Qualifying Exam

Chathrua Gunasekara

December 22, 2014

1.

a. Using the poisson distribution, p(X=0) given lambda = 3.2. probability that a particular month will have no accidents = 4%

```
dpois(0,3.2)

 0.0407622
```

b. mean and the variance of the Poisson distribution are both equal to lambda.

Expected Value = lambda x 1 = 3.2

variance = lambda= 3.2

2.

a) $p(y \leq 2) = p(y = 0) + p(y = 1) + p(y = 2) = 0.005 + 0.010 + 0.035 = 0.05$

b) $p(y = 1, y = 2, y = 3, y = 4) = 1 - p(y = 0) = 1 - 0.005 = 0.995$

c) $E(y) = \sum y.p(y) = 0 \cdot (0.005) + 1 \cdot 0.010 + 2 \cdot 0.035 + 3 \cdot 0.050 + 4 \cdot 0.900 = 3.83$

d) $E(y^2) = \sum y^2 \cdot p(y)$

$=0^2 \cdot (0.005) + 1^2 \cdot 0.010 + 2^2 \cdot 0.035 + 3^2 \cdot 0.050 + 4^2 \cdot 0.900 = 15$

So, the variance, $\sigma^2 = E(y^2) - E(y)^2$

= 15-(3.83)^2 = 0.3311

And standard deviation,

$\sigma = \sqrt{0.3311} = 0.575413$

3.

```
mu = 70
sigma = 3
z socre for 64
(64-70)/3

z = -2

z socre for 76

(76-70)/3

z = 2

#between -2 and +2
pnorm(2) - pnorm(-2)

0.9544997

What % of males will be between 64 and 76 inches tall = 95.44 %
```

4.

```
a)Sample mean
values <-c(13.3,14.5,15.3,15.3,14.3,14.8,15.2,14.9,14.6,14.1)
mean(values)

14.63

b)sample variance
var(values)

0.389

sample standard deviation
sqrt(var(values))

0.6236986



c)

H0 : mu0 = 14.9

H1 : mu0 ≠ 14.9


xbar <- 14.63           # sample mean
mu0 <- 14.90            # hypothesized value
s <- sqrt(0.389)        # sample standard deviation
n <- 10                 # sample size

test_statistic <- (xbar-mu0)/(s/sqrt(n))
test_statistic = -1.368954
```

```
alpha = .01
df <- n-1
t.half.alpha <- qt(1-alpha/2,df=n-1)
c(-t.half.alpha,t.half.alpha)
```

Confidence Interval :

(-3.249836 , 3.249836), If test statistics is between this interval we can not reject the null hypothesis.

Or using P values :

```
pval <- 2*pt(test_statistic,df=n-1)
```

pval = 0.2042047 > alpha therefor can not reject the null hypothesis.

d.

```
e <-qt(0.9,df=n-1)*s # Margin of Error
```

e = 0.8625932

```
c(xbar-e,xbar+e)
```

Confidence Interval is (13.76741 15.49259)

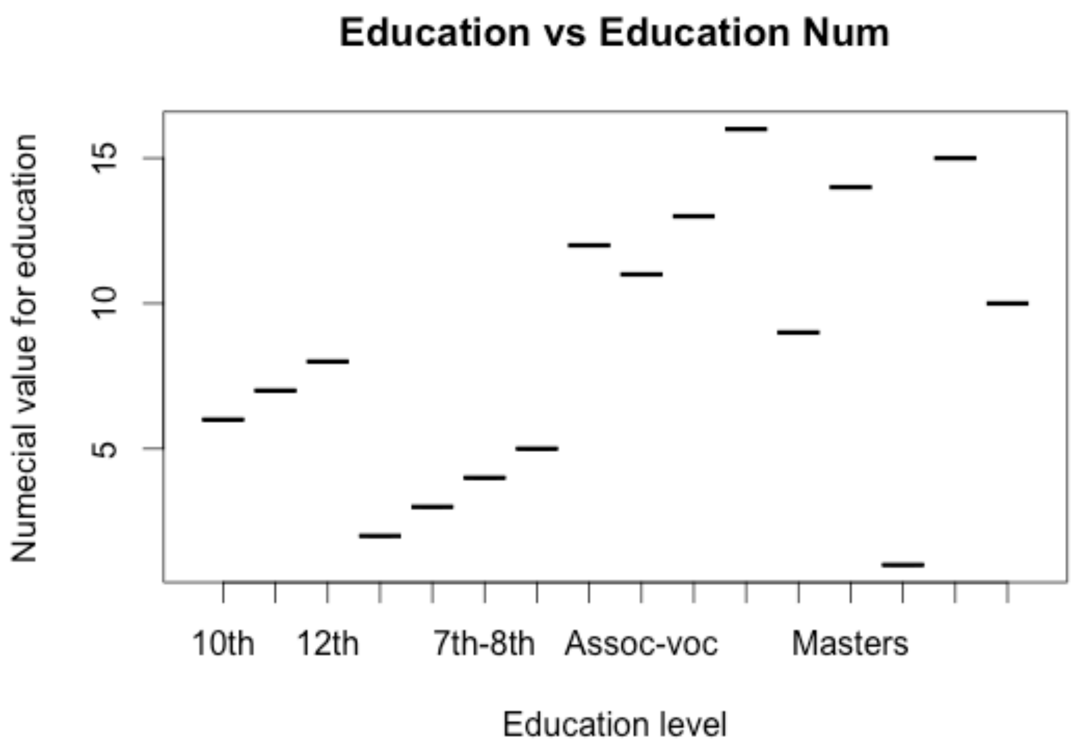
```
#true mean is between this confidence interval so it has not been changed.
```

