Income Prediction Project

INTRODUCTION

The issue of income prediction has long interested researchers and individuals alike. With income tied closely to social/economic class and quality of life, there is clear value in developing a model that can successfully predict an individual's income.

For this project, a clean dataset of 1994 United States census income data curated by Kaggle will be analyzed and utilized in an attempt to develop a successful model. The project goals are two-fold:

1. How accurately can income be predicted from three sets of predictors: one demographic, one achievement-based, and one monetary?
2. Which predictors are most influential in predicting income?

Practical implications of this type of research may suggest that income disparity could be partially addressed by individuals selecting different educational or occupational choices. On the other hand, if demographic predictors are most influential, this research may suggest that larger-scale biases/prejudices need to be more directly confronted in an attempt to address the income gap.

The dataset used for this project has one outcome variable: individual adult income. This variable is binary: income is coded either as less than or equal to $50,000 or as more than $50,000. Essentially, the model will be asked to predict lower-income versus higher-income individuals.

One important note about income distribution in our dataset: over 75% of individuals in our dataset have incomes below $50,000. Although at first glance this is not ideal, this breakdown actually likely mirrors the actual income breakdown in the U.S.

The dataset also contains 10 predictors. The four demographic predictors are: sex (binary), race (categorical), marital status (categorical), and age (continuous). The four achievement-based predictors are: education (categorical), occupation (categorical), work sector (categorical; called workclass), and hours per week worked (continuous). The two monetary predictors are: capital gain (continuous) and capital loss (continuous).

First the data will be wrangled so that similar and/or excessively small categories are combined. Next, various machine learning models will be applied to the dataset in an attempt to predict income, with the most successful models combined into an ensemble. Ultimately, the model that is most successful (considering accuracy, F1 scores, sensitivity, and specificity) will be applied to the validation set to determine final accuracy and F1 scores. This model along with the most influenctial predictors (determined by the variable importance function) will be used to draw tentative/preliminary conclusions about income in the United States.

METHODS/ANALYSIS

To begin, we will import the dataset, wrangle the data to group similar and/or excessively small categories, and create training and validation sets. The training set will be 90% of our original dataset and the validation set will be 10%.

Next, the training set will be split again to create a secondary training set to use for cross-validation and model selection. The primary training set (used to train the models) will be 90% of the training set while the secondary training set (used to cross-validate and select the best model) will be 10% of the training set.

Now exploratory data analysis will be conducted on each of the categorical predictors to see whether they should be utilized in the model. The training set will be grouped by these variables and then the percentage of each group with incomes at or below $50,000 will be displayed.

```
## # A tibble: 2 x 2
##   sex    income_prob
##   <chr>        <dbl>
## 1 Female       0.890
## 2 Male         0.695
```

```
## # A tibble: 5 x 2
##   race               income_prob
##   <chr>                    <dbl>
## 1 Amer-Indian-Eskimo       0.875
## 2 Asian-Pac-Islander       0.725
## 3 Black                    0.880
## 4 Other                    0.919
## 5 White                    0.744
```

```
## # A tibble: 5 x 2
##   marital.status.cond income_prob
##   <chr>                     <dbl>
## 1 Divorced                  0.897
## 2 Married                   0.564
## 3 Never-married             0.955
## 4 Separated                 0.931
## 5 Widowed                   0.925
```

```
## # A tibble: 10 x 2
##    education.cond income_prob
##    <chr>                <dbl>
##  1 Assoc-acdm           0.759
##  2 Assoc-voc            0.746
##  3 Bachelors            0.584
##  4 Below-HS             0.948
##  5 Doctorate            0.267
##  6 HS-grad              0.839
##  7 Masters              0.443
##  8 Prof-school          0.266
##  9 Some-college         0.809
## 10 Some-HS              0.937
```

```
## # A tibble: 14 x 2
##    occupation.cond         income_prob
##    <chr>                         <dbl>
##  1 Adm-clerical                  0.865
##  2 Craft-repair                  0.772
##  3 Exec-managerial               0.517
##  4 Farming-fishing               0.884
##  5 Handlers-cleaners             0.932
##  6 Machine-op-inspct             0.883
##  7 Other-service                 0.958
##  8 Priv-house-serv               0.992
##  9 Prof-specialty                0.558
## 10 Protective-serv               0.674
## 11 Sales                         0.725
## 12 Tech-support                  0.702
```

```
## 13 Transport-moving              0.796
## 14 Unknown-or-Armed-Forces       0.886
```

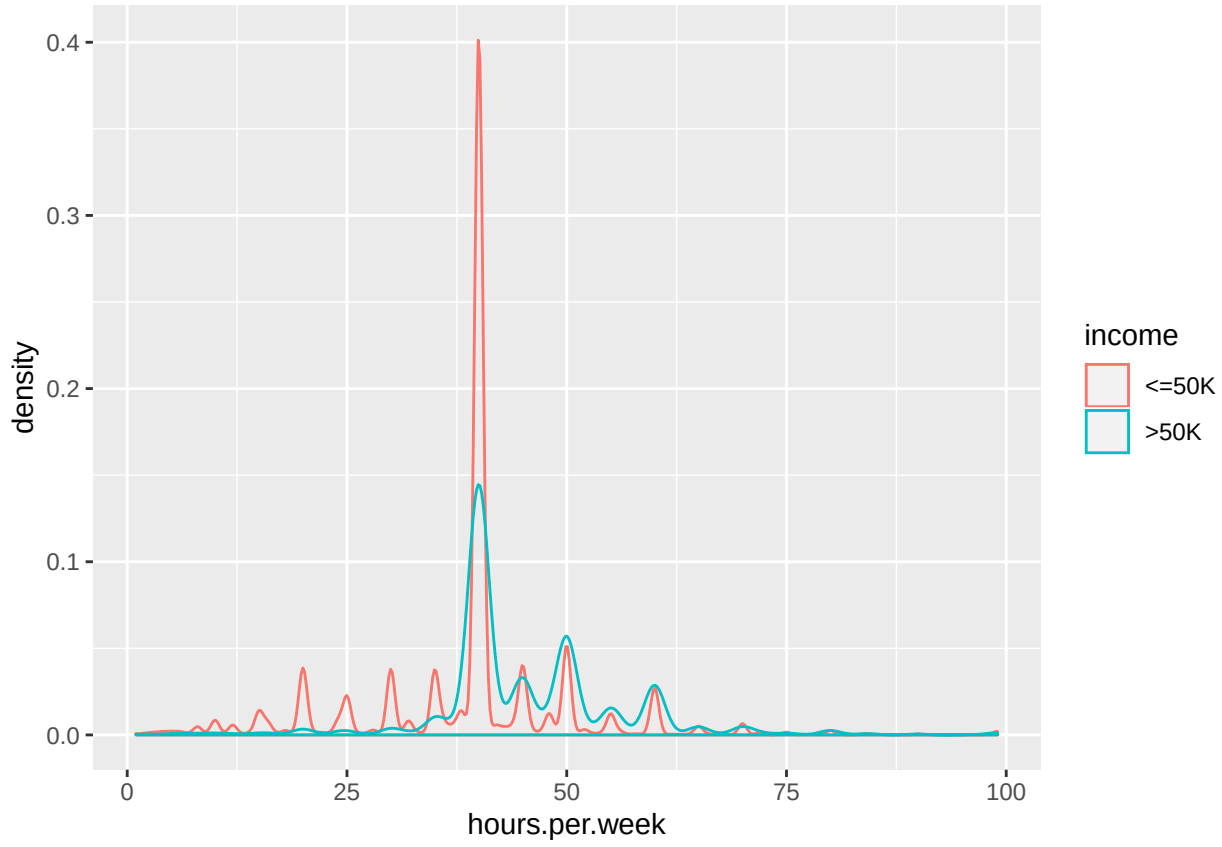
```
## # A tibble: 5 x 2
##   workclass.cond income_prob
##   <chr>                <dbl>
## 1 Government           0.691
## 2 No-Pay               1
## 3 Private              0.781
## 4 Self-employed        0.640
## 5 Unknown              0.886
```

