Mass 211: Decision Trees

Code

This report surveys two popular tree-based additive regression algorithms–Random Forests and Gradient Boosting–and evaluate their performance in predicting the 2-1-1 demands.

Methods

I first filter out ZIP codes with at least 5 calls and 200 people, then regress log number of calls per 1,000 people (`log( # of calls * 1000 / total population)`) on 65 demographic variables.

Calls related to Call2Talk and Suicide Prevention Hotline are excluded since they are more vulnerable to skewness caused by repetitive calls.

Code

Random Forest

Random forests repetitively sample a subset of variables and observations to build a tree, then averages the node values to find the best variable/value for splitting the trees.

By smoothing over different trees, Random Forests is resilient to overfitting. It is robust to an increase of noise variables, and generally performs better with many noise variables than other bagging models such as Gradient Boosting.

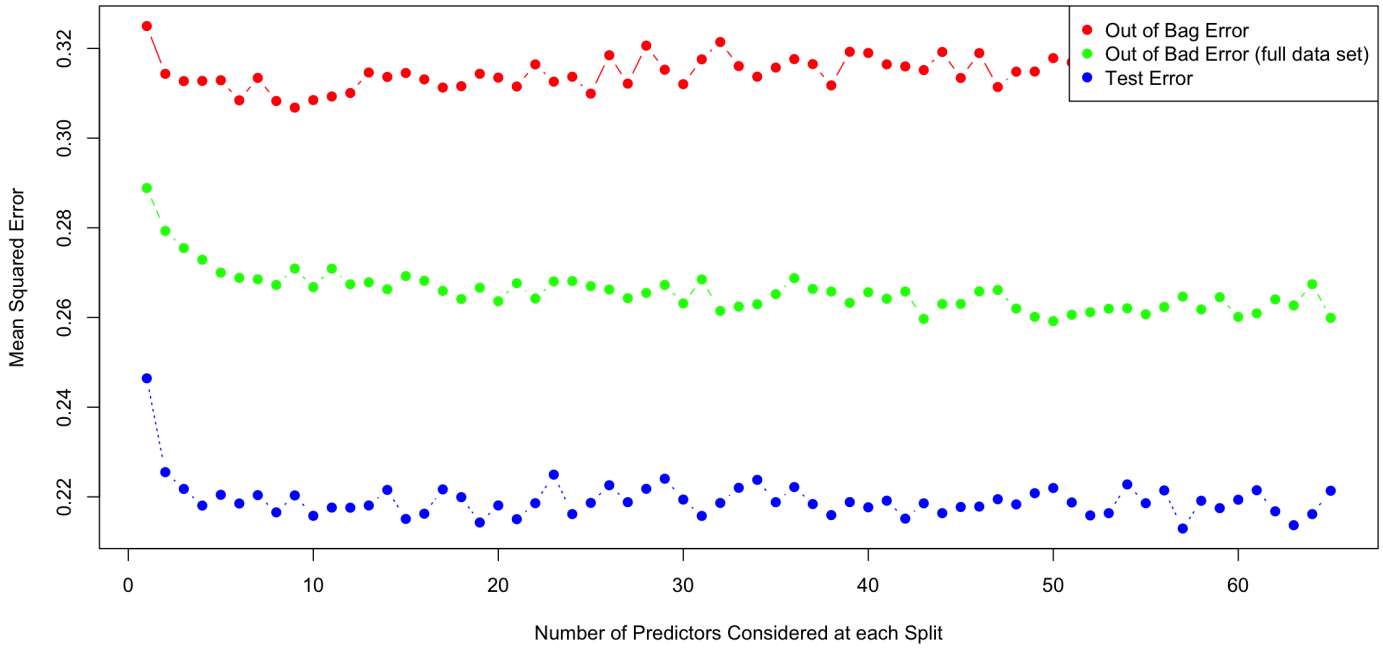
That been said, we still have a few hyperparameters to tune: the size of sample observations used during each iteration, the number of variables to draw, the number of trees to grow (# of iterations), and the minimal/maximal size of terminal nodes (size of the trees).

Number of randomly selected splitting variables m

Among these parameters, perhaps the most imporatant one is the number of random variables (predictors) selected when growing a tree. Following plot shows the Out of Bag Error (for training data) and Test Error (for validation data) for different number of variables selected.