5)

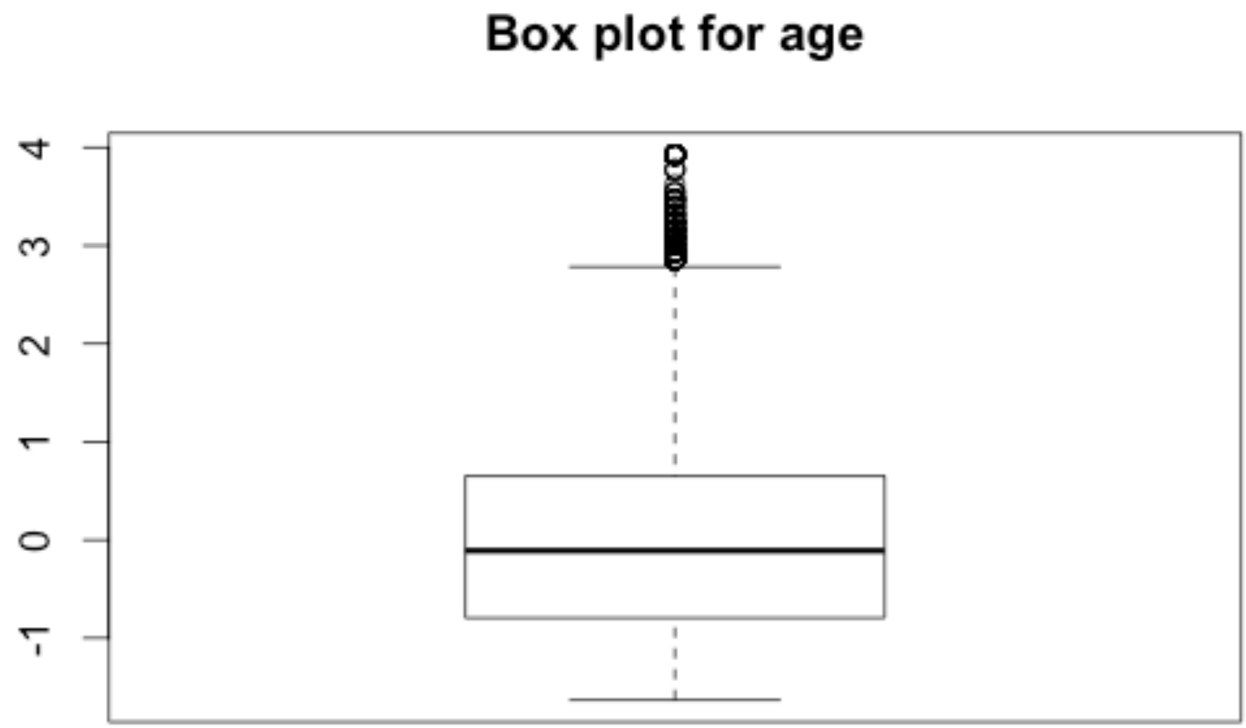
1. c
2. b
3. c
4. c
5. a

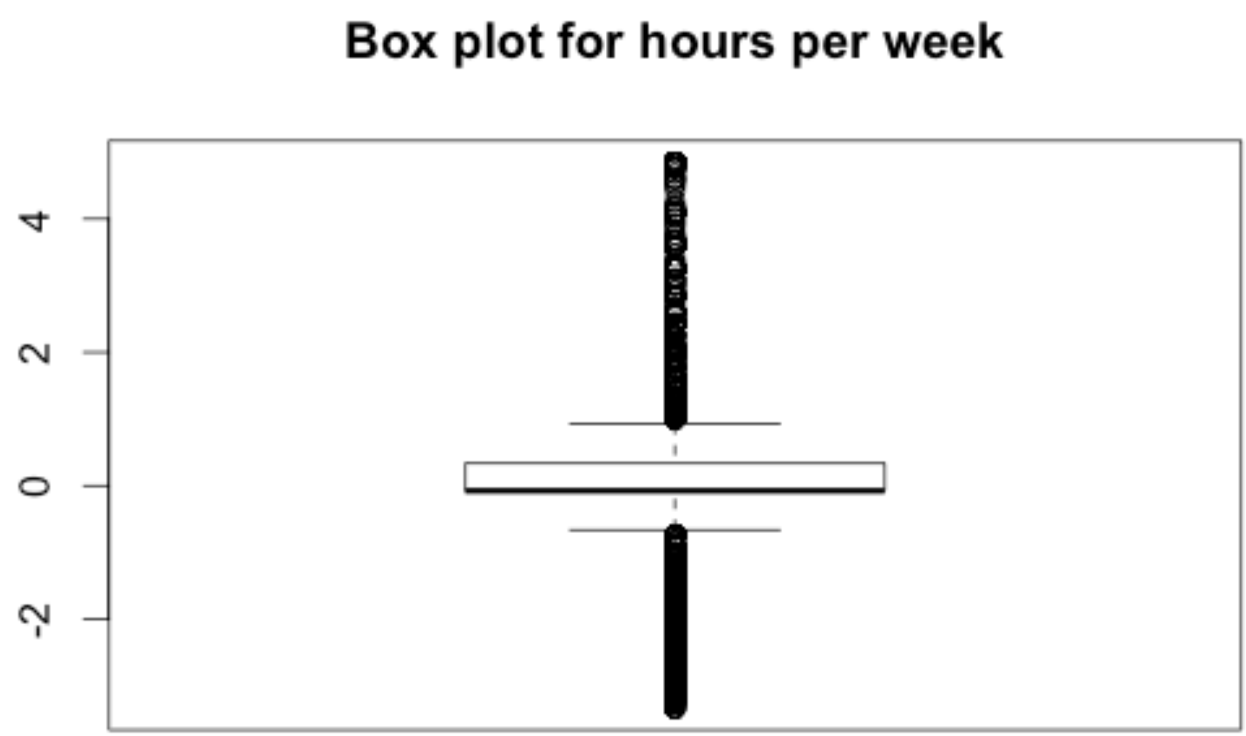
6)

