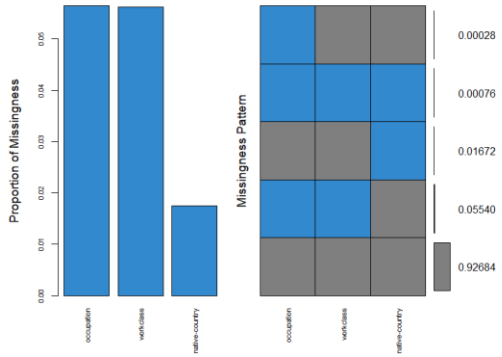
Gregory Powers

Gxp145@case.edu

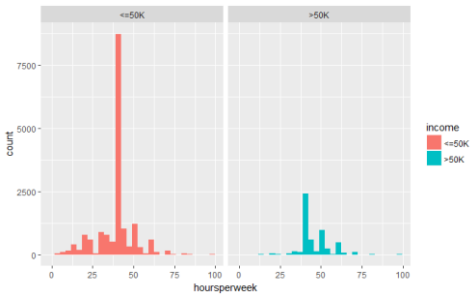
PQHS 471 Midterm



Exploratory Analysis. The data are drawn from the national census. The goal of this analysis was to accurately predict the variable income. Income has two classes: income less than $50,000 (LT50K), or greater than $50,000 (GT50K). Data were split into training and test. Five variables were numeric, the remainder character data that were, for ease of use, turned into factors. In the training set, 24% of those surveyed were in the GT50K category, making it infrequent. About 7% of cases are missing, all in three variables: native-country, occupation, and workclass. The missing values disproportionally fall into the LT50K category; multiple imputation was not considered an option because of the author's unfamiliarity with imputation of string/factor variables (though it was attempted using MICE, it was far too slow). As missing data introduces problems for several procedures, missing cases were deleted listwise. A column containing survey weights was also dropped.

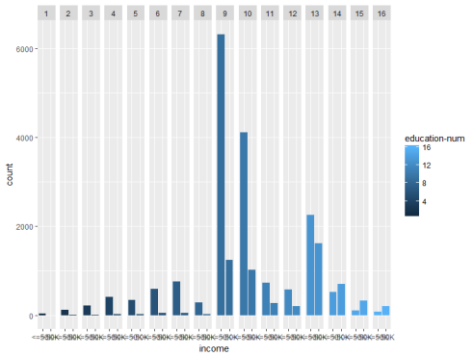
The relationship between variables was initially examined using the hetcor package. Hetor automatically selects the correct correlation method (R, polychoric, tetrachoric and polyserial) based on the level of measurement. Pearson's R will underestimate the association between variables where the data are not measured on an interval level due to restriction of range. None of the variables as initially coded were strongly correlated. Capital-gain, sex, marital-status, education-num (years of education), and age were all moderately correlated with income, race weakly.

Age was numeric that was approximately normally distributed. From the graph, older persons seem to make to make more money. This is confirmed using a t-test: the mean age of the GT50K group is about, on average, 8 years more than in the LT50K group. Also pictured is the distribution of gender by age. Though it is hard to tell from the graphic, female respondents are significantly younger as assessed by a t-test by, on average, over one year. This suggests that age may be a decent classifier of income.

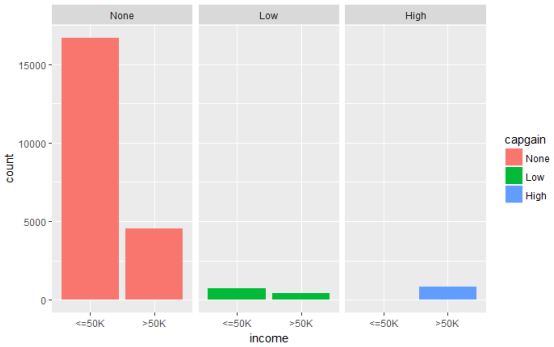


Hours-per-week was also numeric. Not surprisingly, those in the GT50K worked significantly more than those in the LT50K group by on average 6 hours per week. This is visualized in the attached bar chart. Hours-per-week may also be a decent predictor of income class.

