**Performance Assessment: D209- Predictive Analysis**

**A. Research Question**

**1.** For this assessment, the research question is as follows: can we predict readmissions based on the patients survey responses using the random forest method?

**2.** The goal of this analysis is to determine, using the random forest method, if we can predict whether or not a patient was readmitted into the hospital within a month of release based on their responses to the survey questions.

**B. Method Justification**

1. Random Forest works by growing “multiple decision trees which are merged together for a more accurate prediction” (Meltzer et al). Instead of just having a single decision tree, random forest uses bootstrap sampling from the training data set in order to build these multiple decision trees for a single target variable. Because of this we can expect a higher accuracy score for random forests than we would for a single decision tree, and we also expect there to be significantly less overfitting since the results are from multiple trees so less data is crammed into a single one.

**2.** One assumption of random forest is “the predictions from each tree must have very low correlations” (M, Shruti 2023). What this means is that having predictors that are correlated to each other can affect how the model is interpreted overall which can affect the overall accuracy of the model.

**3.** The packages that will be used in R and their respective uses are:

* randomForest: for the random forest analysis
* tidyverse: for data cleaning
* caret: for the random forest analysis
* caTools: for the random forest analysis
* pROC: for calculating the area under the curve

**C. Data Preparation**

**1.** In order to prepare the data for analysis, I will start by removing the unwanted variables from the data set, followed by renaming the item variables into something that is easier to understand. After doing this, I will check for missing data and remove those entries if necessary. The last step to prepare the data is to convert the target ReAdmis variable from No/Yes into 0/1 and then convert to a numeric.

**2.** The following variables were used for the analysis:

* ReAdmis: categorical
* Item1/TimelyAdmission: continuous
* Item2/TimelyTreatment: continuous
* Item3/TimelyVisits: continuous
* Item4/Reliability: continuous
* Item5/Options: continuous
* Item6/Hours\_of\_Treatment: continuous
* Item7/CourteousStaff: continuous
* Item8/Evidence\_of\_Active\_Listening: continuous

**3.** All of the steps to prepare the data for analysis were described above in part **C1** and the code used to prepare the data has been attached as an RScript file alongside this written assessment.

**4.** A copy of the cleaned data set has been attached alongside this written assessment.

**D. Analysis**

**1.** The data was split into training and testing sets using a 70/30 ratio. The respective data sets have been attached alongside this written assessment.

**2.** Below are the steps used to perform the analysis in R:

* Load all necessary packages as described above in part **B3**
* Create a data partition called myIndex on a 70/30 ratio (or p = 0.7) and use this data partition to create training and testing data sets, which were then written into CSV files
* Turn the target ReAdmis variable into a factor for the analysis in each data set
* Create the random forest model using the training data set with ReAdmis as the target variable
* Using this model, create a prediction using the testing data set to predict readmissions with a classification type
* Using this prediction, I then created the confusion matrix to determine how accurate our model is
* Since this is a classification random forest, it is impractical/impossible to determine the MSE of our model. Here it is best to use the area under the curve. Because of this the next step is to create a probabilities prediction using our initial model and the testing set
* Create our ROC Object using these probabilities and use it to find the area under the curve

No other intermediate calculations were used.

**3.** The code to perform the analysis has been attached as a RScript file alongside this written assessment (“How to Perform Random Forest In R, n.d.).

**E. Data Summary and Implications**

**1.** The following screenshot shows the confusion matrix created using the random forest model and the testing data set predictions:

A screenshot of a computer

Description automatically generated

As stated here by the confusion matrix, the accuracy of the prediction is approximately 60.6%. This means that the model can only correctly predict readmissions a little over 60% of the time. This accuracy is not very high, and we can interpret that the model is not any better than simple random guessing of the readmissions variable as the value is significantly closer to 0.5 than it is to 1.

The following are screenshots of the AUC and the plot of the ROC:

A computer screen shot of a computer code

Description automatically generated

A graph showing a curve

Description automatically generated with medium confidence

In a regression random forest model, it would be appropriate to use the MSE to properly evaluate the model, however since the target variable for this assessment is a categorical variable, MSE is not appropriate. This is a classification random forest model, so instead of MSE (would is impossible for classification models) I calculated the area under the curve and displayed it and the plot in the pictures above. The area under the curve came out to 0.4916. Just like the confusion matrix is significantly closer to 0.5 than it is to 1 which suggests that this model is no better than random guessing and is not a reliable way to predict the readmission variable.

**2.** As stated in the previous step, neither the accuracy nor the area under the curve suggest the model is accurate. The model is not any better than simply guessing and is not a reliable way to predict the readmission variable using the survey responses.

**3.** One limitation of this data set is that the patients reason for admittance and readmittance are not included. A patient could have been admitted to the hospital for a reason totally unrelated to their condition, such as an injury, and readmitted later for the same reason (Carlan 2023).

**4.** There is no statistical evidence that suggests we can use the survey responses as a way to predict readmissions. The recommended course of action is to keep treating the patients normally regardless of their survey responses. Further research is needed to make any informed decision regarding the patients survey responses and readmissions.

**F. Sources**

“How to Perform Random Forest in R -.” ProjectPro, www.projectpro.io/recipes/perform-random-forest-r. Accessed 17 Oct. 2023.

Carlan, Christian. “Performance Assessment: D209- Classification Analysis.” 2023. Microsoft Word file.

Meltzer, Rachel, et al. “What Is Random Forest? [Beginner’s Guide + Examples].” CareerFoundry, 31 Aug. 2023, careerfoundry.com/en/blog/data-analytics/what-is-random-forest/#:~:text=2.-,How%20does%20the%20Random%20Forest%20algorithm%20work%3F,group%20than%20they%20do%20alone.

M, Shruti. “Introduction to Random Forest in R.” Simplilearn.Com, Simplilearn, 23 Feb. 2023, www.simplilearn.com/tutorials/data-science-tutorial/random-forest-in-r.