Classification Models for San Francisco OkCupid Dating Profiles- Explorations and Case Studies

Co Sou Statistical Learning S2018

The dataset studied here consists of dating profiles from Okcupid of users from the San Francisco Bay Area. It was scraped, with permission to be shared, on June 30, 2012. It formed the basis for a then published paper "OkCupid Profile Data for Introductory Statistics and Data Science Courses" (2015) by Albert Y. Kim and Adriana Escobedo-Land. This paper examines mostly descriptive procedures and exploratory data analysis of the data. Here, we apply predictive modeling in hopes of adding insights to the existing literature. This study will be done in the various case studies that employ techniques of statistical learning, with emphasis placed on classification goals.

But first, some preliminary observations about the data and its cleaning. On first glance, we notice that while there are many categorical variables with many levels of factors. For the sake of simplicity we have to condense them into a more manageable number of factor levels. Though this will cause some lost of information, the benefits are reduction of redundancy and easier to read regression output. Secondly, a problem with users on a dating website has to do with selection bias. Obviously, users who go on OkCupid would be seeking potential mates but as it turns out many non-single users are also present, for whatever reason they might have, though there aren't as many as single users. This causes an unbalanced ratio of single and non-single users. When using this binary variable, we must be careful about its restriction.

We first focus on Logistic Regression. The goal is to determine whether the numerical variables in the data can predict whether a user is listed as Single or Not Single. We infer that Age, Income, and Height might contribute to the response. For example, it is possible that younger people are more open to non-monogamous relationships? Is there any truth to the stereotype of the rich playboy who dates multiple people simply because he has the money to do so (Income is capped at a half a million dollars for this analysis)? Are taller people, a desirable physical trait, more proned to dating more people along the side?

This is logistic regression simply on the data:

```
##
## Call:
## glm(formula = status ~ income + age + height, family = "binomial",
##     data = Cupid2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8803   0.3948   0.4292   0.4632   0.4915
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.858e+00  6.087e-01   3.053  0.00227 **
## income       3.778e-06  8.853e-07   4.267 1.98e-05 ***
## age          8.407e-03  3.786e-03   2.220  0.02640 *
## height      -3.954e-04  8.658e-03  -0.046  0.96358
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6620.2  on 10982  degrees of freedom
## Residual deviance: 6583.1  on 10979  degrees of freedom
```

```
## AIC: 6591.1
##
## Number of Fisher Scoring iterations: 5
```

```
##   (Intercept)        income           age        height
##  1.858366e+00  3.777670e-06  8.406956e-03 -3.953654e-04
```

```
##
## pred.log     Not Single Single
##   Not Single        983  10000
```

We see that Income and Age are statistically significant predictors while Height is not. This model, however, has an extremely high AIC so it shouldn't be used as a final conclusion. The confusion matrix returned was inconclusive, indicating overfitting of the model to the data.

Instead we gain more accrucacy by splitting the data into training and testing sets.

```
##
## Call:
## glm(formula = status ~ income + age + height, family = binomial,
##     data = Cupid2, subset = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4904   0.3862   0.4370   0.4831   0.6753
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.802e+00  7.562e-01   3.705 0.000211 ***
## income       8.681e-06  1.351e-06   6.424 1.33e-10 ***
## age         -6.608e-02  1.076e-02  -6.139 8.29e-10 ***
## height       1.167e-02  1.018e-02   1.146 0.251787
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4696.1  on 7356  degrees of freedom
## Residual deviance: 4625.5  on 7353  degrees of freedom
## AIC: 4633.5
##
## Number of Fisher Scoring iterations: 5
```

```
##             StatusUnder35
## glm.pred     Not Single Single
##   Not Single        262   3229
##   Single              5    130
```

Like before, the predictors Age and Income are statistically significant while Height is not. This version has a lower AIC, which is expected and desirable. This procedure produces a full confusion matrix. The model correctly predicted 392 out of 3626 observations, for a success rate of about 11%, which is terrible. The selection bias of the status response variable might have been a major contributing factor to this. But at least, we gain some insights: We conclude that the Height predictor is a weak contributor to whether a user is not single. Also, there might be some predictive powers to the Age and Income variables for this type of question. Further exploration might be worth it.

