Group10-HW7

Group 10 – Lince Rumainum

November 14, 2019

Problem 1

Problem 1-(a)

First we need to look at all the data as whole, to be able to see what type of data that are missing data from all of the variables. As shown below, payer code and the medical specialty are missing a lot of its data where race and diagnosis missing a few. Hospital might want to make sure that those huge chunks of missing data are provided for future references.

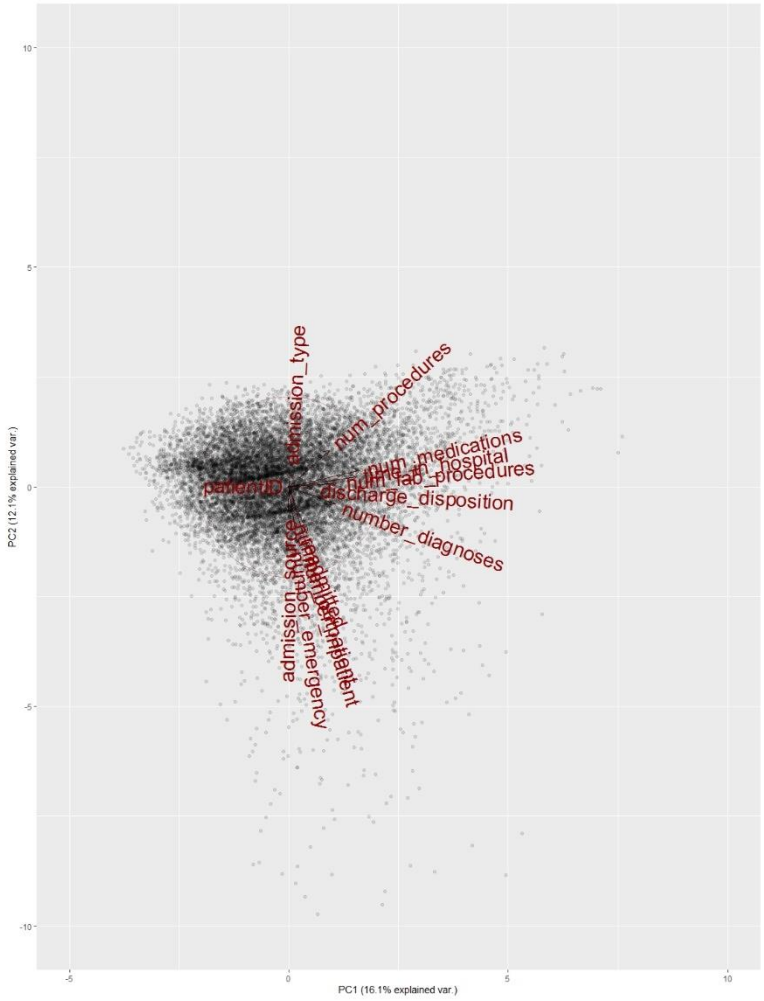
```
Observations: 1
Variables: 45
$ patientID                <dbl> 0
$ race                     <dbl> 0.02269467
$ gender                   <dbl> 0
$ age                      <dbl> 0
$ admission_type           <dbl> 0
$ discharge_disposition    <dbl> 0
$ admission_source         <dbl> 0
$ time_in_hospital         <dbl> 0
$ payer_code               <dbl> 0.371878
$ medical_specialty        <dbl> 0.4806845
$ num_lab_procedures       <dbl> 0
$ num_procedures           <dbl> 0
$ num_medications          <dbl> 0
$ number_outpatient        <dbl> 0
$ number_emergency         <dbl> 0
$ number_inpatient         <dbl> 0
$ diagnosis                <dbl> 0.0001901305
$ number_diagnoses         <dbl> 0
$ max_glu_serum            <dbl> 0
$ A1Cresult                <dbl> 0
$ metformin                <dbl> 0
$ repaglinide              <dbl> 0
$ nateglinide              <dbl> 0
$ chlorpropamide           <dbl> 0
$ glimepiride              <dbl> 0
$ acetohexamide            <dbl> 0
$ glipizide                <dbl> 0
$ glyburide                <dbl> 0
$ tolbutamide              <dbl> 0
$ pioglitazone             <dbl> 0
$ rosiglitazone            <dbl> 0
$ acarbose                 <dbl> 0
$ miglitol                 <dbl> 0
$ troglitazone             <dbl> 0
$ tolazamide               <dbl> 0
$ examide                  <dbl> 0
$ citoglipton              <dbl> 0
$ insulin                  <dbl> 0
$ glyburide.metformin      <dbl> 0
$ glipizide.metformin      <dbl> 0
$ glimepiride.pioglitazone <dbl> 0
$ metformin.rosiglitazone  <dbl> 0
$ metformin.pioglitazone   <dbl> 0
$ diabetesMed              <dbl> 0
$ readmitted               <dbl> 0
```

