STAT 6021 Project 2: Adult

Group 5: Huilin Chang, Andrew Hogue, Alicia Rice and Gavin Wiehl

Adult is a data set from the 1996 Census database that was offered by Barry Becker[1]. The data set is meant for binary classification to predict whether the income of adults will be greater than $50K or not. There is a total of 14 predictors, consisting of eight categorical and six numerical types. The predictors are outlined in detail below (Table1).

Response variable: Income: >$50K, <=$50K.

Table 1 Predictors – full name(abbreviation)(type)

Predictors						
age (numerical)	**workclass** (wc)(categorical)	**fnlwgt** (wgt)(numerical)	**education** (edu)(categorical)	**education-num** (edu_num)(numerical)	**marital-status** (martial)(categorical)	**occupation** (occup)(categorial)
relationship (rp)(categorical)	**race** (categorical)	**sex** (categorical)	**capital-gain** (c_gain) (numerical)	**capital-loss** (c_loss)(numerical)	**hours-per-week** (hours_w)(numerical)	**native-country** (nc)(categorical)

Throughout the course of our analysis, we found **age**, **workclass**, **number of years of education**, **occupation**, **sex**, **hours worked per week**, **marital status** and **race** to be adequate predictors of whether an individual would have an income of greater than $50K. There were a few findings from the analysis that confirmed common knowledge, such as the effect of sex on income. In the model, the odds of a male making over $50K is 1.348 times the odds of a female, holding other variables constant. It was also found that the odds of making over $50K were greater while working in the private sector, as opposed to being self-employed. Lastly, the odds of a married person making over $50K is 7.29 times the odds of a divorced person doing the same, holding other variables constant. The procedures for analyzing data include: (1) data cleaning and manipulation (2) first model and machine learning (3) test hypothesis (4) multicollinearity (5) second model and machine learning (6) model evaluation and (7) conclusions.

1. Data Cleaning and Manipulation:

Initially, the dimension of the adult data using dim() function was checked. There is a total of 32561 rows and 15 columns. At first glance, some cells of the data have "?" present. This NA data was assigned "?" as NA and dropped. This caused the data dimensions to be reduced to 30162 rows with the same number of columns. Since the adult data include many observations, the distribution of each variable has to be checked to see whether skewing of the data occurs. Box plots are used first to check the six numerical variables. Bar charts are used to check eight categorical variables.

Numerical variables: (1) Age: based on the box plot (Figs. 1-2), older people achieve a higher income. The average mean age for people who receive income <=$50K is 37 years old. In contrast, the average mean age for people who receive income > $50K is 45 years old. For younger people their income is significantly lower than for older people. However, when people are reaching retirement age, their income rapidly decline (Fig. 3). (2) fnlwgt: weighting factor

[1] C L Blake, C J Merz. UCI repository of machine learning databases University of California, Irvine, Department of Information and Computer Sciences. 1998 (https://archive.ics.uci.edu/ml/datasets/adult

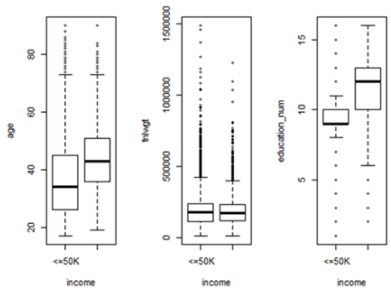


Figure 1Box plots of income vs. age, fnlwgt, education-num

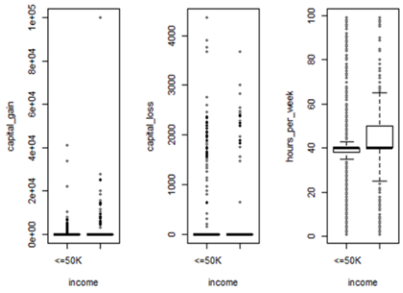


Figure 2 Box plots of income vs. capital-gain, loss and hours-per-week

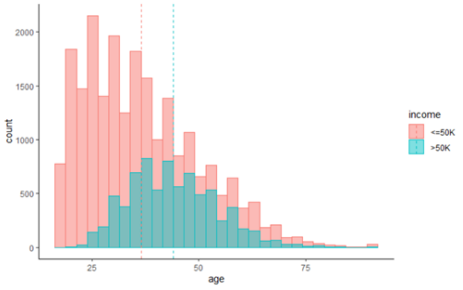


Figure 3 Age vs. count distribution (income)

(3) Education-num: The length of time spent in education has a profound effect on a person's ability to earn above $50K. The more time a person spends in education the greater their earning potential. (4) Capital-gain and (5) Capital-loss: The data was shown to have a significant skew for "capital gain" and "capital loss". For capital gain and capital loss, there are 91% and 95% percent zeros, respectively. When checking the distribution plot for capital gain for non-zero, the data was still skewed at lower bound and had outliners (Figs 4 (a-b)).