Logistic regression might not be appropriate to all model cases. But there are cases where it might perform better. Here we want to predict whether a user is Male or Female based on the same numerical variables as

above.

```
##
## Call:
## glm(formula = sex ~ age + income + height, family = binomial,
##     data = Cupid2, subset = train2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.7704  -0.2711   0.2420   0.5120   6.8862
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.831e+01  1.009e+00 -37.952  < 2e-16 ***
## age         -2.709e-02  9.725e-03  -2.786  0.00534 **
## income       1.449e-05  1.312e-06  11.049  < 2e-16 ***
## height       5.805e-01  1.460e-02  39.761  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8504.7  on 7356  degrees of freedom
## Residual deviance: 4988.8  on 7353  degrees of freedom
## AIC: 4996.8
##
## Number of Fisher Scoring iterations: 6
```

```
##          SexUnder35
## glm.pred3    f    m
##         f  970 2521
##         m   70   65
```

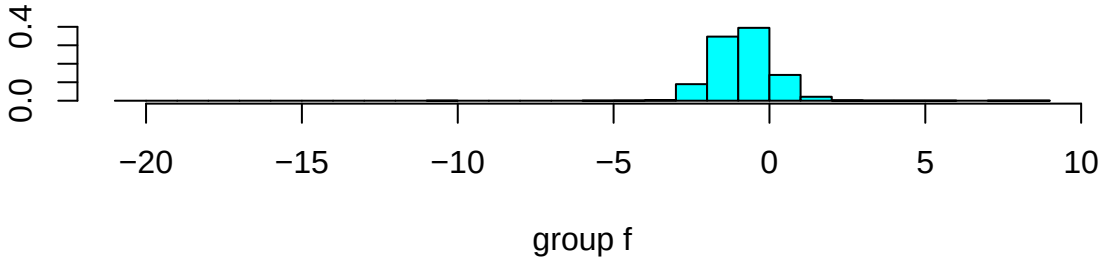
All predictor variables are statistically significant. The confusion matrix for this model gives 1035 correct predictions out of 3626 observations, for a success rate of 28.5%. Better than the above model. Recall that logistic regression is sensitive to the way the data was split for training and testing. That could be rexamined and perhaps, a more evenly balanced training and testing split should be made.

The model above seems more realistic and pausible. We consider using it as the basis for other classification techniques. In particular, Linear Discriminant Analysis is used in hopes of more accurate predictions. Here, LDA has the advantages of dealing with smaller number of observations and better than Logistic Regression for when there is a clear separation between classes of response. We know that Sex is strictly divided with little ambiguity, for the most part.

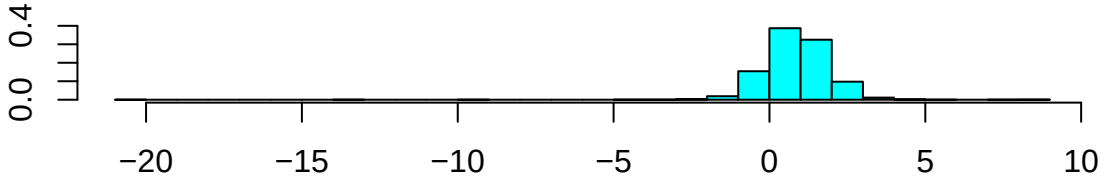
Here is a simple LDA on the data:

```
## Call:
## lda(Cupid2$sex ~ age + income + height, data = Cupid2)
##
## Prior probabilities of groups:
##         f         m
## 0.2720568 0.7279432
##
## Group means:
##        age   income   height
## f 33.19378 48423.69 65.24967
## m 32.42176 66950.59 70.51132
```

```
##
## Coefficients of linear discriminants:
##                  LD1
## age    -1.125630e-02
## income  3.770707e-06
## height  3.128835e-01
```



group f



group m

Notice that there is a 1:3 ratio of female to male, a result of income filtering in the data cleaning. It is possible that women are less likely to share their income rather than men, who understand that (a high) income is a desirable trait in dating.

