Generalized Linear Model

Group # 01

PROJECT: A Predictive Model for Income

Ricardo Castañeda r0731529

Qianli Fan r0775346

Butynets Mariia r0771332

Lieven Govaerts q0152493

Meng Wang r0767603

ZHANG Yanyi r0731121

Kendall Brown r0773111

Ruiman Zhong r0767577

TABLE OF CONTENTS

1.EXPLORATION OF THE DATA

2. CONSTRUCT A LOGISTIC REGRESSION MODEL WITH TRAIN DATA

2.1 Model construction and interpretation

2.2 Goodness of fit tests of the logistic model

2.2.1 Hosmer-Lemeshow Test

2.2.2 Residual Analysis

2.2.3 Dispersion Test

2.2.4 Complete Separation Test

3. MODEL IMPROVEMENT

3.1 Restructure Workclass Variable

3.2 Logistic Regression Based on Test Data

3.2.1 Likelihood Ratio Test on Nested Model

3.2.2 Interpretation

3.3 Model Checking

3.3.1 Hosmer-Lemeshow Test

3.3.2 Complete Separation Test

4. ASSESS THE PERFORMANCE OF THE LOGISTIC MODEL

4.1 Model Assessment of the logistic model with the train data

4.1.1 Receiver Operating Characteristics Curve

4.1.2 Accuracy, Sensitivity and Specificity

4.1.3 Confusion Matrix

4.2 Model Assessment of the logistic model with the test data

4.2.1 Receiver Operating Characteristics Curve

4.2.2 Accuracy, Sensitivity and Specificity

4.2.3 Confusion Matrix

5. ALTERNATIVE SOLUTION

6. CONCLUSION

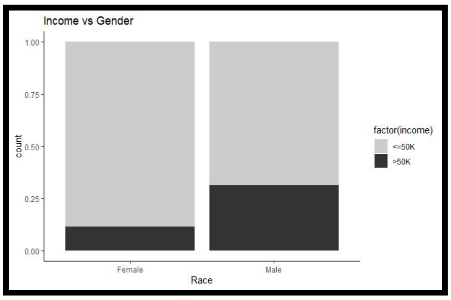
Problem Description: The income dataset contains information on 45743 subjects and 8 variables. The goal is to predict whether the annual income of an individual will exceed $50.000. The data contains the following variables:

- age: age of the individual. Numeric
- edu: Educational level of the individual. Factor.
- mari.sta: Marital status of the individual. Factor.
- gender: Gender of the individual. Factor, i.e. Male or Female
- workclass: Type employment
- hours: number of hours worked per week
- race: The race of the individual
- income: Target variable. Income above or below 50K. Factor i.e. >50K, <=50K

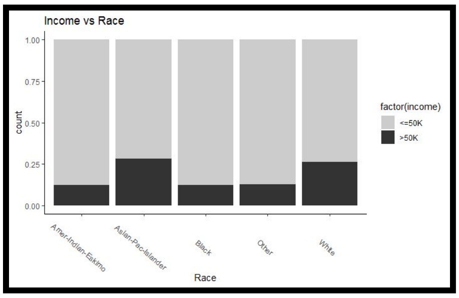
Data preparation. Two continuous covariates (age and hours) from the data set were standardized. The reason is that we want to avoid multicollinearity between the main effect of the covariates and their interactions. The original covariates were deleted and not used in the analysis.

1. EXPLORATION OF THE DATA

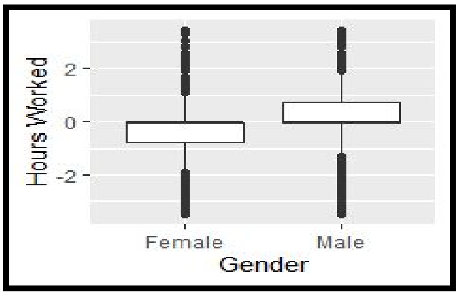
Income versus Gender. We would like to investigate the relationship between gender and income. The dataset contains almost two times more males than females, 30 866 and 14 877 respectively. Among all females, only 11 % have income more than 50 K (1 686), while among males 31 % earn more than 50 K (9 638). The graph below illustrates this. Visually, it seems that gender has an impact on the level of income.