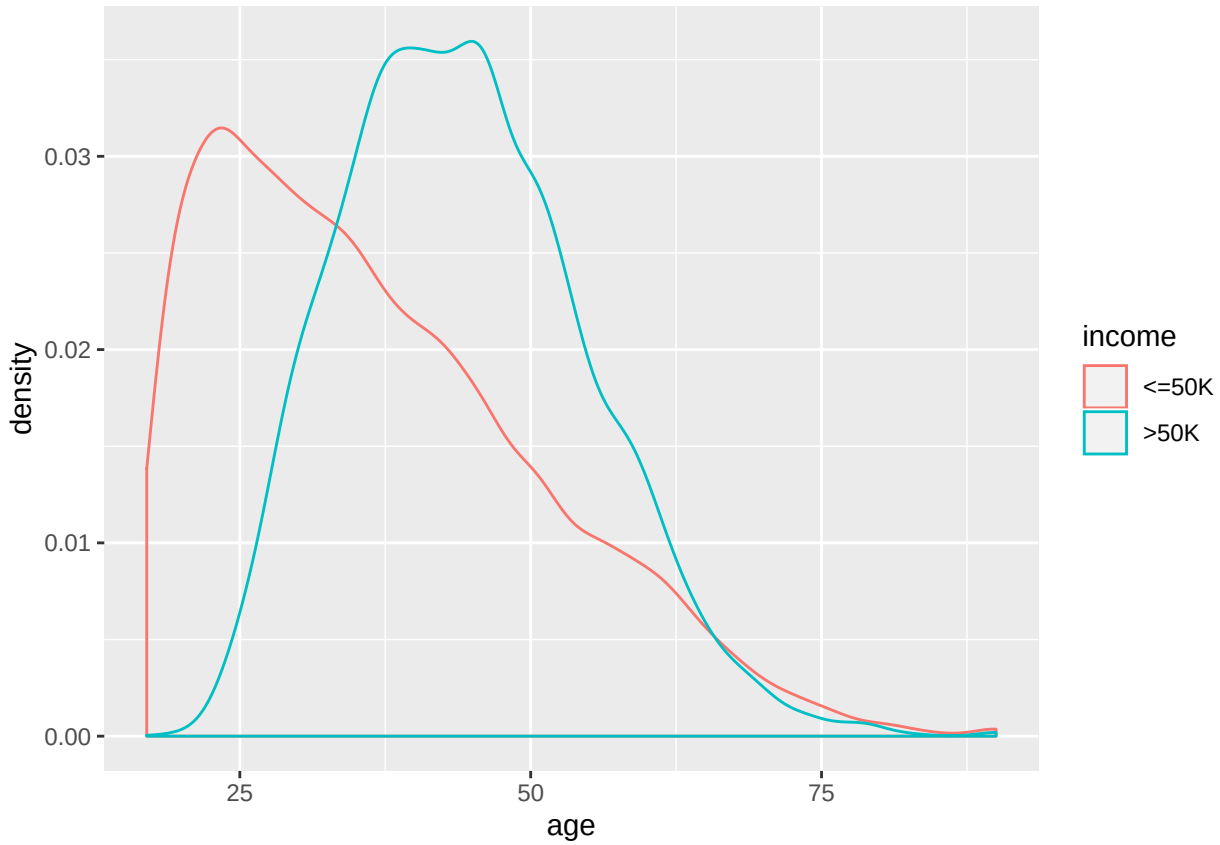
For each of these categorical variables, there are significant group differences in terms of the probability of having an income at or below $50,000.