The group means output suggests that: For both men and women, there is a tendency for Age,Income, and Height to have a positive effect on predicting the respective sex. The coefficients of Linear Discriminants are of interest. They are used to multiply the corresponding predictor values to produce a "score", which is then used to compute the posterior probability to form the decision rule. Here is some interpretation: If age(-1.125630e-02) + income(3.770707e-06) +height(3.128835e-01) is high, then it predict the correct sex and vice versa. The plot is examined. For female group, we see that most of the distribution of probabilities are below zero (low) and thus, is not effective in predicting the correct sex. But for male group, the distribution is skewed above zero and thus, we conclude that the model is effective in predicting the male sex.

Now we use training and testing sets to predict and compare with Logistic Regression. We hope that the success rate is better than that of Logistic Regression.The confusion matrix returned by the model is:

```
##           SexUnder35
## lda.class    f    m
##         f  687  151
##         m  353 2435
```

The success rate is 3122 out of 3626 or 86%. This is a major improvement than that of Logistic Regression using the same training and testing set. We prefer the use of LDA over Logistic Regression for modeling this case.

Finally, we turn our attention to the nonparametric method of classification. In LDA and Logistic Regression, the normal distribution is assumed to be the underlying distribution. Is there any justification of this? By convenience we use the normal distribution to build initial models. By contrast, nonparametric methods

don't rely on such strong assumption.

Here is the confusion matrix for the KNN approach with 1 number of nearest neighbors.

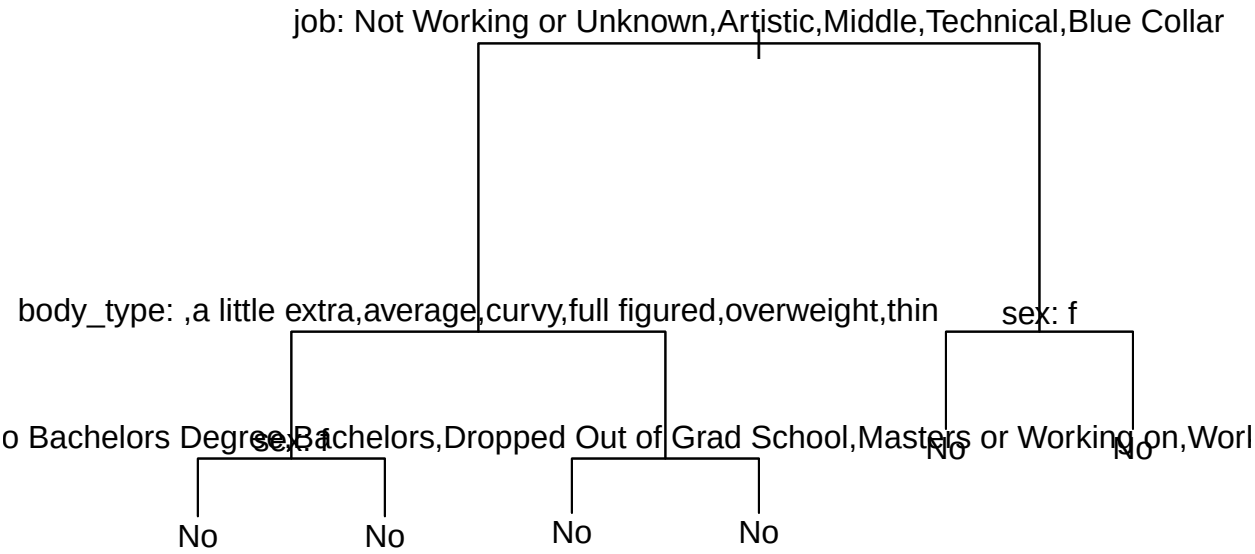
```
##          SexUnder35
## knn.pred    f    m
##        f  728  305
##        m  312 2281
```

The success rate is 3009 out of 3626, or 83%. This is better than Logistic Regression and is slightly worse than that of LDA. Of course the choice of number of nearest neighbors can have an effect on the prediction success.

