Annual Income Classifier

Wenke Huang

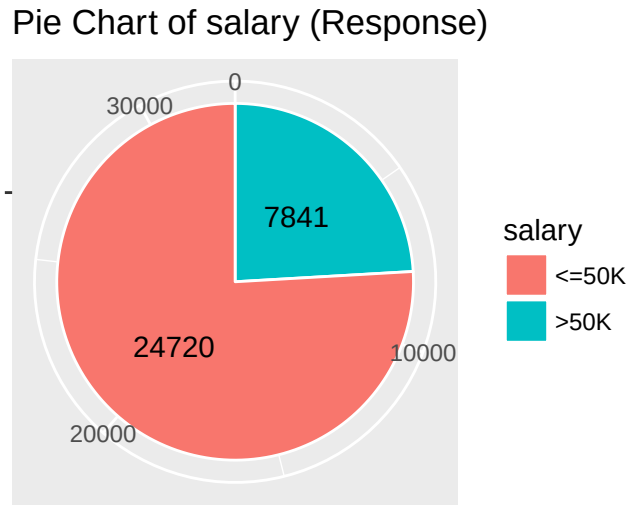
1. Introduction

1.1 Response variable and prediction metric

The Adult data set (URL: https://archive.ics.uci.edu/ml/datasets/Adult) that we have contains 15 variables and 32561 observations in total. The one named "salary" is our response variable. It has two levels, ">= 50K" and "<50K", which stand for annual salary being greater than or equal to 50,000 USD or not, respectively.

The Frequency table and its corresponding Pie chart are shown below.

salary	Freq	Propotion
<=50K	24720	75.92%
>50K	7841	24.08%



We can tell that our response variable does not follow discrete uniform distribution exactly or roughly. This may generate an obvious problem in our following model fitting part if we choose the Overall Accuracy as our metric here. Any model with Overall Accuracy being lower than 75.92% is meaningless. It is because that we can just predict that the salary is lower than or equal to 50,000 dollars, and the Overall Accuracy is going to be 75.92% exactly.

- Hence, we would like to use `ROC curve` and its corresponding `AUC` as our metric in this case.

1.2 Duplicate variables

Among these predictors, we notice that there are two variables which have similar names, `education` and `education_num`. We try to build a contingency table between these two variables.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Preschool	51	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1st−4th	0	168	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5th−6th	0	0	333	0	0	0	0	0	0	0	0	0	0	0	0	0
7th−8th	0	0	0	646	0	0	0	0	0	0	0	0	0	0	0	0
9th	0	0	0	0	514	0	0	0	0	0	0	0	0	0	0	0
10th	0	0	0	0	0	933	0	0	0	0	0	0	0	0	0	0
11th	0	0	0	0	0	0	1175	0	0	0	0	0	0	0	0	0
12th	0	0	0	0	0	0	0	433	0	0	0	0	0	0	0	0
HS−grad	0	0	0	0	0	0	0	0	10501	0	0	0	0	0	0	0
Some−college	0	0	0	0	0	0	0	0	0	7291	0	0	0	0	0	0
Assoc−voc	0	0	0	0	0	0	0	0	0	0	1382	0	0	0	0	0
Assoc−acdm	0	0	0	0	0	0	0	0	0	0	0	1067	0	0	0	0
Bachelors	0	0	0	0	0	0	0	0	0	0	0	0	5355	0	0	0
Masters	0	0	0	0	0	0	0	0	0	0	0	0	0	1723	0	0
Prof−school	0	0	0	0	0	0	0	0	0	0	0	0	0	0	576	0
Doctorate	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	413

The result tells us that variable `education` and variable `education_num` are exactly one-to-one match. Even though `education_num` contains numerical values only, it still should be considered as a categorical variable like `education`. Hence, we would like to keep variable "education" only in case of the potential multicollinearity.

Further detailed information will be discussed in the following parts.

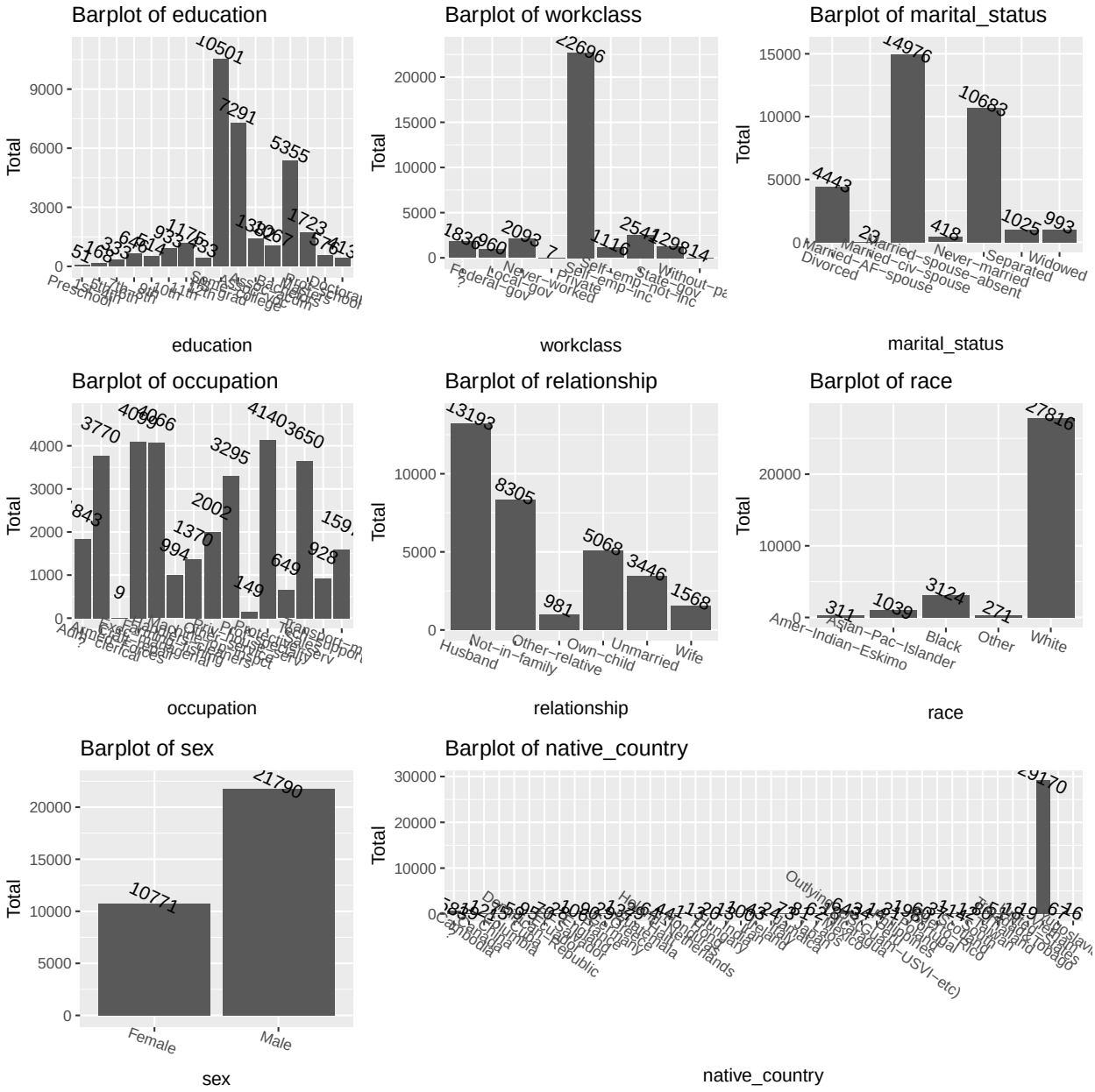
2 Visualization and outlier detection

2.1 Categorical predictor variables

In this case, apart from variable "education_num" that we have deleted already, we have 8 categorical variables. They are `workclass`, `education`, `marital_status`, `occupation`, `relationship`, `race`, `sex` and `native_country`.

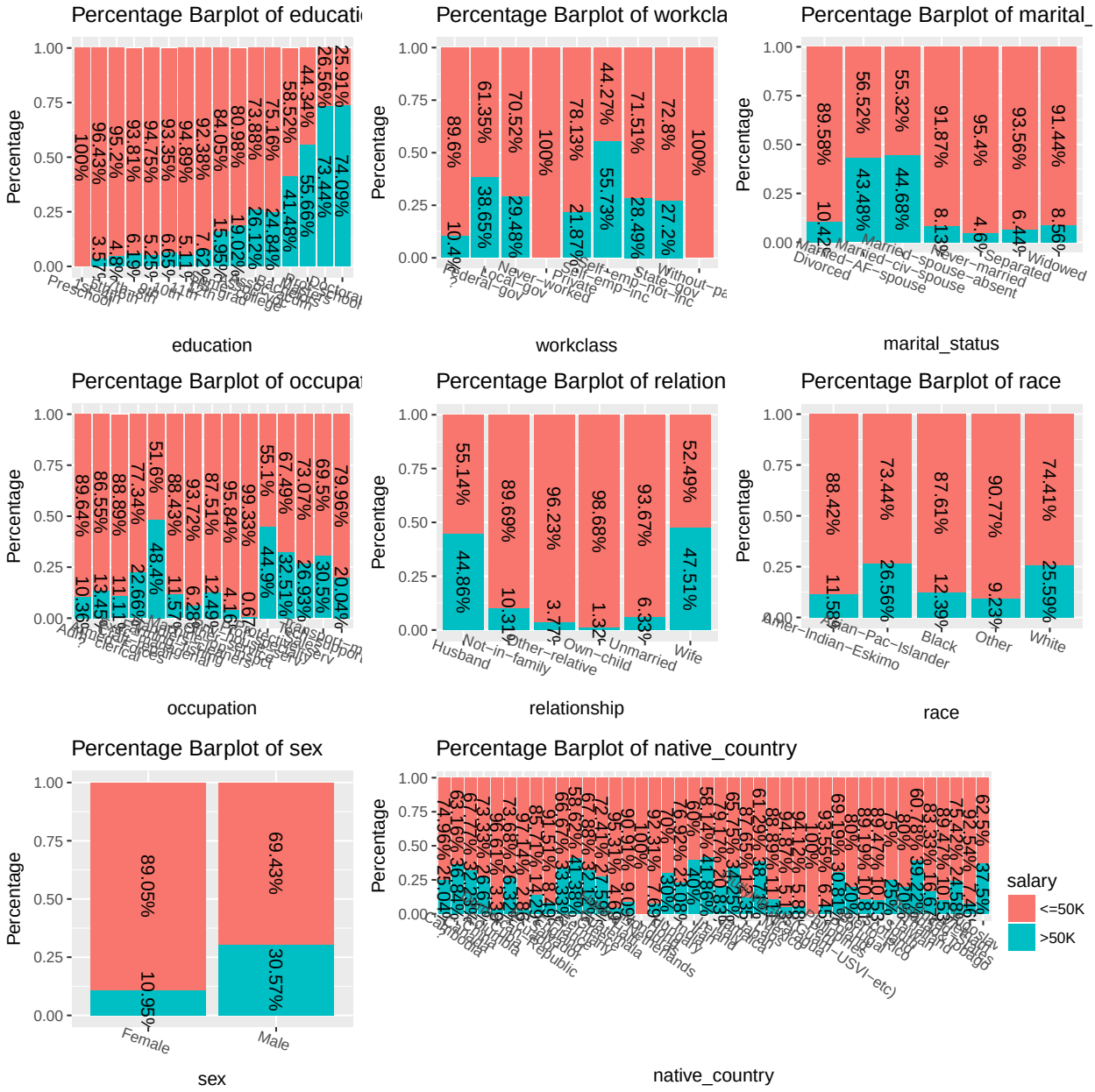
2.1.1 Frequency bar plot

The Frequency bar plots of these 8 categorical variables are shown below to gives us a clearer understanding about univariate distributions of them visually. The frequencies of individual levels within different variables are also labeled in the plots.



2.1.2 Percentage bar plot

Additionally, we would like to see what proportion of people who can earn more than 50,000 dollars are year within different categorical predictors. The Percentage bar plots of these 8 variables will also bn shown below.



Visually speaking, the proportion of people who can make more than 50,000 dollars a year variates within different categorical variables.

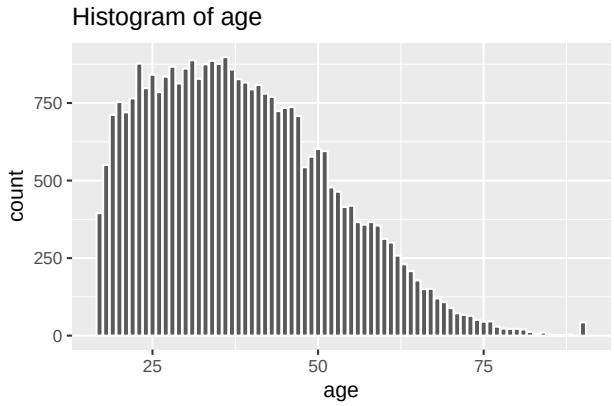
Among these 8 categorical variables, variable `education` is the only one could be considered as a ordinal categorical variable while the rest 7 should be treated as nominal categorical variables. We can also notice from the first of the 8 plots that there is an interesting trend that as education level going higher towards Doctorate Degrees, the proportion of people who can make more than 50,00 dollars a year is also going up gradually.