1) By observing the data set the categorical predictor “education” and the continuous predictor “education_num” represents the same information. So education_num was removed initially.

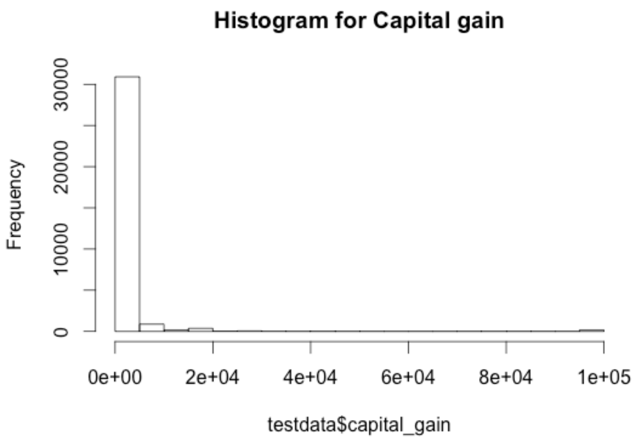


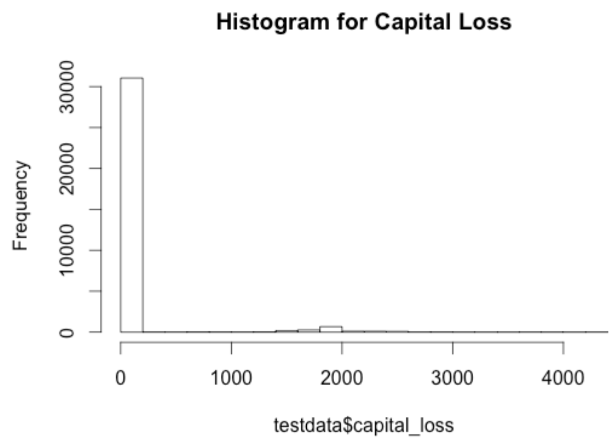
2. age and hr_per_week Continuous predictors were choose to scale. This applies a normal transformation. Each value minus its mean over the sample standard deviation. Following is the boxplots for these continuous predictors.



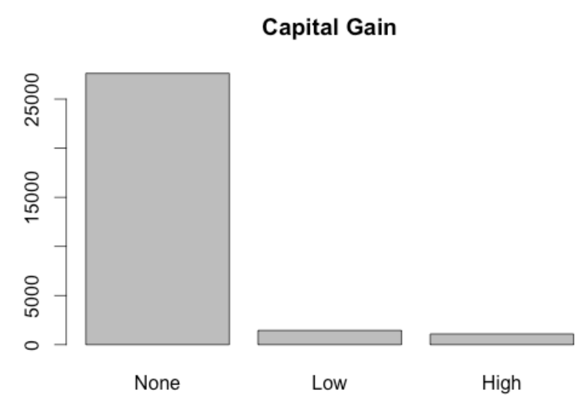


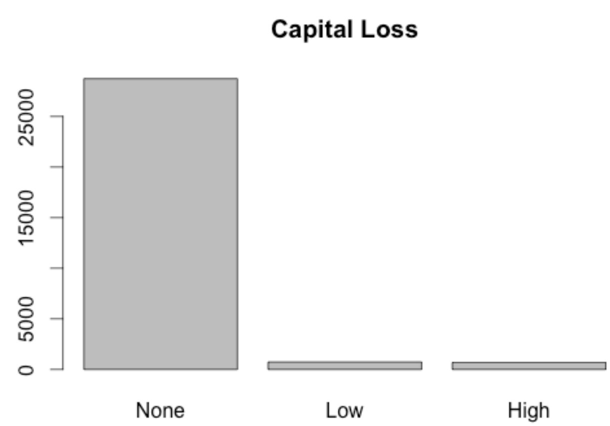
3. The capital gain and capital loss predictors are extremely skewed. No transformation could correct this. So the predictors were coverted to categorical varables with (None,Low,High) factors.