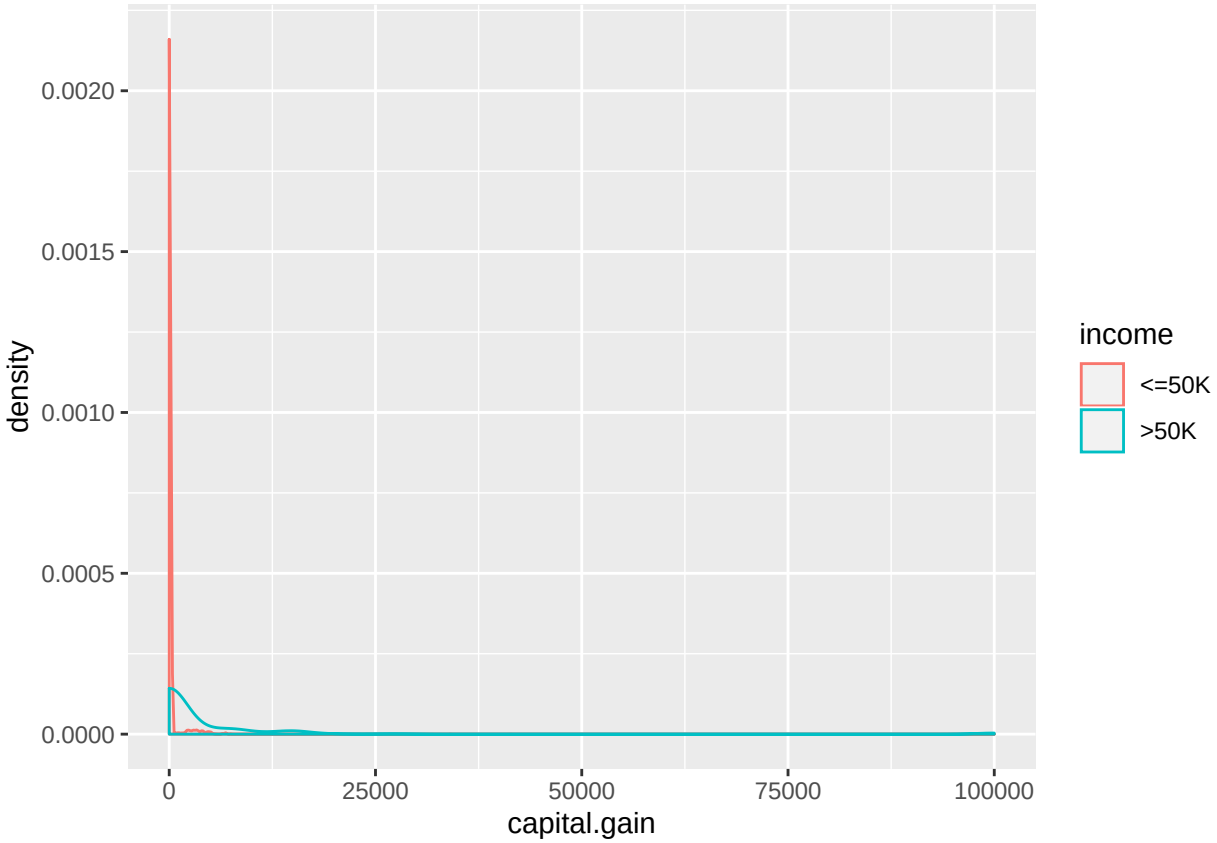
For the demographic variables, the following groups are most likely to have high incomes (above $50,000): 1. Males (predictor: sex) 2. Whites and Asians/Pacific Islanders (predictor: race) 3. Married Individuals (predictor: marital status)

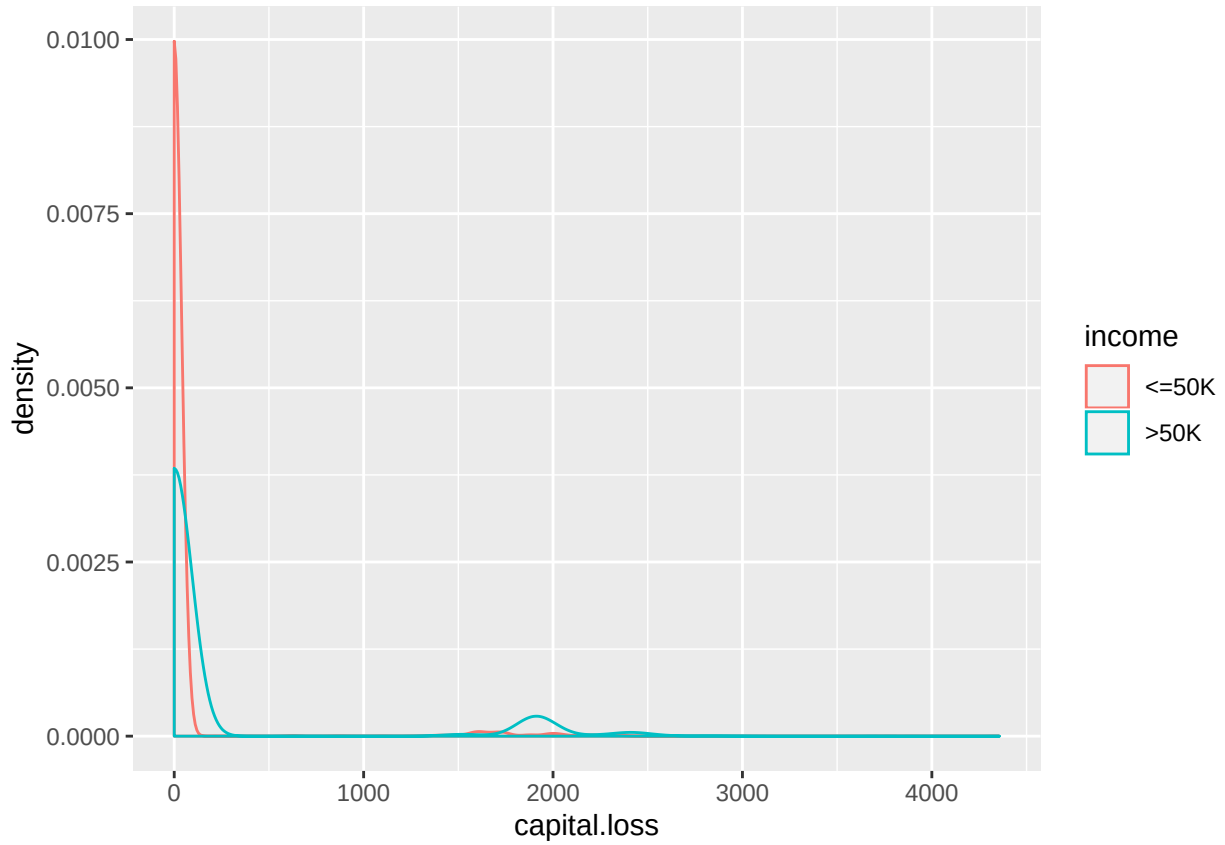
For the achievement-based variables, the following groups are most likely to have high incomes: 1. Individuals with a Bachelors degree or above (predictor: education) 2. Individuals who attend Professional School (predictor: education) 3. Executive/Managerial and Professional-Specialty Occupations (predictor: occupation) 4. Individuals who are self-employed or employed by the government (predictor: workclass)

For continuous variables, some data visualization needs to occur to explore at what values of that variable individuals are more likely to make lower or higher incomes.




density
0.03
0.02
0.01
0.00
25
50
75
age
income
<=50K
>50K


0.0020
0.0015
0.0010
0.0005
0.0000
density
0
25000
50000
75000
100000
capital.gain
income
<=50K
>50K



For age, a demographic continuous variable, those making higher incomes skew much older overall. The exception appears to be with the 70 and older age group, which skews back toward lower incomes. This is not surprisingly since 70 represents a typical retirement age where we would expect individual incomes to drop considerably.