In conclusion, we prefer the use of LDA for this particular case study.

The dataset is rich in categorical variables with a diverse range of factor levels. In the previous case studies, we did not take advantage of them in favor of using only the numerical variables for prediction. Here we employ simple classification tree.

```
##
## Classification tree:
## tree(formula = High ~ education + job + offspring + sex + drugs +
##     body_type + Older + White + diet + drinks + orientation +
##     religion, data = Cupid3)
## Variables actually used in tree construction:
## [1] "job"       "body_type" "sex"       "education"
## Number of terminal nodes:  6
## Residual mean deviance:  0.1571 = 1725 / 10980
## Misclassification error rate: 0.01794 = 197 / 10983
```

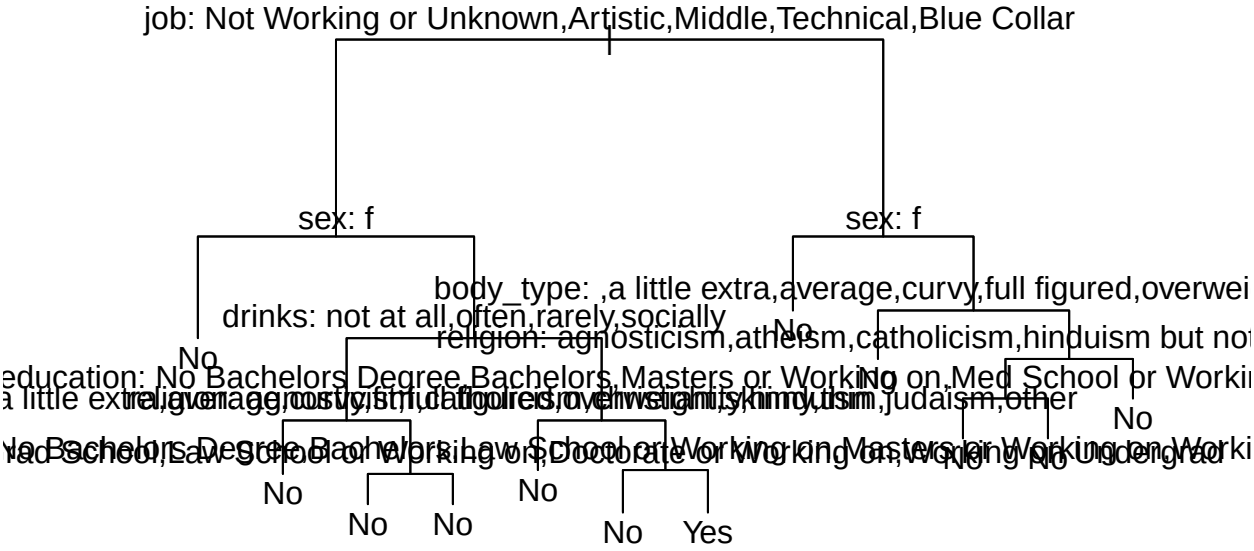


This simple classification tree identifies two prominent predictor variables that form the internal tree structure: job type and body type. In particular, only the strongest factor levels associated with the predictors determine the decision rule. Qualitively, the conclusion is that women who are not working or have non-professional job roles with less than desirable body types and who don't have professional degrees will strongly be predicted as having not high income (high being $100000). Many predictors were used in the initial model but only four predictors were deemed to have strong predictive properties. The misclassification error rate is about 1.7%, which is very desirable.