Code

Code

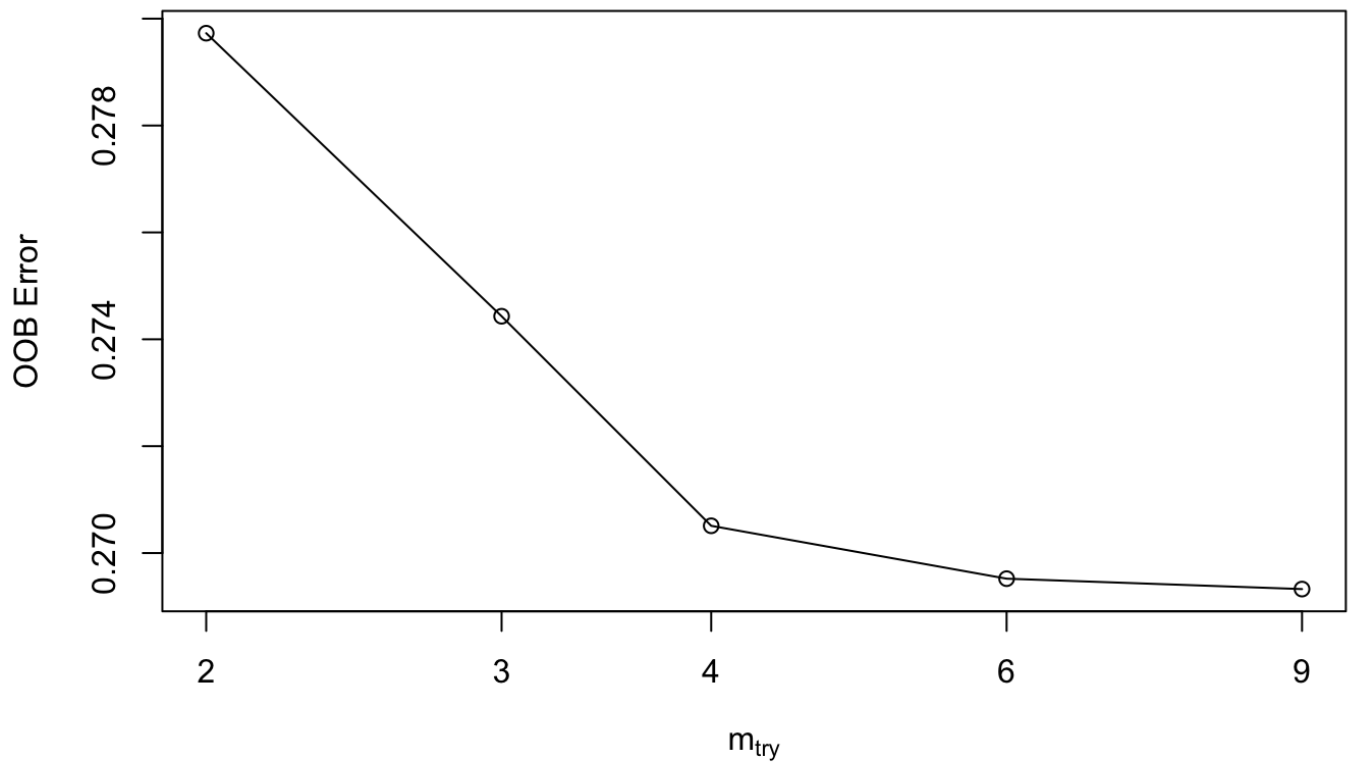


I have run above test several times as well as with different training/validation split, the results are different from time to time, but in general the errors reach their minimum at around 10 variables.

The `randomForest` package actually provides a function `turnRF` for the same purpose.

Code

```
mtry = 2  OOB error = 0.2797281
Searching left ...
Searching right ...
mtry = 3    OOB error = 0.2744315
0.0189346 0.001
mtry = 4    OOB error = 0.2705102
0.01428874 0.001
mtry = 6    OOB error = 0.2695209
0.003657343 0.001
mtry = 9    OOB error = 0.269326
0.0007229654 0.001

Call:
 randomForest(x = x, y = y, mtry = res[which.min(res[, 2]), 1])
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 9

          Mean of squared residuals: 0.2710415
                    % Var explained: 61.72
```