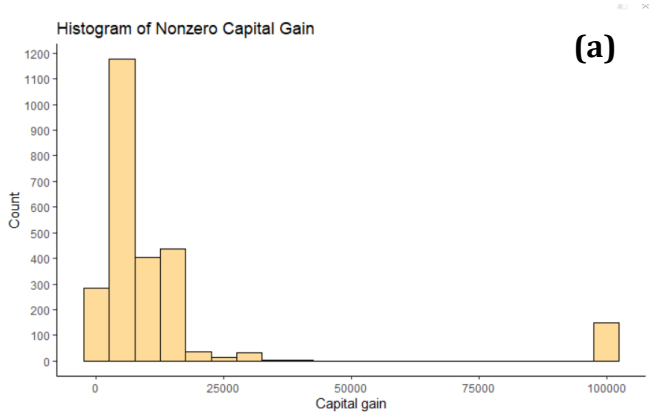


Figure 4(a) Capital-gain count distribution excluding zero

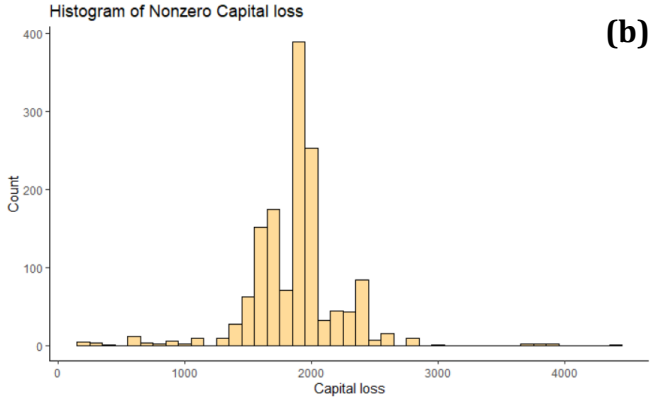
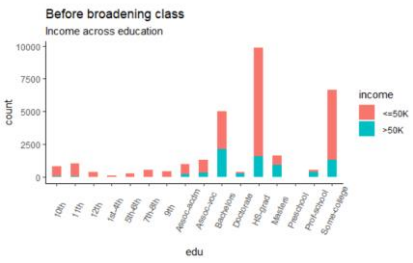


Figure 4(b) Capital-loss count distribution excluding zero

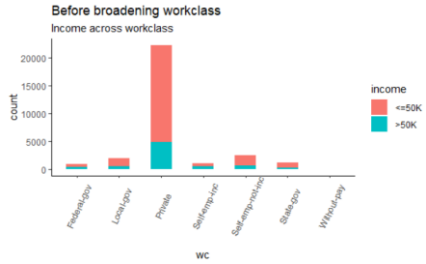
(6) Hours-per-week: the distribution for both groups of earners (<=$50K and>$50K) is wide. However, people on higher income tend to work longer hours.

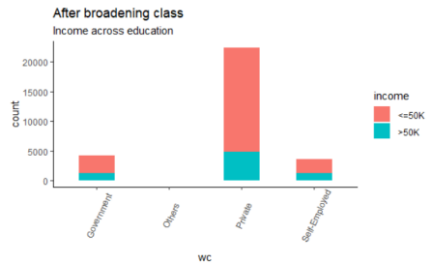
Categorical variables: The classes need to be broadened. A number of the categorical variables have too many classes and needed to be redefined as new ones. (1) Eduction: redefine k<=12 as <HS and associate academic and vocation are combined (2) Work class: redefine government job, self-employment and other (3) Occupation: redefine blue-collar, white-collar, millary, professional, other, sale and service (4) Native country: mutate countries based on geographical locations (5) Marital status: redefine divorced, married, single, seperated and widowed (Figs 5). Three categorical variables (Relationship, Race, Sex) keep the same classes and their corresponding bar plots are shown in Figs. 6(a)-(c).

Education
Before(16)/
After(8)
Fig 5

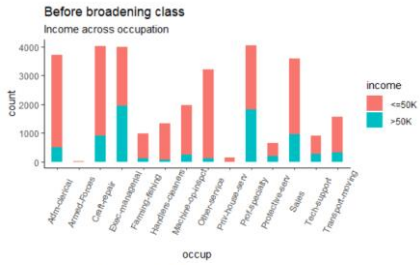


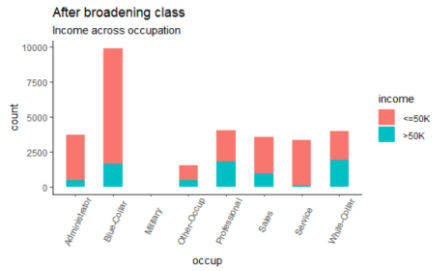
Work class
Before(7)/
After(4)
Fig 5



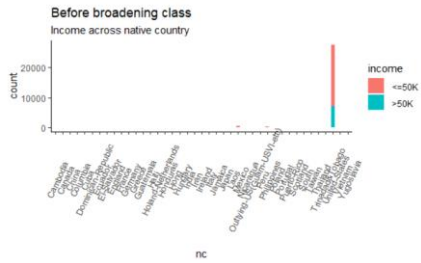


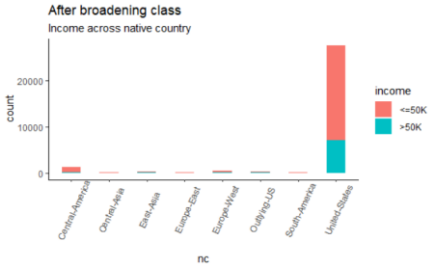
Occupation
Before(14)/
After(8)
Fig 5



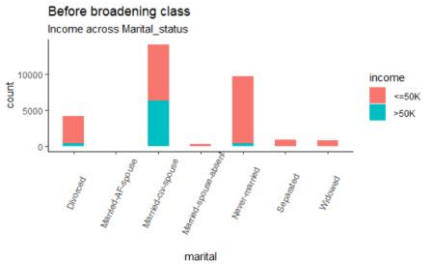


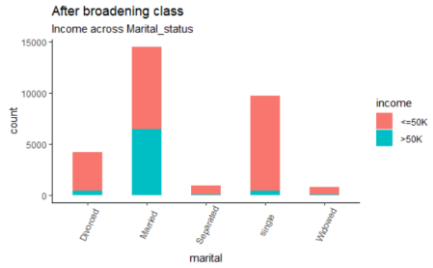
Native country
Before(36)/
After(8)
Fig 5





