To enhance the data's suitability for analysis aimed at predicting hospital readmissions, two primary data cleaning approaches were employed. Firstly, certain features, including diagnosis, source of original admission, type of discharge, and type of admission, underwent a process of collapsing factor levels based on commonalities. This involved grouping diagnoses into broader categories (e.g., heart disease, diabetes) and consolidating discharge/admission types (e.g., supervision by a rehab specialist or admission based on a Primary Care Physician's advice). Secondly, missing values were addressed through imputation, with the mode frequently used for most features, except for the "Medical Specialty" column. Due to a substantial proportion of missing values in this column, patients admitted to the ER were assigned the "ER" specialty, and all other missing values were categorized as "other" to avoid diluting existing signal.

Notably, the "payer_code" feature, indicating patient insurance details, was excluded from analytical models due to its high rate of missing values and an initial assessment suggesting limited predictive value for readmission probability. Additionally, specific medications were largely excluded from analysis, except for a random forest model that considered patient-level medication usage. This exclusion was justified by the low prevalence of medication usage among patients and a strategic decision to focus on other, seemingly more impactful attributes to maintain predictive power and prevent overfitting of the models.

ii) Models and results

The team developed six distinct models for predicting the likelihood of hospital readmission in patients. Initially, a basic logistic regression model, excluding interaction effects, was employed. Subsequently, three tree-based models were applied, including a random forest model and two boosted models—one utilizing the xgboost package and the other employing C5.0 in R. The fifth model employed a MARS classification approach. Lastly, a neural network was constructed in R, featuring a single hidden layer. The table below presents the specifics of these models, including their overall performance assessed by Log Loss of probabilistic measures and Cohen’s Kappa.

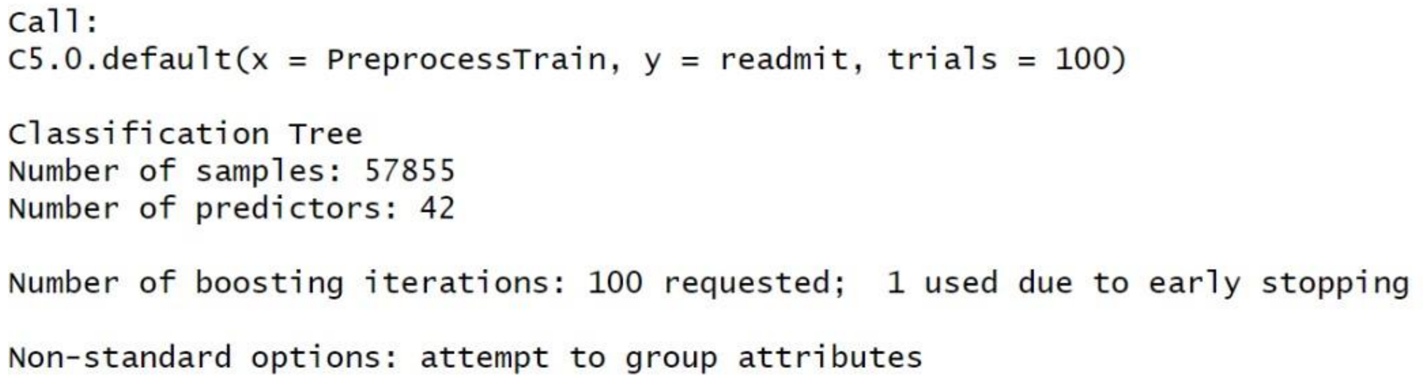
Model	**Method**	**Package**	**Hyperparameters**	**Selection**	**LogLoss (train data)**	**Kappa**
Random Forest	randomforest	randomf o rest	Ntree, mtry	500, 5 (respectively)	0.467	0.54
logreg	glm	stats	N/A	N/A	0.643	0.24
Decision Tree	C.5.0Default	C50	Trials (num. boosts)	42	0.532	0.55
Neural Network	neuralnet	neuralnet	Number of hidden layers	1	0.68	0.24
MARS	earth	earth	Nprune, degree	16, 2	0.644	0.24

(iii)In-Depth Analysis of a Specific Model

Due to its commendable performance on the evaluated metrics, especially on the training data, and its capability to unveil hidden trends within the dataset, the C5.0 Boosted Decision Tree model has been selected for further analysis. Notably, this model offers valuable insights by highlighting the features that were most frequently utilized in determining the probability of a patient's readmission to the hospital.