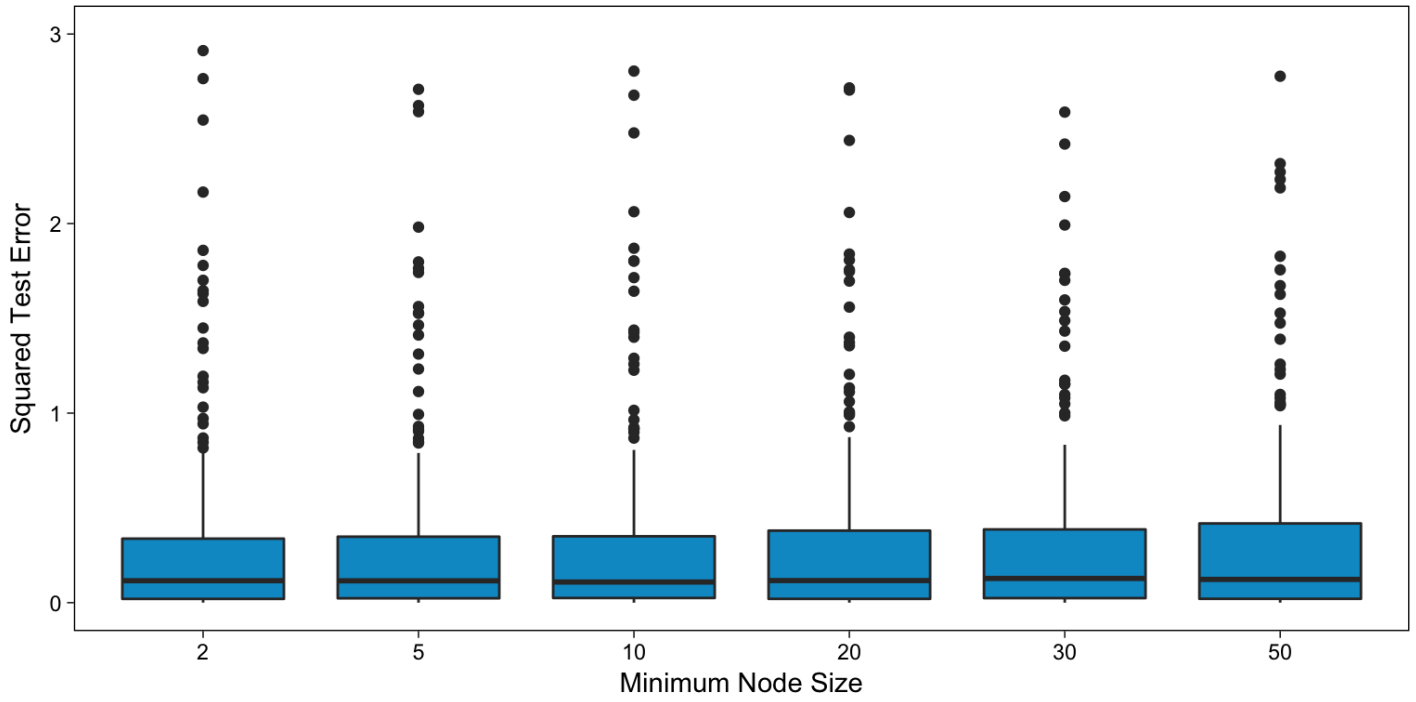


It shows a more or less the same result– `10` seems to be the most reasonable choice.

Minimal/Maximal node size

The minimal and maximal node size defines the depth of the tree. The Elemenets of Statistical Learning (https://web.stanford.edu/~hastie/ElemStatLearn/) recommends let the tree grow to its maximal, because it seldom make much difference in terms of speed and performance.

The minimal node size may have an effect, but the impact is different depending on the dataset.



Above graph shows the squared test error in regards to different minimum node sizes. The larger the node size, the shallower the tree is. It is clear that minimum node size doesn't matter in our case, neither.

Variable Importance

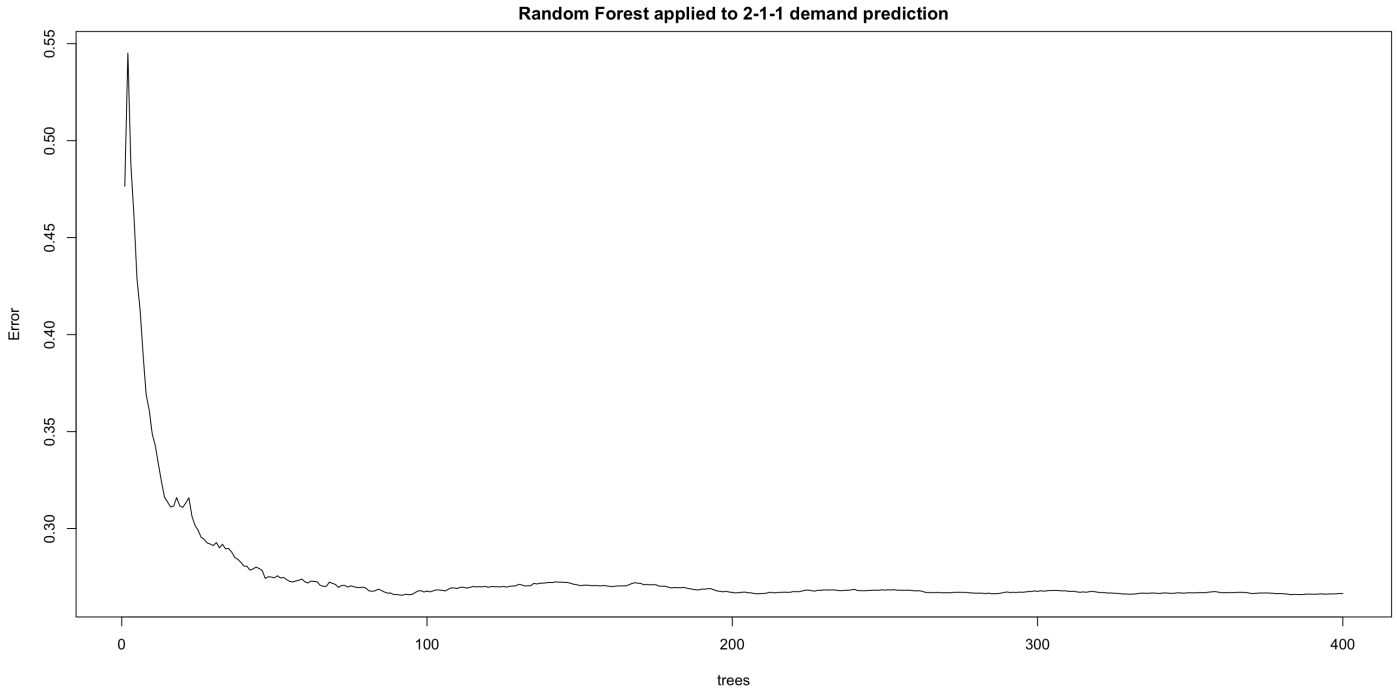
With the parameters recommended by above analysis, we get a regression model that explains 62.38% of the variance and produces an RMSE of 0.266. Note that this is after we removed ZIP codes with NA's in any of the variables, and ZIP codes with no calls at all.

