MITOCW | MIT15_071S17_Session_4.3.11_300k

In the next few videos, we'll be using a data set published by the United States Centers for Medicare and Medicaid Services to practice creating CART models to predict health care cost.

We unfortunately can't use the D2Hawkeye data due to privacy issues.

The data set we'll be using instead, ClaimsData.csv, is structured to represent a sample of patients in the Medicare program, which provides health insurance to Americans aged 65 and older, as well as some younger people with certain medical conditions.

To protect the privacy of patients represented in this publicly available data set, a number of steps are performed to anonymize the data.

So we would need to retrain the models we develop in this lecture on de-anonymized data if we wanted to apply our models in the real world.

Let's start by reading our data set into R and taking a look at its structure.

We'll call our data set Claims, and we'll use the read.csv function to read in the data file ClaimsData.csv.

Make sure to navigate to the directory on your computer containing the file ClaimsData.csv first.

Now let's take a look at the structure of our data frame using the str function.

The observations represent a 1% random sample of Medicare beneficiaries, limited to those still alive at the end of 2008.

Our independent variables are from 2008, and we will be predicting cost in 2009.

Our independent variables are the patient's age in years at the end of 2008, and then several binary variables indicating whether or not the patient had diagnosis codes for a particular disease or related disorder in 2008: alzheimers, arthritis, cancer, chronic obstructive pulmonary disease, or copd, depression, diabetes, heart.failure, ischemic heart disease, or ihd, kidney disease, osteoporosis, and stroke.

Each of these variables will take value 1 if the patient had a diagnosis code for the particular disease and value 0 otherwise.

Reimbursement2008 is the total amount of Medicare reimbursements for this patient in 2008.

And reimbursement2009 is the total value of all Medicare reimbursements for the patient in 2009.

Bucket2008 is the cost bucket the patient fell into in 2008, and bucket2009 is the cost bucket the patient fell into in

2009.

These cost buckets are defined using the thresholds determined by D2Hawkeye.

So the first cost bucket contains patients with costs less than $3,000, the second cost bucket contains patients with costs between $3,000 and $8,000, and so on.

We can verify that the number of patients in each cost bucket has the same structure as what we saw for D2Hawkeye by computing the percentage of patients in each cost bucket.

So we'll create a table of the variable bucket2009 and divide by the number of rows in Claims.

This gives the percentage of patients in each of the cost buckets.

The first cost bucket has almost 70% of the patients.

The second cost bucket has about 20% of the patients.

And the remaining 10% are split between the final three cost buckets.

So the vast majority of patients in this data set have low cost.

Our goal will be to predict the cost bucket the patient fell into in 2009 using a CART model.

But before we build our model, we need to split our data into a training set and a testing set.

So we'll load the package caTools, and then we'll set our random seed to 88 so that we all get the same split.

And we'll use the sample.split function, where our dependent variable is Claims$bucket2009, and we'll set our SplitRatio to be 0.6.

So we'll put 60% of the data in the training set.

We'll call our training set ClaimsTrain, and we'll take the observations of Claims for which spl is exactly equal to TRUE.

And our testing set will be called ClaimsTest, where we'll take the observations of Claims for which spl is exactly equal to FALSE.

Now that our data set is ready, we'll see in the next video how a smart baseline method would perform.