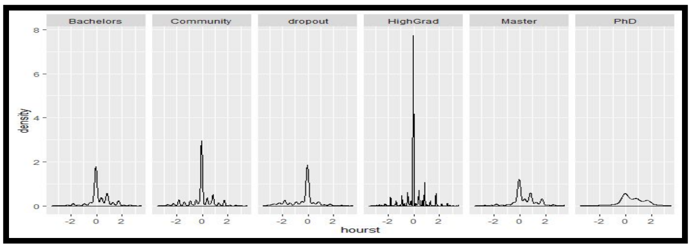
Income versus Race. Now we consider the relationship between income and race. Almost 13 % of Black, Amer-Indian-Eskimo, and other races have an income level of more than 50K. In contrast, 26 % of white people and 28% of Asian-Pac-Islander people have an income level of more than 50 K. The barplot illustrates these differences. Based on visual exploration, there is not much evidence that race has an impact on the level of income: three of the five groups have almost the same proportion of people in both income categories.



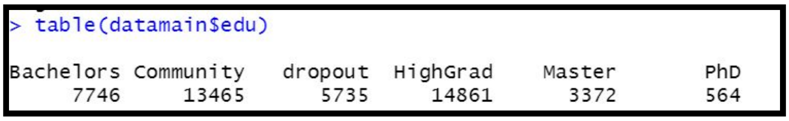
Gender versus working time. Now we investigate the relationship between gender and hours worked. Here we use standardized working hours. On the boxplots, we see the difference in working hours for females and males. It seems that males work more compared to females.



Working hours versus education. Now we explore the relationship between working hours (standardized) and type of education. Below on the graph, we illustrate the densities of working hours for each type of education separately. We notice that the distributions of hours differ for educational groups (assuming the difference in the means of hours for each educational group). This indicates that the working hours might be affected by education.



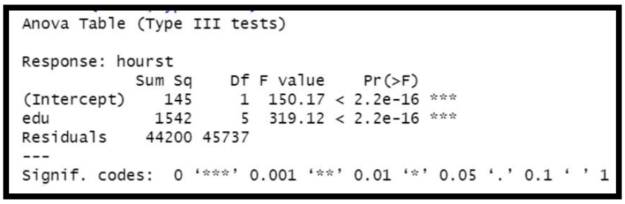
ANOVA analysis of the relationship between working hours and education. Since we noticed the difference in hours across the different types of education, now we apply ANOVA analysis to detect whether education has a significant effect on the level of income. The factor effect model with the Type III Sum of Squares is used for the analysis (with contrast sum in R) because the study is unbalanced (with a different number of people for each level of education, as it is shown below in the table) and we do not take into account the proportion of people in each group.



```
> table(datamain$edu)

Bachelors Community   dropout  HighGrad    Master       PhD
     7746     13465      5735     14861      3372       564
```

The results of the ANOVA analysis below indicate that education has a significant effect on the mean level of income (with p-values less than 0.05, at a 5 % significance level). In other words, there is a significant difference between the means level of income for different educational groups.



```
Anova Table (Type III tests)

Response: hourst
            Sum Sq    Df F value    Pr(>F)    
(Intercept)    145     1  150.17 < 2.2e-16 ***
edu           1542     5  319.12 < 2.2e-16 ***
Residuals    44200 45737                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

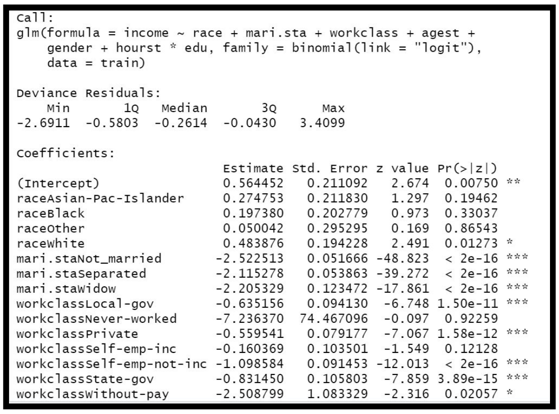
2. CONSTRUCT A LOGISTIC REGRESSION MODEL WITH TRAIN DATA

2.1 Model construction and interpretation

Firstly, we divide our data into two data sets, one data set is train data (containing 80% of the data) and the other one is test data(containing 20% of the data). By constructing them, we can check our model and its predictive ability better.

In our data, income is a binary variable. And logistics regression is always built to predict binary responses. Therefore, we try to explore the relationships between income and other variables by constructing logistic regression. We set variables race, edu, mari.sta, workclass, agest, gender, hourst and hourst*edu(based on the ANOVA analysis above) as independent variables.

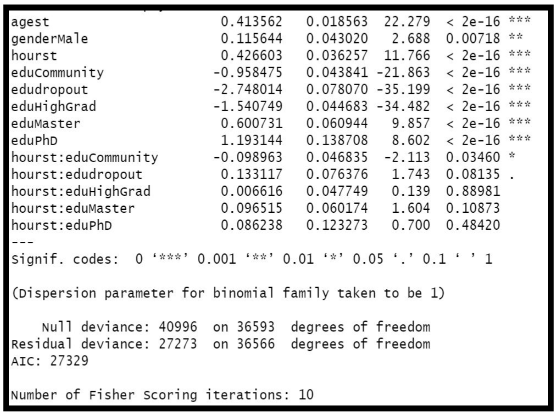
The theoretical model can be expressed as: *logit[P(income|X)] = β*X*
Based on the train data, the Logistic model was fitted, and the resulting summary statistics are found below.