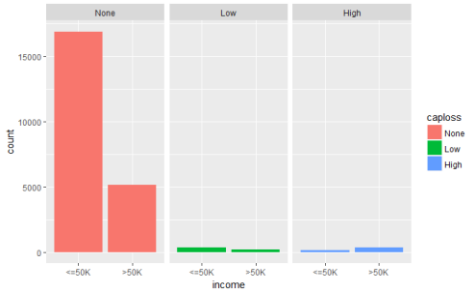
Education-num was the numeric version of the categorical "education" variable. Initially, the two variables weakly correlated; however once binned and ordered by level of education, education-num and education were almost perfectly correlated. Education-num was retained and education dropped from the analysis. There is a statistically significant relationship with income and education, with the GT50K group having on average two years more of education. Professional and white collar workers also on average had about 2.5 years more education than blue collar and service employees. There too was also differences by racial/ethnic group, albeit comparatively small. Education-num then, is may also perform well as a predictor of income.



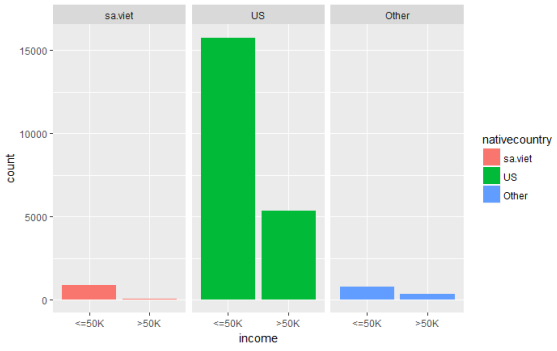
Capitol-gain is another numeric variable. In its original form, it was mostly zero (most people do not earn capitol gains). These zeros presented a problem for my initial test-the-water logistic regression models as they resulted in complete separation. As such, this variable was binned into three ordered categories: none, low and high (the code was borrowed from someone's "Titanic" example). Though the proportion is small, high capital gains is strongly associated with the GT50K income group. This makes sense: in the US, it is mostly the wealthy who pay capital gains taxes.



Capitol-loss was similarly sparse and was binned as well. Though the association between the high income group and GT50K income seems apparent, it does not appear as strong as capitol gain (as evidenced by its weaker correlation). To claim capital loss one must have investments, a sign of relative wealth.

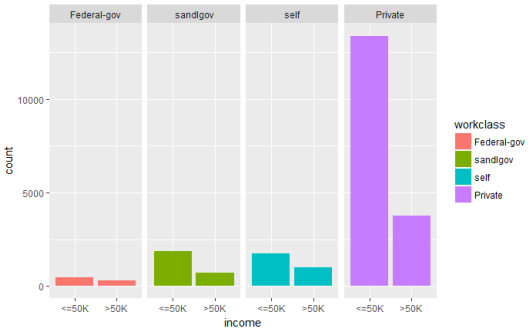


Native-country originally had over 40 categories, dominated by the US born. This variable too had potential to cause separation due to having many sparse levels. To collapse this variable, a logistic regression was run and identified 10 countries that has significantly lower odds of being in the GT50K group; 9 of these nations were in Latin American and the Caribbean in addition to Vietnam. Those reporting a native country of India were the only to have significantly greater odds of being the GT50K group, but this involved less than 150 respondents. To retain as much information as possible but also prevent perfect prediction, this variable was binned into three categories: the Latin countries identified and Vietnam, US born, and Other (because other has a disproportionately higher rate of being in the GT50K group on aggregate, placing India here minimizes information loss). Compared to the Latin America group, US born respondents were about 5 times as likely to be in the GT50K group, Other 6 times.