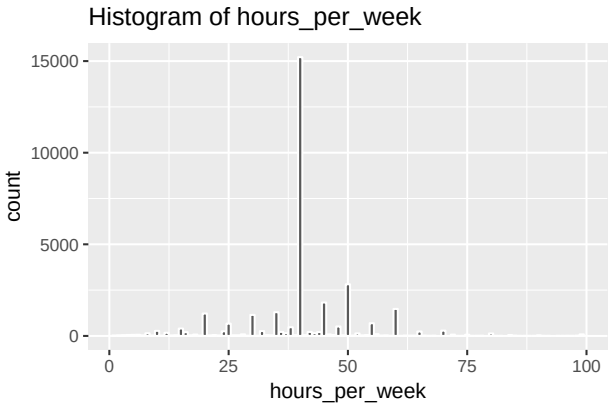
2.2 Numerical variables

We have 5 numerical variables in this case. They are `age`, `capital_gain`, `capital_loss` and `hours_per_week`. As what we have discussed in he previous part, `education_num` will not be considered as one of numerical variables and visualized here even though it has numerical values.

2.2.1 Discrete variables

Among these 5 numerical variables, "age" and "hours_per_week" should be considered as discrete numerical variables because the all values of both variables in this dataset are integers. So we choose Histogram to visualize the univariate distribution of both of them.



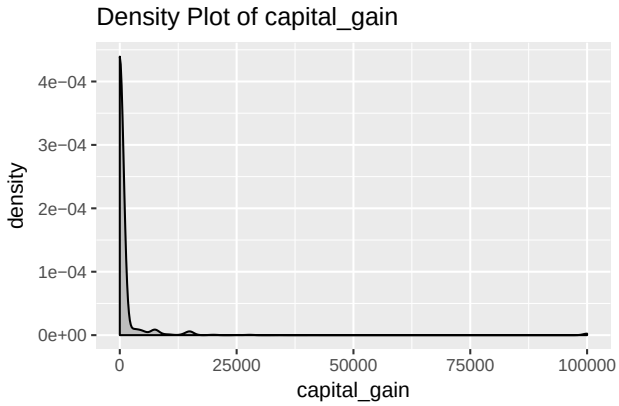


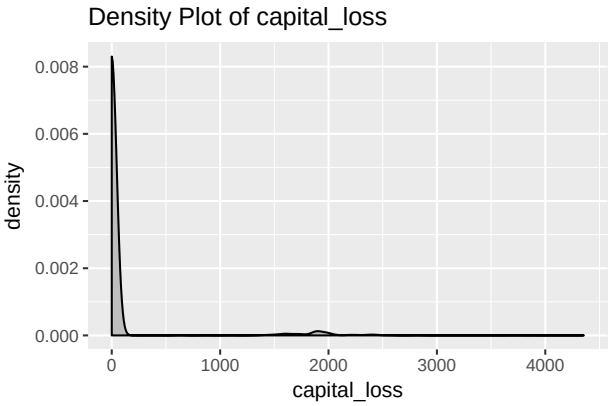
From the Histogram of "age", we can find a very interesting fact that a small proportion of people still work after their 95th birthday, which is a little beyond what we can imagine. We can consider these observations to be outliers, and they will be discussed in the following part.

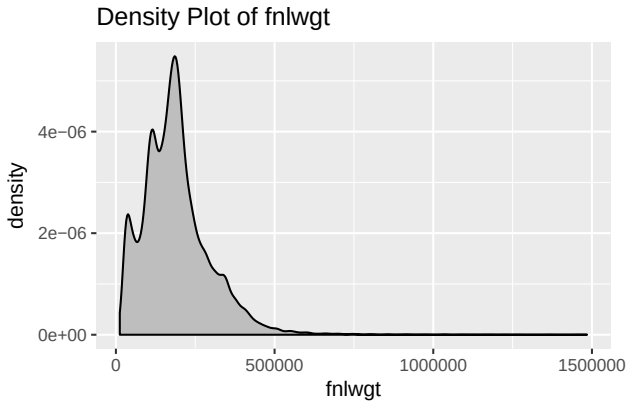
From the Histogram of "hours_per_week", we notice that most people work exactly 40 hours a week. Also at the same time, we still can read that some people work as a part-time employee (lower than 20 hours), or do serious overtime.

2.2.2 Continuous variables

In this dataset, we would like to treat variables `fnlwgt`, `capital_gain` and `capital_loss` as continuous numerical variables here because of their wide ranges. Hence, the Density plot is selected here to visualize these three.

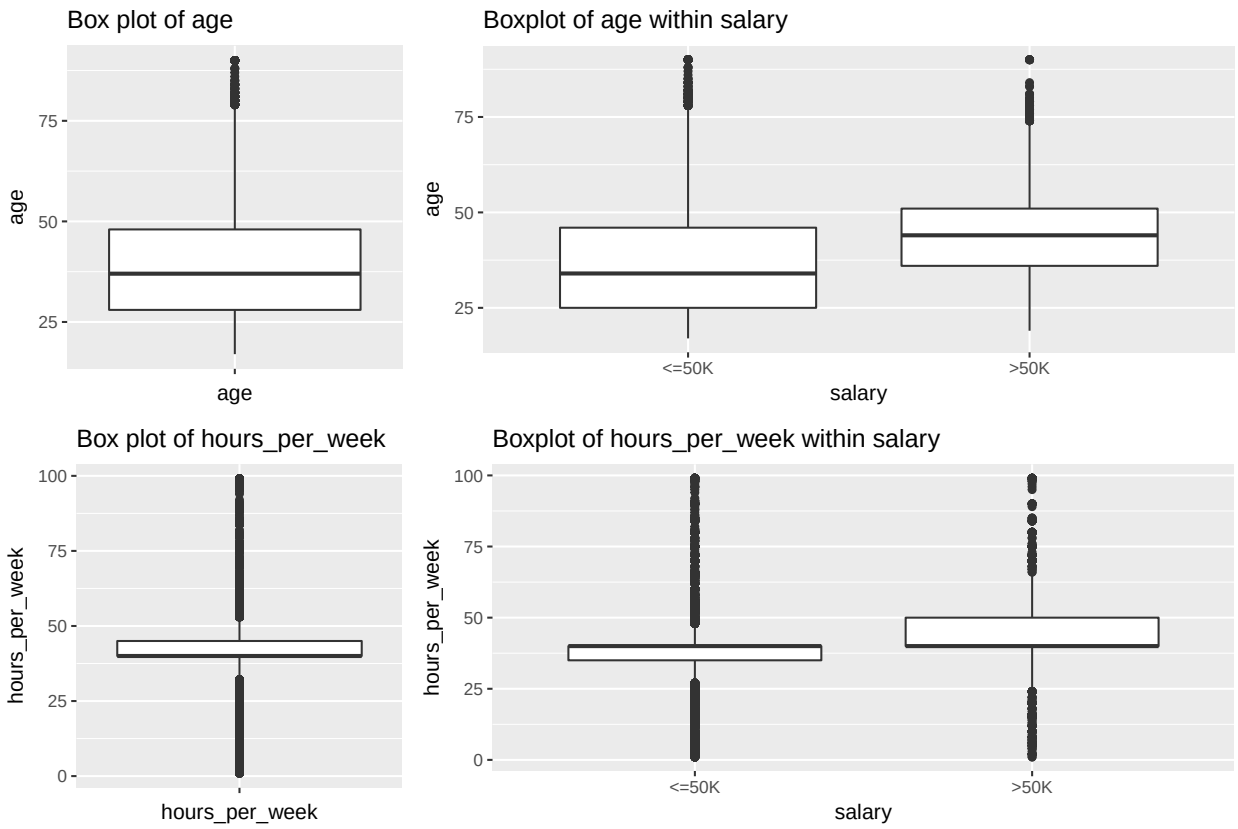


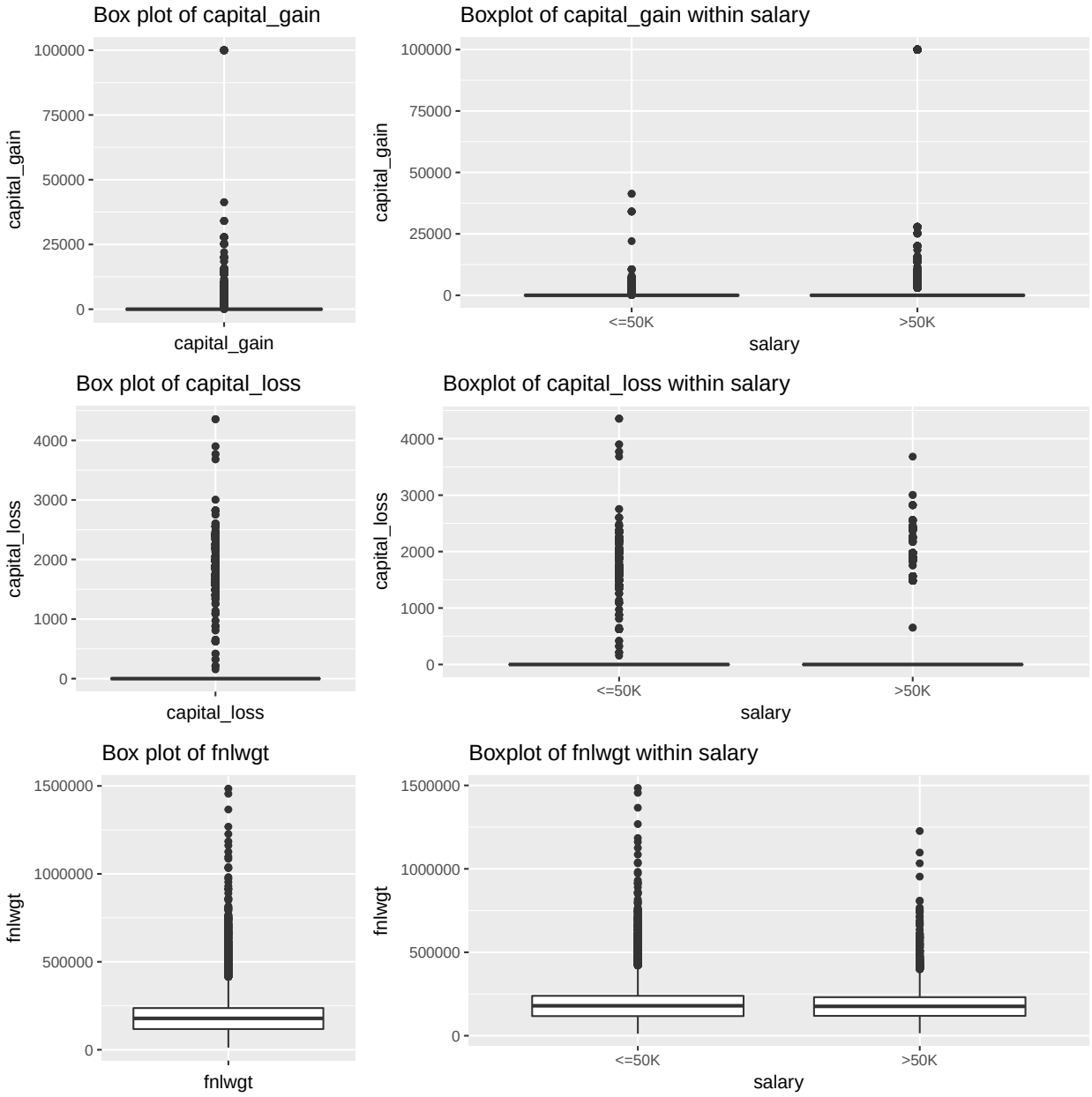




For the first two variables `capital_gain` and `capital_loss`, we notice that most of its values are 0 (91.67% and 95.33%, respectively). It shows that very few people gain or lose their money through capital ways. And all these three variables are highly skewed to the right.

2.2.3 Box plot and outlier detection





We notice that all these five numerical variables contain several points which locate out of the interval $(Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR)$, where Q1, Q3, and IQR stand for 25th percentile, 75th percentile and Q3-Q1, respectively. Here, we would like to treat those points as outliers.

For variables `capital_gain`, `capital_loss`, and `hours_per_week`, I would like to choose to keep those points which do not fall in the above interval. Because those observations actually stand for real things. A person with positive `capital_gain` tend to have higher incomes, which may result in being classified into a different group. Variable `capital_loss` has the same influence but in a different direction. For `hours_per_week`, people who work over time tend to make more money annually while those who work part-time tend to not get paid that much.