For hours per week worked, an achievement-based continuous variable, those making higher incomes skew towards higher hours worked, with very few higher income individuals working fewer than 40 hours and with over 75 hours per week worked being populated almost exclusively by the higher income crowd.

For capital gain and capital loss, continuous monetary variables, most individuals (of any income brakcet) report 0 in both categories. However, individuals that did report any gains or losses overwhelmingly look to be in the higher income crowd. Again, this is not surprisingly since individuals with disposable incomes are far more likely to be involved in activities that could potentially create these gains or losses.

Based on this exploratory data analysis and visualization, all of the above variables should be included in the model.

We'll start with an rpart model, which will allow us to visualize which predictors are primarily useful in predicting income and in which ways.

To visualize the three different categories of predictors, we'll make three different rpart models with each category before combining all categories into one final rpart model.

DEMOGRAPHIC RPART MODEL

We'll start with creating an rpart model based only on the demographic predictors:

```
## n= 26373
##
## node), split, n, loss, yval, (yprob)
```

```
##       * denotes terminal node
##
##  1) root 26373 6350 <=50K (0.75922345 0.24077655)
##    2) marital.status.condMarried< 0.5 13835  879 <=50K (0.93646549 0.06353451) *
##    3) marital.status.condMarried>=0.5 12538 5471 <=50K (0.56364651 0.43635349)
##      6) age< 33.5 2968  813 <=50K (0.72607817 0.27392183) *
##      7) age>=33.5 9570 4658 <=50K (0.51327064 0.48672936)
##       14) age>=59.5 1330  462 <=50K (0.65263158 0.34736842) *
##       15) age< 59.5 8240 4044 >50K (0.49077670 0.50922330)
##         30) age< 35.5 763  301 <=50K (0.60550459 0.39449541) *
##         31) age>=35.5 7477 3582 >50K (0.47906915 0.52093085)
##           62) raceWhite< 0.5 815  350 <=50K (0.57055215 0.42944785) *
##           63) raceWhite>=0.5 6662 3117 >50K (0.46787751 0.53212249) *
```

```
##                                      Overall
## age                                1077.73796
## marital.status.condMarried         1828.41131
## marital.status.condNever-married    977.69241
## raceBlack                           124.98278
## raceOther                            35.68971
## raceWhite                            66.48186
## sexMale                             448.35407
## `raceAsian-Pac-Islander`              0.00000
## `marital.status.condNever-married`    0.00000
## marital.status.condSeparated          0.00000
## marital.status.condWidowed            0.00000
```

Accuracy

```
## [1] 0.7877857
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction <=50K >50K
##      <=50K  1925  322
##      >50K    300  384
##
##                Accuracy : 0.7878
##                  95% CI : (0.7725, 0.8025)
##     No Information Rate : 0.7591
##     P-Value [Acc > NIR] : 0.0001288
##
##                   Kappa : 0.4135
##
##  Mcnemar's Test P-Value : 0.3997749
##
##             Sensitivity : 0.8652
##             Specificity : 0.5439
##          Pos Pred Value : 0.8567
##          Neg Pred Value : 0.5614
##              Prevalence : 0.7591
##          Detection Rate : 0.6568
```

```
##    Detection Prevalence : 0.7666
##       Balanced Accuracy : 0.7045
##
##        'Positive' Class : <=50K
##
```

The accuracy for the demographic model is 0.7878, which is a good start. However, it is important to note that there is a significant gap between sensitivity (the model correctly predicting lower incomes) and specificity (the model correctly predicting higher incomes).

Since we observe this difference and we also observe that lower incomes have a much higher probability of occuring in the dataset, we'll want to use something more nuanced that just accuracy to determine model success. We'll use the F1 score as a measure of balanced accuracy.

F1 Score

```
## [1] 0.8609123
```

The F1 score for the demographic rpart model is 0.8609, with being married, being white, and having an age between 35.5 and 59.5 serving as the key predictors for higher incomes. In fact, the model only predicts an individual to make more than $50,000 if all of these predictors apply.

ACHIEVEMENT RPART MODEL

We'll now compare this to the achievement model. In an equitable world, we would expect the achievement model (which includes educational attainment as well as occupation and work sector choice) to be a much more successful model than the demographic one.

```
## n= 26373
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 26373 6350 <=50K (0.7592234 0.2407766)
##    2) hours.per.week< 43.5 18971 3351 <=50K (0.8233620 0.1766380) *
##    3) hours.per.week>=43.5 7402 2999 <=50K (0.5948392 0.4051608)
##      6) occupation.condExec-managerial< 0.5 5851 2066 <=50K (0.6468980 0.3531020)
##       12) occupation.condProf-specialty< 0.5 4611 1374 <=50K (0.7020169 0.2979831) *
##       13) occupation.condProf-specialty>=0.5 1240  548 >50K (0.4419355 0.5580645) *
##      7) occupation.condExec-managerial>=0.5 1551  618 >50K (0.3984526 0.6015474) *
```

```
##                                       Overall
## education.condBachelors              464.60825
## education.condDoctorate               45.49296
## education.condHS-grad                173.21602
## education.condMasters                381.71058
## education.condProf-school            196.77457
## hours.per.week                       556.12013
## occupation.condExec-managerial       594.11976
## occupation.condProf-specialty        441.96346
## `education.condAssoc-voc`              0.00000
## `education.condBelow-HS`               0.00000
## `education.condHS-grad`                0.00000
## `education.condProf-school`            0.00000
## `education.condSome-college`           0.00000
```

```
## `education.condSome-HS`                    0.00000
## `occupation.condCraft-repair`              0.00000
## `occupation.condExec-managerial`           0.00000
## `occupation.condFarming-fishing`           0.00000
## `occupation.condHandlers-cleaners`         0.00000
## `occupation.condMachine-op-inspct`         0.00000
## `occupation.condOther-service`             0.00000
## `occupation.condPriv-house-serv`           0.00000
## `occupation.condProf-specialty`            0.00000
## `occupation.condProtective-serv`           0.00000
## occupation.condSales                       0.00000
## `occupation.condTech-support`              0.00000
## `occupation.condTransport-moving`          0.00000
## `occupation.condUnknown-or-Armed-Forces`   0.00000
## `workclass.condNo-Pay`                     0.00000
## workclass.condPrivate                      0.00000
## `workclass.condSelf-employed`              0.00000
## workclass.condUnknown                      0.00000
```

Accuracy

```
## [1] 0.7922211
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction <=50K >50K
##      <=50K  2115  499
##      >50K    110  207
##
##                Accuracy : 0.7922
##                  95% CI : (0.7771, 0.8068)
##     No Information Rate : 0.7591
##     P-Value [Acc > NIR] : 1.147e-05
##
##                   Kappa : 0.3002
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9506
##             Specificity : 0.2932
##          Pos Pred Value : 0.8091
##          Neg Pred Value : 0.6530
##              Prevalence : 0.7591
##          Detection Rate : 0.7216
##    Detection Prevalence : 0.8918
##       Balanced Accuracy : 0.6219
##
##        'Positive' Class : <=50K
##
```

F1 Score

```
## [1] 0.8741476
```

The influenctial predictors here for predicting incomes over $50,000 are working at least 43.5 hours per week and being employed in Executive/Managerial or Professional-Specialty occupations. Having higher levels of education (high school graduate or bachelors or above) is also included in the model's variable importance function.