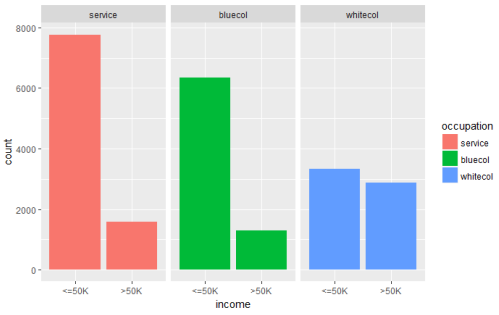


Education was dropped due to being redundant with the numeric equivalent.

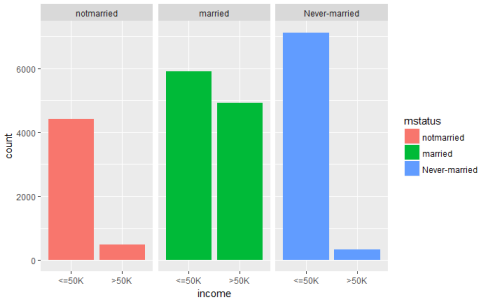
Work class classified the employment type of the respondent. This item was collapsed to aggressively due to the author failing to properly diagnose the cause of the separation issue. All groups were less likely than the privately employed to be classified as GT50K.



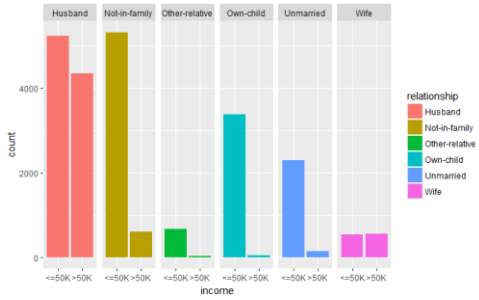
Occupation specified the specific job category of the respondent. Again, due to miss diagnosing the separation problem, this variable too was over binned and the author did not have time to rework it. White collar empolyees are 4.2 times as likely to be classified in the GT50K group vs. service; blue collar and service do not differ.

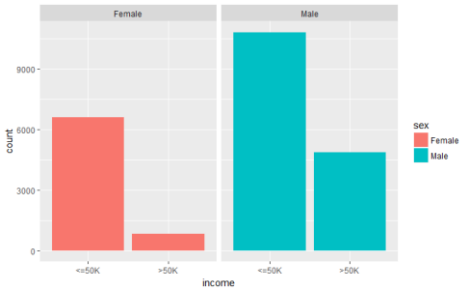


Marital status was collapsed into three categories: never married, married and not marred (e.g. divorced, separated, or widowed). Those reporting being married were 7.65 times as likely to be classified in the GT50K group. This makes sense: having two incomes makes it easier to break through that barrier.



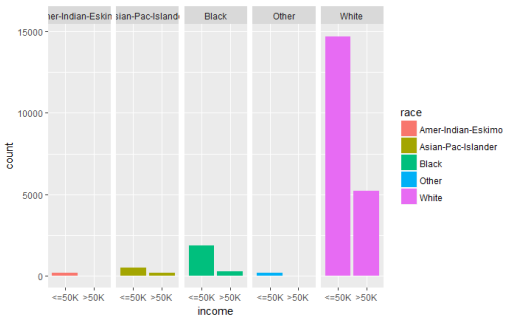
Relationship is, as far as the author can tell, the relation the respondent has to the primary resident. This variable was thus retained as it contained information relating to age, gender, marriage status, etc. and their odds of being in either category are statistically different. This variable correlated strongly with marital status but, as noted below, did not cause collinearity problems.





Sex the sample was predominately male (15,705 vs. 7,466). Gender, as noted earlier, is associated with income class, males being 3.5 times as likely to be in the GT50K group.

Race the sample mostly identified as white. White respondents were also more likely to be in the GT50K group than other racial/ethnic groups.