But for variable `age`, we have 43 observations which stand for 43 people still choose to work after their 90th birthday. These points must be resulted from typo or some other data entry mistakes.

How to deal with these points will be discussed in part `3.2`.

2.2.4 Correlation coefficient matrix

Here, we would like to build a correlation coefficient matrix among all the four numerical variables to see if the potential multicollinearity exists here.

	age	fnlwgt	capital_gain	capital_loss	hours_per_week
age	1	−0.0766	0.0777	0.0578	0.0688
fnlwgt	−0.0766	1	4e−04	−0.0103	−0.0188
capital_gain	0.0777	4e−04	1	−0.0316	0.0784
capital_loss	0.0578	−0.0103	−0.0316	1	0.0543
hours_per_week	0.0688	−0.0188	0.0784	0.0543	1

We notice all coefficients are very small. Only coefficient of group `capital_gain` and `capital_loss`, and group `capital_loss` and `fnlwgt` are slight negative while the rest all are slight positive. Hence, we think that there is no obvious multicollinearity existing among our numerical predictors.

3 Data cleaning, outliers and dataset spliting

3.1 Missing values and Data cleaning

In this case, we notice that there are some observations containing logical mistakes, such as `sex` being Male but `relationship` being Wife at the same time. Part of missing records are shown as examples and listed below, and only related variables are included here.

	workclass	marital_status	occupation	relationship	sex
576	Private	Married−civ−spouse	Exec−managerial	Wife	Male
5362	Never−worked	Never−married	?	Own−child	Male
7110	Private	Married−civ−spouse	Sales	Husband	Female
10846	Never−worked	Divorced	?	Not−in−family	Male
14773	Never−worked	Never−married	?	Own−child	Male
20338	Never−worked	Never−married	?	Own−child	Female
23233	Never−worked	Never−married	?	Own−child	Male
27142	Private	Married−civ−spouse	Sales	Wife	Male
32305	Never−worked	Married−civ−spouse	?	Wife	Female
32315	Never−worked	Never−married	?	Own−child	Male

- Step 1:

For observations 576, 7110 and 27142, variable `sex` and `relationship` are obviously in conflict with each other. A person cannot be wife and male, or husband and female at the same time. There is no way for us to detect which variable value was entried correctly and which was not. Hence, we would like to discard these three observations.

- Step 2:

For the rest observations on the list, variable `occupation` is recorded as a question mark `?` because there is no corresponding level in the variable `workclass`. Hence we would like to add a new level `Unemployed` and change these question marks `?` into it.

- Step 3:

After correcting those points, we notice that the rest question mark `?` values in predictors `occupation` and `workclass` are exactly one-to-one match. It means that if we find variable `occupation` of one observation is recorded as `?`, `workclass` of the same observation is recorded as `?`, too. And if we find variable `occupation` of one observation is not recorded as `?`, `workclass` of the same observation is not `?`, either. Hence, we would like to treat the these question marks as a new level `other`.

- Step 4:

Variable `native_country` still contains question mark `?`. Here, we cannot find a proper way for us to correct those observations. Here, we would like to impute those missing values. Decision tree is used here.

3.2 Dataset spliting and Outliers

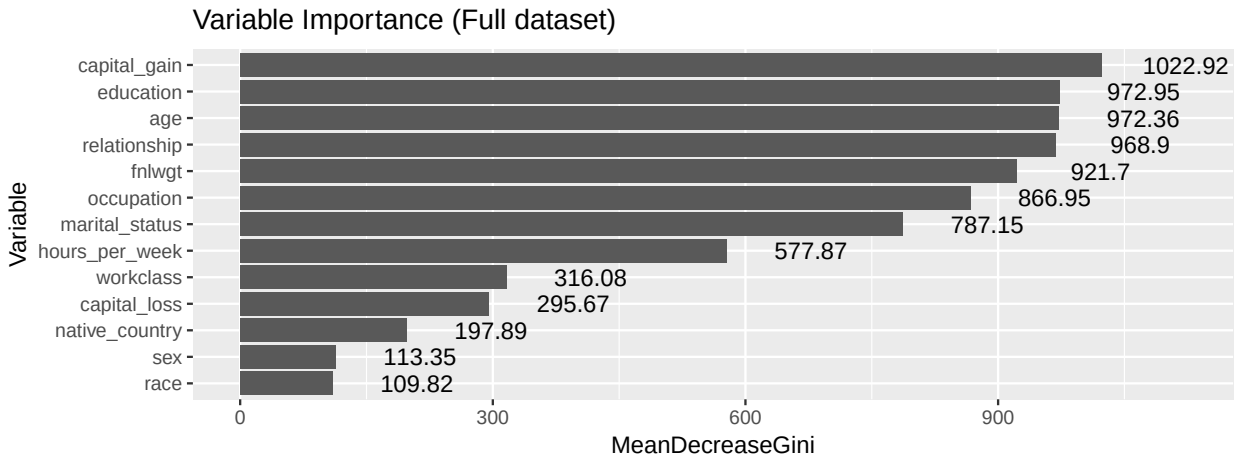
Here, we would like to split our datasets into training and testing sets. We will use training datasets only to fit our statistical models, and both training and testing prediction will be used to validated our models. Like we discussed in the previous part, we would like to have two dataets, one being the original full set while the other one being the subset without outliers.

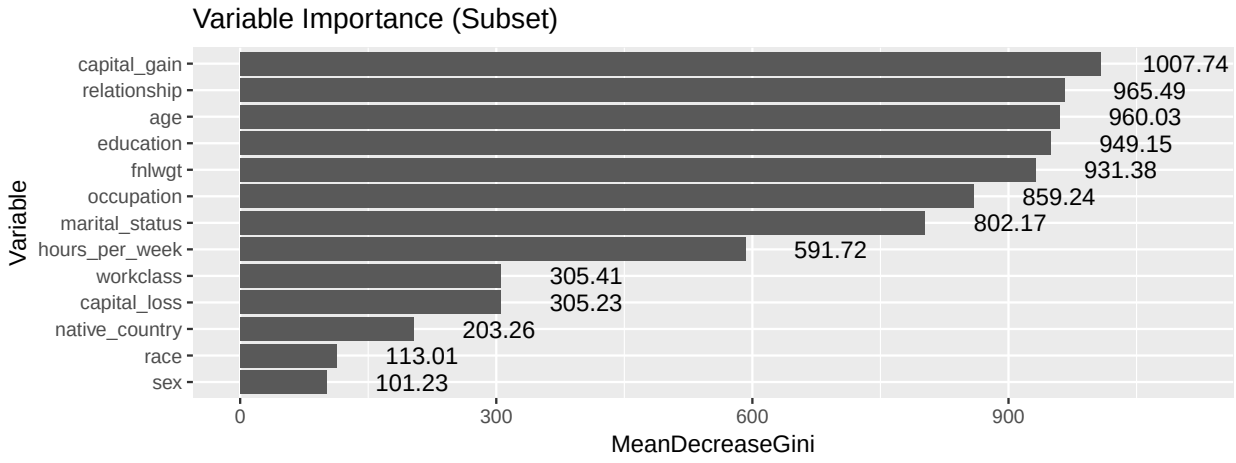
Here, for variable `fnlwgt`, `capital_gain`, `capital_loss`, and `hours_per_week`, we would like to keep the outliers because those points actually stand for some real world issues here. Pairwise deleting may cause serious problems and make our models less trustable.

However, for `age`, we would like to do pairwise delete all observations with `age` value being greater han or equal to 78. 78 is the upper bound of interval $(Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR)$, and there is no value less than the lower bound of this interval.

3.3 Feature selection

We would like to use variable importance from `Random Forest` here to help us select the features that we will include in our models. Both original datasets and datasets without outliers will be calculated here. We visualize the quantitive output in the following graphs.





From both graphs, we can read that the results of our two models are basically the same as each other. Based on the result, we would like to keep `capital_gain`, `education`, `relationship`, `age`, `fnlwgt`, `occupation` and `marital_status` into our final models as our predictors.

Additionaly, for logistic regression models, backward elimination will also be applied in case of insignificant predictors being existing.

4 Model fitting and Prediction

In this part, the five algorithms that we are about to use are `Logistic Regression`, `CART`, `Naive Bayes`, `Bagging` and `Random Forest`. For each algorithm, we would like to build at least two models. One is based on training set obtained from the full dataset, and the other one is from the training set obtained from the dataset without outliers.