Because of the high skewers numerical transformation would not have been appropriate so Converted to Categorical variables. For both variables, none means they don't play the market. Low means they have some investments. High means they have significant investments.





4. nearzerovar() function return following predictors are degenerate.

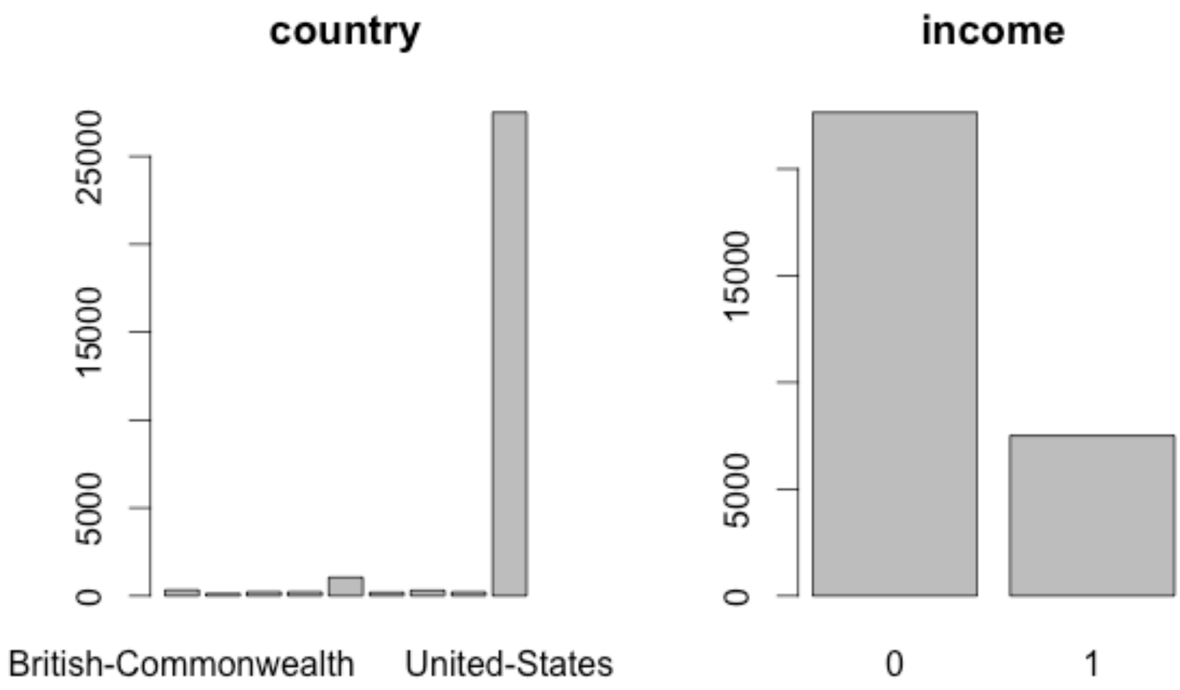
[1] "capital_gain"

[1] "capital_loss"

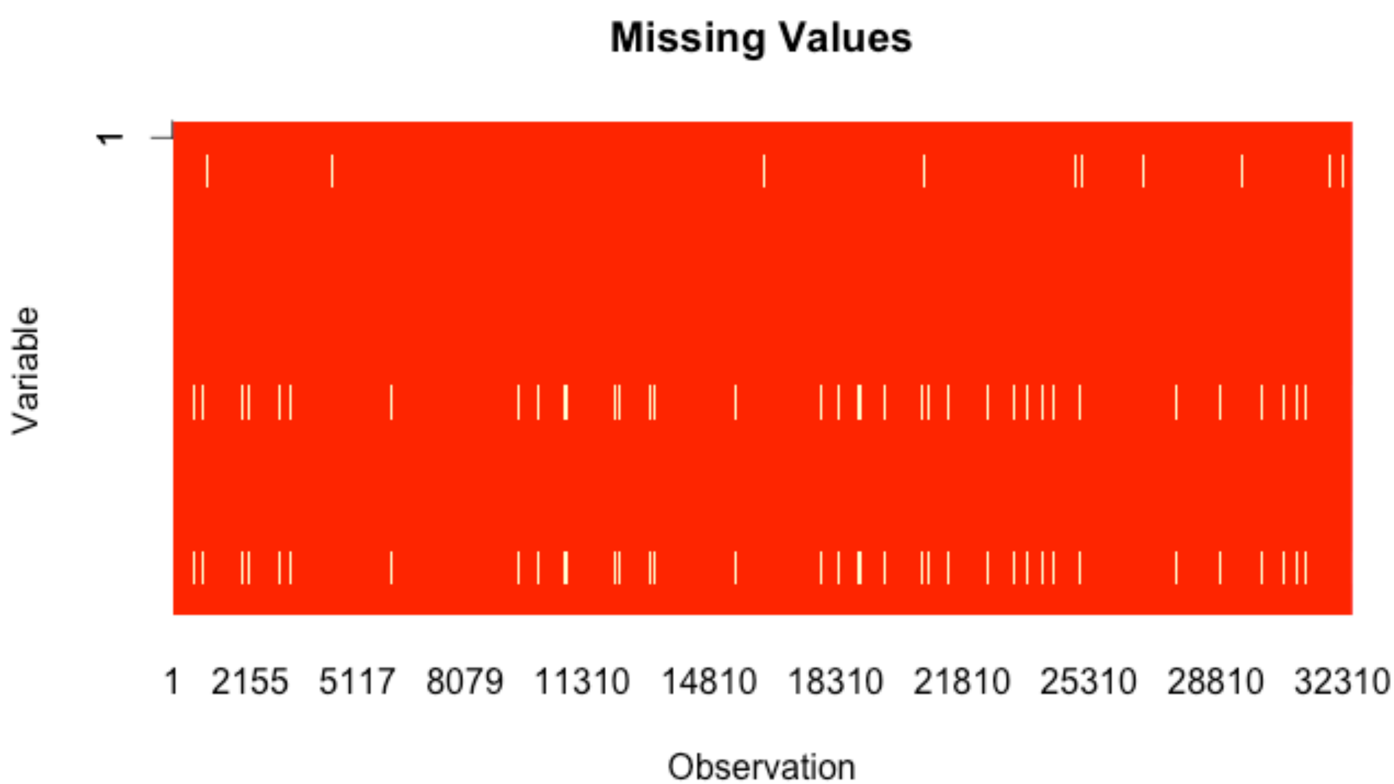
[1] "country"

