Homework 3 (Random Forests)

Zin Min Thant

2024-10-27

This assignment is due Sunday October 27. Once you have provided all the code and answered all the questions, knit your document to an html file and then save the file as a pdf.

Upload your pdf file that contains your code and answers to these questions to the Homework 3 assignment link in Module 7 in Brightspace.

This assignment is worth 5 points. The amount of points/marks each question is worth is provided in each question.

Before you begin: Make sure you install the caret, rpart, rpart.plot, and randomForest packages.

#Scenario and Data Description

You are working at consumer advocacy agency, Equitable Ernest. Equitable Ernest is interested in providing a service that allows an individual to estimate their own credit score (a continuous measure used by banks and insurance companies that determines whether or not to loan money or to provide insurance premium quotes).

In the workspace folder of this project there is an Excel file name creditscore.csv that contains information on several thousand individuals. The features in this data set are:

BureauInquiries - number of inquiries about an individuals credit

CreditUsage - percent of an individual's credit used

TotalCredit - total amount of credit available to an individual

CollectedReports - number of times an unpaid bill was reported to a collection agency

MissedPayments - number of missed payments

HomeOwner - 1 if individual is homeowner, 0 if not

CreditAge - average age of individual's credit

TimeOnJob - how long the individual has been continuously employed (in years)

The CreditScore variable is the target variable. CreditScore is a number between 300 and 850 with larger numbers representing increased credit worthiness.

Import the data

1. Create an R data frame named creditscore by converting the creditscore.csv file into a R data frame. 1 point

```
creditscore <-read.csv("/cloud/project/data/creditscore.csv")
```

Build a Random Forest model

2. Using 5-fold cross validation build a model to predict individual's credit scores using the creditscore data frame. 2 points

Follow the steps in the code chunk to complete this task.

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(rpart)
library(rpart.plot)
library(randomForest)
```

```
## randomForest 4.7-1.2
```

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

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
#Step 1
# set random number seed for bootstrap samples
# different seed values will generate different partitions of the data
#using the same seed will generate the same model each time code is run
#run the code below
set.seed(1234)

#Step 2
#specify the cross validation method
# method = k-fold cross validation
# number = number of folds
#create an R object named cv_method that will be used to specify 5-fold cross valdiation


cv_method <- trainControl(method="cv",number = 5)

#Step 3
# create a grid of parameter values to assess in k-fold cross-validation
# for random forest, a hyperparameter is mtry, the number of randomly
# selected candidate features considered at each split in tree
# testing values of mtry from 1 to 8 (the number of features)
#run the code below
grid <-expand.grid(.mtry=c(1:8))

#Step 4
#Build a random forest to predict an individual's credit score using a random forest
#use 5-fold cross validation
#use the grid object created above for the tuneGrid argument (just like we did in class)
#name the R object that will hold the results of the random forest Credit_Forest
#In the train() function, specify an argument ntree=10 so that the random forest will construct 10 deci
```

```
Forest <- train(CreditScore ~., data = creditscore, method = "rf",trControl=cv_method,tuneGrid=grid,ntr
```

```
#Optional
#If you want to look at the output of the random forest, type Credit_Forest on the line below and you w
```

3. In the data folder there is a .csv file named new_individual_creditscore. This is an individual with certain values of the features for BureauInquiries, CreditUsage, etc... This individual's record did not participate in the creation of the random forest model from #2. 2 points

Apply the random forest built from #2 to predict the credit score of this new individual's record. Round the credit score to 0 decimal places.

What is the predicted credit score for the individual that is in the new_individual_creditscore file?

<581>

```
new_data <-read.csv("/cloud/project/data/new_individual_creditscore.csv")

ForestPredict <- predict(Forest, newdata=new_data)
ForestPredict
```

```
##        1
## 580.5117
```