Predictive Modeling

As this was a classification problem, logistic regression, linear discriminant analysis, quadratic discriminant analysis, random forest and boosted regression tree models were fit (the latter as they are considered superior to the more interpretable classification tree). Each model was trained using caret's train function, utilizing 10 fold cross validation; both tree models optimized their tuning parameters (i.e. number of trees, shrinkage, depth, etc.) via cross validation. These steps were taken to enhance the accuracy of the models. For the non-caret random forest model: the number of variables to try at each branch was taken from cross validation and was coded in manually to save time. Carat's wrapper predict.train was preferred as it simplified testing.

Part of the difficulty with this assignment—aside from the laborious wrangling of factors-- is that fitting the above—particularly the tree models—takes and enormous amount of time. That this assignment was completed was due to the utilization of the packages parallel and doParallel which can be, once initialized, used with caret's resampling procedures. These two packages brought CPU utilization to 93% (once core should be held out e.g. on a hyper threaded quad core 7/8 will be available to R) from 15% and memory use into 6-8GBs. Importantly, no garbage collection will occur unto the CPU "cluster" has been disabled, resulting in ugly crashes if left unattended. The following code was used to train and test the models for this assignment:

```
cluster <- makeCluster(detectCores() - 1)

registerDoParallel(cluster)

train_control <- trainControl(method="cv", number=10, allowParallel = TRUE)

model.gbm = train(income ~ ., trControl = train_control , method = "gbm", data = mt.train.ren, distribution = "bernoulli", tuneLength=5);

model.gbm2 = train(income ~ . -caploss -relationship, trControl = train_control , method = "gbm", data = mt.train.ren, distribution = "bernoulli", tuneLength=5);

model.gbm3 = train(income ~ hoursperweek + mstatus + eduyrs + capgain + occupation + sex, trControl = train_control , method = "gbm", data = mt.train.ren, distribution = "bernoulli", tuneLength=5);

model.qda <- train(income~ ., data=mt.train.ren, trControl=train_control, method="qda", na.action = na.omit); model.qda

model.lda <- train(income~ ., data=mt.train.ren, trControl=train_control, method="lda", na.action = na.omit); model.qda

model.for <- train(income~., data=mt.train.ren, method="rf", trControl=train_control); model.for

rand.for <- randomForest(income ~ ., data = mt.train.ren, ntree=1000, ntry = 2, importance=TRUE)

stopCluster(cluster)
```

registerDoSEQ()

predict.gbm <- predict.train (object=model.gbm, mt.test, type="raw")

predict.glm <- predict.train(object=model.glm, mt.test, type="raw")

predict.qda <- predict.train (object=model.qda, mt.test, type="raw")

predict.lda <- predict.train (object=model.lda, mt.test, type="raw")

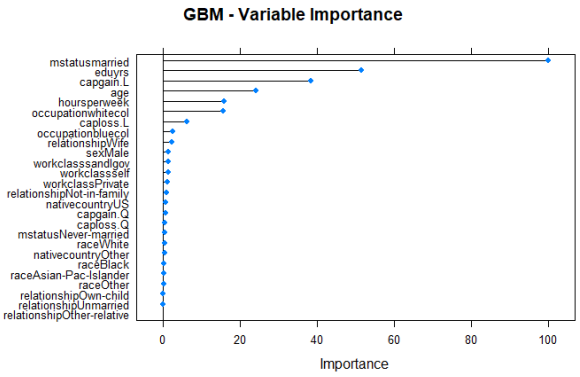
confusionMatrix(predict.gbm, mt.test$income)

confusionMatrix(predict.glm, mt.test$income)

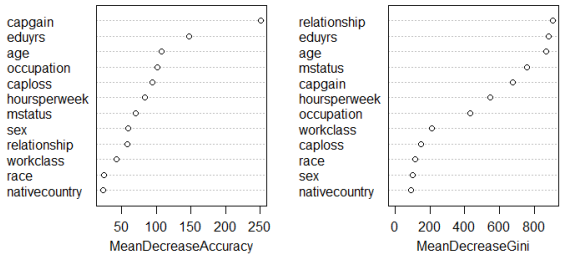
confusionMatrix(predict.qda, mt.test$income)

confusionMatrix(predict.lda, mt.test$income)