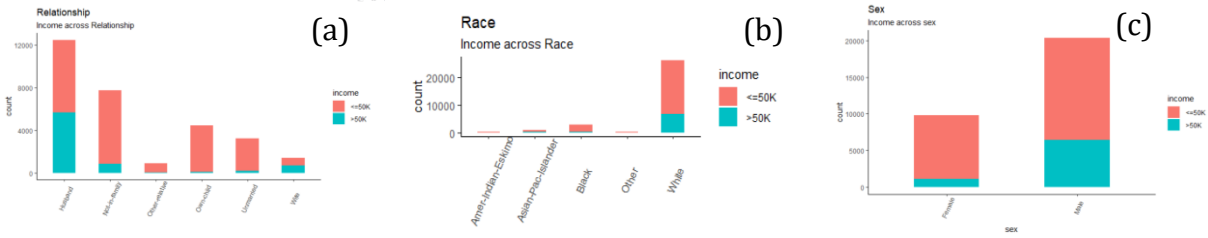
Marital status
Before(7)/
After(5)
Fig 5





Figures 5 categorical classes before and after broadening: education, work class, occupation, native country, marital status.

Relationship(6)
and Race(5)
Fig 6(a), (b),
(c)



Figures 6 Bar plots (a) relationship (b) race (c) sex

Since majority capital-loss and capital-gain are zeros, these were dropped. In addition, "fnlwgt" is the weighting factor and is not necessary for the model, and so it was dropped. The "native-country" data is also skewed because the majority of people were from the United States. Therefore native-country was dropped. With four predictors dropped, the final total number of predictors was 10.

2. First model

Since the data is recorded at the individual level, it is classed as ungrouped data. Fitting the logistic regression model by modeling the log odds of income against 10 predictors. The argument "family" has to be specified as "binominal" for a logistic regression. The 10 predictors are age, workclass, education-num, marital-status, occupation, relationship race, sex, hours-per-week, education. The regression equation is shown in Eq. 1 with Null deviance: 33851 and Residual deviance: 21806 (Table A-1).

Result <-glm(income ~., family = "binomial")

$$log\left(\frac{\pi}{1-\pi}\right) = -7.14719 + \beta_1 * \text{age} + \beta_2 * \text{workclass} + \beta_3 * \text{education} - \text{num} + \beta_4 * \text{occupation} +$$

$$\beta_5 * sex + \beta_6 * hours - per - week + \beta_7 * education + \beta_8 * relationship + \beta_9 * marital - status + \beta_{10} * race \quad \text{(Eq.1)}$$

- β_1: 0.02839
- β_2: 0 if government, -12.0651 if Others, 0.0965 if Private, -0.2146 if Self-Employed
- β_3: 0.2132
- β_4: 0 if Administrator, -0.2479 if Blue-Collar, -0.2654 if Military, 0.4988 if Other-Occup, 0.4618 if Professional, 0.2820 if Sales, -1.0136 if Service, 0.8062 if White-Collar
- β_5: 0 if Female, 0.8968 if male
- β_6: 0.02912
- β_7: 0 if <HS, 0.2904 if Assoc, 0.5773 if Bachelors, 0.9703 if Doctorate, 0.2694 if Hs-grad, 0.7351 if Masters, 1.2539 if Prof-school, 0.4088 if Some-college
- β_8: 0 if Husband, -0.943345 Not-in-family, -1.337961 Other-relative, -2.062235 if Own-child, -1.168674 if Unmarried, 1.325758 if Wife
- β_9 : 0 if Divorced, 0.5728 if Married, -0.0674 if Separated, -0.4940 if single, 0.1662 if Widowed
- β_{10}: 0 if Amer-Indian-Eskimo, 0.4087 if Asian-Pac-Islander, 0.4309 if Black, -0.2238 if Other, 0.5270 if White

A few interpretations: for an increase in age by one year, the estimated log odds of income > $50K increases by 0.02839 while holding workclass, education-num, occupation, sex, hours-per-week, education, marital-status and race constant. Considering a male at age 40 years old with workclass = government, education number = 16 years, occupation is White-Collar, hours-per-week is 40 hrs, education is Doctoral, relationship is Wife, marital-status is married, race is White:

The estimated odds:

$$log\frac{\pi}{1-\pi} = -7.14719 + 0.02839(40) + 0 + 0.2132 * 16 + 0.8062 +$$

$$0.8968 + 0.02912 * 40 + 1.2539 + 1.3258 + 0.5728 + 0.5270$$

The estimated odds calculated from equation is 3.6633 and the probability (p(X)=Pr(Y=1|X)) of obtaining an income > $50K is 97%