From the summary of the MARS model, it shown that the readmitted values are highly correlated with the admission_source, number_emergency, number_inpatient, and number_inpatient who are within the age 70 and 80 years old.

```
Call: earth(formula=readmitted~., data=dfTr_mars, degree=3, nfold=3)

                                                                  coefficients
(Intercept)                                                         0.64240671
h(admission_source-6)                                               0.18986684
h(7-admission_source)                                               0.02020567
h(admission_source-7)                                              -0.18822677
h(2-number_emergency)                                              -0.03650560
h(2-number_inpatient)                                              -0.14200706
h(number_inpatient-2)                                               0.02479780
age[70-80) * h(2-number_inpatient)                                  0.02853282
h(2-number_inpatient) * diabetesMedYes                              0.02186759
h(admission_source-7) * h(28-num_lab_procedures)                   -0.00123096
h(7-admission_source) * h(1-num_procedures)                         0.01152832
h(1-number_outpatient) * h(2-number_emergency)                     -0.03691444
age[80-90) * h(2-number_inpatient) * diabetesMedYes                 0.03334683
h(admission_type-3) * h(1-number_outpatient) * h(2-number_emergency) 0.00884475
h(2-discharge_disposition) * h(2-number_inpatient) * h(9-number_diagnoses) -0.01015202
h(16-num_medications) * h(1-number_outpatient) * h(2-number_emergency)     -0.00345477
h(num_medications-16) * h(1-number_outpatient) * h(2-number_emergency)     -0.00087186

Selected 17 of 25 terms, and 13 of 77 predictors
Termination condition: RSq changed by less than 0.001 at 25 terms
Importance: number_inpatient, discharge_disposition, number_diagnoses, number_outpatient, number_emergency, admission_source, num_lab_procedures, ...
Number of terms at each degree of interaction: 1 6 5 5
GCV 0.2266485  RSS 12794.09  GRSq 0.09095743  RSq 0.09224343  CVRSq 0.08465148

Note: the cross-validation sd's below are standard deviations across folds

Cross validation:   nterms 15.67 sd 3.06    nvars 10.00 sd 1.73

     CVRSq    sd     ClassRate    sd     MaxErr  sd
     0.085 0.002         0.623 0.004       -1.2 1.3
> sqrt(mean(marsFit$residuals^2)) #RSME
[1] 0.4757309
```

Although it is not easy to see from the graph below, it is support that conclusion also. The Principal Component Analysis graph also shown that the readmitted

variable is highly correlated with admission_source, number_emergency, and number inpatient, with the addition of number_outpatient.



