Homework 7: Team 8

Noah Bridges
Bhavya Reddy Kanuganti
Sujata Sahu

Problem 1

Tasked with the goal of determining which patients are most likely to be readmitted to the hospital, it was first necessary to clean the data to make it more amenable to any method of future analysis. The two primary forms this cleaning took were A) collapsing the factor levels of certain features (namely, diagnosis, source of original admission, type of discharge, and type of admission) and B) imputing missing values where needed.

The factor levels of the features listed above were collapsed based on commonalities -- whether by broad categorizations of diagnoses (e.g., heart disease, diabetes), or similar categories of discharge / admission (e.g., leaving the hospital under the supervision of a rehab specialist or being admitted based on the advice of a Primary Care Physician). The provided data contained relatively minimal numbers of missing values. As such, these were imputed using point estimates -- most frequently, the mode of a given feature. The chief exception to this rule was in imputing the missing values in the "Medical Specialty" columns. A large proportion of these values were missing, and creating an effective model to accurately predict the specialty would be a challenge unto itself. Thus, if a patient was admitted to the ER, the "medical specialty" was assigned to the level "ER". All other missing values were assigned to a category "other", to prevent confusing the signal that may exist among already declared specialties. The other feature with a high level of missingness was "payer_code" -- that is, whether a patient had insurance and what kind. Nearly half of these values were missing, and based on a preliminary assessment comparing payer code to the probability of readmission, it seemed like a poor predictor. Thus, it was excluded from all analytical models.

Additionally, the group decided to exclude specific medications from the majority of the analysis (although one random forest model was implemented which utilized the patient-level data on their usage of specific medications). This decision was made for two reasons. First, although these factors did not contain many missing values, the majority of patients were not on most of the medications, meaning that these might not offer population-level insights. Second, in

an effort to reduce the complexity of the models up-front, it was determined that excluding these features from analysis and focusing on the other attributes, many of which appeared to be much more significant, would help maintain predictive power without the risk of overfitting the models.

ii) Models and results

The group produced six different models to predict the probability of a patient being readmitted to the hospital. A basic logistic regression model was implemented (this model did not include interaction effects). Next, three tree-based models were used. A random forest model was produced on the data, as were two boosted models -- one using the package xgboost and the other using C5.0 in R. Fifth, a MARS classification model was fit to the data. Finally, a neural network was created in R using a single hidden layer. The details of these models (along with their overall performance, as measured by the Log Loss of their probabilistic measures and Cohen's Kappa, are shown in the table below.

Model	Method	Package	Hyperparameters	Selection	LogLoss (train data)	Kappa
Random Forest	randomforest	randomforest	Ntree, mtry	500, 5 (respectively)	0.467	0.54
logreg	glm	stats	N/A	N/A	0.643	0.24
Decision Tree	C.5.0Default	C50	Trials (num. boosts)	42	0.532	0.55
Neural Network	neuralnet	neuralnet	Number of hidden layers	1	0.68	0.24
Boosted Trees	xgboost	xgboost	Eta, max_depth, nrounds, gamma, subsample, minchildweight, colsample_bytree	N/A	0.70	0.47
MARS	earth	earth	Nprune, degree	16, 2	0.644	0.24

iii) In-Depth Analysis of a Specific Model

Given its reasonably high level of performance on the above metrics considered (at least on the training data) and its ability to provide additional insights into trends hidden within the data, the model which will be analyzed further is the C5.0 Boosted Decision Tree. A particularly helpful feature of this model reveals which features were used most frequently to determine the probability of a patient being readmitted to the hospital.

```
# C50 Tree
actual <- as.factor(Meds$readmitted)
c50Tree <- C5.0.default(noMeds[,-21], actual, trials = 100)
c50Tree
summary(c50Tree)
```

```
Call:
C5.0.default(x = noMeds[, -21], y =
 actual, trials = 100)

Classification Tree
Number of samples: 57855
Number of predictors: 20

Number of boosting iterations: 100 requested;  42 used due to early stopping
Average tree size: 518.1

Non-standard options: attempt to group attributes
```

The C5.0 tree revealed that nearly all the considered characteristics were used consistently to determine the probability of a patient being readmitted in the future. Of these features, those that were isolated as most important from a single iteration of a tree (rather than a boosted tree) were the number of times in a previous month that a patient had visited either the ER or the inpatient facilities of the hospital. Thus, patients who had a history of high hospital utilization were most likely to continue this high utilization.

While the above result may be somewhat intuitive, in contrast to the expectations of the team, the “indicator level” of a patient was rated as far less important a predictor by both the

boosted trees and the single iterations of trees. This seems to indicate that the severity of a patient's illness is not a strong predictor of their likelihood of re-entering the hospital.

Finally, when assessing the complexity of a single tree (rather than its boosted cousin), it reveals that particular features only possess reasonably strong predictive power when combined with other variables. If one assesses the pure probabilities of almost any single feature, the split between those who will be readmitted and those who will not is nearly 50-50. This fact illustrates the difficulty of creating accurate forecasts of a patient's probability of readmission -- at least with the data provided here. This fact also provides some insight into why a basic logistic regression model (which did not include interaction effects) did not produce particularly accurate predictions of patients' likelihood to be readmitted to the hospital.