4.1 Logistic regression

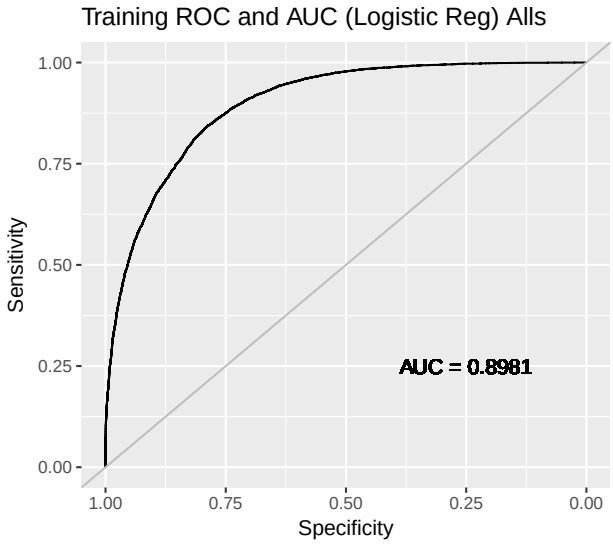
4.1.1 Model with full dataset

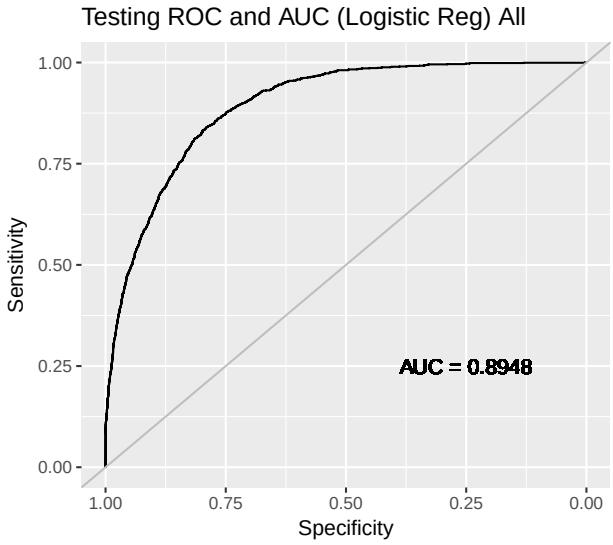
First we would like to build a Logistic Regression by using training dataset obtained from original dataset (without removing outliers).

```
##
## Call:
## glm(formula = salary_num ~ relationship + capital_gain + age +
##     education + occupation + marital_status + fnlwgt, family = binomial(link = "logit"),
##     data = Train_Adult)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.0324  -0.5407  -0.2073  -0.0384   3.6972
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -2.370e+01  1.037e+02  -0.228   0.8193
## relationshipNot-in-family        4.933e-02  2.929e-01   0.168   0.8663
## relationshipOther-relative      -1.064e+00  2.643e-01  -4.027 5.65e-05
## relationshipOwn-child           -1.356e+00  2.925e-01  -4.638 3.52e-06
```

```
## relationshipUnmarried               -3.751e-01  3.068e-01  -1.223   0.2214
## relationshipWife                     3.383e-01  7.429e-02   4.554 5.27e-06
## capital_gain                         3.079e-04  1.120e-05  27.476  < 2e-16
## age                                  2.096e-02  1.727e-03  12.137  < 2e-16
## education1st-4th                     1.806e+01  1.037e+02   0.174   0.8618
## education5th-6th                     1.849e+01  1.037e+02   0.178   0.8585
## education7th-8th                     1.833e+01  1.037e+02   0.177   0.8598
## education9th                         1.856e+01  1.037e+02   0.179   0.8580
## education10th                        1.880e+01  1.037e+02   0.181   0.8562
## education11th                        1.875e+01  1.037e+02   0.181   0.8565
## education12th                        1.924e+01  1.037e+02   0.185   0.8528
## educationHS-grad                     1.959e+01  1.037e+02   0.189   0.8502
## educationSome-college                1.996e+01  1.037e+02   0.192   0.8474
## educationAssoc-voc                   2.016e+01  1.037e+02   0.194   0.8459
## educationAssoc-acdm                  2.011e+01  1.037e+02   0.194   0.8463
## educationBachelors                   2.080e+01  1.037e+02   0.201   0.8410
## educationMasters                     2.110e+01  1.037e+02   0.203   0.8388
## educationProf-school                 2.175e+01  1.037e+02   0.210   0.8339
## educationDoctorate                   2.187e+01  1.037e+02   0.211   0.8330
## occupationArmed-Forces              -1.151e+01  2.903e+02  -0.040   0.9684
## occupationCraft-repair               1.914e-01  8.495e-02   2.253   0.0242
## occupationExec-managerial            9.068e-01  8.099e-02  11.196  < 2e-16
## occupationFarming-fishing           -9.073e-01  1.499e-01  -6.053 1.43e-09
## occupationHandlers-cleaners         -6.402e-01  1.559e-01  -4.106 4.03e-05
## occupationMachine-op-inspct         -1.634e-01  1.092e-01  -1.496   0.1346
## occupationOther                     -8.867e-01  1.272e-01  -6.971 3.16e-12
## occupationOther-service             -8.764e-01  1.250e-01  -7.012 2.35e-12
## occupationPriv-house-serv           -3.715e+00  1.822e+00  -2.039   0.0415
## occupationProf-specialty             5.269e-01  8.601e-02   6.126 8.99e-10
## occupationProtective-serv            5.728e-01  1.327e-01   4.317 1.58e-05
## occupationSales                      4.090e-01  8.594e-02   4.759 1.94e-06
## occupationTech-support               7.901e-01  1.194e-01   6.619 3.62e-11
## occupationTransport-moving           1.105e-01  1.069e-01   1.034   0.3012
## occupationUnemployed                -1.136e+01  2.927e+02  -0.039   0.9690
## marital_statusMarried-AF-spouse      2.879e+00  6.462e-01   4.456 8.35e-06
## marital_statusMarried-civ-spouse     2.013e+00  2.969e-01   6.781 1.19e-11
## marital_statusMarried-spouse-absent  4.004e-02  2.424e-01   0.165   0.8688
## marital_statusNever-married         -4.535e-01  9.451e-02  -4.798 1.60e-06
## marital_statusSeparated             -7.927e-02  1.773e-01  -0.447   0.6549
## marital_statusWidowed               -2.009e-01  1.642e-01  -1.224   0.2210
## fnlwgt                               4.107e-07  1.813e-07   2.265   0.0235
##
## (Intercept)
## relationshipNot-in-family
## relationshipOther-relative          ***
## relationshipOwn-child               ***
## relationshipUnmarried
## relationshipWife                    ***
## capital_gain                        ***
## age                                 ***
## education1st-4th
## education5th-6th
## education7th-8th
## education9th
```

```
## education10th
## education11th
## education12th
## educationHS-grad
## educationSome-college
## educationAssoc-voc
## educationAssoc-acdm
## educationBachelors
## educationMasters
## educationProf-school
## educationDoctorate
## occupationArmed-Forces
## occupationCraft-repair              *
## occupationExec-managerial           ***
## occupationFarming-fishing           ***
## occupationHandlers-cleaners         ***
## occupationMachine-op-inspct
## occupationOther                     ***
## occupationOther-service             ***
## occupationPriv-house-serv           *
## occupationProf-specialty            ***
## occupationProtective-serv           ***
## occupationSales                     ***
## occupationTech-support              ***
## occupationTransport-moving
## occupationUnemployed
## marital_statusMarried-AF-spouse     ***
## marital_statusMarried-civ-spouse    ***
## marital_statusMarried-spouse-absent
## marital_statusNever-married         ***
## marital_statusSeparated
## marital_statusWidowed
## fnlwgt                              *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28853  on 26045  degrees of freedom
## Residual deviance: 17331  on 26001  degrees of freedom
## AIC: 17421
##
## Number of Fisher Scoring iterations: 13
```

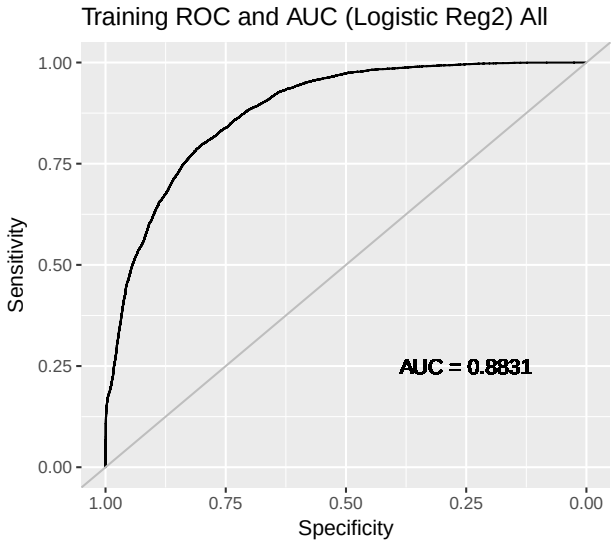


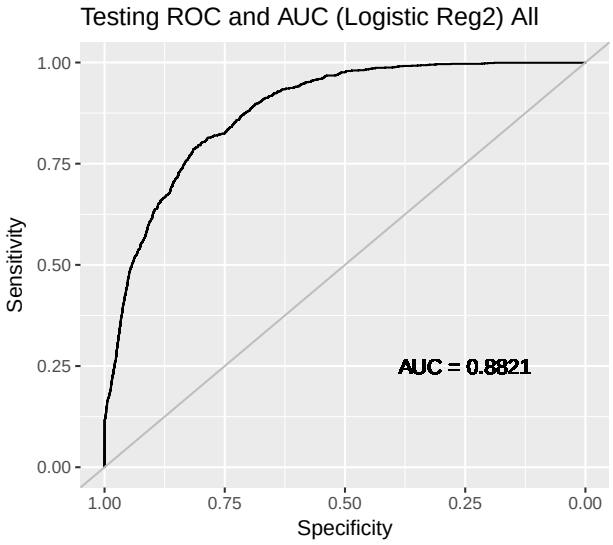


The predictions are also made here. We notice that `education` is not significant here. We will build another model without considerding `education` as one predictor.

```
##
## Call:
## glm(formula = salary_num ~ relationship + capital_gain + age +
##     occupation + marital_status + fnlwgt, family = binomial(link = "logit"),
##     data = Train_Adult)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.7059  -0.5805  -0.2286  -0.0631   3.5056
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -3.451e+00  3.135e-01 -11.011  < 2e-16
## relationshipNot-in-family          4.860e-02  2.876e-01   0.169 0.865809
## relationshipOther-relative        -1.280e+00  2.625e-01  -4.879 1.07e-06
## relationshipOwn-child             -1.559e+00  2.875e-01  -5.421 5.93e-08
## relationshipUnmarried             -4.740e-01  3.011e-01  -1.574 0.115411
## relationshipWife                   1.950e-01  7.196e-02   2.709 0.006748
## capital_gain                       3.047e-04  1.072e-05  28.433  < 2e-16
## age                                1.758e-02  1.630e-03  10.783  < 2e-16
## occupationArmed-Forces            -1.100e+01  1.732e+02  -0.064 0.949349
## occupationCraft-repair            -1.505e-01  8.160e-02  -1.845 0.065058
## occupationExec-managerial          1.183e+00  7.777e-02  15.217  < 2e-16
## occupationFarming-fishing         -1.239e+00  1.460e-01  -8.488  < 2e-16
## occupationHandlers-cleaners       -1.088e+00  1.515e-01  -7.180 6.98e-13
## occupationMachine-op-inspct       -6.440e-01  1.047e-01  -6.152 7.66e-10
## occupationOther                   -9.175e-01  1.210e-01  -7.584 3.36e-14
## occupationOther-service           -1.203e+00  1.214e-01  -9.911  < 2e-16
## occupationPriv-house-serv         -3.529e+00  1.384e+00  -2.551 0.010740
## occupationProf-specialty           1.307e+00  7.809e-02  16.735  < 2e-16
## occupationProtective-serv          4.987e-01  1.293e-01   3.856 0.000115
## occupationSales                    4.713e-01  8.295e-02   5.682 1.33e-08
## occupationTech-support             9.276e-01  1.165e-01   7.963 1.68e-15
```

```
## occupationTransport-moving          -3.524e-01  1.027e-01  -3.432 0.000600
## occupationUnemployed                -1.091e+01  1.869e+02  -0.058 0.953422
## marital_statusMarried-AF-spouse      2.813e+00  6.389e-01   4.403 1.07e-05
## marital_statusMarried-civ-spouse     1.982e+00  2.918e-01   6.793 1.10e-11
## marital_statusMarried-spouse-absent  9.364e-02  2.356e-01   0.397 0.691071
## marital_statusNever-married         -3.537e-01  9.205e-02  -3.843 0.000122
## marital_statusSeparated             -1.157e-01  1.735e-01  -0.667 0.504864
## marital_statusWidowed               -3.223e-01  1.602e-01  -2.012 0.044245
## fnlwgt                               2.988e-07  1.756e-07   1.702 0.088750
##
## (Intercept)                         ***
## relationshipNot-in-family
## relationshipOther-relative          ***
## relationshipOwn-child               ***
## relationshipUnmarried
## relationshipWife                    **
## capital_gain                        ***
## age                                 ***
## occupationArmed-Forces
## occupationCraft-repair              .
## occupationExec-managerial           ***
## occupationFarming-fishing           ***
## occupationHandlers-cleaners         ***
## occupationMachine-op-inspct         ***
## occupationOther                     ***
## occupationOther-service             ***
## occupationPriv-house-serv           *
## occupationProf-specialty            ***
## occupationProtective-serv           ***
## occupationSales                     ***
## occupationTech-support              ***
## occupationTransport-moving          ***
## occupationUnemployed
## marital_statusMarried-AF-spouse     ***
## marital_statusMarried-civ-spouse    ***
## marital_statusMarried-spouse-absent
## marital_statusNever-married         ***
## marital_statusSeparated
## marital_statusWidowed               *
## fnlwgt                              .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28853  on 26045  degrees of freedom
## Residual deviance: 18392  on 26016  degrees of freedom
## AIC: 18452
##
## Number of Fisher Scoring iterations: 12
```





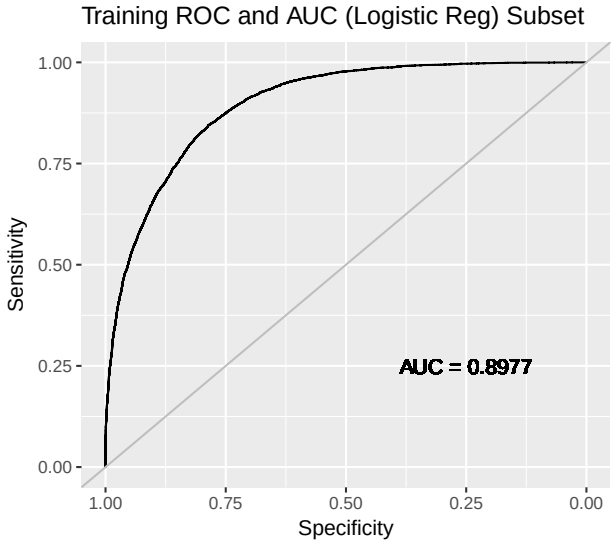
4.1.2 Model with subset (Without outliers)