The accuracy and F1 scores are slightly higher than the scores for the demographic model, but not significantly so. What is most interesting is that the sensitivity of the model increased significantly (from 0.8652 to 0.9506) while the specificity score, which was already low in our demographics model, plummeted (from 0.5439 to 0.2932). While this model is excellent at predicting when a low income individual is low income, it is terrible at correctly predicting that a high income individual is high income.

MONETARY RPART MODEL

Lastly, we will examine the monetary model for comparative purposes:

```
## n= 26373
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
## 1) root 26373 6350 <=50K (0.7592234 0.2407766)
##   2) capital.gain< 5119 25114 5155 <=50K (0.7947360 0.2052640)
##     4) capital.loss< 1820.5 24309 4570 <=50K (0.8120038 0.1879962) *
##     5) capital.loss>=1820.5 805  220 >50K (0.2732919 0.7267081) *
##   3) capital.gain>=5119 1259   64 >50K (0.0508340 0.9491660) *
```

```
##                Overall
## capital.gain 1337.2990
## capital.loss  844.3999
```

Accuracy

```
## [1] 0.8150802
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction <=50K >50K
##      <=50K  2189  506
##      >50K     36  200
##
##                Accuracy : 0.8151
##                  95% CI : (0.8005, 0.829)
##     No Information Rate : 0.7591
##     P-Value [Acc > NIR] : 1.797e-13
##
##                   Kappa : 0.3457
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9838
##             Specificity : 0.2833
##          Pos Pred Value : 0.8122
##          Neg Pred Value : 0.8475
```

```
##              Prevalence : 0.7591
##          Detection Rate : 0.7468
##    Detection Prevalence : 0.9195
##       Balanced Accuracy : 0.6336
##
##        'Positive' Class : <=50K
##
```

F1 Score

```
## [1] 0.8898374
```

The accuracy and F1 scores are higher than for either of the other two models. However, while sensitivity has risen even higher when compared to the achievement model (from 0.9506 to 0.9838), specificity has declined again (from 0.2932 to 0.2833).

The variables used to predict when income will be above $50,000 is either that capital gains are more than $5119 or capital losses are more than $1820. This model's lack of specificity is probably not surprising since it is unable to account for the fact that many higher income individuals do not necessarily have capital gains or capital losses.

COMBINED RPART MODEL

WIth the demographic model performing better with prediciting higher income individuals and the other two models performing better at predicting lower income individuals, we will combine all three into one rpart model:

```
## n= 26373
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 26373 6350 <=50K (0.75922345 0.24077655)
##    2) marital.status.condMarried< 0.5 13835  879 <=50K (0.93646549 0.06353451)
##      4) capital.gain< 7139.5 13595  648 <=50K (0.95233542 0.04766458) *
##      5) capital.gain>=7139.5 240    9 >50K (0.03750000 0.96250000) *
##    3) marital.status.condMarried>=0.5 12538 5471 <=50K (0.56364651 0.43635349)
##      6) capital.gain< 5095.5 11561 4507 <=50K (0.61015483 0.38984517)
##       12) capital.loss< 1846 10928 3967 <=50K (0.63698755 0.36301245)
##         24) education.condBachelors< 0.5 9087 2893 <=50K (0.68163310 0.31836690)
##           48) education.condMasters< 0.5 8471 2475 <=50K (0.70782670 0.29217330)
##             96) occupation.condProf-specialty< 0.5 7877 2128 <=50K (0.72984639 0.27015361) *
##             97) occupation.condProf-specialty>=0.5 594  247 >50K (0.41582492 0.58417508) *
##           49) education.condMasters>=0.5 616  198 >50K (0.32142857 0.67857143) *
##         25) education.condBachelors>=0.5 1841  767 >50K (0.41662140 0.58337860)
##           50) hours.per.week< 34.5 173   57 <=50K (0.67052023 0.32947977) *
##           51) hours.per.week>=34.5 1668  651 >50K (0.39028777 0.60971223) *
##       13) capital.loss>=1846 633   93 >50K (0.14691943 0.85308057) *
##      7) capital.gain>=5095.5 977   13 >50K (0.01330604 0.98669396) *
```

Accuracy

```
## [1] 0.8474923
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction <=50K >50K
##      <=50K  2092  314
##      >50K    133  392
##
##                Accuracy : 0.8475
##                  95% CI : (0.834, 0.8603)
##     No Information Rate : 0.7591
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.543
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9402
##             Specificity : 0.5552
##          Pos Pred Value : 0.8695
##          Neg Pred Value : 0.7467
##              Prevalence : 0.7591
##          Detection Rate : 0.7137
##    Detection Prevalence : 0.8209
##       Balanced Accuracy : 0.7477
##
##        'Positive' Class : <=50K
##
```

F1 Score

```
## [1] 0.9034766
```

This model is much improved over the previous three models. Accuracy and F1 scores are higher than all other models. Sensitivity is slightly below the achievement and monetary models, but specificity is vastly improved over these two models and slightly improved over the demographics model.

The rpart tree provides a nuanced picture that includes predictors from all three categories, with being married, having significant capital gains or losses, having a Bachelor's or Master's degree, working more than 34.5 hours in a week, and being employed in a Professional-Specialty occupation all contributing to the model predicting someone to have an income above $50,000.

