Homework 7

Summary of Model Performances

Model	Method	Packages	Hyperparameters	Selection	Accuracy	Kappa
Decision Tree	rpart	rpart	cp	9.183411e-4	0.6229	0.2350
GLM	glm	stats	NA	NA	0.6178	0.2177
LDA	lda	mass	NA	NA	0.6136	0.2091
Random Forest	rf	randomforest	mtry, ntree	3, 20	0.6196	0.2255
Neural Net	nnet	nnet	size, decay	3, 0.2	0.6227	0.2333
MARS	earth	earth	degree, nprune	2, 10	0.6222	0.2312
KNN	knn	class	k	45	0.6143	0.2166

*10-fold cross validation was used for every model above

Insights from Decision Tree Models

Two decision trees were initially created using numerical attributes and non-numerical attributes. The resulting node dividing attributes from each were then selected to develop an optimal set of features.

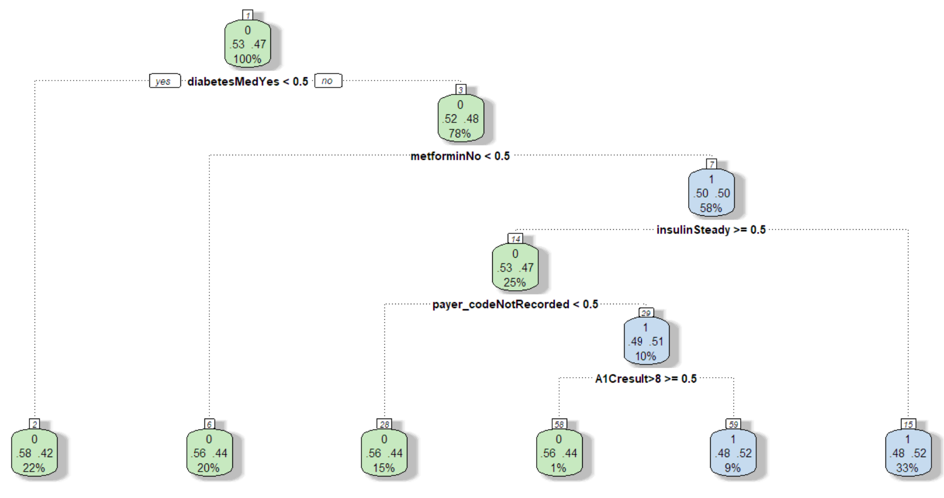

Figure 1 – Decision tree using non-numeric attributes.

Out of 33 non-numerical attributes, the following five resulted in the best node splits for the non-numeric decision tree: diabetesMed, metformin, insulin, payer_code, and A1Cresults.

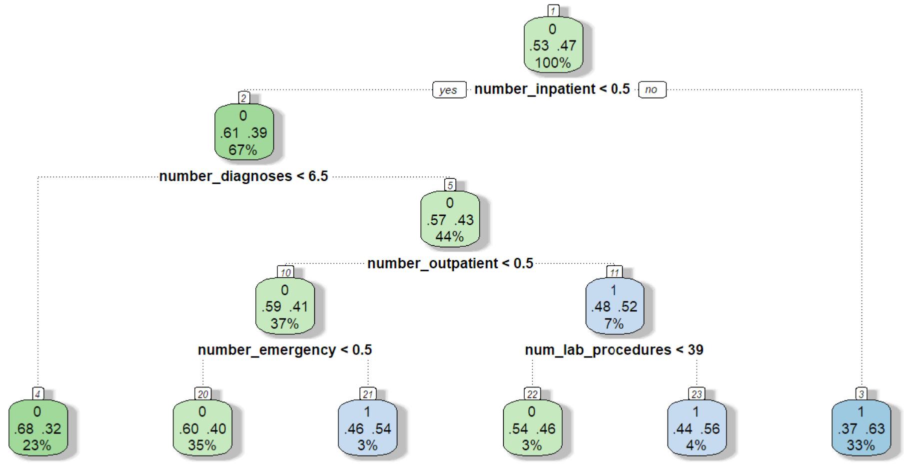

Figure 2 - Decision tree using numeric attributes.