Some predictors were recotagorized to include in to a broader category to reduce the number of factores in a predictor. (occupation,marital status, ect)


type_employer
15000
0
Federal-Govt
Private
education
6000
0
Associates
Dropout
Masters
martial
8000
0
Married
Not-Married
occupaion
6000
0
Admin
Professional
race
20000
0
Amer-Indian
Black
White
sex
15000
0
Female
Male
capital gain
20000
0
None
Low
High
capital loss
20000
0
None
Low
High



5. Missing value visualization



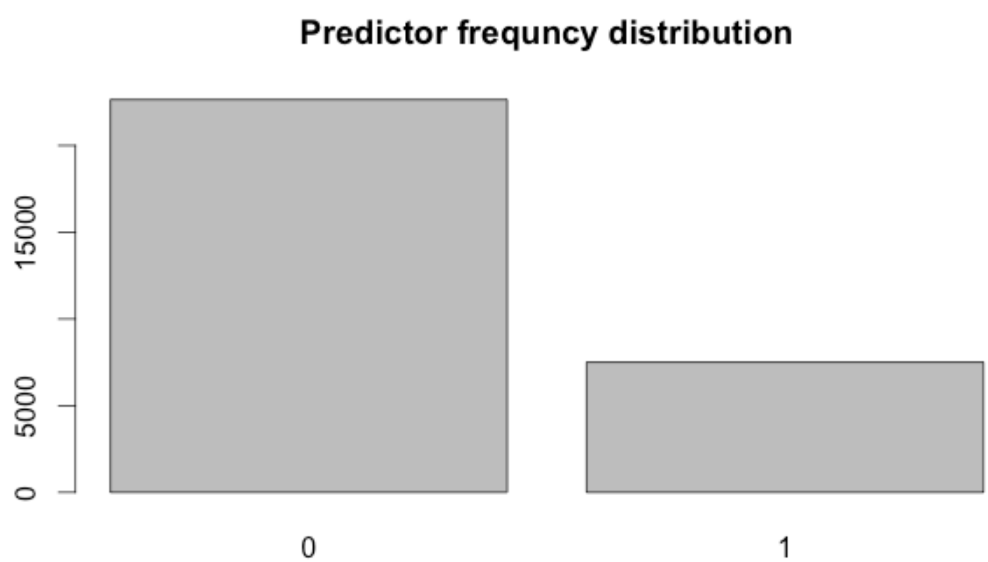
Predictor	Overall Number of missing values
Type_employer	1836
Occupation	1843
country	583

Yes. In low income (<=50,000) data points the number of missing values are high. Because low income people may not be able to give information more confidently as high income people. So there can be many missing values in a low in come data point.

6. Na.omit() function is used to omit the data with missing values beause it does not affet much as there are still more than 30,000 datapoint with complete data. Only about 2000 datapoint had to be removed because the missing data.

7. ROC or Kappa statistic should be used. Accuracy may not be the best statistic because the income class imbalance.

8. Response class is imbalanced.



Stratified sampling should be used to split the dataset in to testing and training sets. The createDataPartition() function is used with p=0.75. which split the data set in to 75% training and 25% testing set.

9. Preprocessing steps:

i. Remove high correlated predictors

ii. Merge factors in categorical predictors so it will reduced the number of dummy variables created. For example the country predictors can be re categorized according the larger geographical region such as (Asia, Europe, South America, Africa, ect)

iii. Remove data points with missing values

iv. Remove no information predictors (fnlwgt) is just a number with no relevance to the income, such as an ID number.

v. Remove zero variance predictors

30162 rows and 10 categorical Predictors are remaining after all the preprocessing. This dataset was split to following training and testing set.

