PREDICTING MEDICAL COSTS WITH CLUSTER-THEN-PREDICT

In the second lecture sequence this week, we heard about cluster-then-predict, a methodology in which you first cluster observations and then build cluster-specific prediction models. In the lecture sequence, we saw how this methodology helped improve the prediction of heart attack risk. In this assignment, we'll use cluster-then-predict to predict future medical costs using medical claims data.

In Week 4, we discussed the importance of high-quality predictions of future medical costs based on information available in medical claims data. In this problem, you will predict future medical claims using part of the DE-SynPUF dataset (http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/SynPUFs/DE_Syn_PUF.html), published by the United States Centers for Medicare and Medicaid Services (CMS). This dataset, available in reimbursement.csv (/c4x/MITx/15.071x/asset/reimbursement.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 dataset, CMS performs a number of steps to anonymize the data, so we would need to re-train the models we develop in this problem on de-anonymized data if we wanted to apply our models in the real world.

The observations in the dataset represent a 1% random sample of Medicare beneficiaries in 2008, limited to those still alive at the end of 2008. The dependent variable, **reimbursement2009**, represents the total value of all Medicare reimbursements for a patient in 2009, which is the cost of the patient's care to the Medicare system. The following independent variables are available:

- **age**: The patient's age in years at the beginning of 2009
- **alzheimers:** Binary variable for whether the patient had diagnosis codes for Alzheimer's disease or a related disorder in 2008
- **arthritis**: Binary variable for whether the patient had diagnosis codes for rheumatoid arthritis or osteoarthritis in 2008
- **cancer**: Binary variable for whether the patient had diagnosis codes for cancer in 2008
- **copd**: Binary variable for whether the patient had diagnosis codes for Chronic Obstructive Pulmonary Disease (COPD) in 2008
- **depression**: Binary variable for whether the patient had diagnosis codes for depression in 2008
- **diabetes**: Binary variable for whether the patient had diagnosis codes for diabetes in 2008
- **heart.failure**: Binary variable for whether the patient had diagnosis codes for heart failure in 2008
- **ihd**: Binary variable for whether the patient had diagnosis codes for ischemic heart disease (IHD) in 2008
- **kidney**: Binary variable for whether the patient had diagnosis codes for chronic kidney disease in 2008
- **osteoporosis**: Binary variable for whether the patient had diagnosis codes for osteoporosis in 2008
- **stroke**: Binary variable for whether the patient had diagnosis codes for a stroke/transient ischemic attack (TIA) in 2008
- **reimbursement2008**: The total amount of Medicare reimbursements for this patient for 2008

PROBLEM 1.1 - PREPARING THE DATASET (1 point possible)

Load reimbursement.csv into a data frame called claims.

How many Medicare beneficiaries are included in the dataset?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 1.2 - PREPARING THE DATASET (1 point possible)

What proportion of patients have at least one of the chronic conditions described in the independent variables alzheimers, arthritis, cancer, copd, depression, diabetes, heart.failure, ihd, kidney, osteoporosis, and stroke?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 1.3 - PREPARING THE DATASET (1 point possible)

What is the maximum correlation between independent variables in the dataset?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 1.4 - PREPARING THE DATASET (1 point possible)

Plot the histogram of the dependent variable. What is the shape of the distribution?

- ○ Skew right -- there are a large number of observations with a small value, but only a small number of observations with a large value.
- ○ Balanced -- there are roughly the same number of observations with an unusually large and unusually small value.
- ○ Skew left -- there are a large number of observations with a large value, but only a small number of observations with a small value.

Show Answer *You have used 0 of 1 submissions*

PROBLEM 1.5 - PREPARING THE DATASET (1 point possible)

To address the shape of the data identified in the previous problem, we will log transform the two reimbursement variables with the following code:

```
claims$reimbursement2008 = log(claims$reimbursement2008+1)

claims$reimbursement2009 = log(claims$reimbursement2009+1)
```

Why did we take the log of the reimbursement value plus 1 instead of the log of the reimbursement value? Hint -- What happens when a patient has a reimbursement cost of $0?