One of the most useful application of Random Forest is to find variable importance. It can be evaluated by the increase in MSE when values of a variable are shuffled (%IncMSE), or the increase in node purities after using a variable for splitting the trees (IncNodePurity). We often only consider the former metric while evaluating the variables.

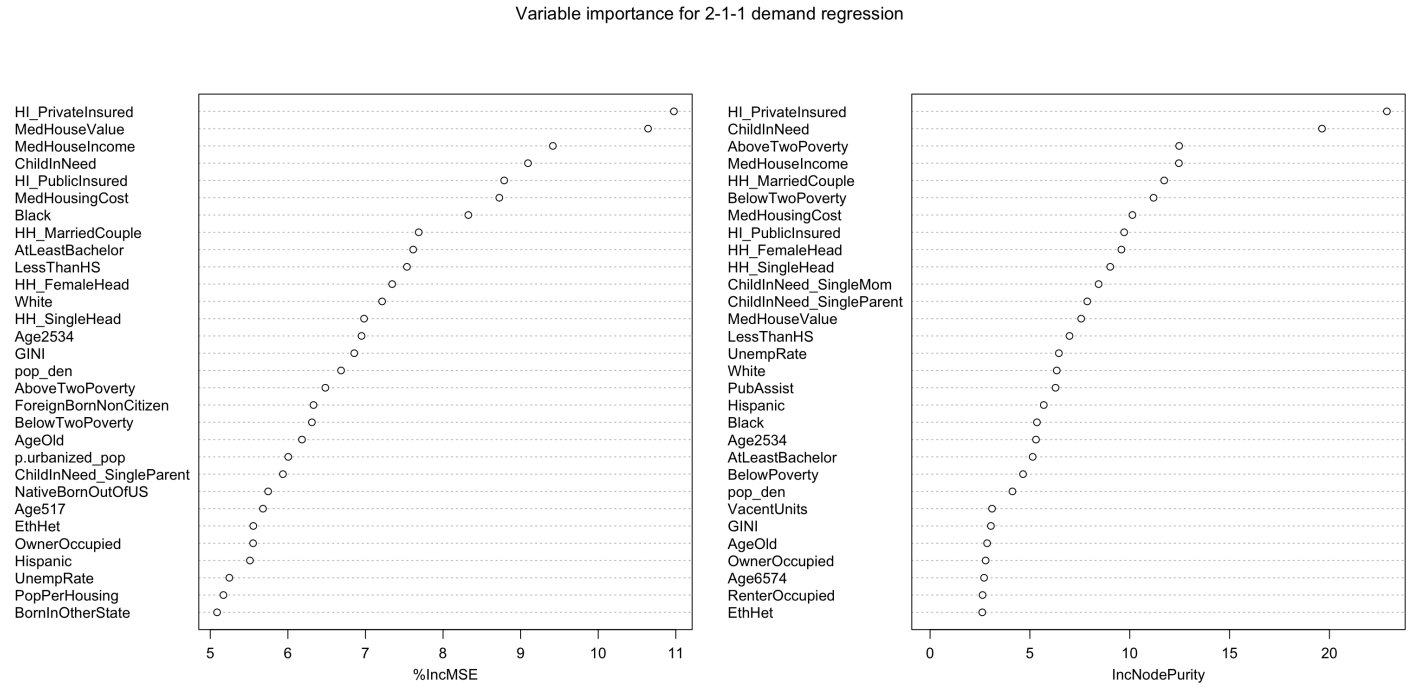
Code

```
Call:
 randomForest(formula = calls ~ ., data = dat, ntree = 400, mtry = 10,      importance =
TRUE)
               Type of random forest: regression
                     Number of trees: 400
No. of variables tried at each split: 10

          Mean of squared residuals: 0.2664179
                    % Var explained: 62.38
```

Code



Code



Surprisingly, the proportion of people who have private health insurance turns out to be the most important variable.

Possible explanations for the top variables are:

HI_PrivateInsured

This is an indicator for “quality employment”. People with private health insurance mostly get their health insurances from employer as an employment benefit. Employment with health care benefits tend to be those high-paying permanent jobs. Therefore high-paying jobs is a reflection of many other factors–such as good education, rich neighborhood, being of specific race, etc.

MedHouseValue

Median value of the property. This variable indicates how “rich” a neighborhood is. This seems to be, according to the model, a better indicator than the actual income of people who live in the neighborhood. One explantion is that some people with low income and high demand in services live in places with relatively high property values, too—namely, poor people who rent and live in the inner city.

MedHouseIncome

This one is straightforward.

ChildInNeed

Percentage of children living in a household that is receiving income support.

MedHousingCost

Median monthly housing cost, including morgages and gross rent. This variable also bears a sense of urbanity in it.

In addition to above two variables, we have also MedRentAsIncomePct—the median value of rent as percentage of house income. Surprisingly, this variable didn’t even make it to the Top 30.

Black

THe proportion of the residents being black. It is not a surprise that race is a huge factor in shaping the landscape of service needs.

Age2534

Age group 23 to 34 year old. Young parents are more likely to use one of the major services 2-1-1 provides: tracking the status of their application for child care assistance.