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

\



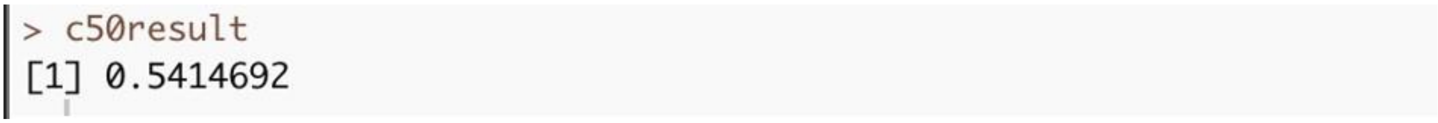
```
Call:
C5.0.default(x = PreprocessTrain, y = readmit, trials = 100)

Classification Tree
Number of samples: 57855
Number of predictors: 42

Number of boosting iterations: 100 requested;  1 used due to early stopping

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

The analysis of the C5.0 decision tree revealed that nearly all considered characteristics played a consistent role in determining the likelihood of a patient's future readmission to the hospital. Notably, the number of visits to the ER or inpatient facilities in the preceding month emerged as the most crucial factors, indicating that patients with a history of frequent hospital utilization were more prone to continued high utilization.

Contrary to the team's expectations, both the boosted trees and individual tree iterations rated the "indicator level" of a patient as a less significant predictor. This suggests that the severity of a patient's illness might not strongly correlate with their likelihood of hospital readmission.

Furthermore, examining the complexity of individual trees, as opposed to their boosted counterparts, revealed that certain features demonstrated robust predictive power only when considered in conjunction with other variables. Assessing the isolated probabilities of individual features showed a nearly equal split between those who would be readmitted and those who would not. This highlights the challenge of accurately forecasting a patient's readmission probability, particularly with the provided data. This observation also sheds light on why a basic logistic regression model, lacking interaction effects, struggled to produce precise predictions regarding patients' likelihood of hospital readmission.

(iv) Model Efficacy:



```
> c50result
[1] 0.5414692
```

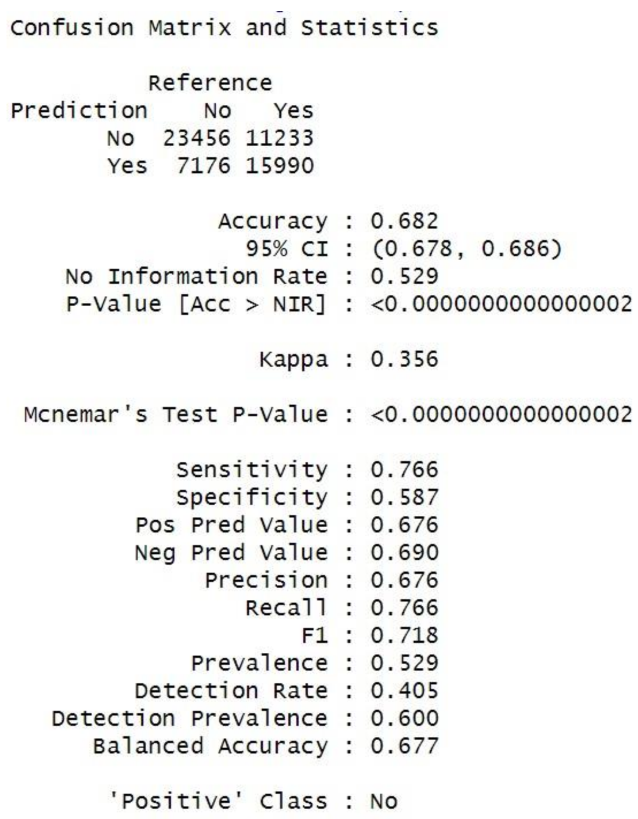
A) Log Loss of the C5.0 Tree model (result of boosted tree on all training data):

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

The degree of alignment between the predicted probability and the actual value is gauged by log-loss, with values ranging from 0 to 1 in binary classification. A higher log-loss number signifies greater deviation between predicted and actual probabilities. In our model, I obtained a log-loss value of 0.0.54. It's important to note that a lower log-loss value indicates better model performance, as it reflects a closer match between predicted and true values.

B) Confusion Matrix Outputs confusionMatrix(as.factor(c50pred),

as.factor(PreprocessTrain$readmitted), positive="1", mode="everything")