```
Call:
glm(formula = income ~ race + mari.sta + workclass + agest +
    gender + hourst * edu, family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6911  -0.5803  -0.2614  -0.0430   3.4099

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)              0.564452   0.211092   2.674  0.00750 **
raceAsian-Pac-Islander   0.274753   0.211830   1.297  0.19462
raceBlack                0.197380   0.202779   0.973  0.33037
raceOther                0.050042   0.295295   0.169  0.86543
raceWhite                0.483876   0.194228   2.491  0.01273 *
mari.staNot_married     -2.522513   0.051666 -48.823  < 2e-16 ***
mari.staSeparated       -2.115278   0.053863 -39.272  < 2e-16 ***
mari.staWidow           -2.205329   0.123472 -17.861  < 2e-16 ***
workclassLocal-gov      -0.635156   0.094130  -6.748 1.50e-11 ***
workclassNever-worked   -7.236370  74.467096  -0.097  0.92259
workclassPrivate        -0.559541   0.079177  -7.067 1.58e-12 ***
workclassSelf-emp-inc   -0.160369   0.103501  -1.549  0.12128
workclassSelf-emp-not-inc -1.098584   0.091453 -12.013  < 2e-16 ***
workclassState-gov      -0.831450   0.105803  -7.859 3.89e-15 ***
workclassWithout-pay    -2.508799   1.083329  -2.316  0.02057 *
```



```
agest                    0.413562   0.018563  22.279  < 2e-16 ***
genderMale               0.115644   0.043020   2.688  0.00718 **
hourst                   0.426603   0.036257  11.766  < 2e-16 ***
eduCommunity            -0.958475   0.043841 -21.863  < 2e-16 ***
edudropout              -2.748014   0.078070 -35.199  < 2e-16 ***
eduHighGrad             -1.540749   0.044683 -34.482  < 2e-16 ***
eduMaster                0.600731   0.060944   9.857  < 2e-16 ***
eduPhD                   1.193144   0.138708   8.602  < 2e-16 ***
hourst:eduCommunity     -0.098963   0.046835  -2.113  0.03460 *
hourst:edudropout        0.133117   0.076376   1.743  0.08135 .
hourst:eduHighGrad       0.006616   0.047749   0.139  0.88981
hourst:eduMaster         0.096515   0.060174   1.604  0.10873
hourst:eduPhD            0.086238   0.123273   0.700  0.48420
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 40996  on 36593  degrees of freedom
Residual deviance: 27273  on 36566  degrees of freedom
AIC: 27329