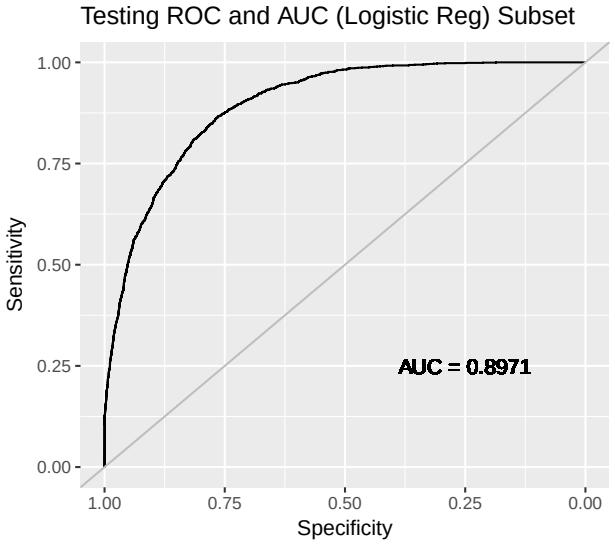
Then, we would like to fit the model by using dataset without outliers.

```
##
## Call:
## glm(formula = salary_num ~ relationship + capital_gain + age +
##     education + occupation + marital_status + fnlwgt, family = binomial(link = "logit"),
##     data = Train_Adult1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.0191  -0.5432  -0.2079  -0.0403   3.6554
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -2.408e+01  1.106e+02  -0.218  0.82760
## relationshipNot-in-family        1.648e-01  2.882e-01   0.572  0.56736
## relationshipOther-relative      -9.192e-01  2.645e-01  -3.475  0.00051
## relationshipOwn-child           -1.246e+00  2.875e-01  -4.334 1.47e-05
## relationshipUnmarried           -2.571e-01  3.027e-01  -0.850  0.39560
## relationshipWife                 3.650e-01  7.394e-02   4.936 7.97e-07
## capital_gain                     3.022e-04  1.099e-05  27.505  < 2e-16
## age                              2.145e-02  1.726e-03  12.427  < 2e-16
## education1st-4th                 1.833e+01  1.106e+02   0.166  0.86835
## education5th-6th                 1.881e+01  1.106e+02   0.170  0.86488
## education7th-8th                 1.835e+01  1.106e+02   0.166  0.86816
## education9th                     1.855e+01  1.106e+02   0.168  0.86672
## education10th                    1.902e+01  1.106e+02   0.172  0.86342
## education11th                    1.912e+01  1.106e+02   0.173  0.86272
## education12th                    1.934e+01  1.106e+02   0.175  0.86111
## educationHS-grad                 1.977e+01  1.106e+02   0.179  0.85810
## educationSome-college            2.016e+01  1.106e+02   0.182  0.85533
## educationAssoc-voc               2.034e+01  1.106e+02   0.184  0.85399
## educationAssoc-acdm              2.032e+01  1.106e+02   0.184  0.85414
```

```
## educationBachelors                    2.095e+01  1.106e+02   0.189  0.84974
## educationMasters                      2.127e+01  1.106e+02   0.192  0.84744
## educationProf-school                  2.193e+01  1.106e+02   0.198  0.84278
## educationDoctorate                    2.205e+01  1.106e+02   0.199  0.84189
## occupationArmed-Forces                4.437e-02  1.343e+00   0.033  0.97365
## occupationCraft-repair                1.869e-01  8.482e-02   2.203  0.02758
## occupationExec-managerial             9.351e-01  8.111e-02  11.529  < 2e-16
## occupationFarming-fishing            -7.416e-01  1.453e-01  -5.105 3.30e-07
## occupationHandlers-cleaners          -6.113e-01  1.534e-01  -3.986 6.72e-05
## occupationMachine-op-inspct          -1.655e-01  1.096e-01  -1.510  0.13109
## occupationOther                      -8.417e-01  1.253e-01  -6.715 1.88e-11
## occupationOther-service              -9.097e-01  1.261e-01  -7.214 5.44e-13
## occupationPriv-house-serv            -3.722e+00  1.720e+00  -2.163  0.03052
## occupationProf-specialty              5.775e-01  8.591e-02   6.722 1.80e-11
## occupationProtective-serv             5.601e-01  1.311e-01   4.272 1.94e-05
## occupationSales                       4.351e-01  8.592e-02   5.064 4.11e-07
## occupationTech-support                7.242e-01  1.188e-01   6.098 1.08e-09
## occupationTransport-moving            8.565e-02  1.065e-01   0.804  0.42139
## occupationUnemployed                 -1.175e+01  3.612e+02  -0.033  0.97404
## marital_statusMarried-AF-spouse       2.644e+00  5.895e-01   4.485 7.30e-06
## marital_statusMarried-civ-spouse      2.140e+00  2.926e-01   7.314 2.58e-13
## marital_statusMarried-spouse-absent -3.125e-04  2.482e-01  -0.001  0.99900
## marital_statusNever-married          -4.548e-01  9.532e-02  -4.771 1.83e-06
## marital_statusSeparated              -1.210e-01  1.788e-01  -0.677  0.49848
## marital_statusWidowed                -1.267e-01  1.613e-01  -0.785  0.43217
## fnlwgt                                5.904e-07  1.821e-07   3.242  0.00119
##
## (Intercept)
## relationshipNot-in-family
## relationshipOther-relative          ***
## relationshipOwn-child               ***
## relationshipUnmarried
## relationshipWife                    ***
## capital_gain                        ***
## age                                 ***
## education1st-4th
## education5th-6th
## education7th-8th
## education9th
## education10th
## education11th
## education12th
## educationHS-grad
## educationSome-college
## educationAssoc-voc
## educationAssoc-acdm
## educationBachelors
## educationMasters
## educationProf-school
## educationDoctorate
## occupationArmed-Forces
## occupationCraft-repair              *
## occupationExec-managerial           ***
## occupationFarming-fishing           ***
```

```
## occupationHandlers-cleaners         ***
## occupationMachine-op-inspct
## occupationOther                     ***
## occupationOther-service             ***
## occupationPriv-house-serv           *
## occupationProf-specialty            ***
## occupationProtective-serv           ***
## occupationSales                     ***
## occupationTech-support              ***
## occupationTransport-moving
## occupationUnemployed
## marital_statusMarried-AF-spouse     ***
## marital_statusMarried-civ-spouse    ***
## marital_statusMarried-spouse-absent
## marital_statusNever-married         ***
## marital_statusSeparated
## marital_statusWidowed
## fnlwgt                              **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28773  on 26102  degrees of freedom
## Residual deviance: 17339  on 26058  degrees of freedom
## AIC: 17429
##
## Number of Fisher Scoring iterations: 13
```



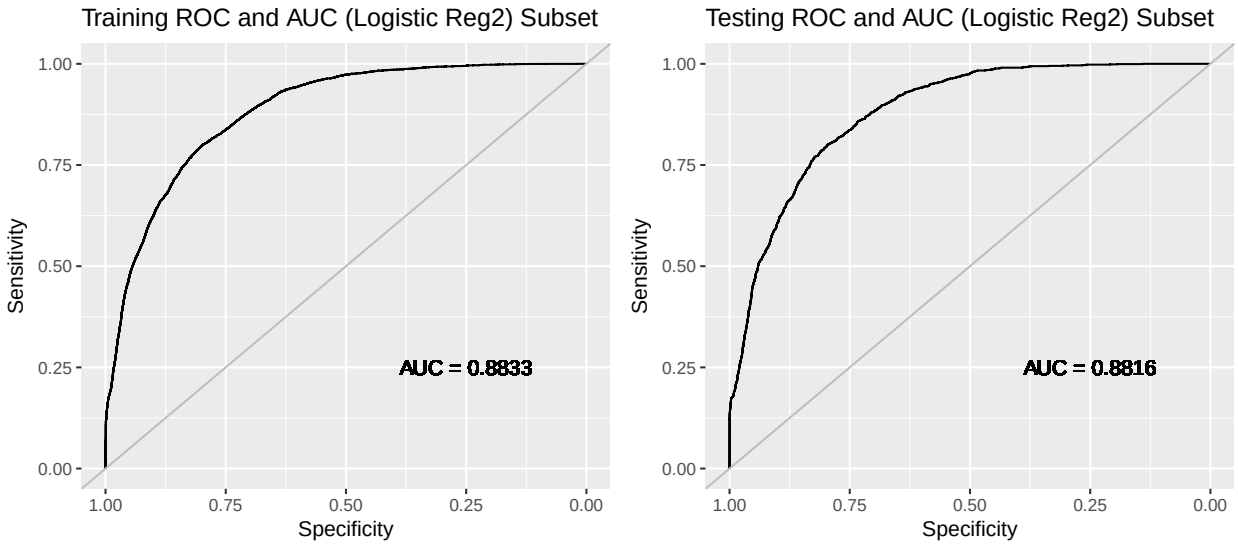


In the second model, we are facing the same problem that `education` is not statistically significant in our model. Like the previous part, we would like to fit another model without `education` variable.

```
##
## Call:
## glm(formula = salary_num ~ relationship + capital_gain + age +
##     occupation + marital_status + fnlwgt, family = binomial(link = "logit"),
##     data = Train_Adult1)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.7181  -0.5846  -0.2296  -0.0618   3.5258
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -3.681e+00  3.102e-01 -11.866  < 2e-16
## relationshipNot-in-family          1.763e-01  2.839e-01   0.621 0.534663
## relationshipOther-relative        -1.147e+00  2.646e-01  -4.335 1.46e-05
## relationshipOwn-child             -1.434e+00  2.829e-01  -5.070 3.98e-07
## relationshipUnmarried             -3.344e-01  2.979e-01  -1.122 0.261714
## relationshipWife                   2.245e-01  7.184e-02   3.125 0.001778
## capital_gain                       3.017e-04  1.055e-05  28.603  < 2e-16
## age                                1.813e-02  1.635e-03  11.092  < 2e-16
## occupationArmed-Forces             1.459e-01  1.187e+00   0.123 0.902175
## occupationCraft-repair            -1.480e-01  8.165e-02  -1.813 0.069858
## occupationExec-managerial          1.213e+00  7.801e-02  15.550  < 2e-16
## occupationFarming-fishing         -1.053e+00  1.409e-01  -7.473 7.84e-14
## occupationHandlers-cleaners       -1.050e+00  1.488e-01  -7.052 1.76e-12
## occupationMachine-op-inspct       -6.265e-01  1.054e-01  -5.943 2.80e-09
## occupationOther                   -8.721e-01  1.198e-01  -7.281 3.30e-13
## occupationOther-service           -1.216e+00  1.226e-01  -9.918  < 2e-16
## occupationPriv-house-serv         -3.606e+00  1.338e+00  -2.694 0.007056
## occupationProf-specialty           1.353e+00  7.812e-02  17.314  < 2e-16
## occupationProtective-serv          4.989e-01  1.282e-01   3.893 9.89e-05
## occupationSales                    4.905e-01  8.316e-02   5.898 3.67e-09
## occupationTech-support             8.783e-01  1.157e-01   7.589 3.23e-14
## occupationTransport-moving        -3.375e-01  1.025e-01  -3.292 0.000994
## occupationUnemployed              -1.016e+01  1.358e+02  -0.075 0.940388
## marital_statusMarried-AF-spouse    2.528e+00  5.891e-01   4.292 1.77e-05
## marital_statusMarried-civ-spouse   2.124e+00  2.885e-01   7.363 1.80e-13
## marital_statusMarried-spouse-absent 2.977e-02 2.403e-01   0.124 0.901387
## marital_statusNever-married       -3.410e-01  9.282e-02  -3.674 0.000239
## marital_statusSeparated           -1.965e-01  1.769e-01  -1.111 0.266554
## marital_statusWidowed             -2.451e-01  1.577e-01  -1.554 0.120091
## fnlwgt                             4.655e-07  1.766e-07   2.636 0.008396
##
## (Intercept)                       ***
## relationshipNot-in-family
## relationshipOther-relative        ***
## relationshipOwn-child             ***
## relationshipUnmarried
## relationshipWife                  **
## capital_gain                      ***
## age                               ***
## occupationArmed-Forces
## occupationCraft-repair            .
## occupationExec-managerial         ***
## occupationFarming-fishing         ***
## occupationHandlers-cleaners       ***
## occupationMachine-op-inspct       ***
## occupationOther                   ***
## occupationOther-service           ***
```

```
## occupationPriv-house-serv           **
## occupationProf-specialty            ***
## occupationProtective-serv           ***
## occupationSales                     ***
## occupationTech-support              ***
## occupationTransport-moving          ***
## occupationUnemployed
## marital_statusMarried-AF-spouse     ***
## marital_statusMarried-civ-spouse    ***
## marital_statusMarried-spouse-absent
## marital_statusNever-married         ***
## marital_statusSeparated
## marital_statusWidowed
## fnlwgt                              **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28773  on 26102  degrees of freedom
## Residual deviance: 18360  on 26073  degrees of freedom
## AIC: 18420
##
## Number of Fisher Scoring iterations: 11
```



4.1.3 Summary of Logistic Regression

We notice that for both of the datasets, we built two models. The first one is based on all predictors selected from Random Forest, and the second one is based on the result of the first one.

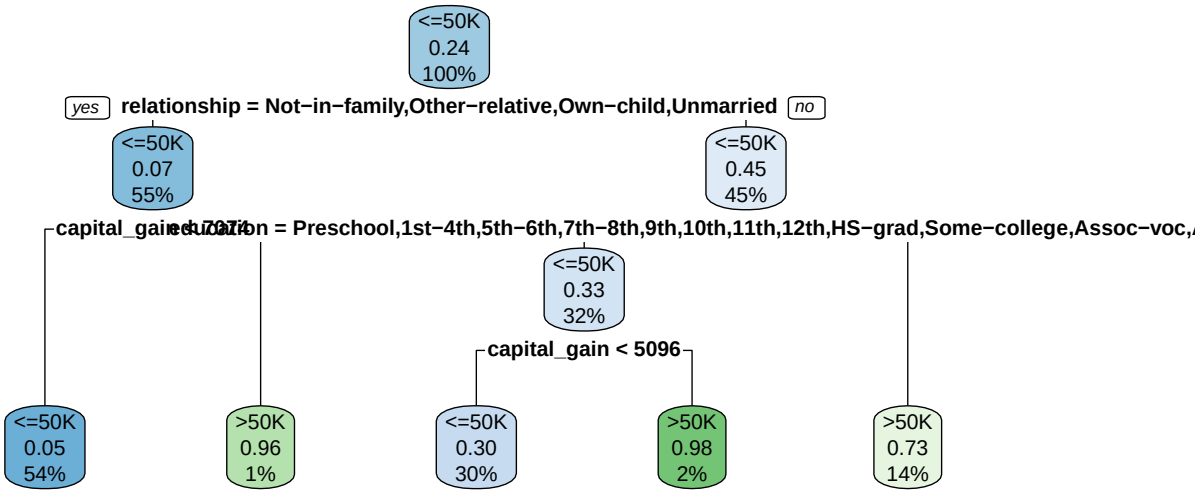
For the `4.1.1` part, even though the second model does not contain any statistical insignificant predictors any more, the overall AIC is going up and the corresponding prediction AUCs are going down, too. The exactly same issue happens to the `4.1.2`.