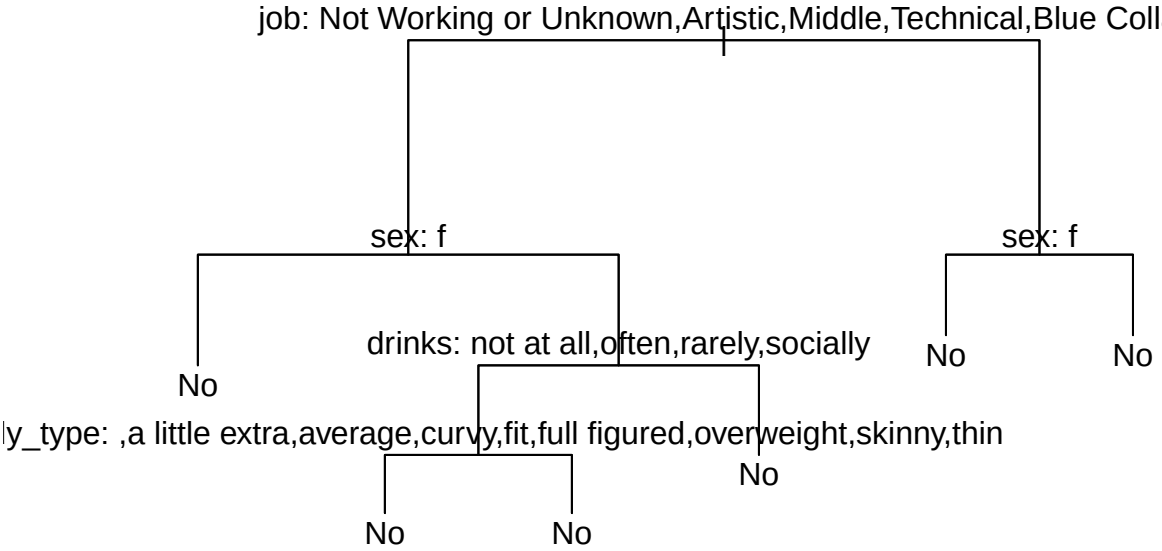
The classification tree needs to be used in conjunction with training and testing sets to overcome overfitting. Here we fit regression models on the tree doing just that. Notice that new predictor variables reveal themselves to be significant:

```
##
## Classification tree:
## tree(formula = High ~ education + job + offspring + sex + drugs +
##     body_type + Older + White + diet + drinks + orientation +
##     religion, data = Cupid3, subset = train.tree)
## Variables actually used in tree construction:
## [1] "job"       "sex"       "drinks"    "body_type" "education" "religion"
## Number of terminal nodes:  12
## Residual mean deviance:  0.134 = 734.2 / 5479
## Misclassification error rate: 0.01748 = 96 / 5491
```

```
##           High.test
## tree.pred2   No  Yes
##        No  5395   97
##        Yes    0    0
```



Just like before job type, body type, sex are prominent predictors but this tree also suggests that drinking behavior and religion as significant. The regression tree looks cluttered. We consider pruning it to get an optimal tree with weaker predictors removed.



The above is a more desirable visualization of the decision ruling.

A remark about this procedure: This technique reveals some strong insights into the data. In particular, it focuses only on female and their relationship to having high income or not but reveals nothing discriminatory about males at all. More analysis needs to be done if we want predictive results about males. Also, since it only gives a glimpse into scenarios of possible models, it should be best used as aide for exploratory data analysis.