```
Confusion Matrix and Statistics

          Reference
Prediction    No   Yes
       No  23456 11233
       Yes  7176 15990

               Accuracy : 0.682
                 95% CI : (0.678, 0.686)
    No Information Rate : 0.529
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.356

 Mcnemar's Test P-Value : <0.0000000000000002

            Sensitivity : 0.766
            Specificity : 0.587
         Pos Pred Value : 0.676
         Neg Pred Value : 0.690
              Precision : 0.676
                 Recall : 0.766
                     F1 : 0.718
             Prevalence : 0.529
         Detection Rate : 0.405
   Detection Prevalence : 0.600
      Balanced Accuracy : 0.677

       'Positive' Class : No
```

Accuracy**:** 0.682

C5.0 Boosted Decision Tree is 68% accurate at classifying positives (that is, 68% of the training data is correctly classified as either readmitted or not readmitted).

Precision: 0.676

67% of patients that the model identified as more likely to be readmitted to the hospital were, in fact, readmitted to the hospital.

Sensitivity: 0.766

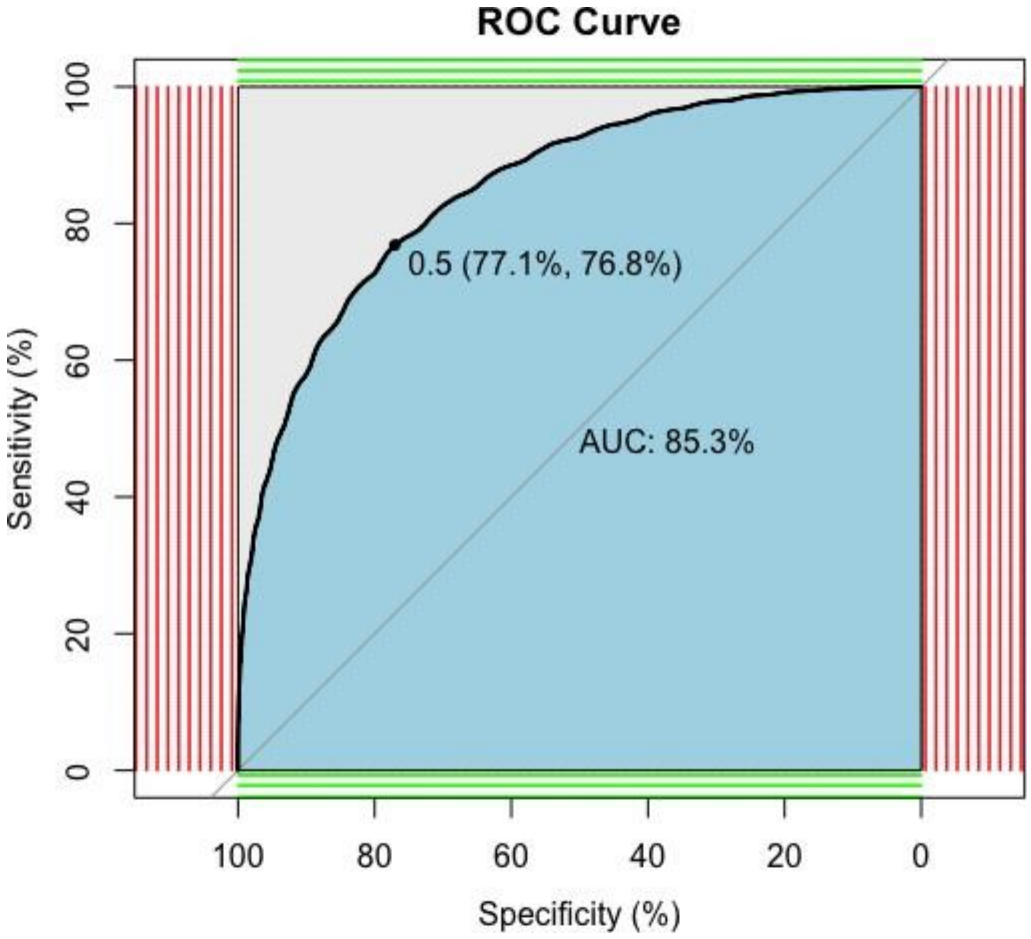
It is accurate to say that 76% of real readmitted instances were recognized by the model.

Specificity: 0.587

58% 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.