Findings: for the logistic regression model, no multicollinearity was evidenced by the variance inflation parameters. For each model, variable importance was assessed (pictured, the variable importance plot for the boosted regression model and the random forest). The variable importance plots and tables more or less confirm many of the inklings arrived at during the exploratory phase: married status, education, capital gains, hours worked, gender and white collar occupation are the most important variables in the boosted tree model. Similar also is the variable importance in the logistic regression model and the random forest model.



rand.for



As the full boosted model was, out of the gate, the best performing, gbm models were fit removing unimportant parameters. This did not improve model fit. Thus, the models discussed include all of the variables in the dataset, save those that were dropped prior to analysis (education(cat) and survey weight).

Assessed were the confusion tables for each model. The worst performing model was the QDA with an test accuracy of 0.7481, kappa of 0.4585 a sensitivity of 0.7132 and specificity of 0.8512. The performance of the logistic regression model and the LDA were, unsurprisingly, very similar, though the logistic model was slightly more accurate. Both had a test accuracy of about 0.835, kappa of ~0.55, sensitivity of 0.9282, and specificity of ~0.56. The best performing, albeit by a small margin was the boosted tree model. The boosted model had a test accuracy of 0.8425, kappa of 0.5510, sensitivity of 0.9284, and specificity of 0.5888. The random forest model was not tested due to a computational issue; however, on the training data, it performed second only to the boosted tree approach. All models saw a drastic decrease in specificity test vs. train and were, unsurprisingly, better at classifying respondents as LT50K than GT50K.

Full Test Performance Output

```
GBM Model Confusion Matrix and Statistics

          Reference
Prediction <=50K >50K
     <=50K  4849  727
     >50K    374 1041

               Accuracy : 0.8425
                 95% CI : (0.8338, 0.851)
    No Information Rate : 0.7471
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5538
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9284
            Specificity : 0.5888
         Pos Pred Value : 0.8696
         Neg Pred Value : 0.7357
             Prevalence : 0.7471
         Detection Rate : 0.6936
   Detection Prevalence : 0.7976
      Balanced Accuracy : 0.7586

       'Positive' Class : <=50K

Logistic Model Confusion Matrix and Statistics

          Reference
Prediction <=50K >50K
     <=50K  4848  778
     >50K    375  990

               Accuracy : 0.8351
                 95% CI : (0.8262, 0.8437)
    No Information Rate : 0.7471
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.528
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9282
            Specificity : 0.5600
         Pos Pred Value : 0.8617
         Neg Pred Value : 0.7253
             Prevalence : 0.7471
         Detection Rate : 0.6935
   Detection Prevalence : 0.8047
      Balanced Accuracy : 0.7441
```

```
      'Positive' Class : <=50K

QDA Model Confusion Matrix and Statistics

          Reference
Prediction <=50K >50K
     <=50K  3725  263
     >50K   1498 1505

               Accuracy : 0.7481
                 95% CI : (0.7378, 0.7582)
    No Information Rate : 0.7471
    P-Value [Acc > NIR] : 0.4299

                  Kappa : 0.4585
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7132
            Specificity : 0.8512
         Pos Pred Value : 0.9341
         Neg Pred Value : 0.5012
             Prevalence : 0.7471
         Detection Rate : 0.5328
   Detection Prevalence : 0.5704
      Balanced Accuracy : 0.7822

      'Positive' Class : <=50K

LDA Model Confusion Matrix and Statistics

          Reference
Prediction <=50K >50K
     <=50K  4884  828
     >50K    339  940

               Accuracy : 0.8331
                 95% CI : (0.8241, 0.8417)
    No Information Rate : 0.7471
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5138
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9351
            Specificity : 0.5317
         Pos Pred Value : 0.8550
         Neg Pred Value : 0.7349
             Prevalence : 0.7471
         Detection Rate : 0.6986
```

```
  Detection Prevalence : 0.8171
     Balanced Accuracy : 0.7334

      'Positive' Class : <=50K
```