Number of Fisher Scoring iterations: 10
```

From this table the intercept and regression coefficient are obtained. Using these results the fitted model may be expressed as:

Logit[P(Income>=50K|X)]=0.56+0.27*Asian+0.20*Black+0.05*Other+0.48*White-2.52*not_married-2.11*separated-2.20*widow-0.6*local-gov-0.83*state-gov-2.51*without-pay-7.2*never-worked-0.55*private-0.16*selfemp-inc-1.10*selfemp-not-inc+0.41*agest+0.13*male+0.42*hourst-0.96*community-2.7*droupout-1.54*highgrad+0.57*master+1.11*PhD-0.11hourst*community+0.13*hourst*dropout-0.01*hourst*Highgrad+0.10*hourst*Master+0.09hourst*PhD

Each level of education, marriage status, age, working hours, and gender is significant.

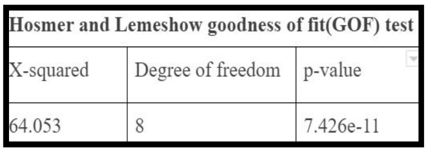
Based on the result, higher degree may help one get higher income; race and workclass has no significant effect on income; the probability of getting higher income than 50K are highly related with age and hours one worked, which are older people are more likely to earn more and the one who works longer has more chance to get higher income than 50K.

2.2 Goodness of fit tests of the logistic model

2.2.1 Hosmer-Lemeshow Test

After the model is constructed, it is necessary to check the goodness-of-fit of this model. As there are continuous variables in the model, the Hosmer-Lemeshow test should be used.

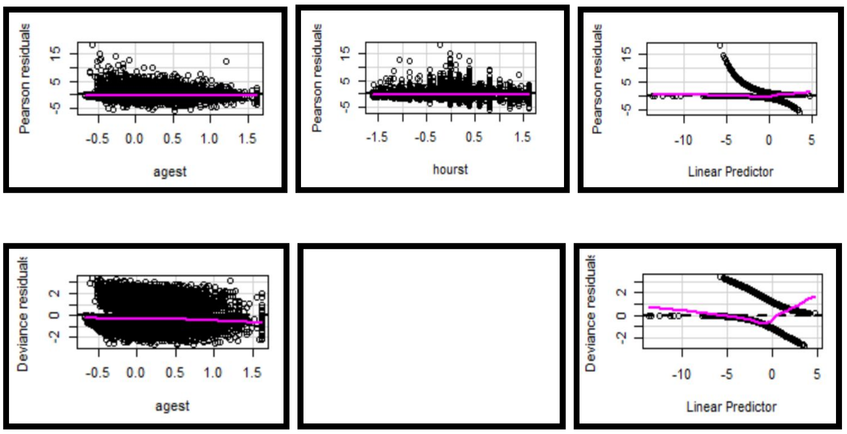
depict that the GOF test rejects the model. There is a large sample size and the GOF test is sensitive to the sample size.



Hosmer and Lemeshow goodness of fit(GOF) test		
X-squared	Degree of freedom	p-value
64.053	8	7.426e-11

2.2.2 Residual Analysis

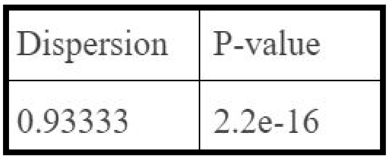
After the Hosmer-Lemeshow test, the residual analysis is performed. With the R program, the Pearson residual plots and Deviance residual plots are obtained as follow:



2.2.3 Dispersion Test

Then the dispersion test is performed and it showed the following results:

The result indicates that the dispersion is equal to 0.93333 and the p-value is almost equal to 0, which means that the underdispersion seems to exist. However, the dispersion value is somehow acceptable, so there is no need to worry about it too much.



Dispersion	P-value
0.93333	2.2e-16

2.2.4 Complete Separation Test

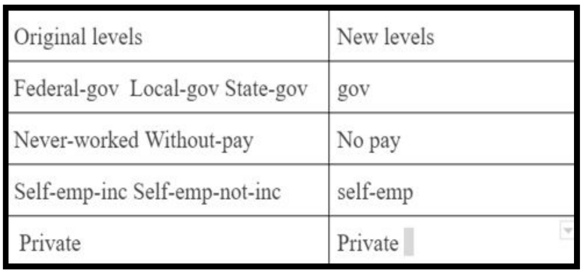
Besides, the complete separation is also tested. The results indicate that the intercept term and workclass terms have an increasing trend, while other terms converge to constants. This means that the intercept and workclass may have infinite maximum likelihood estimates.

3. MODEL IMPROVEMENT

3.1 Restructure Workclass Variable

As the Firth procedure failed to converge and solve the complete separation, we turn to restructure “workclass” from their meaning.

We combine level ‘Federal-gov’, ‘Local-gov’ and ‘State-gov’ to level ‘gov’;‘never-worked’ with ‘without pay’ to ‘gov’; ‘self-emp-inc’ and ‘self-emp-not-inc’ to level ‘self-emp’; ‘private’ to ‘private’. We also noticed that the income of people whose work class is never-worked is 100% less than 50K. This definitely causes complete separation.

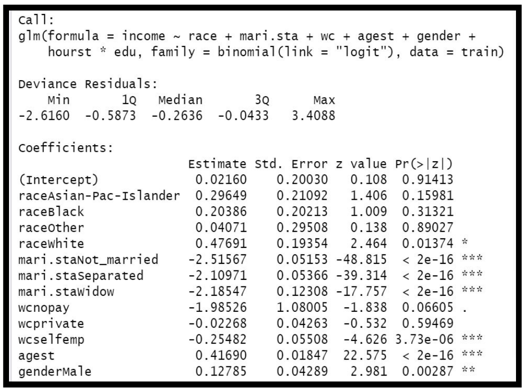


Original levels	New levels
Federal-gov Local-gov State-gov	gov
Never-worked Without-pay	No pay
Self-emp-inc Self-emp-not-inc	self-emp
Private	Private

3.2 Logistic Regression Based on Test Data

3.2.1 Likelihood Ratio Test on Nested Model

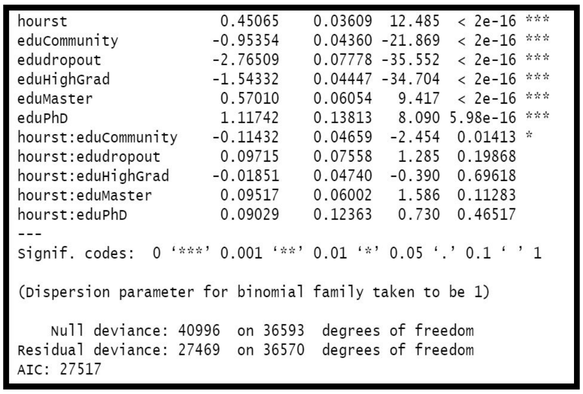
The outputs of the model are as follows. All levels of race are not significant. We simplify the model by deleting it and use the likelihood ratio test to compare.