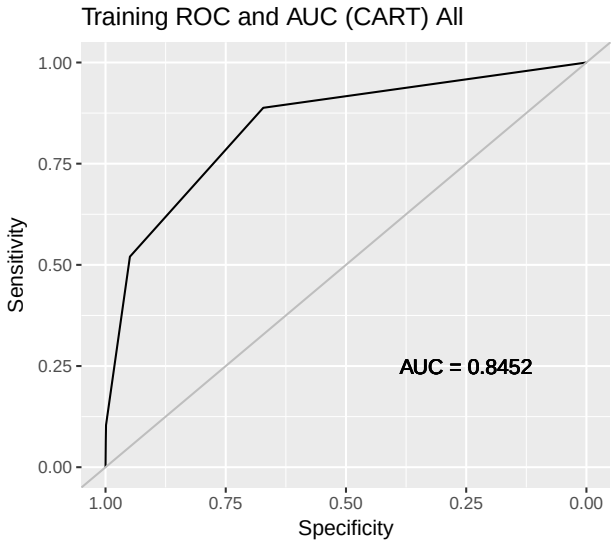
Hence, in this case, I still would like to keep first model of both parts as our final models of this algorithm since their AICs and AUCs are better.

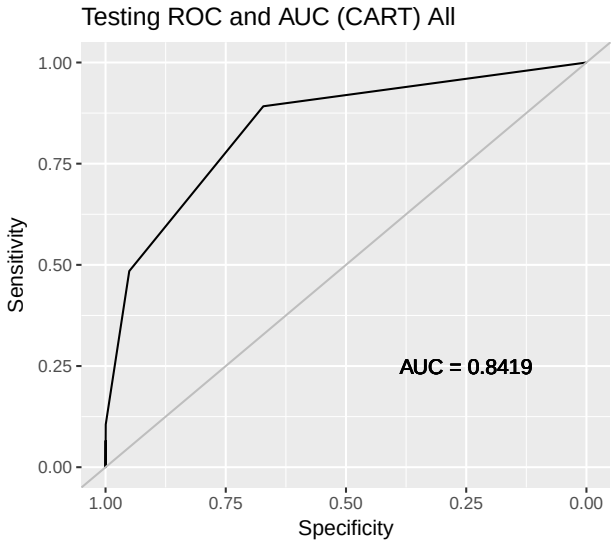
4.2 Classification and Regression Tree

4.2.1 Model with full dataset

Here, we would like to build a CART model by using training dataset obtained from original dataset (without removing outliers).

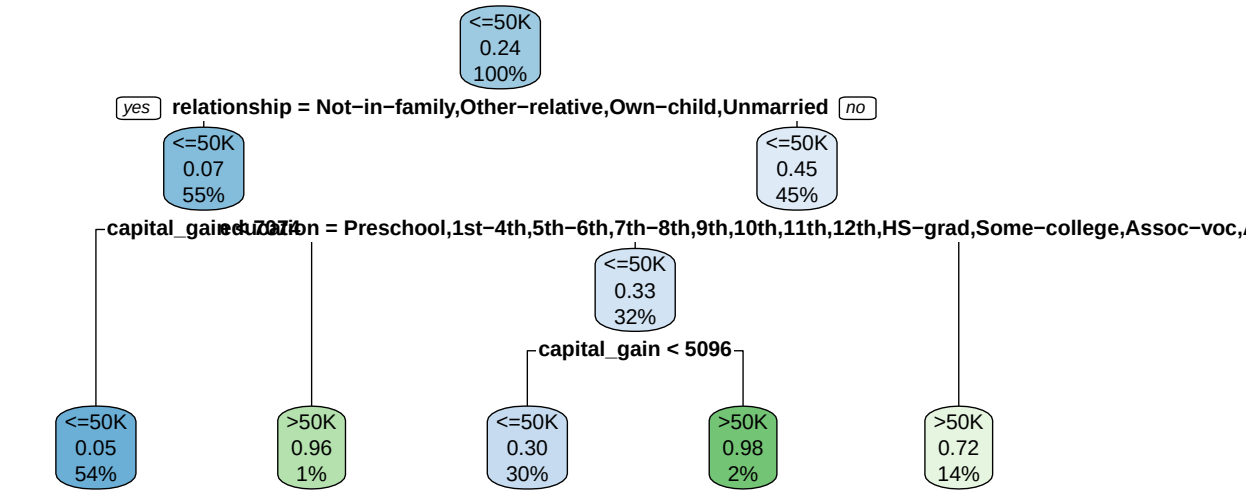


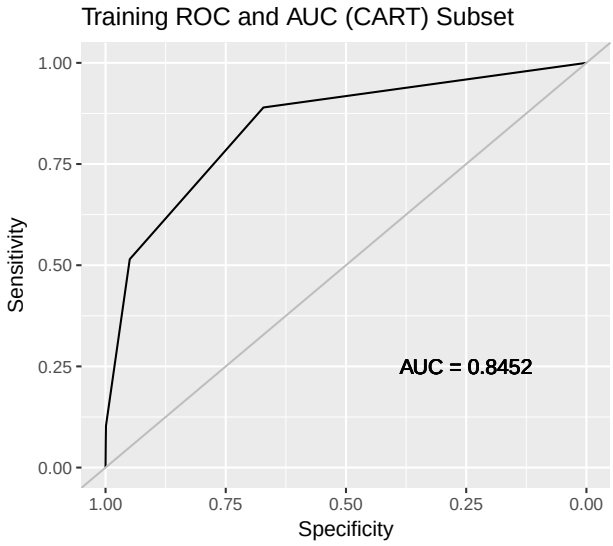


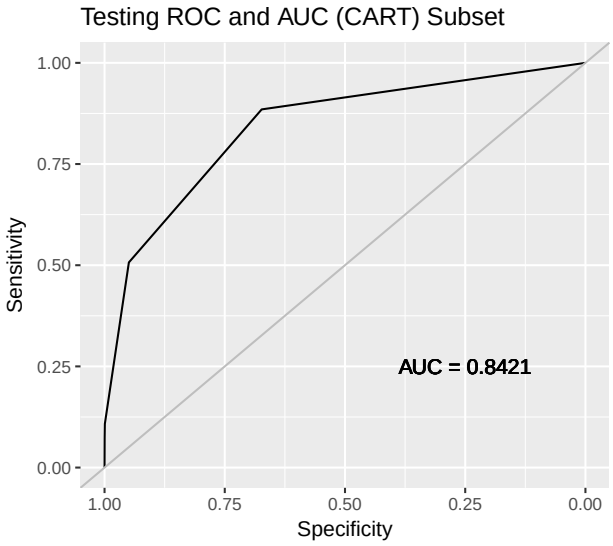


4.2.2 Model with subset (Without outliers)

Following, we would like to build a CART model by using training dataset obtained from dataset with outliers.







4.2.3 Summary of CART

We noticed from the tree plot and ROC curves that these two models are slightly different with each other. This also shows us one of weakness which is its sensitivity. The model may change dramatically by removing or adding observations. Here, we only removed very few of the total observations, which does not result in significant variation.

Bagging and Random Forest models will also be attempt later.

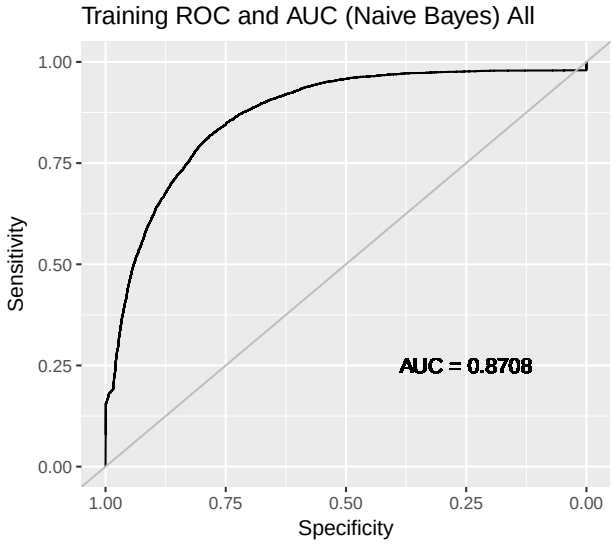
4.3 Naive Bayes classifier

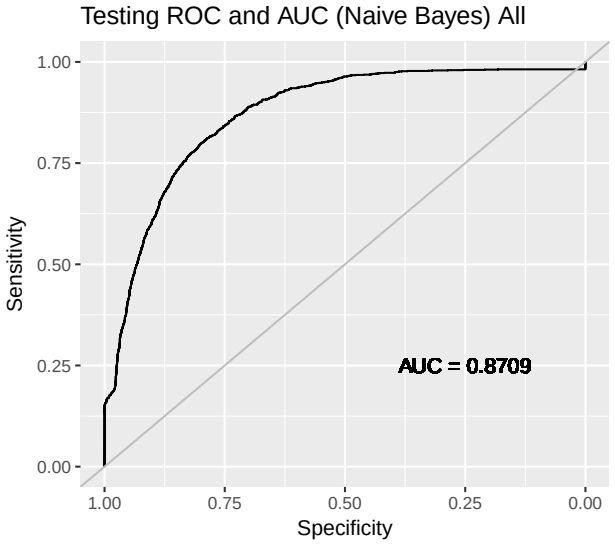
4.3.1 Model with full dataset

Firstly, we would like to build a Naive Bayes model by using training dataset obtained from original dataset (without removing outliers). The summary of this model and the training and testing ROC curves and AUCs are also shown below.

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##     <=50K      >50K
## 0.7575443 0.2424557
##
## Conditional probabilities:
##        relationship
## Y           Husband Not-in-family Other-relative   Own-child   Unmarried
##   <=50K 0.294967310   0.300086159    0.038011251 0.205311439 0.129086210
##   >50K  0.754711006   0.109263658    0.004592241 0.008551069 0.028503563
##        relationship
## Y              Wife
##   <=50K 0.032537631
##   >50K  0.094378464
##
##        capital_gain
## Y            [,1]      [,2]
##   <=50K  146.5552  1005.488
##   >50K  4059.0163 14719.714
##
##        age
## Y          [,1]     [,2]
##   <=50K 36.77269 14.06292
##   >50K  44.29264 10.55898
##
##        education
## Y          Preschool      1st-4th      5th-6th      7th-8th          9th
##   <=50K 0.0017738584 0.0064365719 0.0130758705 0.0256449242 0.0199178957
##   >50K  0.0000000000 0.0006334125 0.0022169438 0.0058590657 0.0039588282
##        education
## Y               10th         11th         12th      HS-grad Some-college
##   <=50K 0.0352744412 0.0443971416 0.0166742689 0.3556839491 0.2404845168
##   >50K  0.0079176564 0.0072842439 0.0042755344 0.2091844814 0.1745051465
##        education
## Y          Assoc-voc   Assoc-acdm    Bachelors      Masters  Prof-school
##   <=50K 0.0399371547 0.0320308145 0.1271096244 0.0311185444 0.0060818002
##   >50K  0.0462391132 0.0311955661 0.2883610451 0.1241488519 0.0546318290
##        education
## Y          Doctorate
##   <=50K 0.0043586235
##   >50K  0.0395882819
##
##        occupation
## Y       Adm-clerical Armed-Forces Craft-repair Exec-managerial
##   <=50K 0.1334955147 0.0004054533 0.1276671228    0.0857533830
##   >50K  0.0638163104 0.0000000000 0.1182897862    0.2514647664
##        occupation
## Y       Farming-fishing Handlers-cleaners Machine-op-inspct        Other
```