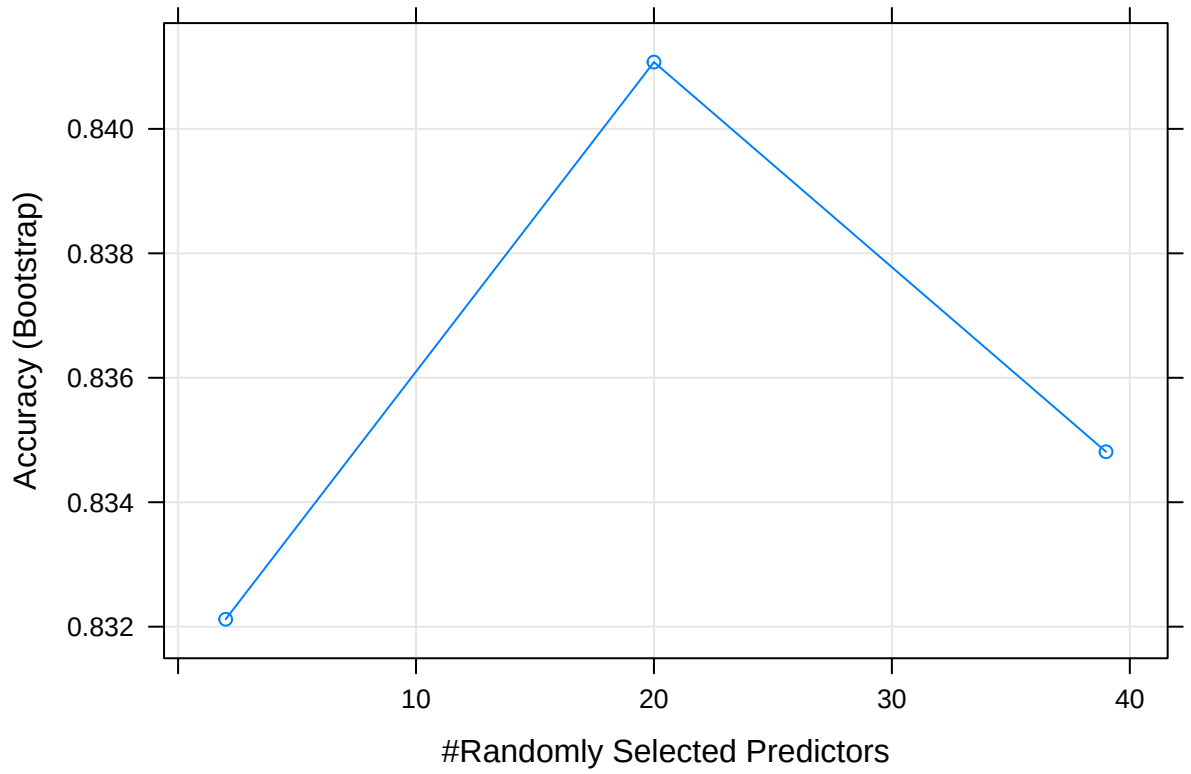
What's most surprising about this combined model is the fact that marital status is the first tree split. According to this model, if an individual is unmarried, their only path to a higher income is to have captial gains above $7140. Marriage is not a predictor that is usually associated with income in common discourse, but there could be multiple possible explanations. Perhaps those who choose to follow a conventional path and get married are also those who choose more traditional, money-oriented career paths. Perhaps marriage provides the stability (especially for those with children) for individuals to seek out improved career and educational opportunities. Although this limited dataset doesn't permit us to answer these questions, this would be an interesting avenue for further exploration.

We'll now create other models using other modeling techniques to see if any of these techniques are effective at capturing the relationships between our outcome and our predictor variables.

RANDOM FOREST MODEL

We'll try out random forest as a natural extension of rpart. However, we'll have to first split the data again to allow the random forest model to be created on a much more manageable chunk of the data.

```
##
## Call:
##  randomForest(x = x, y = y, mtry = param$mtry)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 20
##
##         OOB estimate of  error rate: 15.26%
## Confusion matrix:
##       <=50K >50K class.error
## <=50K  4610  395  0.07892108
## >50K    611  976  0.38500315
```




```
## rf variable importance
##
##   only 20 most important variables shown (out of 39)
##
##                                 Overall
## age                            100.000
## marital.status.condMarried      82.357
## capital.gain                    79.411
## hours.per.week                  55.622
## capital.loss                    24.747
## education.condBachelors         17.405
## marital.status.condNever-married  11.906
## workclass.condPrivate           11.650
```

```
## occupation.condExec-managerial    11.647
## sexMale                           11.593
## occupation.condProf-specialty     10.853
## education.condMasters             10.409
## education.condHS-grad              8.431
## occupation.condCraft-repair        7.713
## occupation.condSales               7.639
## workclass.condSelf-employed        7.493
## education.condSome-college         7.140
## raceWhite                          6.072
## education.condProf-school          5.720
## occupation.condTransport-moving    5.669
```

Accuracy

```
## [1] 0.8444217
```

F1 Score

```
## [1] 0.9004367
```

The random forest model performs decently on accuracy (0.8444) and provides an F1 score over 0.9 (0.9004). The variables identified as being most important in constructing this model are age, marital status, capital gain, hours worked per week, and capital loss. These variables are similar to the variables influential in the combined rpart model, with the notable addition of age.

Combined rpart and random forest models perform well on income prediction, but we'll test out a variety of other models (knn, glm, lda, and qda) to see if any of these can improve prediction accuracy.

K-NEAREST NEIGHBORS MODEL

Next we'll create a k-nearest neighbors (knn) model. We'll need to use the same partioned training set we used for random forest in order for the knn function to run in a reasonable time frame.

```
## 12-nearest neighbor model
## Training set outcome distribution:
##
## <=50K  >50K
##  5005  1587
```

Accuracy

```
## [1] 0.8341863
```

F1 Score

```
## [1] 0.8966397
```

The k-nearest neighbors model does not perform as well as the random forest model, but accuracy and F1 scores are only slightly lower, so we'll include this model in our list of candidates for the ensemble model.

GENERALIZED LINEAR MODEL

Next up we'll try a generalized linear model (glm).

```
## glm variable importance
##
##   only 20 most important variables shown (out of 39)
##
##                                    Overall
## marital.status.condMarried          100.00
## capital.gain                         98.00
## hours.per.week                       59.79
## capital.loss                         55.21
## age                                  48.69
## `education.condBelow-HS`             34.66
## `occupation.condExec-managerial`     31.58
## `education.condSome-HS`              31.41
## `education.condProf-school`          29.43
## education.condDoctorate              27.25
## `occupation.condFarming-fishing`     27.00
## education.condMasters                24.49
## `occupation.condOther-service`       24.38
## education.condBachelors              19.68
## `marital.status.condNever-married`   18.14
## `occupation.condTech-support`        17.01
## `occupation.condProf-specialty`      16.71
## `occupation.condHandlers-cleaners`   16.08
## `education.condHS-grad`              15.18
## `occupation.condMachine-op-inspct`   13.95
```

Accuracy

```
## [1] 0.8529512
```

F1 Score

```
## [1] 0.9057511
```

The glm model seems promising, with a higher accuracy and F1 score than any of the other models that have been tested so far.