- ○ Every patient in Medicare gets at least $1 in reimbursement
- ○ To avoid log-transformed values of negative infinity
- ○ To avoid log-transformed values of infinity
- ○ There was no reason

Show Answer *You have used 0 of 1 submissions*

PROBLEM 1.6 - PREPARING THE DATASET (1 point possible)

Plot the histogram of the log-transformed dependent variable. The distribution is reasonably balanced, other than a large number of people with variable value 0, corresponding to having had $0 in reimbursements in 2009. What proportion of beneficiaries had $0 in reimbursements in 2009?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 2.1 - INITIAL LINEAR REGRESSION MODEL (1 point possible)

In Week 3 when we learned about the sample.split function, we mentioned that you split data into a training and testing set a bit differently when there is a continuous outcome. Run the following commands to randomly select 70% of the data for the training set and 30% of the data for the testing set:

```
set.seed(144)

spl = sample(1:nrow(claims), size=0.7*nrow(claims))

train = claims[spl,]

test = claims[-spl,]
```

Use the train data frame to train a linear regression model (name it lm.claims) to predict reimbursement2009 using all the independent variables.

What is the training set Multiple R-squared value of lm.claims?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 2.2 - INITIAL LINEAR REGRESSION MODEL (1 point possible)

Obtain testing set predictions from lm.claims. What is the testing set RMSE of the model?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 2.3 - INITIAL LINEAR REGRESSION MODEL (1 point possible)

What is the "naive baseline model" that we would typically use to compute the R-squared value of lm.claims?

- Predict 0 for every observation
- Predict mean(train$reimbursement2008) for every observation
- Predict mean(test$reimbursement2008) for every observation
- Predict mean(train$reimbursement2009) for every observation
- Predict mean(test$reimbursement2009) for every observation

Show Answer *You have used 0 of 1 submissions*

PROBLEM 2.4 - INITIAL LINEAR REGRESSION MODEL (1 point possible)

What is the testing set RMSE of the naive baseline model?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 2.5 - INITIAL LINEAR REGRESSION MODEL (1 point possible)

In Week 4, we saw how D2Hawkeye used a "smart baseline model" that predicted that a patient's medical costs would be equal to their costs in the previous year. For our problem, this baseline would predict reimbursement2009 to be equal to reimbursement2008.

What is the testing set RMSE of this smart baseline model?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 3.1 - CLUSTERING MEDICARE BENEFICIARIES (1 point possible)

In this section, we will cluster the Medicare beneficiaries. The first step in this process is to remove the dependent variable using the following commands:

train.limited = train

train.limited$reimbursement2009 = NULL

test.limited = test

test.limited$reimbursement2009 = NULL

Why do we need to remove the dependent variable in the clustering phase of the cluster-then-predict methodology?

- ○ Leaving in the dependent variable might lead to unbalanced clusters
- ○ Removing the dependent variable decreases the computational effort needed to cluster
- ○ Needing to know the dependent variable value to assign an observation to a cluster defeats the purpose of the methodology

Show Answer *You have used 0 of 1 submissions*

PROBLEM 3.2 - CLUSTERING MEDICARE BENEFICIARIES (2 points possible)

In the market segmentation assignment in this week's homework, you were introduced to the preProcess command from the caret package, which normalizes variables by subtracting by the mean and dividing by the standard deviation.

In cases where we have a training and testing set, we'll want to normalize by the mean and standard deviation of the variables in the training set. We can do this by passing just the training set to the preProcess function:

library(caret)

preproc = preProcess(train.limited)

train.norm = predict(preproc, train.limited)

test.norm = predict(preproc, test.limited)

What is the mean of the arthritis variable in train.norm?

What is the mean of the arthritis variable in test.norm?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 3.3 - CLUSTERING MEDICARE BENEFICIARIES (1 point possible)

Why is the mean arthritis variable much closer to 0 in train.norm than in test.norm?