HH_MarriedCouple

Household type: married couple. Families with married couples are more stable and more resilient to financial difficulties.

AtLeastBachelor

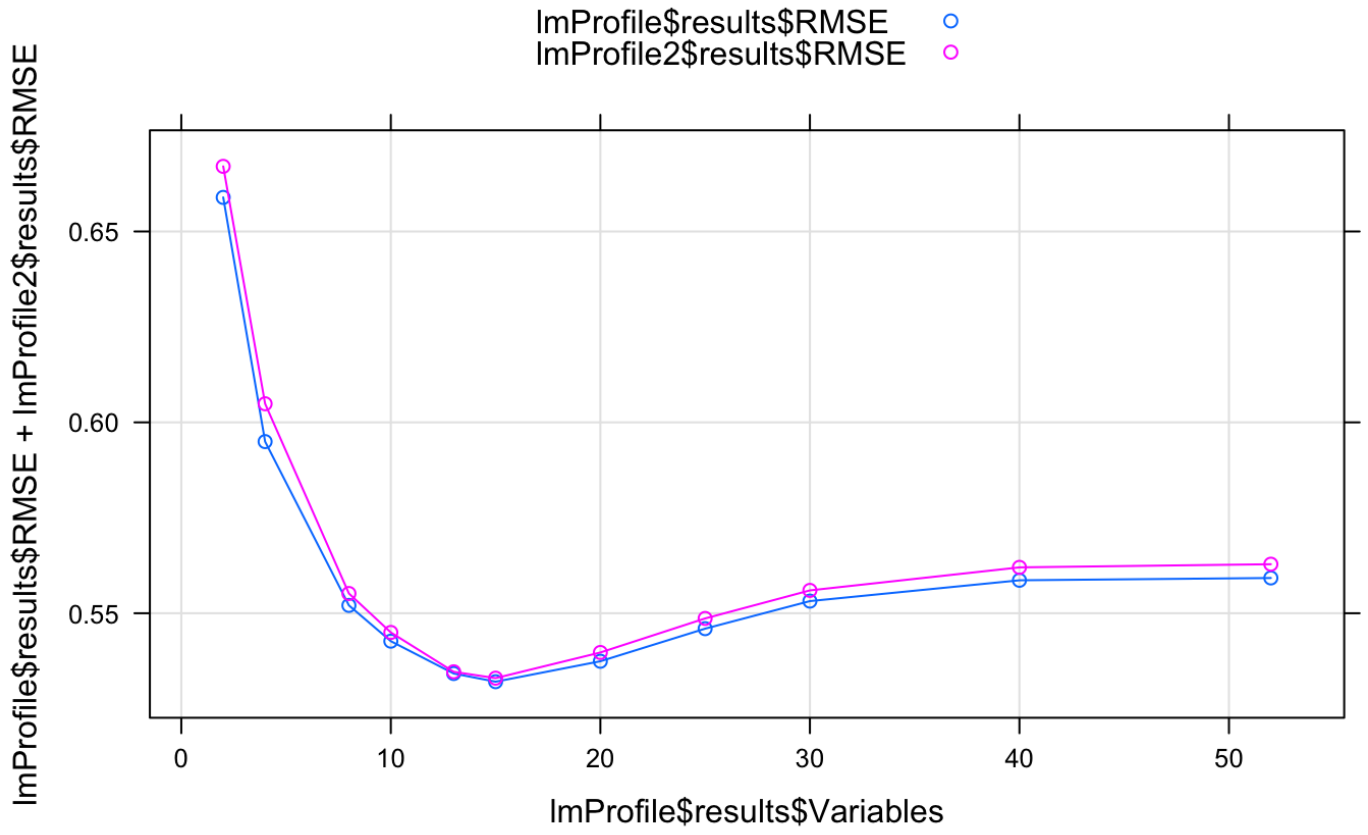
Percentage of people with at least a 4-year Barchelor’s degree for population 25 years and over. Education has implications in income and social stigma of using public services.

Comapre with Linear Regression

For comparison, using a Linear Regression model and Backwards Selection with bootstrap resampling, we identifies the Top 5 variables as: `White`, `Hispanic`, `Black`, `OwnerOccupied`, and `MedHouseValue`.

Code

```
Recursive feature selection

Outer resampling method: Bootstrapped (550 reps)

Resampling performance over subset size:

 Variables   RMSE Rsquared    MAE  RMSESD RsquaredSD   MAESD Selected
         2 0.6589   0.3957 0.5170 0.06505    0.11221 0.05613
         4 0.5950   0.5081 0.4588 0.05196    0.08417 0.04199
         8 0.5521   0.5781 0.4201 0.04718    0.06678 0.02800
        10 0.5427   0.5926 0.4136 0.04542    0.06303 0.02750
        13 0.5342   0.6064 0.4091 0.04432    0.05902 0.02704
        15 0.5321   0.6104 0.4078 0.04563    0.06011 0.02768        *
        20 0.5374   0.6053 0.4128 0.04331    0.05750 0.02702
        25 0.5460   0.5960 0.4198 0.04025    0.05434 0.02638
        30 0.5532   0.5879 0.4247 0.04114    0.05497 0.02687
        40 0.5586   0.5821 0.4289 0.04028    0.05390 0.02706
        52 0.5592   0.5814 0.4293 0.04044    0.05426 0.02703

The top 5 variables (out of 15):
   White, Hispanic, Black, OwnerOccupied, MedHouseValue
```



Race seems to be the most prominent factor, but it actually didn't turn to be conclusive in the best fitted model.