2-1. Machine learning (Fig. 5(a)) for full model

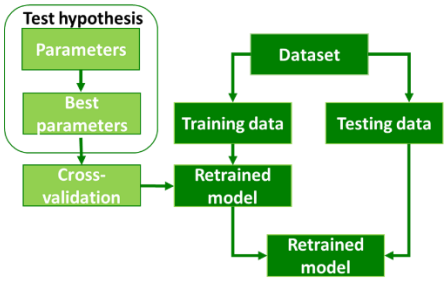


Fig. 5(a) Machine learning workflow

Further, apply seed (2019) and split the data 50:50 into training and testing sets. Fit a logistic regression model using age, workclass, education-num, occupation, sex, hours-per-week, education, relationship, martial-status and race by using glm() function. To produce ROC curve for this logistic regression model, a number of functions from ROCR package is used. The procedure is as follows:

```
set.seed(2019)
##choose the observations to be in the training. I am splitting the dataset into halves
sample<-sample.int(nrow(data), floor(.50*nrow(data)), replace = F)
train<-data[sample, ]
test<-data[-sample, ]
##use training data to fit logistic regression model with 10 predictors
result<-glm(income ~age+wc+edu_num+occup+sex+hours_w+ edu +rp + marital, family = "binomial", data = train)
```

It can be seen from the plot that the ROC curve is above the random guessing line (Fig 5(b)). This means the fitted model is better than random guessing. The AUC is 0.8842 (an ideal case will be 1). This means the area below the curve is greater than 0.5, which corresponds to random guessing. This model performs well in the classification problem.

The histogram shows the distribution of the predicted probabilities of all observations between 0 to 1 (Fig. 5(c)). The confusion matrix with 0.5 as cutoff, the false positive rate is 911/(911+10357) = 8% ; false negative rate is 1696/(1696+2117) = 44% , the model accuracy = 82.72% (Table 2)

The model accuracy = $\frac{TP+TN}{TP+TN+FP+FN} = \frac{10357+2117}{10357+911+1696+2117} = 82.72\ \%$

Figure 5(b) first model ROC curve (c) prediction histogram

Table 2 confusion matrix of first model

	False	True
<=$50K	10357	911
>$50K	1696	2117

3. Test hypothesis

Since the first model has 10 predictors, it is necessary to reduce the number of predictors to simplify the model. The predictors that are considered to be dropped are based on statistical significance and multicollinearity.

Considerations for dropping race (Table A-2)

Adopt test statistic $H_0 : \beta_{race} = \beta_2 = 0$, H_a : at least one of the coefficients in the null is not zero.

The deviance for the full model is $D(\beta) = 21806$, and the deviance for the reduced model to be $D(\beta_1) = 21824$

The difference in deviance between the full and reduced model is computed $D(\beta_2|\ \beta_1) = D(\beta_1) - D(\beta) = 16$

The test statistic is $\Delta G^2 = 21824 - 21806 = 16$ is larger than X^2_{p-1} distribution 9.49 at 95% confidence level with r=4 degree of freedom. The p-value (1-pchisq(16, 4)) is 0.003. Reject the null hypothesis and our data supports going with the more complicated model with the race predictor.

Considerations in dropping hours-per-week (Table A-3):

The histogram plot shows hours-per-week is widely distributed (Fig. 6); this is suspected to check whether this variable is statistically significant. Adopt the Wald test and we are looking to drop hours-per-week predictor, the hypotheses become

$H_0 : \beta_{hours-per-week} = 0$

$H_a : \beta_{hours-per-week} \neq 0$

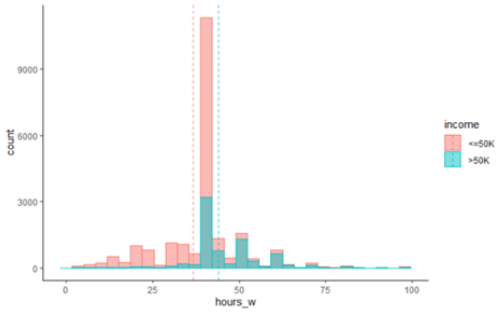


Figure 6 hours-per-week count distribution

The z statistic = $\frac{18.856}{0.001543} = 12220.35$, which is larger than 95% confidence level 1.96. In addition,

the p value from 1 - pnorm(abs(12220.35)))*2 = 0 ,

we reject the null hypothesis, the coefficient is considered to not be zero.

4. Multicollinearity

The relationship and marital-status fall in the same category, there is no necessity to include the martial-status predictor and it was decided to drop it. The education and education-num fall in the same category. Therefore, it was decided to drop education. The second model included eight predictors: age, workclass, education-num, occupation, sex, hours-per-week, marital-status, race.