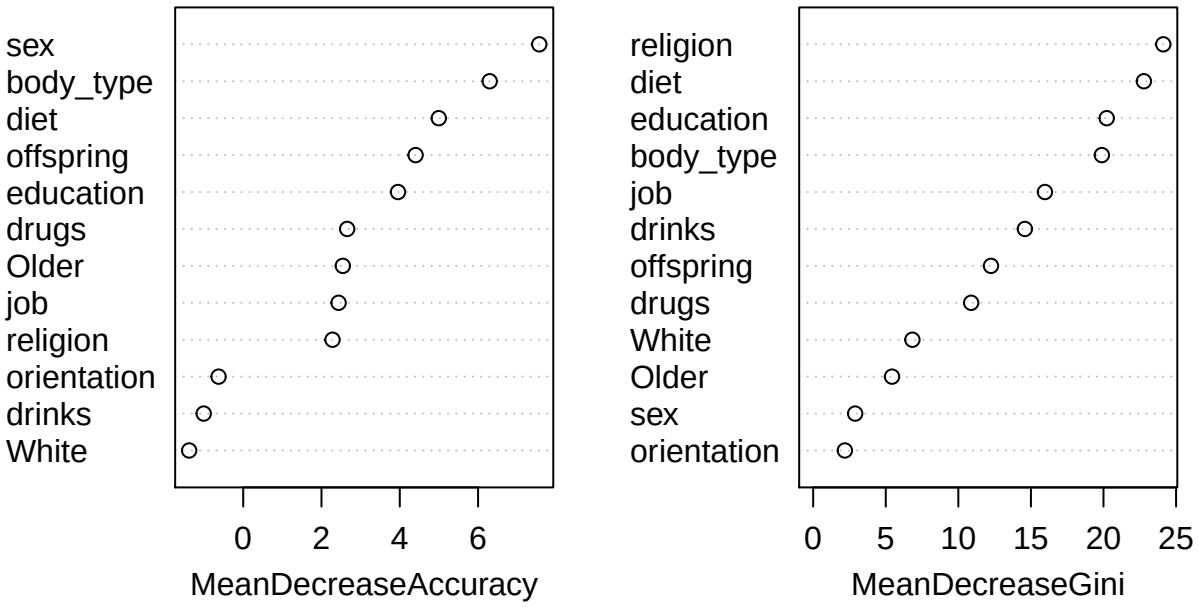
Continuing with this case study we can extend it by performing bagging and boosting. Recall that decision tree methods have high variance and is a by product of the splitting mechanism. Bagging takes advantage of cheap computing power to give lower variance models via bootstrapping. Applying bagging on the same model framework gives the following result:

```
## randomForest 4.6-12

## Type rfNews() to see new features/changes/bug fixes.

##
## Call:
##  randomForest(formula = High ~ education + job + offspring + sex +      drugs + body_type + Older + W
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 1.82%
## Confusion matrix:
##       No Yes class.error
## No  5391   0           0
## Yes  100   0           1

##                     No        Yes MeanDecreaseAccuracy MeanDecreaseGini
## education    0.9423609 11.6230357            3.9544993        20.221723
## job         -1.4949539 14.9420766            2.4364798        15.966432
## offspring    3.6906512  3.4598150            4.4037340        12.249363
## sex          5.1608643 14.5957966            7.5633081         2.891515
## drugs        2.9433248 -0.9055793            2.6554964        10.883893
## body_type    4.3406440  9.6997386            6.2971431        19.885201
## Older        1.9186945  3.0238100            2.5445030         5.432575
## White       -0.2675931 -4.2963917           -1.3806056         6.833711
## diet         4.4222788  3.3856971            4.9957308        22.782558
## drinks      -1.5831991  2.2697417           -1.0092736        14.584094
## orientation -0.8816370  0.8999239           -0.6288446         2.183299
## religion     0.5766492  6.3764669            2.2808109        24.120302
```



We see a very small OOB error rate of 1.82% compared to the residual deviance of 0.134 for the regression tree model. The graphical output suggests a similar profile of predictors that have high predictive powers. Once again, it is stressed that this is more useful as an exploratory aide. These predictors should be compared with the decision tree models and we can identify common predictors and use them as the foundations for other questions to be asked.

Recall that Bagging is a special case of the more general Random Forest technique, which uses all predictor variables in the data. This method is not compatible with the question being asked in this case study since we only selected multiple but not all predictors to build the model. We therefore skip this discussion.