Code

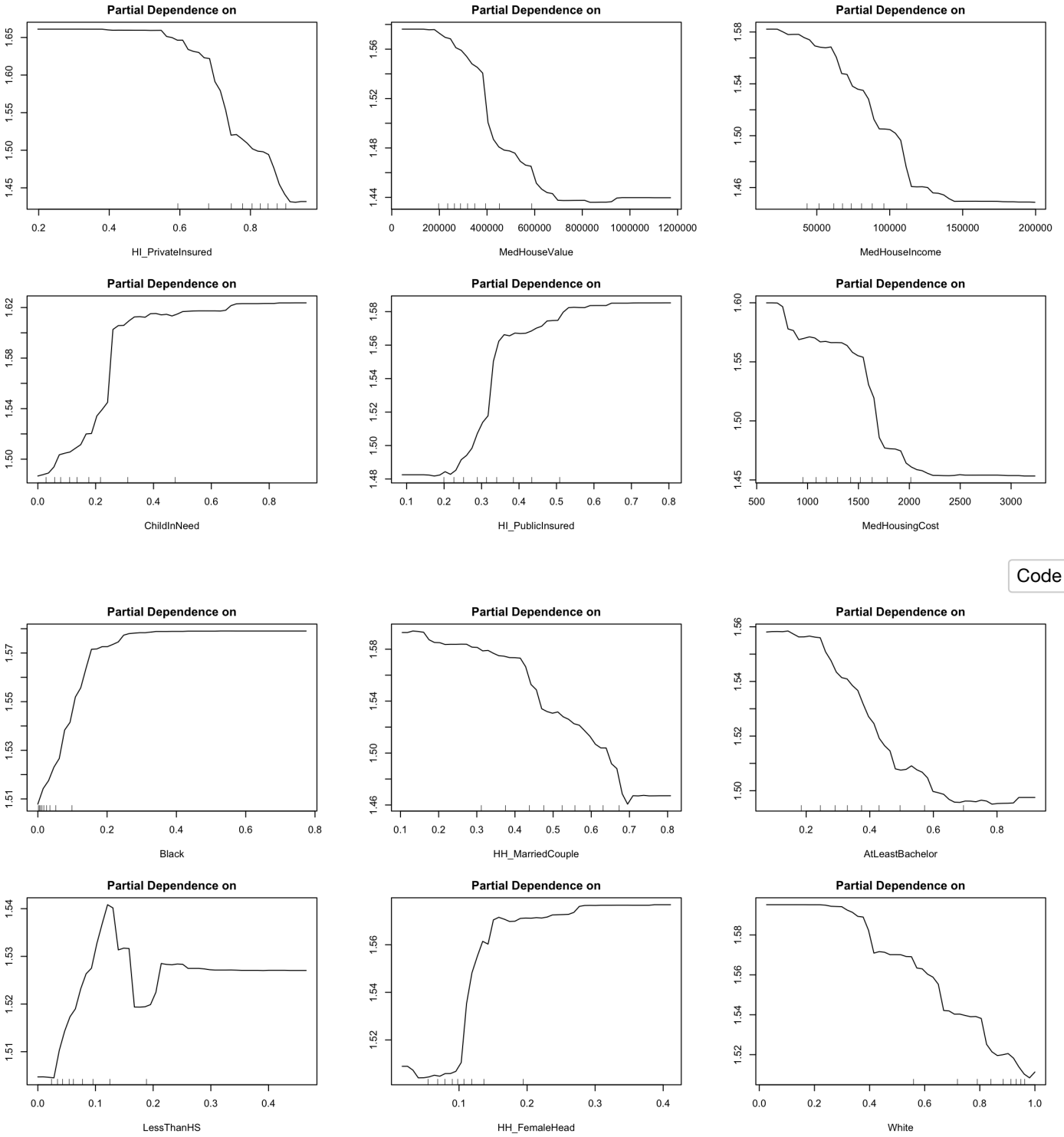
```
Call:
lm(formula = y ~ ., data = tmp)

Residuals:
     Min       1Q   Median       3Q      Max
-1.58493 -0.27802 -0.01279  0.30032  1.34309

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.529792   0.023011  66.480  < 2e-16 ***
White            -0.227939   0.292218  -0.780 0.435829
Hispanic         -0.179677   0.172504  -1.042 0.298227
Black            -0.003182   0.124119  -0.026 0.979558
MedHouseValue    -0.277045   0.046361  -5.976 5.02e-09 ***
OwnerOccupied    -0.302235   0.068236  -4.429 1.22e-05 ***
AtLeastBachelor  -0.233340   0.065371  -3.569 0.000401 ***
BornInState      -0.069511   0.048073  -1.446 0.148964
Asian            -0.119920   0.092844  -1.292 0.197223
HH_LiveAlone      0.248851   0.069730   3.569 0.000402 ***
HH_MarriedCouple  0.268484   0.104674   2.565 0.010678 *
Age6064           0.218490   0.028329   7.713 9.66e-14 ***
LessThanHS       -0.178061   0.057455  -3.099 0.002076 **
EthHet            0.159168   0.063508   2.506 0.012593 *
HH_SingleHead     0.266207   0.070988   3.750 0.000203 ***
p.urbanized_pop   0.165489   0.029690   5.574 4.56e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4722 on 405 degrees of freedom
Multiple R-squared:  0.6971,    Adjusted R-squared:  0.6859
F-statistic: 62.15 on 15 and 405 DF,  p-value: < 2.2e-16
```

The model explains 68.59% of the variance and achieves an RMSE of 0.4722.