5. Second model

Second model regression equation (Table A-4):

$$log\left(\frac{\pi}{1-\pi}\right) = -8.8236 + \beta_1 * \text{age} + \beta_2 * \text{workclass} + \beta_3 * \text{education} - \text{num} + \beta_4 * \text{occupation} + \beta_5 * \text{sex} + \beta_6 * \text{hours} - \text{per} - \text{week} + \beta_7 * \text{marital} - \text{status} + \beta_8 * \text{race} \quad ---------- \text{(Eq. 2)}$$

Where;

β_1: 0.02972

β_2: 0 if government, -12.1619 if Others, 0.1054 if Private, -0.1698 if Self-Employed

β_3: 0.2989

β_4: 0 if Administrator, -0.3007 if Blue-Collar, -0.3766 if Military, 0.4669 if Other-Occup, 0.5051 if Professional, 0.2375 if Sales, -1.026 if Service, 0.78734 if White-Collar

β_5: 0 if Female, 0.2989 if Male

β_6: 0.02910

β_7 : 0 if Divorced, 1.9870 if Married, -0.0619 if Separated, -0.4906 if single, 0.0267 if Widowed

β_8: 0 if Amer-Indian-Eskimo, 0.3528 if Asian-Pac-Islander, 0.4358 if Black, -0.2800 if Other, 0.5551 if White

5-1. Machine learning for second model

Further, apply seed (2019) and split the data 50:50 into training and testing sets. It can be seen from the plot that the ROC curve is above the random guessing line (Fig. 7). This means the fitted model is better than random guessing. The AUC is 0.8784 (an ideal case will be 1). This means the area below the curve is greater than 0.5, which corresponds to random guessing. This model performs well in the classification problem.

The confusion matrix with 0.5 as cutoff, the false positive rate is 949/(949+10319) = 8% ; false negative rate is 1721/(1721+2092) = 45%. The model accuracy = $\frac{TP+TN}{TP+TN+FP+FN} = \frac{10319+2092}{10319+949+1721+2092} = 82.30\%$

Table 3 confusion matrix of second model with cutoff 0.5

	False	True
<=$50K	10319	949
>$50K	1721	2092

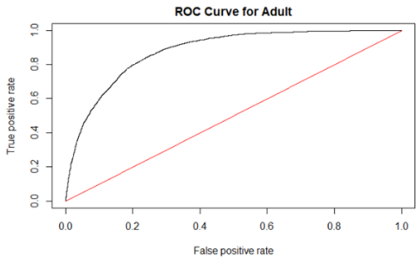


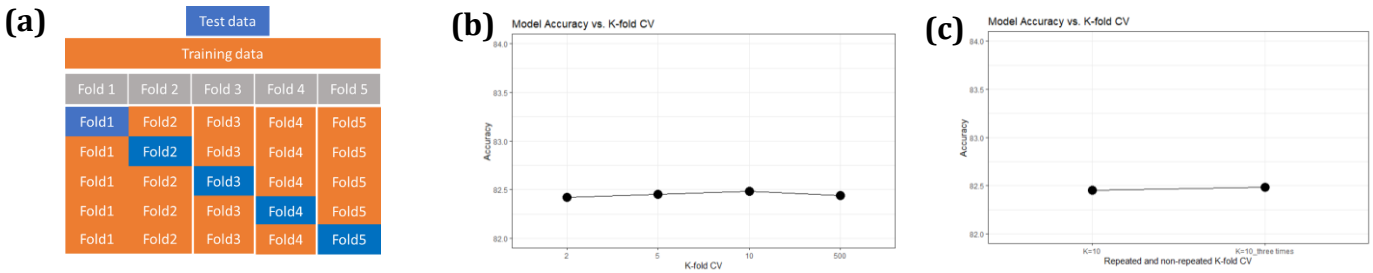
Figure 7 ROC curve of second mode

The accuracy of our second model is 0.8230 based on the cut-off of 0.5.

6. Model Evaluation

6-1. K-fold cross validation

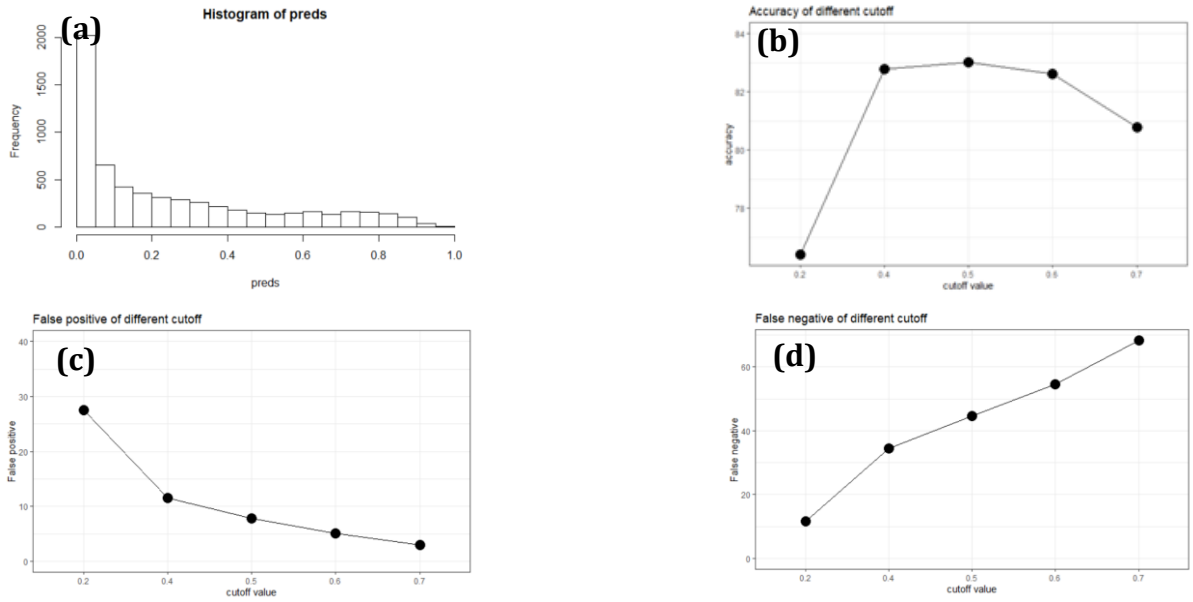
Adopt K-fold cross validation method (Fig. 8(a)) to evaluate the model performance on different subsets. K -fold cross validation is partitioning the original data into K equal-sized pieces. Repeat the holdout process K times. There is a bias-variance trade-off associated with the choice of K in K-fold cross-validation. Higher K can reduce the bias, and increase variance and running time, and vice versa. Use seed (2019) and caret package from R, where the K values are set to K=2, K=5, K=10, K=10 repeated three times and K= 500, to evaluate model accuracy. The model accuracy is very close to 82.5% for K=2, 5 ,10, and 500 (Fig. 8(a)). Comparison between K=10 and K=10 repeated three times, there is negligible difference. (Fig. 8(b)).