Now, we are going to take a look at the performances evaluation techniques, quantify the predictive quality of the models in this problem. For the decision tree, below is the summary of the hospital model

```
> summary(fit_st)
Call:
rpart(formula = readmitted ~ ., data = dfTr_new)
  n= 56531

          CP nsplit rel error    xerror        xstd
1 0.1838814      0 1.0000000 1.0000000 0.004431630
2 0.0100000      1 0.8161186 0.8161186 0.004322345

Variable importance
number_inpatient number_emergency
              94               6

Node number 1: 56531 observations,    complexity param=0.1838814
  predicted class=0  expected loss=0.4738816  P(node) =1
    class counts: 29742 26789
   probabilities: 0.526 0.474
  left son=2 (37611 obs) right son=3 (18920 obs)
  Primary splits:
      number_inpatient  < 0.5 to the left,  improve=1389.4100, (0 missing)
      number_emergency  < 0.5 to the left,  improve= 456.4287, (0 missing)
      number_outpatient < 0.5 to the left,  improve= 353.7418, (0 missing)
      number_diagnoses  < 5.5 to the left,  improve= 353.5034, (0 missing)
      admission_source  < 6.5 to the left,  improve= 205.2241, (0 missing)
  Surrogate splits:
      number_emergency  < 0.5 to the left,  agree=0.686, adj=0.062, (0 split)
      number_outpatient < 4.5 to the left,  agree=0.667, adj=0.004, (0 split)
      admission_source  < 21  to the left,  agree=0.665, adj=0.000, (0 split)
      acarbose          splits as  RLLL,    agree=0.665, adj=0.000, (0 split)
      miglitol          splits as  LLLR,    agree=0.665, adj=0.000, (0 split)

Node number 2: 37611 observations
  predicted class=0  expected loss=0.3952567  P(node) =0.6653164
    class counts: 22745 14866
   probabilities: 0.605 0.395

Node number 3: 18920 observations
  predicted class=1  expected loss=0.3698203  P(node) =0.3346836
    class counts:  6997 11923
   probabilities: 0.370 0.630
```

Below you can also see the accuracy and kappa value of the decision tree:

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 22745 14866
         1  6997 11923

               Accuracy : 0.6133
                 95% CI : (0.6092, 0.6173)
    No Information Rate : 0.5261
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2129

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7647
            Specificity : 0.4451
         Pos Pred Value : 0.6047
         Neg Pred Value : 0.6302
             Prevalence : 0.5261
         Detection Rate : 0.4023
   Detection Prevalence : 0.6653
      Balanced Accuracy : 0.6049

       'Positive' Class : 0
```

Since readmitted only have two levels, the pruning using the decision tree still gives the same accuracy and kappa value. Below is the summary of using the random forest model from the caret package. In the model below, only 25% of the training data was used. Using different grid values and 3 cross-validation, it shows that having mtry = 6, gives the best model.

```
#########################################
######### RANDOM FOREST #################
#########################################
#let's tune across mtry = 2,3,...,6
rfGrid <-  expand.grid(mtry = 2:6)
# use 25% of training data
dfTr_quarter <- dfTr_new %>% sample_frac(0.25)
# random forest model
rf_model<-train(readmitted~.,data=dfTr_quarter,
                method="rf",  #random forest
                trControl=trainControl(method="cv",number=3),  #cross-validation
                tuneGrid=rfGrid,   #hyper-parameter tuning
                allowParallel=TRUE)

#prints out the CV accuracy and kappa for each mtry value
print(rf_model)

#the best model (based on tuning grid)
print(rf_model$finalModel)
```

```
> print(rf_model)
Random Forest

14464 samples
   36 predictor

No pre-processing
Resampling: Cross-Validated (3 fold)
Summary of sample sizes: 9642, 9643, 9643
Resampling results across tuning parameters:

  mtry  RMSE       Rsquared    MAE
  2     0.4864603  0.08830549  0.4831319
  3     0.4801135  0.09444717  0.4726554
  4     0.4775105  0.09408564  0.4661171
  5     0.4764592  0.09374790  0.4622465
  6     0.4756414  0.09469024  0.4591837

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 6.
> #the best model (based on tuning grid)
> print(rf_model$finalModel)