```
Call:
glm(formula = income ~ race + mari.sta + wc + agest + gender + 
    hourst * edu, family = binomial(link = "logit"), data = train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.6160  -0.5873  -0.2636  -0.0433   3.4088  

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)    
(Intercept)             0.02160    0.20030   0.108  0.91413    
raceAsian-Pac-Islander  0.29649    0.21092   1.406  0.15981    
raceBlack               0.20386    0.20213   1.009  0.31321    
raceOther               0.04071    0.29508   0.138  0.89027    
raceWhite               0.47691    0.19354   2.464  0.01374 *  
mari.staNot_married    -2.51567    0.05153 -48.815  < 2e-16 ***
mari.staSeparated      -2.10971    0.05366 -39.314  < 2e-16 ***
mari.staWidow          -2.18547    0.12308 -17.757  < 2e-16 ***
wcnopay                -1.98526    1.08005  -1.838  0.06605 .  
wcprivate              -0.02268    0.04263  -0.532  0.59469    
wcselfemp              -0.25482    0.05508  -4.626 3.73e-06 ***
agest                   0.41690    0.01847  22.575  < 2e-16 ***
genderMale              0.12785    0.04289   2.981  0.00287 ** 
```

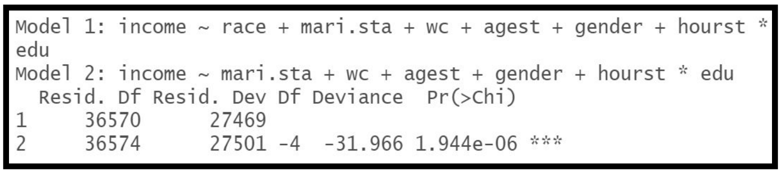


```
hourst                  0.45065    0.03609  12.485  < 2e-16 ***
eduCommunity           -0.95354    0.04360 -21.869  < 2e-16 ***
edudropout             -2.76509    0.07778 -35.552  < 2e-16 ***
eduHighGrad            -1.54332    0.04447 -34.704  < 2e-16 ***
eduMaster               0.57010    0.06054   9.417  < 2e-16 ***
eduPhD                  1.11742    0.13813   8.090 5.98e-16 ***
hourst:eduCommunity    -0.11432    0.04659  -2.454  0.01413 *  
hourst:edudropout       0.09715    0.07558   1.285  0.19868    
hourst:eduHighGrad     -0.01851    0.04740  -0.390  0.69618    
hourst:eduMaster        0.09517    0.06002   1.586  0.11283    
hourst:eduPhD           0.09029    0.12363   0.730  0.46517    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 40996  on 36593  degrees of freedom
Residual deviance: 27469  on 36570  degrees of freedom
AIC: 27517
```

The result of the likelihood ratio test is as follows:



```
Model 1: income ~ race + mari.sta + wc + agest + gender + hourst * 
edu
Model 2: income ~ mari.sta + wc + agest + gender + hourst * edu
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)    
1     36570      27469                           
2     36574      27501 -4  -31.966 1.944e-06 ***
```

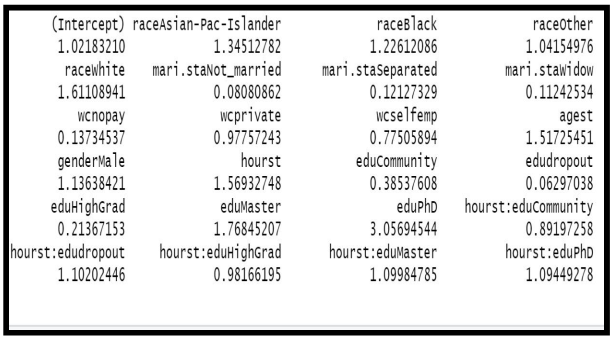
The two models are different statistically. The result of the likelihood ratio test shows that we can not delete the variable ‘race’.

3.2.2 Interpretation

Final model:

Logit[P(Income>=50K|X)]=0.02+0.29*Asian+0.20*Black+0.04*Other+0.48*White-2.5*1not_married-2.11*separated-2.19*widow-1.99*nopay-0.02*private-0.25*selfemp+0.41*agest+0.13*male+0.45*hourst-0.95*community-2.7*droupout-1.54*highgrad+0.57*master+1.11*PhD-0.11*hourst*community+0.097*hourst*dropout-0.01*hourst*Highgrad+0.09*hourst*Master+0.09*hourst*PhD

The odds ratios are as follows.



```
      (Intercept) raceAsian-Pac-Islander              raceBlack              raceOther
       1.02183210             1.34512782             1.22612086             1.04154976
        raceWhite    mari.staNot_married      mari.staSeparated          mari.staWidow
       1.61108941             0.08080862             0.12127329             0.11242534
          wcnopay              wcprivate              wcselfemp                  agest
       0.13734537             0.97757243             0.77505894             1.51725451
       genderMale                 hourst           eduCommunity             edudropout
       1.13638421             1.56932748             0.38537608             0.06297038
      eduHighGrad              eduMaster                 eduPhD    hourst:eduCommunity
       0.21367153             1.76845207             3.05694544             0.89197258
hourst:edudropout     hourst:eduHighGrad       hourst:eduMaster         hourst:eduPhD
       1.10202446             0.98166195             1.09984785             1.09449278
```

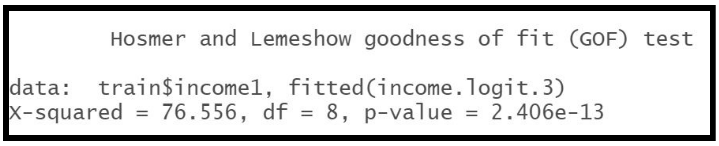
Generally speaking, each level of education, marriage status, age, working hours, and gender is significant. Interaction between education and work worth watching. For example, the odds of the income being higher than 50K of Ph.D. is 3 times higher than that of bachelors. The odds of the income being higher than 50K of employers dropped out of school is 0.06 times higher than that of bachelors.

It also implies that an integrated marriage benefits income. The odds of the income being higher than 50K of married people is 7 times higher than that of other marital status. As for gender, the odds of the income being higher than 50K of the male is 1.13 times higher than that of the female. Part of levels in workclass and race are significant. For example, The odds of the income being higher than 50K of working in government is 1.28 times higher than that of self-employers. And The odds of the income being higher than 50K of white people is 1.61 times higher than that of indian and eskimo employers.

The probability of getting income being higher than 50K rises with age and working hours per week. An increase of 13 year in age is associated with a 50% increase in the odds of getting a high income. However, the increase of odds in working hours per week depends on the education level. For example, an increase of 11 working hours is associated with a 50% increase in the odds of getting a high income in bachelor level, while it is only a 40% increase in Community level.

3.3 Model Checking

3.3.1 Hosmer-Lemeshow Test