The top five variables that are most influential in this model are: marital status, age, hours worked per week, capital gain, and capital loss. Again, this model shows that variables from all three categories (demographic, achievement, and monetary) are essential for creating a robust and accurate model.

LINEAR DISCRIMINANT ANALYSIS MODEL

Next we'll try out a linear discriminant analysis (lda) model:

```
## Call:
## lda(x, grouping = y)
##
## Prior probabilities of groups:
##     <=50K      >50K
## 0.7592234 0.2407766
##
## Group means:
##       education.condAssoc-voc education.condBachelors
```

```
## <=50K              0.04225141               0.1263047
## >50K               0.04535433               0.2831496
##       education.condBelow-HS education.condDoctorate education.condHS-grad
## <=50K            0.045947161             0.004394946             0.3557908
## >50K             0.007874016             0.037952756             0.2148031
##       education.condMasters education.condProf-school
## <=50K            0.03101433               0.00604305
## >50K             0.12299213               0.05259843
##       education.condSome-college education.condSome-HS hours.per.week
## <=50K                  0.2377766            0.11741497       38.78874
## >50K                   0.1774803            0.02472441       45.43402
##       occupation.condCraft-repair occupation.condExec-managerial
## <=50K                   0.1276033                      0.08510213
## >50K                    0.1188976                      0.25086614
##       occupation.condFarming-fishing occupation.condHandlers-cleaners
## <=50K                     0.03550916                        0.05228987
## >50K                      0.01464567                        0.01196850
##       occupation.condMachine-op-inspct occupation.condOther-service
## <=50K                       0.07071867                    0.1277031
## >50K                        0.02960630                    0.0176378
##       occupation.condPriv-house-serv occupation.condProf-specialty
## <=50K                   0.0059931079                     0.09314289
## >50K                    0.0001574803                     0.23259843
##       occupation.condProtective-serv occupation.condSales
## <=50K                     0.01807921            0.1062778
## >50K                      0.02755906            0.1269291
##       occupation.condTech-support occupation.condTransport-moving
## <=50K                  0.02611996                      0.05218998
## >50K                   0.03496063                      0.04204724
##       occupation.condUnknown-or-Armed-Forces raceAsian-Pac-Islander
## <=50K                             0.06667333             0.03151376
## >50K                              0.02708661             0.03763780
##        raceBlack   raceOther raceWhite   sexMale
## <=50K 0.10872497 0.010138341 0.8384857 0.6128952
## >50K  0.04661417 0.002834646 0.9078740 0.8494488
##       marital.status.condMarried marital.status.condNever-married
## <=50K                  0.3529441                       0.41047795
## >50K                   0.8615748                       0.06125984
##       marital.status.condSeparated marital.status.condWidowed      age
## <=50K                  0.037706637                0.036757729 36.85597
## >50K                   0.008818898                0.009448819 44.29811
##       workclass.condNo-Pay workclass.condPrivate
## <=50K         0.0008989662             0.7170254
## >50K          0.0000000000             0.6344882
##       workclass.condSelf-employed workclass.condUnknown capital.gain
## <=50K                  0.09439145            0.06607401     143.7081
## >50K                   0.16724409            0.02692913    3889.2797
##       capital.loss
## <=50K     53.28892
## >50K     196.25291
##
## Coefficients of linear discriminants:
##                                              LD1
## education.condAssoc-voc                2.898406e-02
```

```
## education.condBachelors                 4.935460e-01
## education.condBelow-HS                 -8.287892e-01
## education.condDoctorate                 1.413093e+00
## education.condHS-grad                  -2.787991e-01
## education.condMasters                   8.722249e-01
## education.condProf-school               1.291809e+00
## education.condSome-college             -5.394846e-02
## education.condSome-HS                  -4.053139e-01
## hours.per.week                          1.400962e-02
## occupation.condCraft-repair            -1.263508e-01
## occupation.condExec-managerial          6.438600e-01
## occupation.condFarming-fishing         -6.033425e-01
## occupation.condHandlers-cleaners       -2.874360e-01
## occupation.condMachine-op-inspct       -3.103259e-01
## occupation.condOther-service           -1.684985e-01
## occupation.condPriv-house-serv         -5.704396e-02
## occupation.condProf-specialty           2.715609e-01
## occupation.condProtective-serv          2.796719e-01
## occupation.condSales                    2.102133e-01
## occupation.condTech-support             3.264624e-01
## occupation.condTransport-moving        -2.317579e-01
## occupation.condUnknown-or-Armed-Forces  9.587153e-02
## raceAsian-Pac-Islander                  7.579645e-02
## raceBlack                               1.633841e-01
## raceOther                              -8.940715e-02
## raceWhite                               2.475178e-01
## sexMale                                 1.913424e-01
## marital.status.condMarried              1.381661e+00
## marital.status.condNever-married       -1.859793e-03
## marital.status.condSeparated            1.045220e-01
## marital.status.condWidowed              4.386332e-02
## age                                     1.119685e-02
## workclass.condNo-Pay                   -5.305852e-01
## workclass.condPrivate                   2.444542e-02
## workclass.condSelf-employed            -1.240435e-01
## workclass.condUnknown                  -2.479259e-01
## capital.gain                            4.024912e-05
## capital.loss                            4.629595e-04
```

Accuracy

```
## [1] 0.8396452
```

F1 Score

```
## [1] 0.8983124
```

The lda model does not perform as well as the glm model, but its accuracy and F1 scores are in line with the other tested models (including combined rpart, random forest, and k-nearest neighbors).

This model provides some additional insight as to why being married is such an influential predictor in previous models. Of individuals in the training set with incomes above $50,000, 86% are married. There is also some interesting insight here in regards to hours worked per week. On average, those with lower

incomes work 38.79 hours per week while those with higher incomes work 45.43 hours per week. Both of these observations are in line with the exploratory data analysis we performed earlier, but this different method of calculating group means provides an additional way of looking at the data.

In this model, education and marital status emerge as the predictors having the strongest effect on the model, with the coefficients for doctorate level education and being married having the highest absolute value.