22622 for training set

7540 for testing set

To be used in some models, Dummy variables were created for all the categorical predictors using the same dataframe(30162 rows and 10 column) mentioned above. Then number of predictors increased to 28 predictors.

10. Linear Models

Model	Tuning Para.	AUC	Sensitivity	Specificity
Logistic reg	------No---	0.87996	0.91717	0.55448
LDA	-------No---	0.86991	0.91458	0.54237
PLSDA	Ncome = 10	0.86886	0.9301	0.50191
GLMNet	alpha = 0 and lambda = 0.1	0.86384	0.93842	0.45082
Nearest S C	Threshold =0	0.84545	0.94131	0.39741

Important predictors from Linear model - PLSDA

age 0.08234

marital.Married 0.07472

relationship.Husband 0.06400

marital.Never-Married 0.04421

occupation.Blue-Collar 0.03361

occupation.White-Collar 0.03196

education.Bachelors 0.03138

education.HS-grad 0.03087

sex.Male 0.02983

sex.Female 0.02983

relationship.Not-in-family 0.02857

education.Dropout 0.02854

occupation.Professional 0.02814

relationship.Own-child 0.02513

marital.Not-Married 0.02473

occupation.Service 0.02161

education.Masters 0.01998

relationship.Unmarried 0.01714

type_employer.Private 0.01630

education.HS-Graduate 0.01203

11.

Model	Tuning Para	ROC	Sensitivity	Specificity
MDA	Subclasses = 1	0.87640	0.91652	0.55499
NNet	Size = 1, decay =0.1	0.8872	0.90826	0.57705
FDA	degree = 1 and nprune = 17.	0.8763	0.9184	0.5478
SVM	C=8	0.8793	0.9235	0.5818
KNN	K=9	0.8809	0.8974	0.5715
NaiveBayes	Laplace = 2	0.8723	0.8652	0.5253

12. Best models based on AUC

1. Neural Network , 2. KNN 3. SVM – from Non Linear models

1. Logistic Reg 2. LDA 3 . PLSDA – from Linea modelss

13.

From Linear best models are Logisted ,LDA, PLSDA

From Non linear best models are NeuralNet,

model	Accuracy	Sensitivity	Specificit	Kappa
Logistic	0.8309	0.9202	0.5615	0.5158
LDA	0.8256	0.9163	0.5519	0.501
PLSDA	0.8236	0.9302	0.5019	0.4781
Nnet	0.8524	0.9129	0.5992	0.6212
KNN	0.8413	0.9135	0.5832	0.5419
SVM	0.8245	0.9161	0.5124	0.524

Based on the Kappa statistic Neural Network model is the best to classify the income.

14. Important predictors

maritalNever-Married 100.000

maritalNot-Married 83.698

educationBachelors 61.242

educationMasters 61.242

educationProf-School 54.926

age 48.643

educationDoctorate 42.061

occupationBlue-Collar 39.170

occupationWhite-Collar 36.761

maritalWidowed 30.950

hr_per_week 22.221

educationDropout	14.719
sexMale	11.332
relationshipWife	8.994
raceAsian	0.000
occupationMilitary	0.000
raceWhite	0.000
occupationSales	0.000
occupationOther-Occupations	0.000
relationshipNot-in-family	0.000

Generalized Linear Model

22622 samples

9 predictor

2 classes: '0', '1'

Resampling: Repeated Train/Test Splits Estimated (25 reps, 0.75%)

Summary of sample sizes: 16968, 16968, 16968, 16968, 16968, 16968, ...

Resampling results

ROC	Sens	Spec	ROC SD	Sens SD	Spec SD
0.8799647	0.9171745	0.5544847	0.003771528	0.003718161	0.01107291

Linear Discriminant Analysis

22622 samples

28 predictor

2 classes: '0', '1'

Resampling: Repeated Train/Test Splits Estimated (25 reps, 0.75%)

Summary of sample sizes: 16968, 16968, 16968, 16968, 16968, 16968, ...

Resampling results

ROC	Sens	Spec	ROC SD	Sens SD	Spec SD
0.8699113	0.9145844	0.5423738	0.004330018	0.0034694	0.01161766

Partial Least Squares

22622 samples

28 predictor

2 classes: '0', '1'

Resampling: Repeated Train/Test Splits Estimated (25 reps, 0.75%)

Summary of sample sizes: 16968, 16968, 16968, 16968, 16968, 16968, ...

Resampling results across tuning parameters:

ncomp	ROC	Sens	Spec	ROC SD	Sens SD	Spec SD
1	0.8210978	0.9254815	0.3017200	0.005172884	0.004257385	0.011524530
2	0.8484582	0.9169108	0.5011230	0.004623632	0.003693251	0.009662471
3	0.8556016	0.9229951	0.5017484	0.005319495	0.002918410	0.011519707
4	0.8613771	0.9274217	0.4864534	0.005105521	0.002653747	0.012626050
5	0.8632581	0.9286461	0.4915139	0.005045462	0.003048491	0.012534031
6	0.8648489	0.9273558	0.5007249	0.005025836	0.003253502	0.011862911
7	0.8666984	0.9302190	0.4949538	0.004959947	0.003485078	0.012873540
8	0.8679223	0.9303132	0.4993035	0.005013923	0.002984934	0.012224660
9	0.8681823	0.9301342	0.4976262	0.005143483	0.003351018	0.012741545
10	0.8688623	0.9301436	0.5019190	0.005076225	0.003489609	0.013201262

ROC was used to select the optimal model using the largest value.