Figures 8 K-fold cross-validation (a) K-fold cross validation (b) Accuracy vs. K-fold CV for K=2, 5, 10 and 500 (c) negligible differences between K =10 and K=10 repeated three times

6-2. Cutoff ratio adjusting

The histogram plot shows the probability of prediction lies between 0 to 1 (Fig. 9(a)). The cutoff values are 0.2, 0.4, 0.5, 0.6, 0.7 and have a model accuracy of 76.41%, 82.79%, 83.01%, 82.61%, 80.79%, respectively (Fig. 9(b)). The greatest accuracy is obtained at a cutoff value of 0.5, which is consistent with the distribution of probability of prediction. The trade-off between false positive and negative rates can be seen at various cutoffs (Figs 9(c)-(d)). Decide to use cutoff = 0.5



Figures 9 (a) prediction histogram, cutoff vs. (b) accuracy (c) false positive rate (d) false negative rate

The second model using train/test splits 80%/20% The confusion matrix with 0.5 as cutoff, the false positive rate is 355/(355+4174) = 7.83% ; false negative rate is 670/(670+834) = 44.54%. (Table 4). The AUC is 0.8865 (an ideal case will be 1) (Fig.10)

Table 4 confusion matrix of second model with cutoff 0.5

	False	True
<=$50K	4174	355
>$50K	670	834

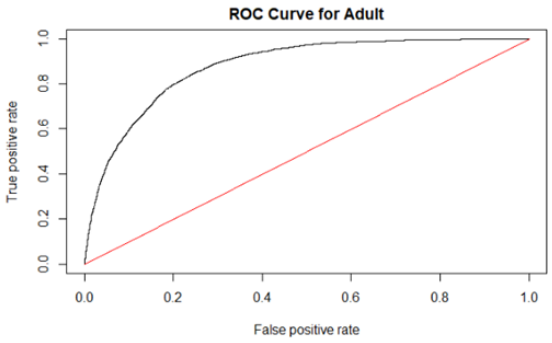


Figure 10 ROC curve of second mode

7. Conclusions

The Adult data set from the 1996 Census data was used to predict whether an individual will earn an income >50K or <=50K. The data processing included omitting missing data and broadening classes. Some predictors showed that the data was skewed. Therefore, these predictors were excluded from the model. Multicollinearity was considered and the predictors which have similar

meanings were consider redundant. For example, with relationship and marital status, relationship was removed. The test statistic and Wald test were adopted to decide whether some potential statistical insignificant predictors can be excluded.

The response variable is binary classification, this means using ungrouped logistic regression. The first and second models have an accuracy of 82-83%. However, since the second model is much simpler than the first model, it was decided to use the second model as the final one. The model evaluations include K-fold cross validation and adjusting cut-off ratio. K-fold cross validation shows negligible differences between K=2, K=5, K=10 and K=500 and the given model accuracy is in the range of 82-83%. The exercise aimed to precisely obtain the highest prediction rate as to whether an individual obtains an income greater than $50K or not. The optimized cutoff value of 0.5 was chosen due to the fulfillment of false positive and negative rates. The confusion matrix of the regression equation gives the false positive rate at 7.83% and the false negative rate at 44.54%, meaning the model predicts someone who actually fails to achieve an income of >$50K as doing so 7.83% of the time, and also wrongly predicts someone who achieves an income of >$50K as not doing so 44.54% of the time. The given regression equation is $log\left(\frac{\pi}{1-\pi}\right) = -8.8236 + \beta_1 * \text{age} + \beta_2 * \text{workclass} + \beta_3 * \text{education} - \text{num} + \beta_4 * \text{occupation} + \beta_5 * \text{sex} + \beta_6 * \text{hours} - \text{per} - \text{week} + \beta_7 * \text{marital} - \text{status} + \beta_8 * \text{race}$, Where;

β_1: 0.02972

β_2: 0 if government, -12.1619 if Others, 0.1054 if Private, -0.1698 if Self-Employed

β_3: 0.2989

β_4: 0 if Administrator, -0.3007 if Blue-Collar, -0.3766 if Military, 0.4669 if Other-Occup, 0.5051 if Professional, 0.2375 if Sales, -1.026 if Service, 0.78734 if White-Collar

β_5: 0 if Female, 0.2989 if Male

β_6: 0.02910

β_7 : 0 if Divorced, 1.9870 if Married, -0.0619 if Separated, -0.4906 if single, 0.0267 if Widowed

β_8: 0 if Amer-Indian-Eskimo, 0.3528 if Asian-Pac-Islander, 0.4358 if Black, -0.2800 if Other, 0.5551 if White

Consider a female at age 25 years old with workclass = government, education number is 16 years, occupation is White-Collar, hours-per-week is 40 hours, marital status is single, race is White

$log\frac{\pi}{1-\pi} = -8.8236 + 0.02972 * 25 + 0 + 0.2989 * 16 + 0.7873 + 0.02910 * 40 - 0.4906 + 0.5551$, the log odds is -0.688, the estimated odd is 0.502 and the probability to earn an income >$50K is 33.44%.