Partial Plot

Partial dependence plot shows the marginal effect of a variable on the class response, i.e., the effect of adding another variable to a model of existing variables. We apply this plotting technique with our Random Forest model.

Code



Code

We shall see a steady linear relationship if a variable is very important. The ticks at the base of the plots are deciles of the x variables, so we can infer the distribution of the x variables, too. The direction of the lines explains whether a variable is positively or negatively correlated with the target variable.

The shape of the lines also bears some implications. For example, the plateau in the line of `MedHousingCost` (monthly housing cost), in the range between 1,000 USD to 1,500 USD per month, implies that a neighborhood's probability of in high demand of human services decreases sharply once its median rent reached above 1,500 per month. The same can be said for `HI_PublicInsured` and `ChildInNeed`, if the proportion of people who relies on Medicaid or other public health insurance options for their health care needs is more than 30%, or more than 20% of the children live in a family that requires income assistance, then the demand for 2-1-1 services increases dramatically.

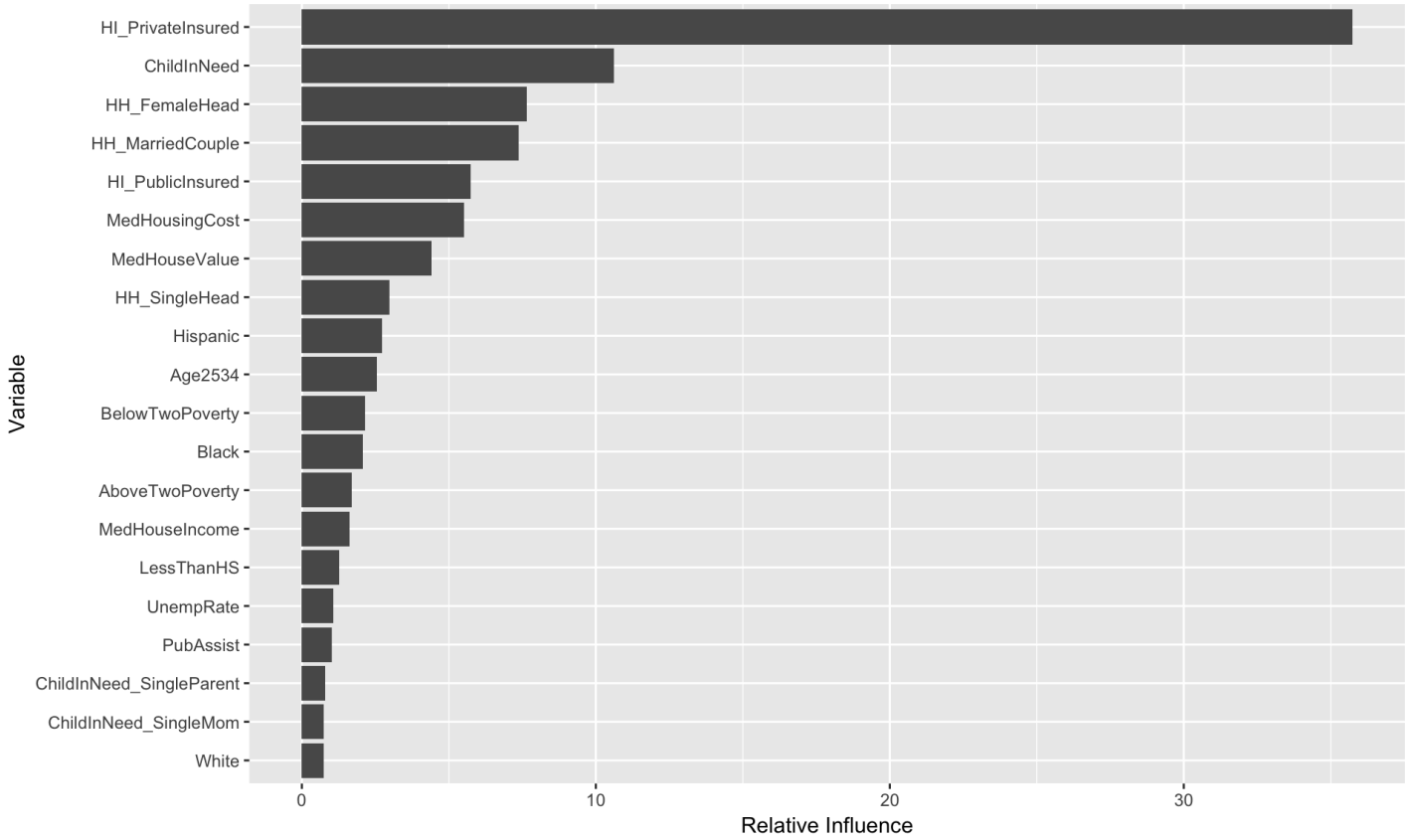
Race, income, education give relatively straight lines—which is maybe why we identified these variables as the most significant ones when running linear regression models.