Call:
 randomForest(x = x, y = y, mtry = param$mtry, allowParallel = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 6

          Mean of squared residuals: 0.2257881
                    % Var explained: 9.35
>
```

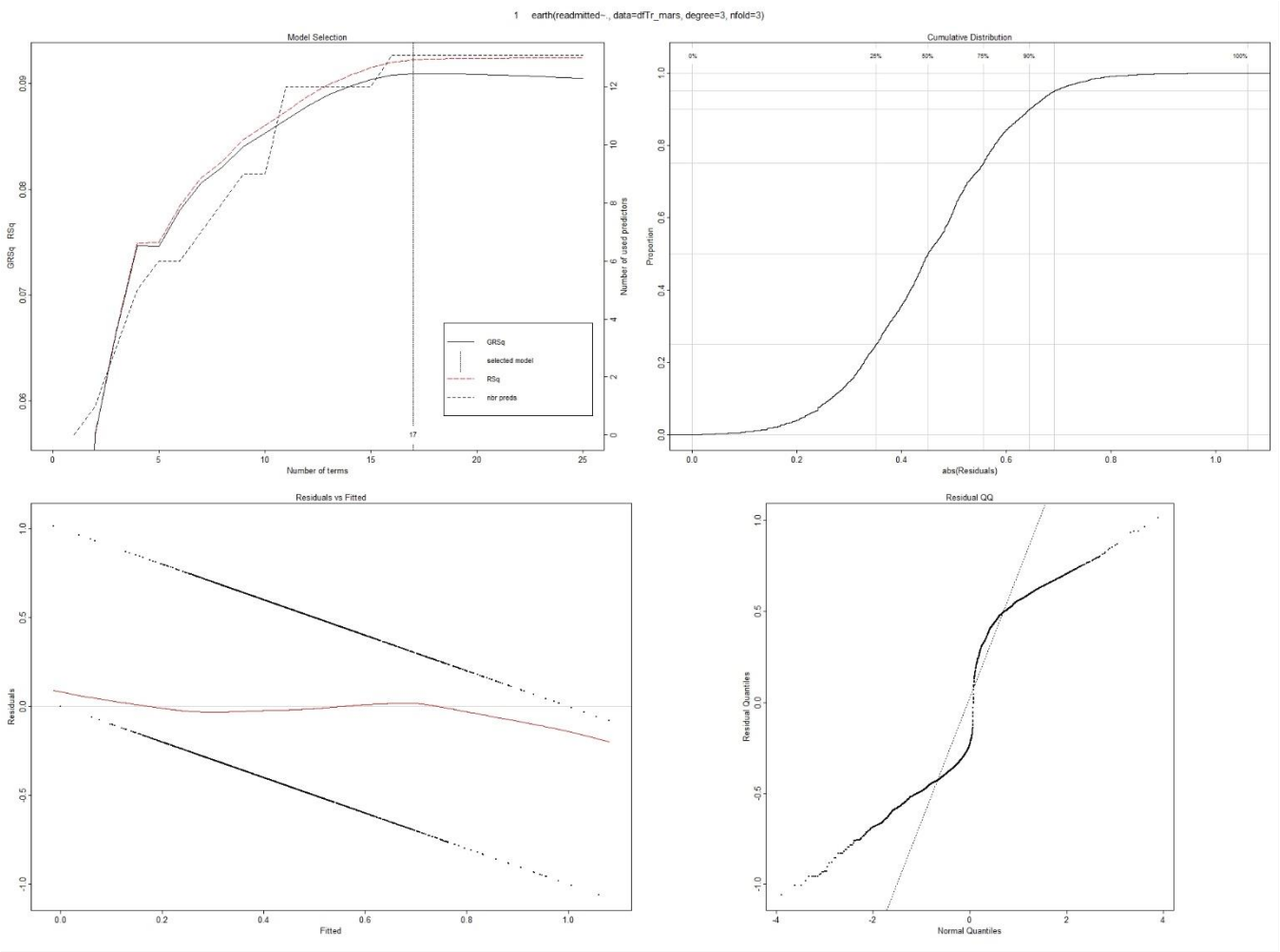
Using a different random forest model from random forest function, and specifying 10 trees and mtry = 3, we get an accuracy of 0.6401 and kappa value of 0.2651.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 24399 15003
         1  5343 11786

               Accuracy : 0.6401
                 95% CI : (0.6361, 0.644)
    No Information Rate : 0.5261
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2651

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8204
            Specificity : 0.4400
         Pos Pred Value : 0.6192
         Neg Pred Value : 0.6881
             Prevalence : 0.5261
         Detection Rate : 0.4316
   Detection Prevalence : 0.6970
      Balanced Accuracy : 0.6302

       'Positive' Class : 0
```

