edX (https://www.edx.org)

MITx: 15.071x The Analytics Edge

zhushun0008 (/dashboard) ▾

Courseware (/courses/MITx/15.071x/1T2014/courseware) **Course Info (/courses/MITx/15.071x/1T2014/info)**

Discussion (/courses/MITx/15.071x/1T2014/discussion/forum) **Progress (/courses/MITx/15.071x/1T2014/progress)**

Syllabus (/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/)

Schedule (/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/)

PREDICTING EARNINGS FROM CENSUS DATA

The United States government periodically collects demographic information by conducting a census.

In this problem, we are going to use census information about an individual to predict how much a person earns -- in particular, whether the person earns more than $50,000 per year. This data comes from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets/Adult).

The file census.csv (/c4x/MITx/15.071x/asset/census.csv) contains 1994 census data for almost 32,000 individuals in the United States.

The available variables include:

- *age* = the age of the individual in years
- *workclass* = the classification of the individual's working status (does the person work for the federal government, work for the local government, work without pay, and so on)
- *education* = the level of education of the individual (e.g., 5th-6th grade, high school graduate, PhD, so on)
- *maritalstatus* = the marital status of the individual
- *occupation* = the type of work the individual does (e.g., administrative/clerical work, farming/fishing, sales and so on)
- *relationship* = relationship of individual to his/her household
- *race* = the individual's race
- *sex* = the individual's sex
- *capitalgain* = the capital gains of the individual in 1994 (from selling an asset such as a stock or bond for more than the original purchase price)
- *capitalloss* = the capital losses of the individual in 1994 (from selling an asset such as a stock or bond for less than the original purchase price)
- *hoursperweek* = the number of hours the individual works per week
- *nativecountry* = the native country of the individual
- *over50k* = whether or not the individual earned more than $50,000 in 1994

Help

PROBLEM 1.1 - A LOGISTIC REGRESSION MODEL (1 point possible)

As we did in lecture, let's begin by building a logistic regression model to predict whether an individual's earnings are above $50,000 using the other variables. First, read the dataset census.csv into R.

Then, split the data randomly into a training set and a testing set, setting the seed to 2000 before creating the split. Split the data so that the training set contains 60% of the observations, while the testing set contains 40% of the observations.

Next, build a logistic regression model using all of the independent variables to predict the dependent variable "over50k", and use the training set to build the model.

Which variables are significant, or have factors that are significant? (Use 0.1 as your significance threshold, so variables with a period or dot in the stars column should be counted too. You might see a warning message here - ignore it.)

□ age
□ workclass
□ education
□ maritalstatus
□ occupation
□ relationship
□ race
□ sex
□ capitalgain
□ capitalloss
□ hoursperweek
□ nativecountry

Show Answer *You have used 0 of 3 submissions*

PROBLEM 1.2 - A LOGISTIC REGRESSION MODEL (1 point possible)

What is the accuracy of the model on the testing set? Use a threshold of 0.5. (You might see a warning message when you make predictions on the test set - you can safely ignore it.)

Show Answer *You have used 0 of 3 submissions*

PROBLEM 1.3 - A LOGISTIC REGRESSION MODEL (1 point possible)

What is the baseline accuracy for the testing set?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 1.4 - A LOGISTIC REGRESSION MODEL (1 point possible)

What is the area-under-the-curve (AUC) for this model on the test set?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 2.1 - A CART MODEL (1 point possible)

We have just seen how the logistic regression model for this data achieves a high accuracy. Moreover, the significances of the variables give us a way to gauge which variables are relevant for this prediction task. However, it is not immediately clear which variables are more important than the others, especially due to the large number of factor variables in this problem.

Let us now build a classification tree for this model. Using the same training set, fit a CART model, and plot the tree. Use the default parameters, so don't set a value for minbucket or cp. Remember to specify method="class" as an argument to rpart, since this is a classification problem.

How many splits does the tree have in total?

Show Answer *You have used 0 of 5 submissions*

PROBLEM 2.2 - A CART MODEL (1 point possible)

Which variable does the tree split on at the first level (the very first split of the tree)?

- ○ age
- ○ workclass
- ○ education
- ○ maritalstatus
- ○ occupation
- ○ relationship
- ○ race
- ○ sex
- ○ capitalgain
- ○ capitalloss
- ○ hoursperweek
- ○ nativecountry

Show Answer *You have used 0 of 2 submissions*

PROBLEM 2.3 - A CART MODEL (1 point possible)

Which variables does the tree split on at the second level (immediately after the first split of the tree)?

- □ age
- □ workclass
- □ education
- □ maritalstatus
- □ occupation
- □ relationship
- □ race
- □ sex
- □ capitalgain

□ capitalloss
□ hoursperweek
□ nativecountry

Show Answer *You have used 0 of 3 submissions*

PROBLEM 2.4 - A CART MODEL (1 point possible)

What is the accuracy of the model on the testing set? (Use a threshold of 0.5, so add the argument type="class".)

This highlights a very regular phenomenon when comparing CART and logistic regression. CART often performs a little worse than logistic regression in out-of-sample accuracy. However, as is the case here, the CART model is often much simpler to describe and understand.

Show Answer *You have used 0 of 3 submissions*

PROBLEM 2.5 - A CART MODEL (1 point possible)

Let us now consider the ROC curve and AUC for the CART model. Plot the ROC curve for the CART model you have estimated. Observe that compared to the logistic regression ROC curve, the CART ROC curve is less smooth than the logistic regression ROC curve. Which of the following explanations for this behavior is most correct? (HINT: Think about what the ROC curve is plotting and what changing the threshold does.)

○ The number of variables that the logistic regression model is based on is larger than the number of variables used by the CART model, so the ROC curve for the logistic regression model will be smoother.