In conclusion, Adult data is successful in using the logistic regression method to predict whether an individual will have an income exceeding $50K or not. The model predictors are age, workclass, education-number, occupation, sex, hours-per-week, marital-status and race and the response variable is income, which has a precision rate of 82-83%. The concept could be extendable by obtaining a dataset that has predictors for earning an income > $1 billion and < =$1 billion. Nevertheless, this is a good starting point on the machine learning journey.

A-1 Model 1 logistic regression fuction

```
glm(formula = income ~ age + wc + edu_num + occup + sex + hours_w +
    edu + rp + marital + race, family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7209  -0.5889  -0.2297  -0.0229   3.3771

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -7.471918   0.405648 -18.420  < 2e-16 ***
age                      0.028390   0.001591  17.844  < 2e-16 ***
wcOthers               -12.065173 119.636904  -0.101 0.919671
wcPrivate                0.096513   0.050212   1.922 0.054592 .
wcSelf-Employed         -0.214622   0.064835  -3.310 0.000932 ***
edu_num                  0.213156   0.043430   4.908 9.20e-07 ***
occupBlue-Collar        -0.247904   0.069482  -3.568 0.000360 ***
occupMilitary           -0.265374   1.296768  -0.205 0.837852
occupOther-Occup         0.498761   0.088834   5.615 1.97e-08 ***
occupProfessional        0.461820   0.076333   6.050 1.45e-09 ***
occupSales               0.281975   0.077174   3.654 0.000258 ***
occupService            -1.013563   0.112069  -9.044  < 2e-16 ***
occupWhite-Collar        0.806173   0.072285  11.153  < 2e-16 ***
sex Male                 0.896780   0.073168  12.256  < 2e-16 ***
hours_w                  0.029123   0.001567  18.581  < 2e-16 ***
eduAssoc                 0.290400   0.264660   1.097 0.272531
eduBachelors             0.577348   0.326916   1.766 0.077388 .
eduDoctorate             0.970276   0.476384   2.037 0.041675 *
eduHS-grad               0.269412   0.163245   1.650 0.098871 .
eduMasters               0.735076   0.373338   1.969 0.048961 *
eduProf-school           1.253929   0.428663   2.925 0.003442 **
eduSome-college          0.408761   0.203328   2.010 0.044394 *
rp Not-in-family        -0.943345   0.159611  -5.910 3.42e-09 ***
rp Other-relative       -1.337961   0.214903  -6.226 4.79e-10 ***
rp Own-child            -2.062235   0.198591 -10.384  < 2e-16 ***
rp Unmarried            -1.168674   0.176157  -6.634 3.26e-11 ***
rp Wife                  1.325758   0.097640  13.578  < 2e-16 ***
maritalMarried           0.572832   0.163933   3.494 0.000475 ***
maritalSeparated        -0.067375   0.150089  -0.449 0.653506
maritalsingle           -0.494008   0.080825  -6.112 9.83e-10 ***
maritalWidowed           0.166163   0.142698   1.164 0.244247
race Asian-Pac-Islander  0.408697   0.235270   1.737 0.082362 .
race Black               0.430887   0.223627   1.927 0.054003 .
race Other              -0.223874   0.347034  -0.645 0.518857
race White               0.527018   0.213638   2.467 0.013630 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 33851  on 30161  degrees of freedom
Residual deviance: 21806  on 30127  degrees of freedom
AIC: 21876
```

A-2 Drop race from Model 1

```
glm(formula = income ~ age + wc + edu_num + occup + sex + hours_w +
    edu + rp + marital, family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7612  -0.5889  -0.2306  -0.0225   3.3923

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -6.970043   0.345408 -20.179  < 2e-16 ***
age                      0.028630   0.001590  18.008  < 2e-16 ***
wcOthers               -12.049759 119.503380  -0.101 0.919684
wcPrivate                0.102136   0.050056   2.040 0.041307 *
wcSelf-Employed         -0.206782   0.064616  -3.200 0.001374 **
edu_num                  0.214655   0.043307   4.957 7.18e-07 ***
occupBlue-Collar        -0.248263   0.069424  -3.576 0.000349 ***
occupMilitary           -0.290128   1.283847  -0.226 0.821215
occupOther-Occup         0.501929   0.088773   5.654 1.57e-08 ***
occupProfessional        0.462510   0.076225   6.068 1.30e-09 ***
occupSales               0.285632   0.077097   3.705 0.000212 ***
occupService            -1.026949   0.111958  -9.173  < 2e-16 ***
occupWhite-Collar        0.810031   0.072224  11.216  < 2e-16 ***
sex Male                 0.898467   0.073111  12.289  < 2e-16 ***
hours_w                  0.029226   0.001567  18.654  < 2e-16 ***
eduAssoc                 0.290642   0.264085   1.101 0.271086
eduBachelors             0.575141   0.326107   1.764 0.077790 .
eduDoctorate             0.963722   0.475230   2.028 0.042570 *
eduHS-grad               0.271812   0.162958   1.668 0.095319 .
eduMasters               0.732132   0.372402   1.966 0.049301 *
eduProf-school           1.244394   0.427611   2.910 0.003613 **
eduSome-college          0.407814   0.202924   2.010 0.044464 *
rp Not-in-family        -0.969909   0.159293  -6.089 1.14e-09 ***
rp Other-relative       -1.377141   0.214238  -6.428 1.29e-10 ***
rp Own-child            -2.088491   0.198291 -10.532  < 2e-16 ***
rp Unmarried            -1.204266   0.175632  -6.857 7.04e-12 ***
rp Wife                  1.321093   0.097565  13.541  < 2e-16 ***
maritalMarried           0.544962   0.163515   3.333 0.000860 ***
maritalSeparated        -0.079193   0.149680  -0.529 0.596749
maritalsingle           -0.496743   0.080723  -6.154 7.57e-10 ***
maritalWidowed           0.164877   0.142603   1.156 0.247601
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 33851  on 30161  degrees of freedom
Residual deviance: 21824  on 30131  degrees of freedom
AIC: 21886
```