Now, by using MARS from the earth package, we can see the correlation between the variables. Using hyperparameter of degree = 3, gives the best result, since higher degree does not change the RSME value of the model.

```
Call: earth(formula=readmitted~., data=dfTr_mars, degree=3, nfold=3)

                                                                                                  coefficients
(Intercept)                                                                                        0.64240671
h(admission_source-6)                                                                              0.18986684
h(7-admission_source)                                                                              0.02020567
h(admission_source-7)                                                                             -0.18822677
h(2-number_emergency)                                                                             -0.03650560
h(2-number_inpatient)                                                                             -0.14200706
h(number_inpatient-2)                                                                              0.02479780
age[70-80) * h(2-number_inpatient)                                                                 0.02853282
h(2-number_inpatient) * diabetesMedYes                                                             0.02186759
h(admission_source-7) * h(28-num_lab_procedures)                                                  -0.00123096
h(7-admission_source) * h(1-num_procedures)                                                        0.01152832
h(1-number_outpatient) * h(2-number_emergency)                                                    -0.03691444
age[80-90) * h(2-number_inpatient) * diabetesMedYes                                                0.01334683
h(admission_type-1) * h(1-number_outpatient) * h(2-number_emergency)                                0.00884475
h(2-discharge_disposition) * h(2-number_inpatient) * h(9-number_diagnoses)                         -0.01015202
h(16-num_medications) * h(1-number_outpatient) * h(2-number_emergency)                             -0.00345477
h(num_medications-16) * h(1-number_outpatient) * h(2-number_emergency)                             -0.00087186

Selected 17 of 25 terms, and 13 of 77 predictors
Termination condition: RSq changed by less than 0.001 at 25 terms
Importance: number_inpatient, discharge_disposition, number_diagnoses, number_outpatient, number_emergency, admission_source, num_lab_procedures, ...
Number of terms at each degree of interaction: 1 6 5 5
GCV 0.2266485  RSS 12794.09  GRSq 0.09095743  RSq 0.09224343  CVRSq 0.08465148

Note: the cross-validation sd's below are standard deviations across folds

Cross validation:   nterms 15.67 sd 3.06    nvars 10.00 sd 1.73

    CVRSq    sd     ClassRate    sd     MaxErr  sd
    0.085 0.002         0.623 0.004       -1.2 1.3
> sqrt(mean(marsFit$residuals^2)) #RSME
[1] 0.4757309
```

Below is the plot of MARS model with degree hyperparameter of 3 and 3 number of folds.



Here are all data of the five different models:

Table 1 *Accuracy and Kappa Data for Different Models*

					CV performance	
Model	Method	Package	Hyperparameter	Selection	Accuracy	Kappa
random forest	rf	randomForest	ntree, mtry	10, 3	0.6401	0.2651
decision tree	rpart	rpart	cp	0.043	0.6133	0.2129
random forest	rf	caret	mtry, CV	6,3	0.6441	0.2468
Bag tree	ridge	glmnet	lambda	0.00029	0.5483	0.2836
MARS	earth	earth	degree	3	0.6487	0.6882