QUADRATIC DISCRIMINANT ANALYSIS MODEL

Lastly, we will test out a quadratic discriminant analysis (qda) model.

Accuracy

```
## [1] 0.7792562
```

F1 Score

```
## [1] 0.8409929
```

One note about the qda model: it refused to accept workclass as a predictor, even after workclass was wrangled and re-categorized in multiple ways. N/As were elminated, small categories were combined, etc., but qda refused to run with workclass. As a result, the qda model uses only the other nine predictors.

The accuracy and F1 scores of the qda model are significantly below any of the other combined models we've previously tested, so this model should be excluded from the ensemble.

ENSEMBLE MODEL

Afer examining the results of all the models tested out, five appear to be clearly superior, as judged by accuracy and F1 scores. As a result, the following five models will be included in the ensemble model: combined rpart, random forest, k-nearest neighbors, generalized linear model, and linear discriminant analysis. Since there are five models being assembled, whatever three or models predict for a given individual is what the ensemble will use as its prediction.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction <=50K >50K
##      <=50K  2112  289
##      >50K    113  417
##
##                Accuracy : 0.8628
##                  95% CI : (0.8499, 0.8751)
##     No Information Rate : 0.7591
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5901
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9492
##             Specificity : 0.5907
##          Pos Pred Value : 0.8796
##          Neg Pred Value : 0.7868
##              Prevalence : 0.7591
##          Detection Rate : 0.7206
```

```
##    Detection Prevalence : 0.8192
##       Balanced Accuracy : 0.7699
##
##        'Positive' Class : <=50K
##
```

Accuracy

```
## # A tibble: 1 x 6
##     glm   lda rpart    rf   knn ensemble
##   <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl>
## 1 0.853 0.840 0.847 0.844 0.834    0.863
```

F1 Score

```
## # A tibble: 1 x 6
##     glm   lda rpart    rf   knn ensemble
##   <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl>
## 1 0.906 0.898 0.903 0.900 0.897    0.913
```

Specificity, accuracy, and F1 score are all optimized using the ensemble model. Sensitivity is slightly lower than in several other models, but still very high (at 0.9492).

The ensemble model has therefore been selected and will be used against the validation set to determine final accuracy and F1 scores.

RESULTS

The ensemble model has been judged to be the best model to use for predicting income. Now this ensemble model will be tested against the validation set to determine final accuracy numbers.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction <=50K >50K
##      <=50K  2343  336
##      >50K    129  449
##
##                Accuracy : 0.8572
##                  95% CI : (0.8447, 0.8691)
##     No Information Rate : 0.759
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5712
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9478
##             Specificity : 0.5720
##          Pos Pred Value : 0.8746
##          Neg Pred Value : 0.7768
##              Prevalence : 0.7590
##          Detection Rate : 0.7194
##    Detection Prevalence : 0.8225
```

```
##       Balanced Accuracy : 0.7599
##
##        'Positive' Class : <=50K
##
```

Final Accuracy

```
## [1] 0.8572306
```

Final F1 Score

```
## [1] 0.9097263
```

Final F1 and accuarcy scores are not as high as with the training test set which is not surprising since the training test set was used for cross-validation and model selection. However, the F1 score is still above 0.9, which is a good sign for the robustness and accuracy of the developed model.

As seen in the confusion matrix, sensitivity and specificity scores are also slightly lower than they were with the training test set.

Unfortunately, specificity remains a real concern, with a value just below 0.6. While our model is highly successful in correctly identifying those earning $50,000 or less, it is much less successful at identifying those who earn more than $50,000. It is important to note that the demographic predictors are most responsible for the improvements in specificity, since the monetary and achievement models had specificity levels at strikingly low levels (below 0.3). It appears that, while predictors such as education, occupation, and capital gains are effective at telling us when someone is not making much money, additional factors such as age and marital status, and to a lesser extent race and gender, are better equipped to alert us to the fact that an individual is making a higher income. This could make sense given that those advancing in age (up until retirement) are more likely to simultaneously be advancing their careers (and thus their incomes). Being married may, as discussed above, also be an indicator for traditional life and career progression or it may simply provide a support net that makes it easier for individuals to make certain educational and occupational decisions.

Overall, after examining variable importance for multiple models used in the ensemble, the most influential variables appear to include: age, marital status, occupation, education, hours worked per week, capital gains, and capital losses. Sex, race, and workclass did not appear as the most influential predictors in any of the combined models. The fact that workclass was minimally influential makes sense since the identified work sectors are such broad categories and, as shown in the developed models, the specific occupation done within a certain sector is much more correlated with income. The fact that sex and race are not key variables in the combined models is great news in terms of societal equity. However, exploratory data analysis (and the rpart demographic model) did show significant gender and racial differences in income. More study is needed to see what impact sex and race have on income when other variables such as education and occupation are held steady.

CONCLUSION

The developed model is reasonably effective at predicting income and provides compelling evidence that demographic, achievement, and monetary predictors are essential to successful predictions. The model's final F1 score is impressive, but unfortunately the specificity score remains low.

One reason for this, which is also a key limitation of this model, is that no information is provided in the dataset in regards to state or region. Economies (including average salaries, housing costs, etc.) vary greatly from region to region and even from urban areas to rural areas within the same region. It may be that the specificity rate (correctly predciting higher-income individuals) which we were unable to boost above 0.6 could be greatly improved by adding one or more geographic predictors. It would be interesting to

explore how much of the still unexplained variability in the dataset can actually be explained by geographical predictors.

Another limitation to this model is that the data is 25 years old, which, especially for social and economic data, is considerably out of date. Since a new census is about to be conducted in the U.S., it would be interesting to use this new data (once generated) to build a new model and then compare/contrast these models to see key similarities and differences. I would be interested to know whether or not the same predictors are still the most influential. A comparable study using up-to-date data could provide interesting insight into societal/economic trends.

One avenue for future study would be looking deeper into the connection between marital status and income. As discussed above, marital status could be an indicator of more traditional life and educational choices which may in turn yield higher incomes. Non-married individuals may also tend to be more vulnerable members of society: LGBTQ individuals, unmarried parents, etc. More research could reveal a clearer picture of the correlation between being married and having a higher income.

Finally, further study should examine income as a continuous, rather than binary, variable. ALthough the binary analysis conducted here does reveal valuable information, it would be interesting to look at the complexities of income in more depth. Someone making $20,000 is very different from someone making $49,000. In the same manner, someone making $51,000 is not nearly as high income as someone making $200,000. The next test for our model would be examining how well it can be extrapolated to predict exact income levels.