A-3 Considering to drop hour-per-week from Model 1

```
glm(formula = income ~ age + wc + edu_num + occup + sex + marital +
    race + hours_w, family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6947  -0.5936  -0.2526  -0.0380   3.4782

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -8.823594   0.262487 -33.615  < 2e-16 ***
age                      0.029720   0.001543  19.261  < 2e-16 ***
wcOthers               -12.161906 119.171734  -0.102  0.91871
wcPrivate                0.105419   0.049306   2.138  0.03251 *
wcSelf-Employed         -0.169846   0.063807  -2.662  0.00777 **
edu_num                  0.298867   0.008887  33.631  < 2e-16 ***
occupBlue-Collar        -0.300743   0.067679  -4.444 8.84e-06 ***
occupMilitary           -0.376567   1.279174  -0.294  0.76847
occupOther-Occup         0.466906   0.087178   5.356 8.52e-08 ***
occupProfessional        0.505090   0.072322   6.984 2.87e-12 ***
occupSales               0.237541   0.075274   3.156  0.00160 **
occupService            -1.026116   0.110663  -9.272  < 2e-16 ***
occupWhite-Collar        0.787345   0.070027  11.243  < 2e-16 ***
sex Male                 0.298893   0.048944   6.107 1.02e-09 ***
maritalMarried           1.986955   0.061327  32.399  < 2e-16 ***
maritalSeparated        -0.061873   0.146342  -0.423  0.67244
maritalsingle           -0.490623   0.075780  -6.474 9.52e-11 ***
maritalWidowed           0.026656   0.139630   0.191  0.84860
race Asian-Pac-Islander  0.352767   0.231220   1.526  0.12709
race Black               0.435840   0.220187   1.979  0.04777 *
race Other              -0.280052   0.343516  -0.815  0.41493
race White               0.555098   0.210463   2.638  0.00835 **
hours_w                  0.029100   0.001543  18.856  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 33851  on 30161  degrees of freedom
Residual deviance: 22270  on 30139  degrees of freedom
AIC: 22316
```

A-4 Model 2 Logistic regression function

```
glm(formula = income ~ age + wc + edu_num + occup + sex + hours_w +
    marital + race, family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6947  -0.5936  -0.2526  -0.0380   3.4782

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -8.823594   0.262487 -33.615  < 2e-16 ***
age                      0.029720   0.001543  19.261  < 2e-16 ***
wcOthers               -12.161906 119.171734  -0.102  0.91871
wcPrivate                0.105419   0.049306   2.138  0.03251 *
wcSelf-Employed         -0.169846   0.063807  -2.662  0.00777 **
edu_num                  0.298867   0.008887  33.631  < 2e-16 ***
occupBlue-Collar        -0.300743   0.067679  -4.444 8.84e-06 ***
occupMilitary           -0.376567   1.279174  -0.294  0.76847
occupOther-Occup         0.466906   0.087178   5.356 8.52e-08 ***
occupProfessional        0.505090   0.072322   6.984 2.87e-12 ***
occupSales               0.237541   0.075274   3.156  0.00160 **
occupService            -1.026116   0.110663  -9.272  < 2e-16 ***
occupWhite-Collar        0.787345   0.070027  11.243  < 2e-16 ***
sex Male                 0.298893   0.048944   6.107 1.02e-09 ***
hours_w                  0.029100   0.001543  18.856  < 2e-16 ***
maritalMarried           1.986955   0.061327  32.399  < 2e-16 ***
maritalSeparated        -0.061873   0.146342  -0.423  0.67244
maritalsingle           -0.490623   0.075780  -6.474 9.52e-11 ***
maritalWidowed           0.026656   0.139630   0.191  0.84860
race Asian-Pac-Islander  0.352767   0.231220   1.526  0.12709
race Black               0.435840   0.220187   1.979  0.04777 *
race Other              -0.280052   0.343516  -0.815  0.41493
race White               0.555098   0.210463   2.638  0.00835 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 33851  on 30161  degrees of freedom
Residual deviance: 22270  on 30139  degrees of freedom
AIC: 22316
```