- ○ Small rounding errors exist in the normalization procedure
- ○ The distribution of the arthritis variable is different in the training and testing set
- ○ The distribution of the dependent variable is different in the training and testing set

Show Answer *You have used 0 of 1 submissions*

PROBLEM 3.4 - CLUSTERING MEDICARE BENEFICIARIES (1 point possible)

Set the random seed to 144 (it is important to do this again, even though we did it earlier). Run k-means clustering with 3 clusters on train.norm, storing the result in an object called km.

The description "older-than-average beneficiaries with below average incidence of stroke and above-average 2008 reimbursements" uniquely describes which cluster center?

- ○ Cluster 1
- ○ Cluster 2
- ○ Cluster 3

Show Answer *You have used 0 of 1 submissions*

PROBLEM 3.5 - CLUSTERING MEDICARE BENEFICIARIES (1 point possible)

Recall from the recitation that we can use the flexclust package to obtain training set and testing set cluster assignments for our observations (note that the call to as.kcca may take a while to complete):

```
library(flexclust)

km.kcca = as.kcca(km, train.norm)

cluster.train = predict(km.kcca)

cluster.test = predict(km.kcca, newdata=test.norm)
```

How many test-set observations were assigned to Cluster 2?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 4.1 - CLUSTER-SPECIFIC PREDICTIONS (1 point possible)

Using the subset function, build data frames train1, train2, and train3, containing the elements in the train data frame assigned to clusters 1, 2, and 3, respectively (be careful to take subsets of train, not of train.norm). Similarly build test1, test2, and test3 from the test data frame.

Which training set data frame has the highest average value of the dependent variable?

- ○ train1
- ○ train2
- ○ train3

Show Answer *You have used 0 of 1 submissions*

PROBLEM 4.2 - CLUSTER-SPECIFIC PREDICTIONS (1 point possible)

Build linear regression models lm1, lm2, and lm3, which predict reimbursement2009 using all the variables. lm1 should be trained on train1, lm2 should be trained on train2, and lm3 should be trained on train3.

Which variables have a positive sign for the coefficient in at least one of lm1, lm2, and lm3 and a negative sign for the coefficient in at least one of lm1, lm2, and lm3?

- ☐ age
- ☐ alzheimers
- ☐ arthritis
- ☐ cancer
- ☐ copd
- ☐ depression
- ☐ diabetes
- ☐ heart.failure
- ☐ ihd
- ☐ kidney
- ☐ osteoporosis
- ☐ reimbursement2008

Show Answer *You have used 0 of 3 submissions*

PROBLEM 4.3 - CLUSTER-SPECIFIC PREDICTIONS (1 point possible)

Using lm1, make test-set predictions called pred.test1 on data frame test1. Using lm2, make test-set predictions called pred.test2 on data frame test2. Using lm3, make test-set predictions called pred.test3 on data frame test3.

Which vector of test-set predictions has the smallest average predicted reimbursement amount?

- ○ pred.test1
- ○ pred.test2
- ○ pred.test3

Show Answer *You have used 0 of 1 submissions*

PROBLEM 4.4 - CLUSTER-SPECIFIC PREDICTIONS (1 point possible)

Obtain the test-set RMSE for each cluster. Which cluster has the largest test-set RMSE?

- ○ Cluster 1
- ○ Cluster 2
- ○ Cluster 3

Show Answer *You have used 0 of 1 submissions*

PROBLEM 4.5 - CLUSTER-SPECIFIC PREDICTIONS (1 point possible)

To compute the overall test-set RMSE of the cluster-then-predict approach, we can combine all the test-set predictions into a single vector and all the true outcomes into a single vector:

all.predictions = c(pred.test1, pred.test2, pred.test3)

all.outcomes = c(test1$reimbursement2009, test2$reimbursement2009, test3$reimbursement2009)

What is the test-set RMSE of the cluster-then-predict approach?

We see a modest improvement over the original linear regression model, which is typical in situations where the observations do not cluster strongly into different "types" of observations. However, it is often a good idea to try the cluster-then-predict approach on datasets with a large number of observations to see if you can improve the accuracy of your 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