The final value used for the model was ncomp = 9.

glmnet

22622 samples

28 predictor

2 classes: '0', '1'

Resampling: Repeated Train/Test Splits Estimated (25 reps, 0.75%)

Summary of sample sizes: 16968, 16968, 16968, 16968, 16968, 16968, ...

Resampling results across tuning parameters:

alpha	lambda	ROC	Sens	Spec	ROC SD	Sens SD	Spec SD
0.0	0.1	0.8638485	0.9384224	0.450859986	0.005655341	0.0043027220	0.012347593
0.0	0.2	0.8610042	0.9563551	0.375977257	0.005745578	0.0034350049	0.010700393
0.0	20.0	0.8480020	1.0000000	0.000000000	0.005945144	0.0000000000	0.000000000
0.2	0.1	0.8604895	0.9608288	0.340810235	0.005702182	0.0035831254	0.012370141
0.2	0.2	0.8531465	0.9989451	0.009182658	0.005966442	0.0007356016	0.005218438
0.2	20.0	0.5000000	1.0000000	0.000000000	0.000000000	0.0000000000	0.000000000
0.6	0.1	0.8454274	1.0000000	0.000000000	0.006374152	0.0000000000	0.000000000
0.6	0.2	0.7587508	1.0000000	0.000000000	0.005908302	0.0000000000	0.000000000
0.6	20.0	0.5000000	1.0000000	0.000000000	0.000000000	0.0000000000	0.000000000
0.8	0.1	0.8067229	1.0000000	0.000000000	0.016570798	0.0000000000	0.000000000
0.8	0.2	0.7575690	1.0000000	0.000000000	0.005723933	0.0000000000	0.000000000
0.8	20.0	0.5000000	1.0000000	0.000000000	0.000000000	0.0000000000	0.000000000

ROC was used to select the optimal model using the largest value.

The final values used for the model were alpha = 0 and lambda = 0.1.

Nearest Shrunken Centroids

22622 samples

28 predictor

2 classes: '0', '1'

Pre-processing: centered, scaled

Resampling: Repeated Train/Test Splits Estimated (25 reps, 0.75%)

Summary of sample sizes: 16968, 16968, 16968, 16968, 16968, 16968, ...

Resampling results across tuning parameters:

threshold	ROC	Sens	Spec	ROC SD	Sens SD	Spec SD
0	0.8454513	0.9413139	0.397412935	0.004319150	0.0039207738	0.009172483
1	0.8453774	0.9468707	0.370206112	0.004381660	0.0037334138	0.008593492
2	0.8452393	0.9541323	0.338678038	0.004470150	0.0035947885	0.009051073
3	0.8450614	0.9622227	0.299587775	0.004512389	0.0030314698	0.009592952
4	0.8447973	0.9746645	0.230362473	0.004561640	0.0026193308	0.008477597
5	0.8444082	0.9863998	0.140639659	0.004629258	0.0025697198	0.008070591
6	0.8437068	0.9958276	0.045088842	0.004727403	0.0009908936	0.008072573
7	0.8424795	0.9995950	0.002501777	0.004913761	0.0003755082	0.001437952
8	0.8396445	1.0000000	0.000000000	0.005264669	0.0000000000	0.000000000
9	0.8346362	1.0000000	0.000000000	0.005436800	0.0000000000	0.000000000
10	0.8289169	1.0000000	0.000000000	0.005691125	0.0000000000	0.000000000
11	0.8190368	1.0000000	0.000000000	0.006370865	0.0000000000	0.000000000
12	0.8092706	1.0000000	0.000000000	0.006857796	0.0000000000	0.000000000
13	0.8062404	1.0000000	0.000000000	0.005685198	0.0000000000	0.000000000
14	0.8027894	1.0000000	0.000000000	0.006408375	0.0000000000	0.000000000
15	0.7925135	1.0000000	0.000000000	0.007596168	0.0000000000	0.000000000
16	0.7691202	1.0000000	0.000000000	0.008435752	0.0000000000	0.000000000

17	0.7664731	1.0000000	0.000000000	0.005018473	0.0000000000	0.000000000
18	0.7664731	1.0000000	0.000000000	0.005018473	0.0000000000	0.000000000
19	0.7664731	1.0000000	0.000000000	0.005018473	0.0000000000	0.000000000
20	0.7664731	1.0000000	0.000000000	0.005018473	0.0000000000	0.000000000
21	0.7664731	1.0000000	0.000000000	0.005018473	0.0000000000	0.000000000
22	0.7571205	1.0000000	0.000000000	0.005208846	0.0000000000	0.000000000
23	0.7571205	1.0000000	0.000000000	0.005208846	0.0000000000	0.000000000
24	0.7571205	1.0000000	0.000000000	0.005208846	0.0000000000	0.000000000
25	0.7571205	1.0000000	0.000000000	0.005208846	0.0000000000	0.000000000

ROC was used to select the optimal model using the largest value.