```
	Hosmer and Lemeshow goodness of fit (GOF) test

data:  train$income1, fitted(income.logit.3)
X-squared = 76.556, df = 8, p-value = 2.406e-13
```

The Hosmer-Lemeshow test rejects the new model. But since the sample size of the data is too large, the GOF test is sensitive to sample size. So, it always rejects the model. So, we are not worried about the result and choose other criteria for model evaluation.

3.3.2 Complete Separation Test

All levels of Work class started at 1 and after 7 steps, the results the results remained constant (W1 = 2.2023, Wc2 =2.2387, and Wc3 = 2.2264). Therefore, it can be said that the data do not present complete separation issues.

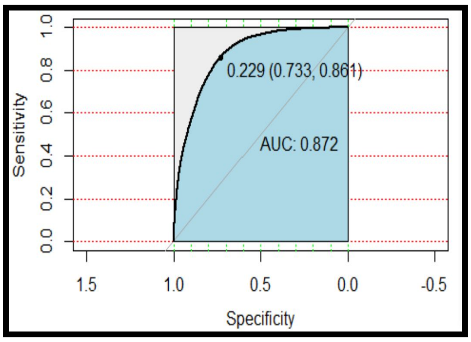
4. ASSESS THE PERFORMANCE OF THE LOGISTIC MODEL

Based on all the analysis and improvement we have done above; we get our model. In this part, the performance of this logistic model will be assessed by applying ROC, sensitivity and specificity etc. for both train data and test data.

4.1 Model assessment of the logistic model with train data

4.1.1 Receiver Operating Characteristic Curve

A receiver operating characteristic curve, or ROC curve, is created by plotting the true positive rate(TPR or sensitivity) against the false positive rate(FPR or 1-specificity) as its discrimination threshold is varied. Basically, it is usually used to illustrate the diagnostic ability of a binary classifier system. Besides, ROC curve can be summarized by the area underneath it (area under ROC curve, namely AUC). In short, ROC is a probability curve and AUC represents degree or measure of separability. It tells how much a model is capable of distinguishing between classes. Normally, the value of AUC is between 0.5 and 1. The higher, the better. In our case, the higher the AUC, the better our model is at distinguishing between income with ">50K" and "<=50K". Note that we denote ">50K" as "1" (positive) and "<=50K" as "0" (negative) in all the calculations below.

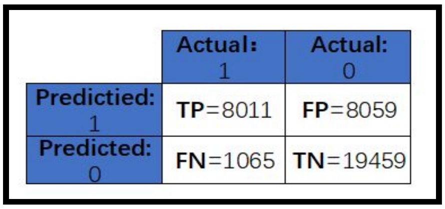


The graph on the left is plotted based using the pROC package in R. It's very clear to see that the AUC of our improved logistic model is 0.872, which means that there is 87.2% chance that our model will be able to distinguish income between positive class and negative class. Or in other words, the probability that a randomly chosen positive instance gets a higher score than a randomly chosen negative instance is 87.2%.

By maximizing the polygon of AUC, an optimal threshold value (0.229) and its corresponding sensitivity and specificity can be calculated (see graph above). Therefore, we set our cutoff value as 0.2 to construct a contingency table later on.

4.1.2 Confusion Matrix

By setting threshold is equal to 0.2, our confusion matrix showing predicted income versus real income is as follows:



	Actual: 1	Actual: 0
Predictied: 1	TP=8011	FP=8059
Predicted: 0	FN=1065	TN=19459

In the table, TP and FP are short for True Positive and False Positive. FN and TN are False Negative and True Negative correspondingly.

4.1.3 Accuracy, Sensitivity, and Specificity

Accuracy = (TP+TN)/(TP+TN+FP+FN) = (8011+19459)/ (8011+8059+1065+19459) = 0.7507

Accuracy is used as a statistical measure of how well a binary classification test correctly identifies a condition. That is, 75.07% of true results including True Positives (income >= 50K) and True Negatives (income < 50K) among the total number of cases are examined by our model.

Sensitivity = TP/(TP+FN) =8011/(8011+1065) = 0.8827

Sensitivity refers to the test's ability of our model to correctly detect instances with income (>=50K) which do have the condition. Therefore, we could see 88.27% of the actual income positive results have been correctly examined by the logistic model.

Specificity = TN/(TN+FP) = 19459/ (19459+8059) = 0.7071

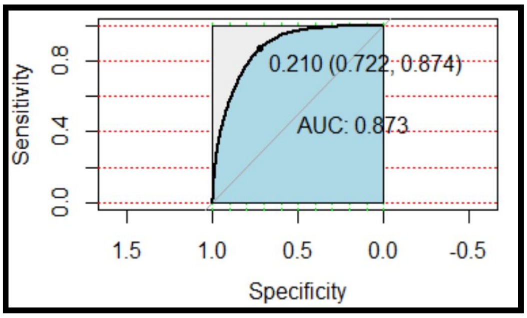
Basically, Specificity measures a test's ability to correctly generate a negative result for people who don't have the condition. Here in our model, the proportion of true negative income results (<50K) being examined correctly is 70.71%.