Boosting, yet another tree based method, can be applied in this framework. Unlike the previous, this one is particularly useful to reduce the effects of overfitting.

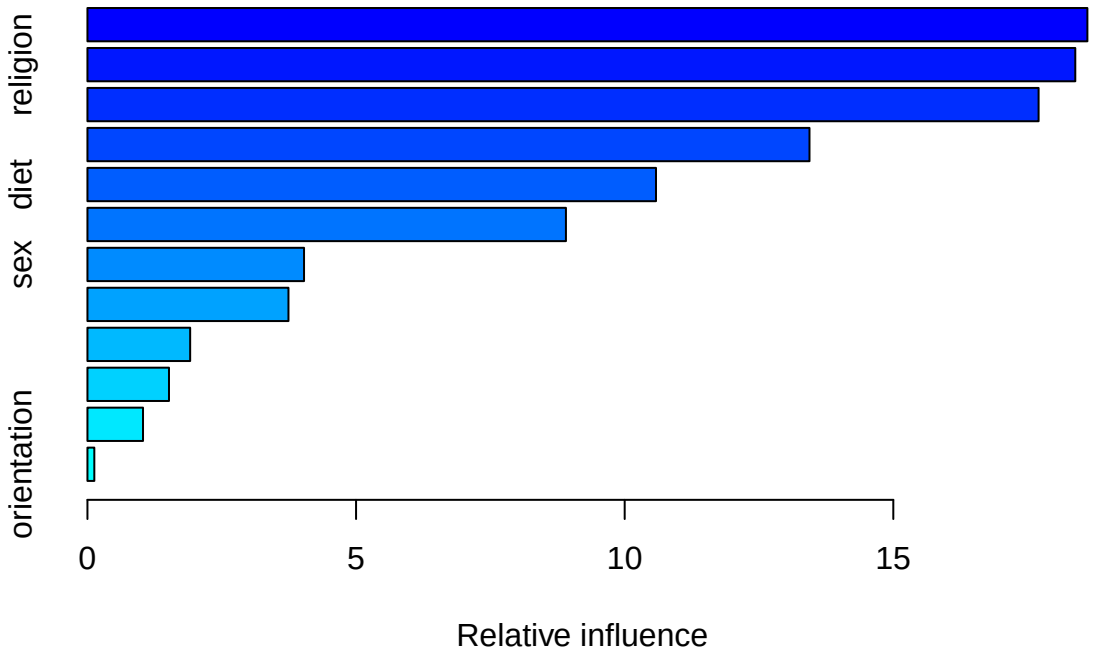
```
## Loading required package: survival

## Loading required package: lattice

## Loading required package: splines

## Loading required package: parallel

## Loaded gbm 2.1.3
```




```
##                     var    rel.inf
## body_type     body_type 18.6127117
## religion       religion 18.3874748
## education     education 17.7044107
## job                 job 13.4399946
## diet               diet 10.5828759
## drinks           drinks  8.9066662
## sex                 sex  4.0315761
## offspring     offspring  3.7415555
## Older             Older  1.9120381
## drugs             drugs  1.5180857
## White             White  1.0345690
## orientation orientation  0.1280418
```

Once again, the results show a similar profile of predictors that are influential to modeling the response. This one reveals a very strong predictive effect for the religion variable, which wasn't as strong in previous models, and demoted the job type variable in terms of significance. Beware of the influence plot shown, as its y-axis is not labeled properly; see the actual summary table instead. Also consider that this method is computationally expensive and can take a long time to render for large data size.

In conclusion, we are privy toward using sex, body type, education, drinking behavior, and religion type as all possible strong predictors to predicting whether someone has a high income or not. The decision tree methods only discriminates against the female sex while providing no decision rule for males. Bagging and Boosting don't have this restriction; though they also treat sex as a whole predictor instead of levels and don't provide insight to the predictive powers of the individual sex types. In future work, we should clean the data to more strongly separate the two into separate variables to see which sex level has more predictive effects on the response.