(iv) Model Efficacy:

A) Log Loss of the C5.0 Tree model (result of boosted tree on all training data):

c50result <- LogLoss(c50pred[,2], hosData$readmitted)

c50result

```
> c50result <- LogLoss(c50pred[,2], hosData$readmitted)
> c50result
[1] 0.5414692
```

How closely the forecast probability matches the associated real or true value is indicated by log-loss (0 or 1 in case of binary classification). The higher the log-loss number, the more the predicted probability deviates from the actual value. The log loss value for our model we got is 0.54 .The lower the log loss the better is the model.

B) Confusion Matrix Outputs

confusionMatrix(as.factor(c50pred), as.factor(hosData$readmitted), positive="1", mode="everything")

```
> confusionMatrix(as.factor(c50pred), as.factor(hosData$readmitted), positive="1", mode="everything")
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 25360  8114
         1  5272 19109

               Accuracy : 0.7686
                 95% CI : (0.7652, 0.7721)
    No Information Rate : 0.5295
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5329

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

            Sensitivity : 0.7019
            Specificity : 0.8279
         Pos Pred Value : 0.7838
         Neg Pred Value : 0.7576
              Precision : 0.7838
                 Recall : 0.7019
                     F1 : 0.7406
             Prevalence : 0.4705
         Detection Rate : 0.3303
   Detection Prevalence : 0.4214
      Balanced Accuracy : 0.7649

       'Positive' Class : 1
```

Accuracy**:** 0.7686

C5.0 Boosted Decision Tree is 76% accurate at classifying positives (that is, 76% of the training data is correctly classified as either readmitted or not readmitted).

Precision: 0.7838

78% of patients that the model identified as more likely to be readmitted to the hospital were, in fact, readmitted to the hospital.

Sensitivity: 0.7019

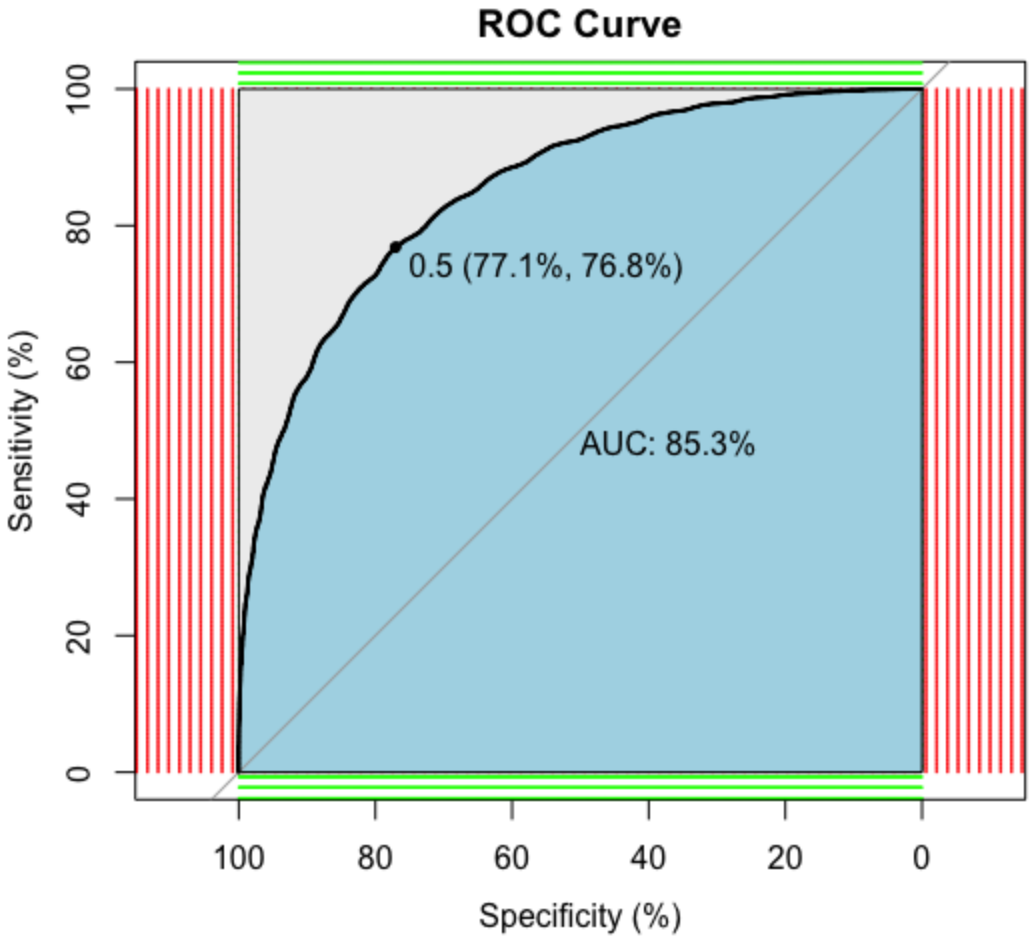
It is accurate to say that 70% of real readmitted instances were recognized by the model.

Specificity: 0.8279

71% of the real non-admitted cases were correctly identified.

C) ROC Curve

A graphical depiction called a Receiver Operator Characteristic (ROC) curve is used to illustrate how well binary classifiers can diagnose problems.It is calculated on predicted scores therefore the accuracy of the ROC is different from the actual accuracy which is calculated on predicted scores.The specificity and sensitivity scores are 77.1% and 76.8 %



D) Gini Index

The Gini coefficient is used in classification problems. It can be calculated as 2*AUC -1 = 2*0.76 -1=0.52

Generally having a gini coefficient above 0.6 is considered to be a strong predictive model; although this model failed to reach that threshold, it is another example of the difficulty of modeling with this data set.