Gradient Boosting

Another popular technique in additive learning is Gradient Boosting. It creates a set of weaker learners and additively use information derived from existing models to improve the choices made for latter models.

We apply Gradient Boosting to the same variables we trained above.

Code



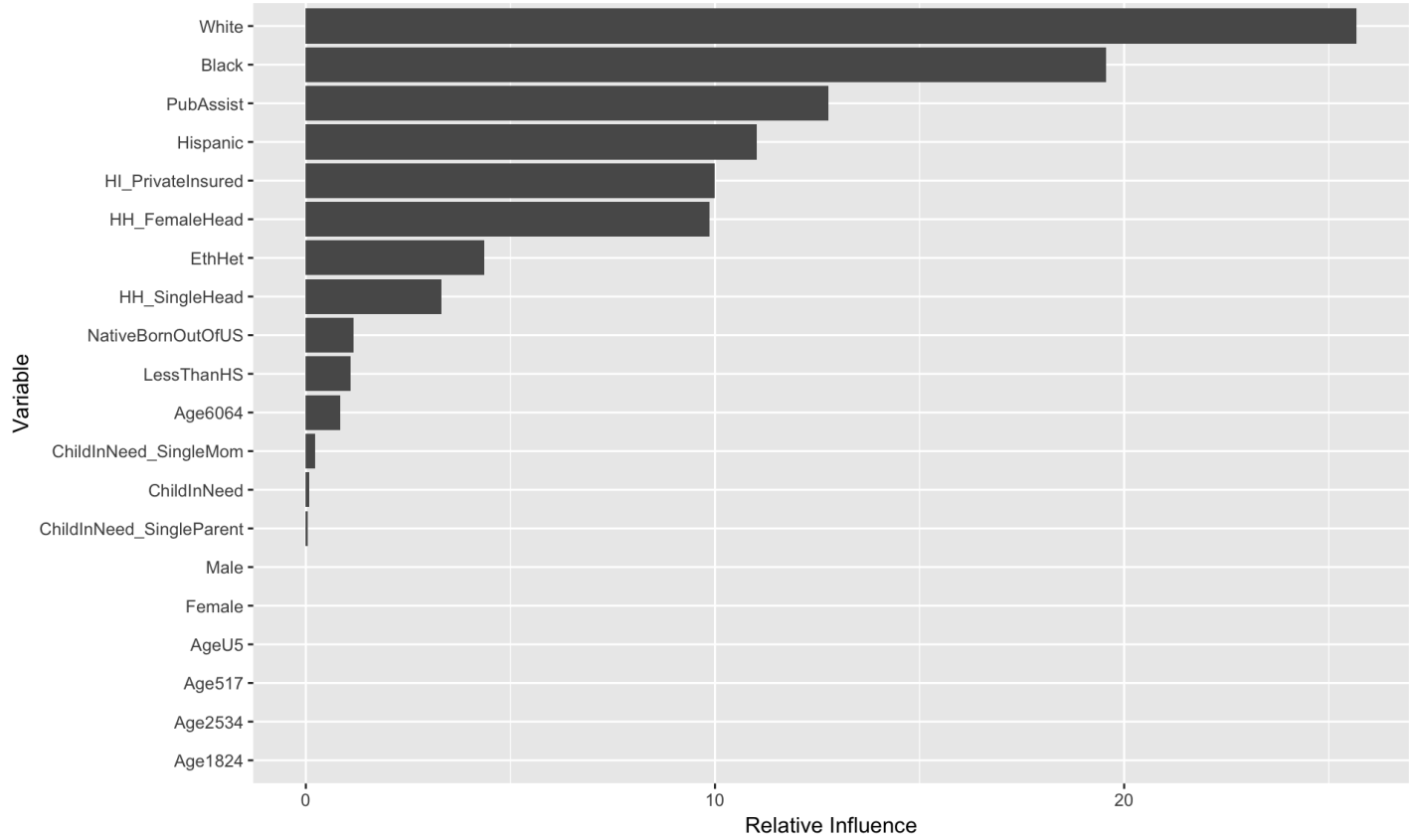
The list of most influential variables is different, but `HI_PrivateInsured` is still the most important one.

Poisson regression with GBM

The `gbm` function from the `gbm3` package allows us to specify the distribution of the response variable. Since the number of calls can be considered as count data, we may apply Poisson distribution here as well.

We use the raw number of calls per ZIP code, together with an offset variable–log total population.

Code



This time race is identified as the most influencial factor again, but insurance status and the proportion of single moms (`HH_FemaleHead` and `HH_SingleHead`) also remained important.

Next Step

Currently all above analyses are based on the full data we have. Samples and test are just splitting ZIP codes into groups. Another way of evaluating models is sampling call records and generate two copies of aggregated metrics for all ZIP codes, on different samples of calls, of course. The sampling process can be completely random, or based on time—i.e., train the model with old data and evaluate the model with the most recent data.