```
##   <=50K    0.0359333029      0.0514925751      0.0710556992 0.0666970757
##   >50K     0.0133016627      0.0109263658      0.0326207443 0.0240696754
##        occupation
## Y       Other-service Priv-house-serv Prof-specialty Protective-serv
##   <=50K  0.1282753028    0.0058283919   0.0917845015    0.0174344939
##   >50K   0.0183689628    0.0001583531   0.2378463975    0.0259699129
##        occupation
## Y              Sales Tech-support Transport-moving   Unemployed
##   <=50K 0.1081546805 0.0247326542     0.0509857585 0.0003040900
##   >50K  0.1266825020 0.0367379256     0.0397466350 0.0000000000
##
##        marital_status
## Y           Divorced Married-AF-spouse Married-civ-spouse
##   <=50K 0.1608636156      0.0004054533       0.3354619634
##   >50K  0.0592240697      0.0012668250       0.8522565321
##        marital_status
## Y       Married-spouse-absent Never-married    Separated      Widowed
##   <=50K          0.0151031372  0.4140692312 0.0376564797 0.0364401196
##   >50K           0.0047505938  0.0636579572 0.0079176564 0.0109263658
##
##        fnlwgt
## Y           [,1]     [,2]
##   <=50K 190813.3 107226.4
##   >50K  187896.0 103360.2
```





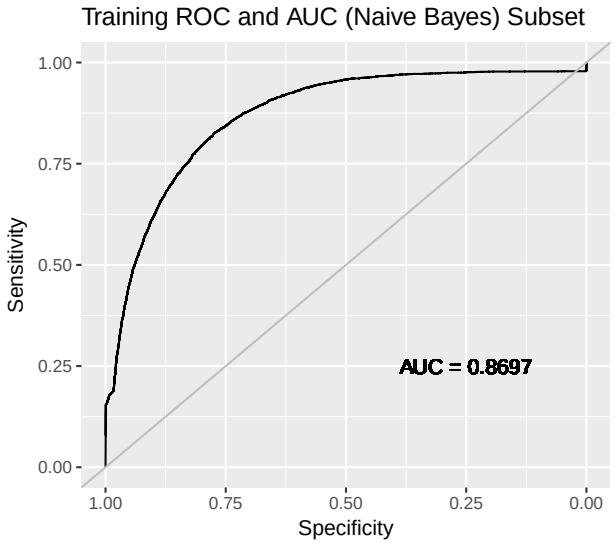
4.3.2 Model with subset (Without outliers)

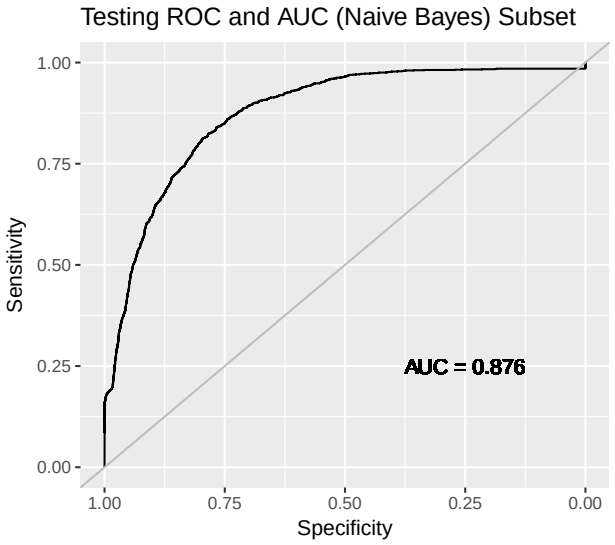
Then, we would like to build another Naive Bayes model by using training dataset obtained from dataset (after removing outliers). The summary of this model and the training and testing ROC curves and AUCs are also shown below, as well.

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
```

```
##
## A-priori probabilities:
## Y
##    <=50K     >50K
## 0.759951 0.240049
##
## Conditional probabilities:
##        relationship
## Y          Husband Not-in-family Other-relative  Own-child  Unmarried
##   <=50K 0.29480264    0.30065030     0.03806019 0.20365983 0.12980793
##   >50K  0.75359081    0.10916055     0.00446856 0.00877753 0.02792850
##        relationship
## Y             Wife
##   <=50K 0.03301911
##   >50K  0.09607405
##
##        capital_gain
## Y            [,1]      [,2]
##   <=50K  152.2676  1012.858
##   >50K  4102.7933 14957.907
##
##        age
## Y           [,1]     [,2]
##   <=50K 36.74366 14.05225
##   >50K  44.17028 10.50472
##
##        education
## Y          Preschool      1st-4th      5th-6th      7th-8th          9th
##   <=50K 0.0022180773 0.0062509452 0.0128043555 0.0239451530 0.0197610526
##   >50K  0.0000000000 0.0007979572 0.0025534631 0.0046281519 0.0031918289
##        education
## Y               10th         11th         12th      HS-grad Some-college
##   <=50K 0.0356404698 0.0466804456 0.0161314715 0.3560518224 0.2393507083
##   >50K  0.0084583466 0.0084583466 0.0041493776 0.2090647941 0.1784232365
##        education
## Y          Assoc-voc   Assoc-acdm    Bachelors      Masters  Prof-school
##   <=50K 0.0406815547 0.0324141755 0.1269345163 0.0307506175 0.0059988910
##   >50K  0.0459623364 0.0330354293 0.2839131823 0.1217682732 0.0563357804
##        education
## Y          Doctorate
##   <=50K 0.0043857438
##   >50K  0.0392594957
##
##        occupation
## Y       Adm-clerical Armed-Forces Craft-repair Exec-managerial
##   <=50K 0.1325301205 0.0003528759 0.1281947875    0.0848918687
##   >50K  0.0643153527 0.0001595914 0.1172997127    0.2481646984
##        occupation
## Y       Farming-fishing Handlers-cleaners Machine-op-inspct        Other
##   <=50K    0.0345314312      0.0532338559      0.0692645057 0.0676009477
##   >50K     0.0150015959      0.0114905841      0.0317586977 0.0247366741
##        occupation
## Y       Other-service Priv-house-serv Prof-specialty Protective-serv
##   <=50K  0.1297071130    0.0056460150   0.0914956899    0.0173917427
```

```
##   >50K   0.0180338334    0.0001595914   0.2417810405    0.0263325886
##        occupation
## Y              Sales Tech-support Transport-moving   Unemployed
##   <=50K 0.1070726420 0.0262640520     0.0516207088 0.0002016434
##   >50K  0.1243217364 0.0360676668     0.0403766358 0.0000000000
##
##        marital_status
## Y           Divorced Married-AF-spouse Married-civ-spouse
##   <=50K 0.1578867772      0.0006049302       0.3351313203
##   >50K  0.0571337376      0.0011171401       0.8531758698
##        marital_status
## Y       Married-spouse-absent Never-married    Separated      Widowed
##   <=50K          0.0160306498  0.4141755306 0.0384130665 0.0377577255
##   >50K           0.0044685605  0.0643153527 0.0084583466 0.0113309927
##
##        fnlwgt
## Y           [,1]     [,2]
##   <=50K 190141.8 106593.1
##   >50K  188371.8 103027.8
```





4.3.3 Summary of Naive Bayes

For the results of both models, the prediction results are not bad, and both of the AUCs are basicly same as the previous `CART` models.

We know that `Naive Bayes` algorithm is based on the Bayes theory with strong independence assumptions between different predictors. In this case, we have tested the correlationship between different numerical variable. However, we did not test that between categorical variables or categorical and numerical variables. This may result in not well model performances.

4.4 Bagging

4.4.1 Model with full dataset

Firstly, we would like to build a `Bagging` model by using training dataset obtained from original dataset (without removing outliers). The training and testing ROC curves and AUCs are also shown below.

Unlike the previous three models, we only include very brief model information here. Because in `Bagging`, we will build many models, and then take the majority vote because this is a classification model.

We would like to tune our parameter `nbagg` here. `10`, `15`, `20` will be considered as our potential choices here.

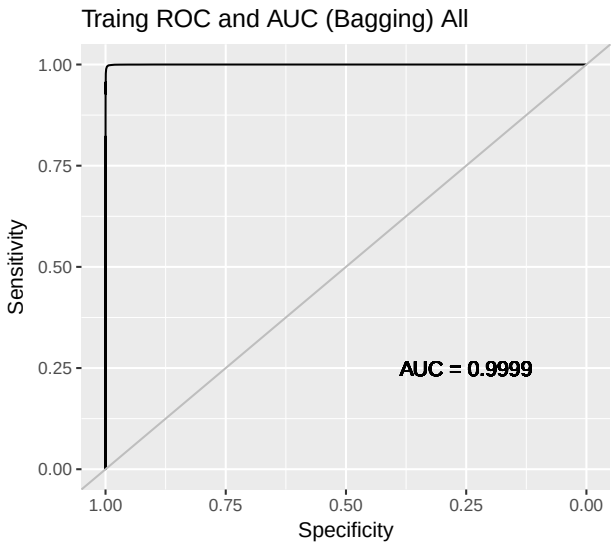
	Training AUC
nbagg=15	0.999576
nbagg=20	0.999808
nbagg=25	0.999901

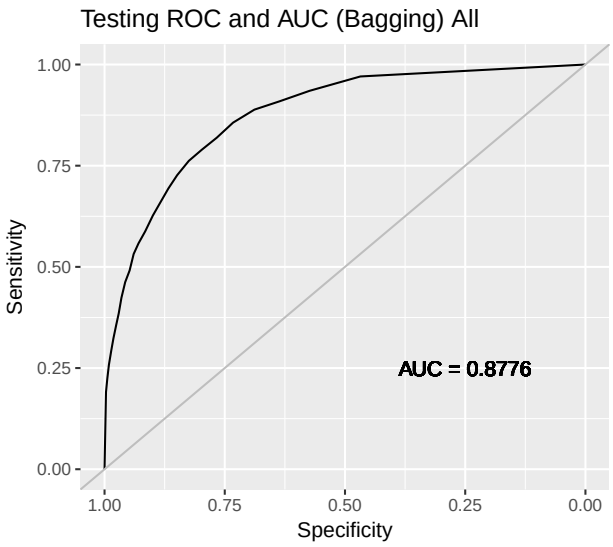
	Testing AUC
nbagg=15	0.872218
nbagg=20	0.875816
nbagg=25	0.877591

The left and right tables stand for training and testing AUC, respectively. We notice that as `nbagg` going higher, the training AUC is keeping going up towards 1. One the right side, the testing AUC first goes up, too. We should take the result seriously, because the training AUCs are almost perfect here.

Hence, we would like to choose `nbagg = 25` as our final parameters since it has the highest testing AUC. The detailed information about this model will be listed below.

```
##
## Bagging classification trees with 25 bootstrap replications
##
## Call: bagging.data.frame(formula = salary ~ relationship + capital_gain +
##     age + education + occupation + marital_status + fnlwgt, data = Train_Adult,
##     nbagg = 25)
```





4.4.2 Model with subset (Without outliers)

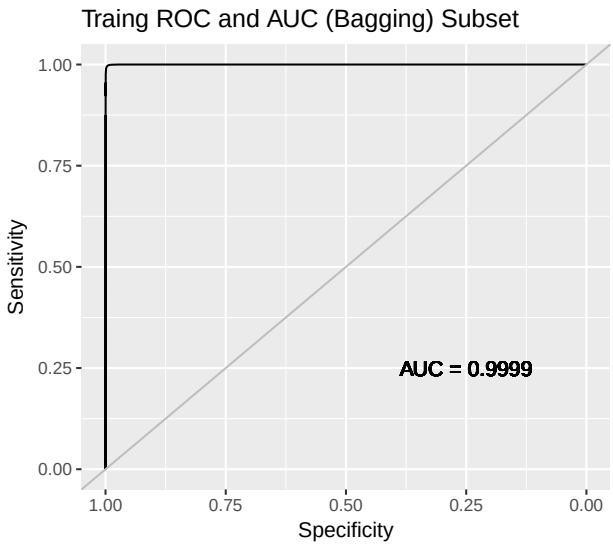
Then, we would like to build a `Bagging` model by using training dataset obtained from new dataset (after removing outliers). The brief model information, training and testing ROC curves and AUCs are also shown below.

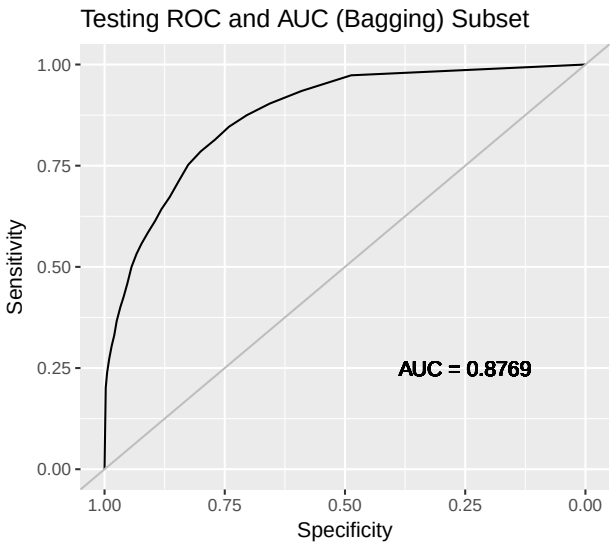
	Training AUC
nbagg=15	0.999632
nbagg=20	0.999858
nbagg=25	0.999928

	Testing AUC
nbagg=15	0.866431
nbagg=20	0.872501
nbagg=25	0.876916

The left and right tables stand for training and testing AUC, respectively, like the previous part. We notice that as `nbagg` going higher, the training AUC is keeping going up towards 1. One the right side, the testing AUC first goes up, too.