4.2 Model assessment of the logistic model with test data

Similarly, the same performance assessment of our logistic model can be done with the test data to see if the model is good enough to generalize.

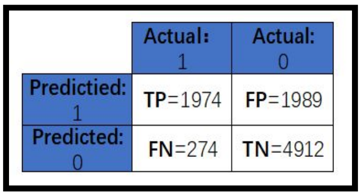
4.2.1 Receiver Operating Characteristic Curve

The AUC on test data is 0.873, which means that there is 87.3% chance that our model will be able to distinguish income between positive class and negative class. And the optimal threshold value here is 0.210, very similar to that based on the train dataset. Hence, we still take the cutoff value as 0.2 to calculate the confusion matrix.



4.2.2 Confusion Matrix

Again, by setting threshold is equal to 0.2, the contingency table showing predicted income versus real income based on test data is as follows:



	Actual: 1	Actual: 0
Predictied: 1	**TP**=1974	**FP**=1989
Predicted: 0	**FN**=274	**TN**=4912

4.2.3 Accuracy, Sensitivity, and Specificity

Accuracy = (TP+TN)/(TP+TN+FP+FN) = (1974+4912)/(1974+4912+1989+274) = 0.7527

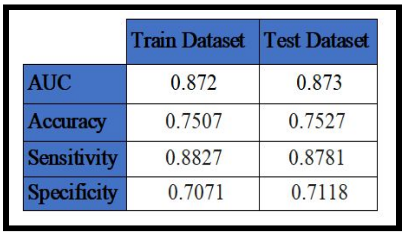
That is, in terms of test data, 75.27% of true results including True Positives (income >= 50K) and True Negatives (income < 50K) among the total number of cases are examined by our model.

Sensitivity = TP/(TP+FN) =1974/ (1974+274) = 0.8781
So, 87.81% of the actual income positive results have been correctly examined by the logistic model.

Specificity = TN/(TN+FP) = 4912/ (4912+1989) = 0.7118
Namely, the proportion of true negative income results (<50K) being examined correctly is 71.18% for the test dataset.

4.3 Comparison of Both Results

From the left table, it's obvious to see that the model gets similar test results from both train data and test data regarding AUC, accuracy, sensitivity and specificity. And hence, we could say our model could be generalized and it does work well in this aspect.



	Train Dataset	Test Dataset
AUC	0.872	0.873
Accuracy	0.7507	0.7527
Sensitivity	0.8827	0.8781
Specificity	0.7071	0.7118

5. ALTERNATIVE SOLUTION

As an alternative to the logistic models presented in this report, two predictive support vector machines were built using both a gaussian and linear kernel. For the following assessments, the positive class were the individuals with an income larger than 50K. the Fitting the linear SVM and assessing its performance against the test data shows comparable results to the logistic models drafted earlier. In respect to accuracy, the SVM achieves a rating of .7512. This rating is on par with the logistic models as are the corresponding sensitivity and specificity ratings. The sensitivity sees a slight decrease to .8383, whereas the specificity of the model can be calculated to be .7200. These measurements imply that the linear SVM appears to be classifying in accordance with the logistic model. This trend is present in the gaussian model as well with the model achieving an accuracy rating of .7694, sensitivity of .8639, and specificity of .7386.

In terms of relative performance to the logistic model, the SVM performed well enough to be considered a viable alternative to the logistic model. When considering a gaussian kernel, the model does bear stronger performance, generally performing about two percent better than the logistic methods via trading off some sensitivity for specificity.

6. CONCLUSION

Initial analysis of the data set showed an issue with complete separation preventing regression coefficients from converging to finite estimates. After grouping similar work classes together this issue was overcome. Building a logistic model which considered this change resulted in a model with decent predictive power as shown by the AUC value of .873 when examining the test data. From this model building process, the claim of the existence of a relationship between an individual's income level and their general demographic data cannot be rejected. Two support vector machines were built to further examine this claim. The results of these predictive models support the conclusions made from the final logistic model presented in this report.