Out of 13 numerical attributes, the following five resulted in the best node splits for the numeric decision tree: number_inpatient, number_diagnoses, number_outpatient, number_emergency, and num_lab_procedures.

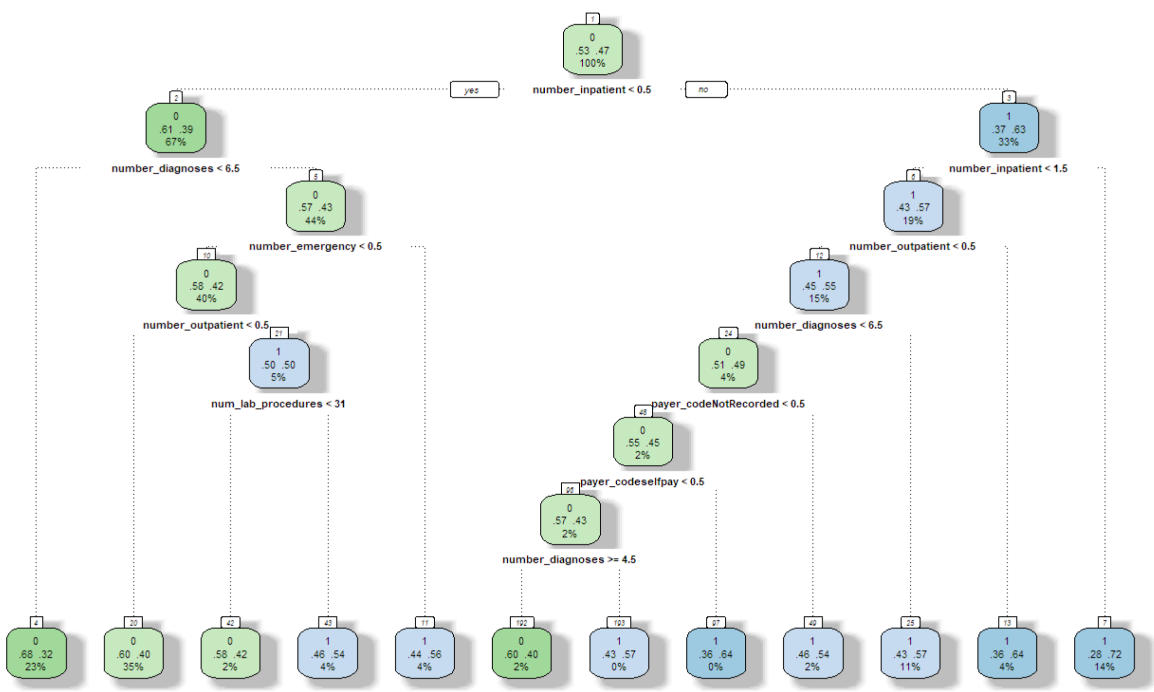

Figure 3 - Decision tree using optimal set of attributes.

Out of 10 previously identified attributes, the following six resulted in the best node splits for the optimal set decision tree: number_inpatient, number_diagnoses, number_emergency, num_lab_procedures, number_outpatient, and payer_code.

Summary of Insights:

- The number of outpatient, inpatient, and emergency visits in the previous year greatly influenced whether patients would be readmitted.
- Similarly, the number of diagnoses and lab procedures a person had also undergone considerably influenced their chances of readmittance.
- Whether the payer code is recorded or if the patient has previously self-paid for treatment has a notable influence on whether the patient was readmitted.

Performance Evaluation of Decision Tree Model

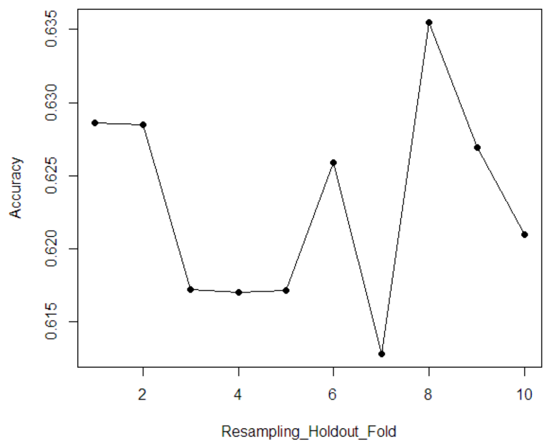

Figure 4 – Accuracy of best performing decision tree model.