The final value used for the model was threshold = 0.

Neural Network

22622 samples

9 predictor

2 classes: '0', '1'

No pre-processing

Resampling: Repeated Train/Test Splits Estimated (25 reps, 0.75%)

Summary of sample sizes: 16968, 16968, 16968, 16968, 16968, 16968, ...

Resampling results across tuning parameters:

size	decay	ROC	Sens	Spec	ROC SD	Sens SD	Spec SD
1	0.000	0.8705316	0.8970944	0.5891969	0.014219352	0.015329727	0.03285999
1	0.001	0.8787454	0.9099788	0.5737313	0.006186793	0.004157792	0.01368856
1	0.100	0.8802671	0.9082647	0.5770576	0.003869122	0.004125598	0.01144904
1	1.000	0.8802317	0.9053826	0.5840512	0.003832171	0.004103522	0.01116490

ROC was used to select the optimal model using the largest value.

The final values used for the model were size = 1 and decay = 0.1.

Flexible Discriminant Analysis

22622 samples

9 predictor

2 classes: '0', '1'

Resampling: Repeated Train/Test Splits Estimated (25 reps, 0.75%)

Summary of sample sizes: 16968, 16968, 16968, 16968, 16968, 16968, ...

Resampling results across tuning parameters:

nprune	ROC	Sens	Spec	ROC SD	Sens SD	Spec SD
2	0.3292055	1.0000000	0.0000000	0.004894496	0.000000000	0.0000000
9	0.8477668	0.9327431	0.4535608	0.005429629	0.009875514	0.0322599
17	0.8763237	0.9184083	0.5478038	0.004233990	0.005295729	0.0110436

Tuning parameter 'degree' was held constant at a value of 1

ROC was used to select the optimal model using the largest value.

The final values used for the model were degree = 1 and nprune = 17.

k-Nearest Neighbors

22622 samples

28 predictor

2 classes: '0', '1'

Resampling: Repeated Train/Test Splits Estimated (25 reps, 0.75%)

Summary of sample sizes: 16968, 16968, 16968, 16968, 16968, 16968, ...

Resampling results across tuning parameters:

k	ROC	Sens	Spec	ROC SD	Sens SD	Spec SD
3	0.8233105	0.8877796	0.5502203	0.005330046	0.005281699	0.01379943
5	0.8428730	0.8939393	0.5613362	0.004777924	0.005070641	0.01581545
9	0.8580986	0.8974523	0.5715991	0.004465693	0.005023055	0.01416908

ROC was used to select the optimal model using the largest value.

The final value used for the model was k = 9.

Naive Bayes Classifier for Discrete Predictors

Call:

naiveBayes.default(x = trainX, y = trainY)

A-priori probabilities:

```
trainY
       0        1 
0.751083 0.248917 
```

Conditional probabilities:

```
      age
trainY       [,1]      [,2]
     0 -0.1348400 1.0285740
     1  0.4130562 0.7764739
```

type_employer

trainY	Federal-Govt	Not-Working	Other-Govt	Private	Self-Employed
0	0.0250720970	0.0005296922	0.1054676005	0.7698781708	0.0990524395
1	0.0458177944	0.0000000000	0.1287515539	0.6515716569	0.1738589948

education

trainY	Associates	Bachelors	Doctorate	Dropout	HS-grad	HS-Graduate	Masters	Prof-School
0	0.075922547	0.127596963	0.003884409	0.155258666	0.361838620	0.237949503	0.031722677	0.005826614
1	0.081690641	0.283786184	0.035872847	0.030190020	0.214526727	0.177943527	0.121115255	0.054874800

marital

trainY	Married	Never-Married	Not-Married	Widowed
0	0.33753163	0.40839268	0.22023424	0.03384145
1	0.85277926	0.06233351	0.07423193	0.01065530

occupation

trainY	Admin	Blue-Collar	Military	Other-Occupations	Professional	Sales	Service	White-Collar
0	0.1421929257	0.3660761580	0.0002354188	0.0460243658	0.0971102348	0.1157083162	0.1430168913	0.0896356895
1	0.0674835731	0.2214526727	0.0001775884	0.0644645711	0.2413425679	0.1250221985	0.0184691884	0.2615876399

relationship

trainY	Husband	Not-in-family	Other-relative	Own-child	Unmarried	Wife
0	0.298040139	0.302925078	0.037137308	0.196633512	0.133305868	0.031958095
1	0.754395312	0.108684070	0.005150062	0.008346652	0.029834843	0.093589061

race

trainY	Amer-Indian	Asian	Black	Other	White
0	0.011005827	0.028073686	0.110999941	0.009475605	0.840444941

```
  1 0.004439709 0.033386610 0.048481620 0.003019002 0.910673060

      sex
trainY    Female      Male
     0 0.3850862 0.6149138
     1 0.1505949 0.8494051

      hr_per_week
trainY       [,1]      [,2]
     0 -0.1347207 1.0011172
     1  0.4015488 0.8871906
```