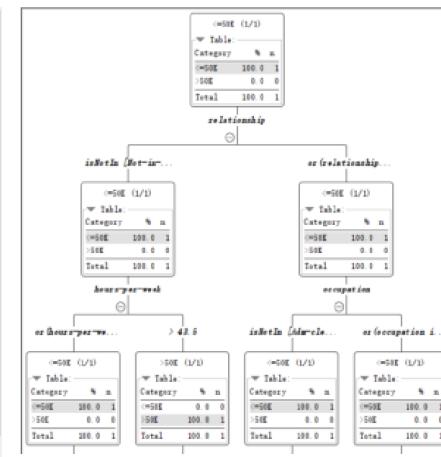
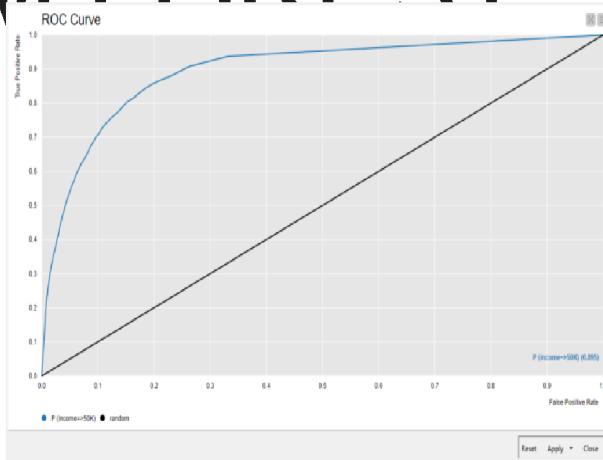
# QUES 12 : MODEL 3 – DECISION TREE IMPORTANT VARIABLES/DECISION RULES

| Row ID  | Condition   | Outcome | Record count | Number of correct |
|---------|---|---------|--------------|-------------------|
| 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        | 4,027             |
| 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    | 3,062        | 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             |
| Row51   | \$workclass\$ = "Private" AND \$occupation\$ = "Exec-managerial" AND \$edu...     | >50K    | 2,788        | 2,563             |
| Row240  | \$marital-status\$ = "Never-married" AND \$education-num\$ = "12" AND \$rel...    | <=50K   | 2,638        | 2,411             |
| Row205  | \$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             |
| Row241  | \$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             |
| Row1495 | \$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             |
| Row1387 | \$education-num\$ = "7" AND \$relationship\$ = "Unmarried"                        | <=50K   | 1,773        | 1,725             |
| Row1446 | \$workclass\$ = "Private" AND \$education-num\$ = "11" AND \$relationship\$ ...   | <=50K   | 1,739        | 1,680             |
| Row497  | \$hours-per-week\$ <= 43.0 AND \$age\$ <= 44.5 AND \$occupation\$ = "Handler..."  | <=50K   | 1,667        | 1,509             |
| Row337  | \$fnlwgt\$ <= 266.0 AND \$hours-per-week\$ <= 46.0 AND \$age\$ <= 43.5 AND \$...  | >50K    | 1,662        | 1,323             |
| Row364  | \$fnlwgt\$ <= 202.5 AND \$age\$ > 32.5 AND \$age\$ <= 65.0 AND \$fnlwgt\$ <= 4... | >50K    | 1,492        | 1,190             |
| Row1453 | \$education-num\$ = "6" AND \$relationship\$ = "Unmarried"                        | <=50K   | 1,444        | 1,374             |
| Row1329 | \$marital-status\$ = "Separated" AND \$relationship\$ = "Own-child"               | <=50K   | 1,374        | 1,374             |



# QUES 12 : MODEL 4 – RANDOM FOREST

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



| RowID   | TruePositives | FalsePositives | TrueNegatives | FalseNegatives | Recall | Precision | Sensitivity | Specificity | F-measure | Accuracy | Cohen's kappa |
|---------|---------------|----------------|---------------|----------------|--------|-----------|-------------|-------------|-----------|----------|---------------|
| >50K    | 121289        | 16944          | 25346         | 8286           | 0.94   | 0.88      | 0.94        | 0.60        | 0.91      |          |               |
| <=50K   | 25346         | 8286           | 121289        | 16944          | 0.60   | 0.75      | 0.60        | 0.94        | 0.67      |          |               |
| Overall |               |                |               |                |        |           |             |             |           | 0.85     | 0.58          |



# QUES 12 : MODEL 4 – RANDOM FOREST IMPORTANT VARIABLES/DECISION RULES

| Row ID         | #splits (level 0) | #splits (level 1) | #splits (level 2) | #candidates (level 0) | #candidates (level 1) | #candidates (level 2) |
|----------------|-------------------|-------------------|-------------------|-----------------------|-----------------------|-----------------------|
| marital-status | 25                | 25                | 40                | 30                    | 47                    | 87                    |
| relationship   | 20                | 30                | 43                | 20                    | 39                    | 88                    |
| education      | 14                | 23                | 55                | 19                    | 36                    | 84                    |
| education-num  | 11                | 23                | 65                | 21                    | 37                    | 95                    |
| age            | 7                 | 20                | 52                | 22                    | 37                    | 86                    |
| workclass      | 5                 | 10                | 19                | 29                    | 44                    | 74                    |
| occupation     | 5                 | 21                | 52                | 16                    | 29                    | 83                    |
| sex            | 5                 | 11                | 12                | 18                    | 38                    | 80                    |
| hours-per-week | 5                 | 22                | 31                | 24                    | 61                    | 93                    |
| native-country | 2                 | 8                 | 17                | 24                    | 40                    | 87                    |
| race           | 1                 | 5                 | 8                 | 23                    | 51                    | 87                    |
| fnlwgt         | 0                 | 2                 | 5                 | 19                    | 47                    | 80                    |
| capital-gain   | 0                 | 0                 | 0                 | 20                    | 51                    | 85                    |
| capital-loss   | 0                 | 0                 | 0                 | 15                    | 43                    | 91                    |



# QUES 13 : COMPARISION!

|                         | 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         |