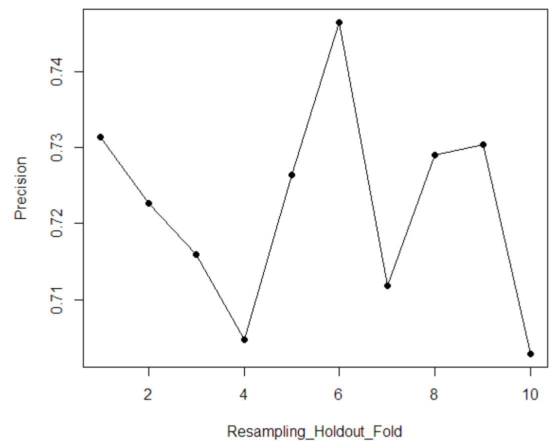

Figure 5 – Precision scores of best performing decision tree model.

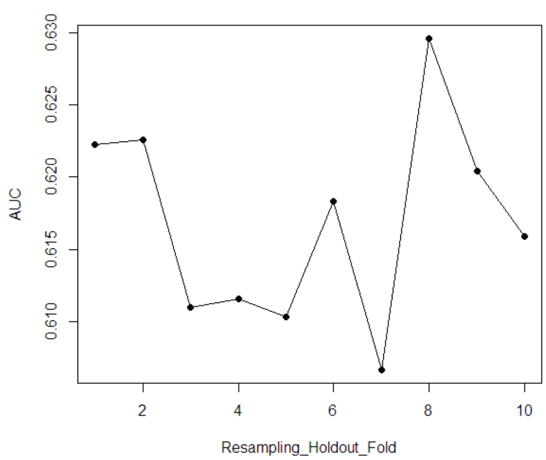

Figure 6 – AUC scores of best performing decision tree model.

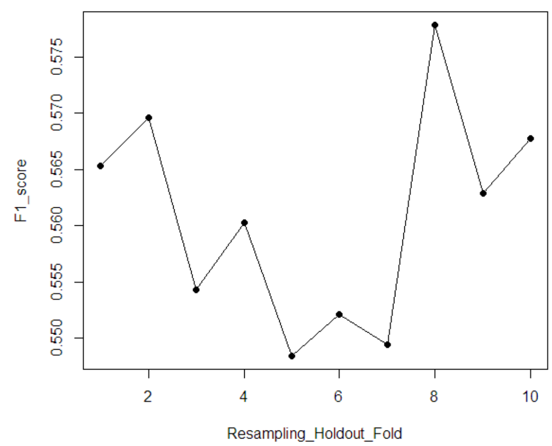

Figure 7 – F-1 scores of best performing decision tree model.

Observations:

- The model's accuracy varied between the different resampling holdout folds. However, the variation seems very little overall with a range between 0.6 – 0.635 and for most folds it held higher than 0.615.
- The precision of model decreased initially but seemed to improve between folds 4-6. However, for the latter of the folds precision of the model continued to decrease. Overall, the model held high precision as it was above 0.7 for all the folds.
- The F1-scores varied between the different folds with the lowest scores between folds 3-7. For fold 10, the last fold, the F1-score was between 0.565 – 0.570 indication classification errors.
- The AUC values followed a similar pattern to the F1-scores. However, the AUC values for all folds are slightly higher than their corresponding F1-scores.

Conclusions:

- Considering the accuracy and precision results, the model struggles to correctly classify the classes but is able to predict admitted patients relatively well.
- The AUC scores range between 0.6 – 0.635, meaning the model is only slightly better at classification than randomly guessing. However, this may be due to there being only two classes.
- The F1-score of the model remained above 0.5 for all folds, meaning the model performed fair but there is significant misclassification of patient readmittance.