Hence, we would like to choose `nbagg = 25` as our final parameters since it has the highest testing AUC. The detailed information about this model will be listed below.





4.4.3 Summary of Bagging

In this case, there is one serious problem which we must take a look at. This both `Bagging` models give us over-fitted results. As we know that, when we increase the complexity of models, our traing metric will go better towards being the best while our testing metric will go down first and then go back worse, which is defined as over-fitting. Over-fitting is kind of a general common result if we choose to apply the complicated tree models. In this case, we can choose to lower our model complexity manually to deal with problem.

Also, more tuning parameters like `nbagg=30`, `nbagg=35` or more that we did not attempt may still be tuned and helped us find out the best model.

4.5 Random Forest

4.5.1 Model with full dataset

Firstly, we would like to build a `Random Forest` model by using training dataset obtained from original dataset (without removing outliers). The training and testing ROC curves and AUCs are also shown below.

Like `Bagging` model, brief summary including basic model information, `ntree`, confusion matrix (cutoff = 0.5), etc will also be listed below.

- Tuning Parameters

Here, we would like to tune the parameters `ntree` and `mtry` here. We will try 500, 1000, and 1500 trees, and 2, 3, 4 numbers of variables at each split. The outputs are listed below. We will find out our best tuning parameters based on their performances.

- The `left` and `right` tables stand for training and testing AUC, respectively.

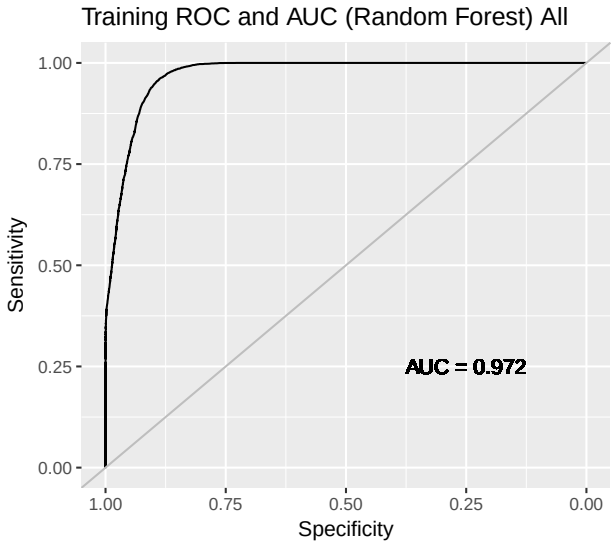
	mtry=2	mtry=3	mtry=4
ntree=500	0.971562	0.994237	0.999886
ntree=1000	0.971997	0.994329	0.999885
ntree=1500	0.972038	0.994474	0.999889

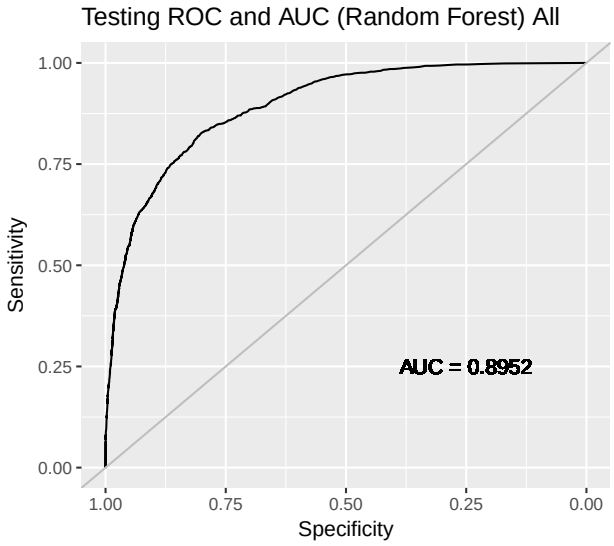
	mtry=2	mtry=3	mtry=4
ntree=500	0.894862	0.893944	0.891967
ntree=1000	0.895195	0.895016	0.892094
ntree=1500	0.895118	0.895056	0.892386

We notice that as `ntree` and `mtry` going higher, the training AUC is keeping going up towards 1. One the right side, the testing AUC first goes up and then drop back down. This results tell us that we need to choose our model parameters wisely, or over-fitting will become a serious problem.

Hence, we would like to choose `mtry=2` and `ntree=1000` as our final parameters based on the tables above. The detailed information about this model will be listed below.

```
##
## Call:
##  randomForest(formula = salary ~ relationship + capital_gain +      age + education + occupation + ma
##                Type of random forest: classification
##                      Number of trees: 1000
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 14.54%
## Confusion matrix:
##       <=50K >50K class.error
## <=50K 18478 1253  0.06350413
## >50K   2533 3782  0.40110847
```





4.5.2 Model with subset (Without outliers)

Then, we would like to build a `Random Forest` model by using training dataset obtained from the dataset without outliers. The brief summary information, training and testing ROC curves and AUCs are also shown below.

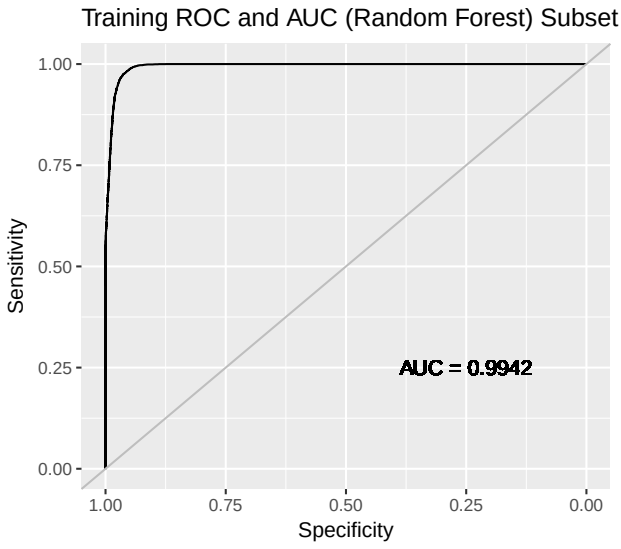
Like the previous part, we will also tune `ntree` and `mtry` here. The choices of these tuning parameters will be set as the same values as previous part.

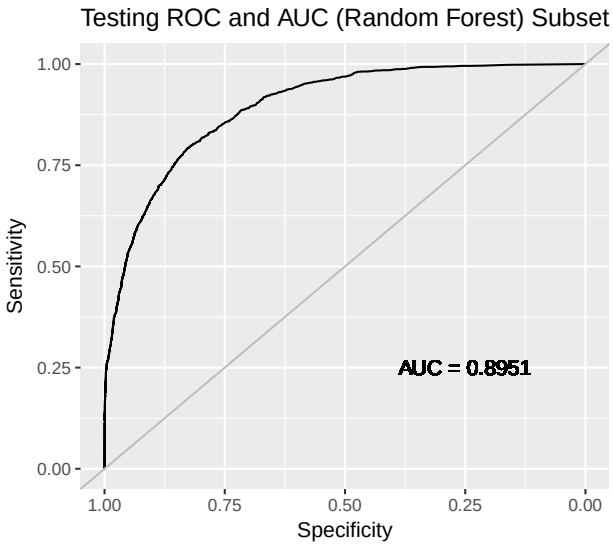
- The `left` and `right` tables stand for training and testing AUC, respectively.

	mtry=2	mtry=3	mtry=4
ntree=500	0.971634	0.994137	0.999861
ntree=1000	0.971514	0.994205	0.999865
ntree=1500	0.971431	0.994202	0.999864

	mtry=2	mtry=3	mtry=4
ntree=500	0.893363	0.894815	0.891098
ntree=1000	0.893966	0.895008	0.891327
ntree=1500	0.894309	0.895068	0.8914

In this case, we notice that potential overfitting still exists like the previous part. Hence, we would like to select `ntree=1500`, and `mtry=3` based on the testing AUC. The detailed result will be shown below.





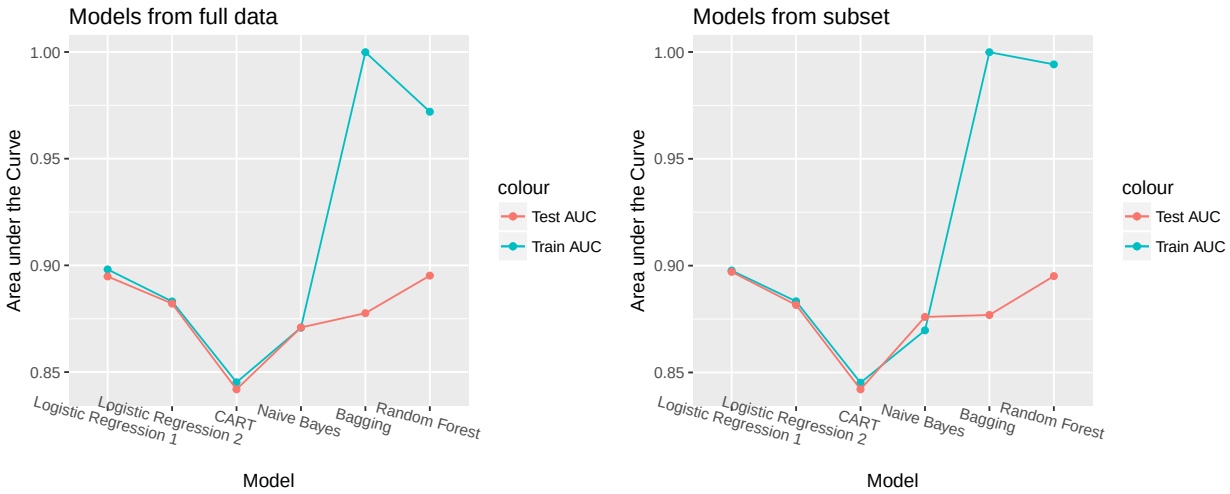
4.5.3 Summary of Random Forest

For complicated tree models, over-fitting is one important issue that we cannot ignore. In this case, we can read that over-fitting issue still exists, but luckily it was not as serious as the model of `bagging`. Like what we mentioned in the previous part, we can choose to lower the complexity of our models manually in order to solve this issue.

Also, more tuning parameters that we did not attempt may still be tuned and helped us find out the best model.

5 Comparison and Conclusion

5.1 Comparison



We noticed that in the the results in both plots, the performances of `CART` are the worst. Because both training and testing AUC are the lowest among different models, respectively. For the `Bagging` model, we have an excellent training AUC but a normal testing AUC, which shows us a very serious overfitting problem. So we decided to think the performance of `Random Forest` is the most satisfactory among all. Because its training AUC is better the rest models except `Bagging`, and its testing AUC is the second best among all five.

5.2 Conclusion

In `5.1` part, we have already compared the AUCs of different methods. The final choice of model should depend on the background and business need of this project.

- Situation 1

If we would like to understand correlationship among predictors and response quantitively and/or make business suggestions based on our model results, `Logistic Regression 1` should be considered as our best choice here. Because we can interpret our results quantitively based on Logistic Regression algorithm, and its performance here is the best among `Logistic Regression`, `CART` and `Naive Bayes` algorithms.

- Situation 2

If we would like to build a black box which can make almost perfect predictions, `Random Forest` should be considered as our best choice here. Because of its most satisfactory overall performance which we discussed in the previous part.

5.3 Further thoughts

5.3.1 Cross-validation

Cross-validation can be applied here to help us build statistical models. Also, it is a very useful tool to help us prevent or lower the potential overfitting issues.

5.3.2 Outliers

In this case, for each algorithm, we chose to build at least one model based on the dataset with and without outliers. The results that we get are basically the same as each other within different methods.

In the second dataset (the one without outliers), we may try to downweight those outliers instead of deleting those points aggressively. Because not all outliers are resulted from data entry error, and some important information may and will be thrown away if we choose to delete those points.

5.3.3 Variable fnlwgt

From the description (URL: https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.names) of the Adult dataset, we can have access to the brief defination of variable `fnlwgt`. After reading it, we can know that this variable is created by considering information including but not limited to demographic characters and different states, which result in potential multicollinearity among predictors, such as `race`, `sex`, `fnlwgt`, etc. For some particular methods, like `Logistic Regression` and `Naive Bayes`, multicollinearity may cause serious problems and make our predictons less trustable.

Hence, variable `fnlwgt` or the transformation of `fnlwgt` being treated as the weight rather than one predictor here may help increase the performance of our new models and make our models more persuasive.