○ CART models require a higher number of observations in the testing set to produce a smoother/more continuous ROC curve; there is simply not enough data.

○ The probabilities from the CART model take only a handful of values (five, one for each end bucket/leaf of the tree); the changes in the ROC curve correspond to setting the threshold to one of those values.

○ The CART model uses fewer continuous variables than the logistic regression model (capitalgain for CART versus age, capitalgain, capitallosses, hoursperweek), which is why the CART ROC curve is less smooth than the logistic regression one.

Show Answer *You have used 0 of 2 submissions*

PROBLEM 2.6 - A CART MODEL (1 point possible)

What is the AUC of the CART model on the test set?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 3.1 - A RANDOM FOREST MODEL (1 point possible)

Before building a random forest model, we'll down-sample our training set. While some modern personal computers can build a random forest model on the entire training set, others might run out of memory when trying to train the model since random forests is much more computationally intensive than CART or Logistic Regression. For this reason, before continuing we will define a new training set to be used when building our random forest model, that contains 2000 randomly selected obervations from the original training set. Do this by running the following commands in your R console (assuming your training set is called "train"):

```
set.seed(1)

trainSmall = train[sample(nrow(train), 2000), ]
```

Let us now try to build a random forest model using the dataset "trainSmall" as the data used to build the model. Go ahead and attempt to build a random forest model. You should get an error that random forest "can not handle categorical predictors with more than 32 categories". This means that we have a factor variable with more than 32 different possible values. Which one of your variables is causing this error?

○ age
○ workclass
○ education
○ maritalstatus
○ occupation
○ relationship
○ race
○ sex
○ capitalgain
○ capitalloss
○ hoursperweek
○ nativecountry

Show Answer *You have used 0 of 3 submissions*

PROBLEM 3.2 - A RANDOM FOREST MODEL (1 point possible)

Now, build your random forest model without the problematic variable identified in the previous problem. Set the seed to 1 before building the model. Remember to use the dataset "trainSmall" to build the model.

Then, make predictions using this model on the entire test set. What is the accuracy of the model on the test set? (Remember that you don't need a "type" argument when making predictions with a random forest model.)

Show Answer *You have used 0 of 3 submissions*

PROBLEM 3.3 - A RANDOM FOREST MODEL (1 point possible)

As we discussed in lecture, random forest models work by building a large collection of trees. As a result, we lose some of the interpretability that comes with CART in terms of seeing how predictions are made and which variables are important. However, we can still compute metrics that give us insight into which variables are important.

One metric that we can look at is the number of times, aggregated over all of the trees in the random forest model, that a certain variable is selected for a split. To view this metric, run the following lines of R code (replace "MODEL" with the name of your random forest model):

```
vu = varUsed(MODEL, count=TRUE)

vusorted = sort(vu, decreasing = FALSE, index.return = TRUE)

dotchart(vusorted$x, names(MODEL$forest$xlevels[vusorted$ix]))
```

This code produces a chart that for each variable measures the number of times that variable was selected for splitting (the value on the x-axis). Which of the following variables is the most important in terms of the number of splits?

○ age
○ maritalstatus
○ capitalgain
○ education

Show Answer *You have used 0 of 2 submissions*

PROBLEM 3.4 - A RANDOM FOREST MODEL (1 point possible)

A different metric we can look at is related to "impurity", which measures how homogenous each bucket or leaf of the tree is. In each tree in the forest, whenever we select a variable and perform a split, the impurity is decreased. Therefore, one way to measure the importance of a variable is to average the reduction in impurity, taken over all the times that variable is selected for splitting in all of the trees in the forest. To compute this metric, run the following command in R (replace "MODEL" with the name of your random forest model):

```
varImpPlot(MODEL)
```

Which one of the following variables is the most important in terms of mean reduction in impurity?

○ workclass
○ occupation
○ sex
○ capitalloss

Show Answer *You have used 0 of 2 submissions*

PROBLEM 4.1 - SELECTING CP BY CROSS-VALIDATION (1 point possible)

We now conclude our study of this data set by looking at how CART behaves with different choices of its parameters.

Let us select the cp parameter for our CART model using k-fold cross validation, with k = 10 folds. Do this by using the train function. Set the seed beforehand to 2. Test cp values from 0.002 to 0.1 in 0.002 increments, by using the following command:

```
cartGrid = expand.grid( .cp = seq(0.002,0.1,0.002))
```

Also, remember to use the entire training set "train" when building this model. The train function might take some time to run.

Which value of cp does the train function recommend?

Show Answer *You have used 0 of 4 submissions*

PROBLEM 4.2 - SELECTING CP BY CROSS-VALIDATION (1 point possible)

Fit a CART model to the training data using this value of cp. What is the prediction accuracy on the test set?

Show Answer *You have used 0 of 4 submissions*

PROBLEM 4.3 - SELECTING CP BY CROSS-VALIDATION (1 point possible)

Compared to the original accuracy using the default value of cp, this new CART model is an improvement, and so we should clearly favor this new model over the old one -- or should we? Plot the CART tree for this model. How many splits are there?

This highlights one important tradeoff in building predictive models. By tuning cp, we improved our accuracy by over 1%, but our tree became significantly more complicated. In some applications, such an improvement in accuracy would be worth the loss in interpretability. In others, we may prefer a less accurate model that is simpler to understand and describe over a more accurate -- but more complicated -- model.

Show Answer *You have used 0 of 3 submissions*

Please remember not to ask for or post complete answers to homework questions in this discussion forum.

Show Discussion New Post

EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.

(http://www.facebook.com/EdxOnline)

(https://twitter.com/edXOnline)

(https://plus.google.com/108235383044095082

(http://youtube.com/user/edxonline)