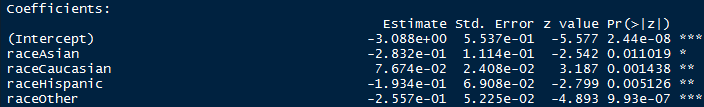
Brian Keafer

Hospital Readmits Classification Problem

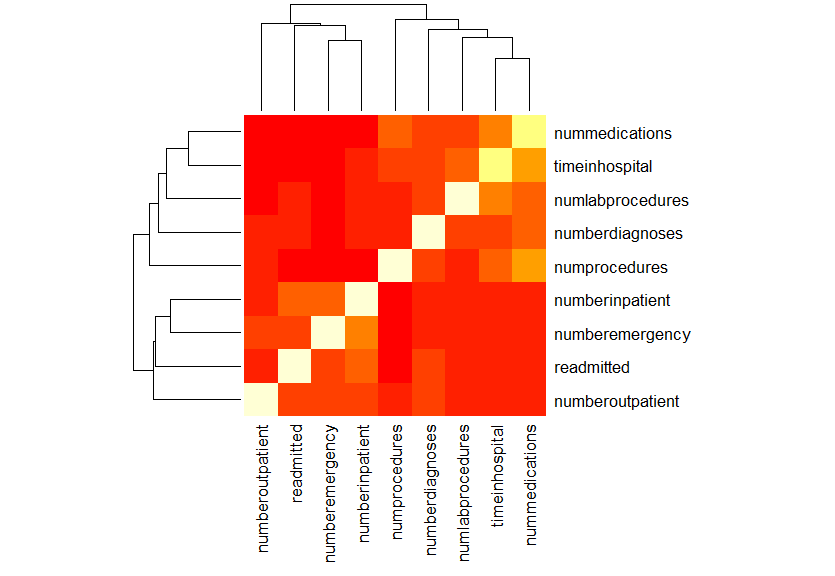
**Part (A) Insights**

**Insight for Insurance Companies:**

****A close up of a sign

Description automatically generatedReadmission rates are logically and intuitively linked to age. This is not breaking news to the insurance industries as insurance rates change depending on age group. This is supports by the significance each age group has when evaluating p-values against the readmitted value (see below). Additionally, the “variable” within the readmitted dataset was significant in predicting a readmittance rate. Although, this variable had the greatest amount of missing data, it was still quite obvious that patients with insurance were far more likely to me readmitted compared to those who payed out of pocket or used medicare. Further, grouping the patients that did not report payercode, were the least likely to be readmitted. A further insight on the demographics of the patients was that the race variable was most significantly influenced by “Other” so those individuals that did not identify with the given choices were likely to be readmitted (Evidence Below).

**Insights for Hospitals/Doctors:**

An analysis in the numeric data available revealed some interesting relationships. Time in hospital, number of procedures, and number of medications are all interrelated. They all also had very significant affects on the readmittance variable across many different modeling types. The greater the counts of each numeric value, the greater the probability of readmittance. Hospitals should expect to see patients again whom have recorded high counts in any of these numeric variables. This is supported by the heatmap seen to the right.

**Insight for Patients**

Six different classification models were built with optimized parameters and an extensive data set, yet the accuracy rate of prediction of readmittance did not exceed 65%. This clearly shows all of the complications of predicting if a patient will be readmitted. The greater number of procedures, diagnoses, and medications increased the chances of readmittance. Further, the single most significant predictor variable for a patient to be aware of is if they take or have had a change in their insulin medication. The extraordinary high z-value and extraordinary low p-value for this variable in each of the classification models, reveled the importance placed upon a patient to follow up with their doctor after a change in medication. Despite the complexity of the models produced, we are able to predict with reasonable accuracy if a patient will return to the hospital, but the question of why continues to remain unclear.

**Part (B)**

**Performance Evaluation Techniques**

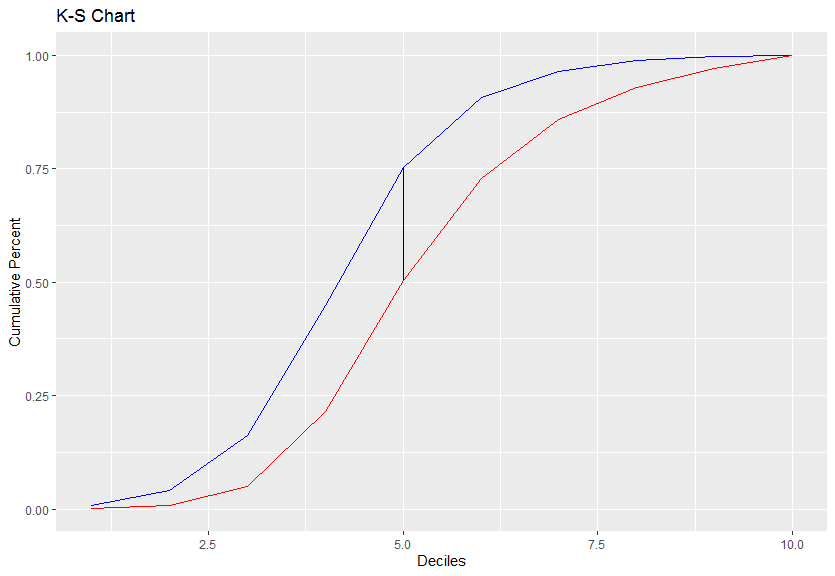
In order to most efficiently and effectively evaluate a model, the logistic regression model was selected as it would run relatively quickly, allowing prompt analysis. The most accurate model was in fact, random forest but due to the computing power needed to re-run the random forest model with small changes depending on the evaluation technique, this was not feasible. However, the logistic regression model still proved through evaluation techniques, to be a feasible model in classification prediction of readmittance.

The logistic regression model was trained using the same variables as the random forest model. The formula build can be seen below as an excerpt from the r-script, including each predictor variable.

***logreg\_form <- readmitted ~ race + gender + age + admissiontype + dischargedisposition + admissionsource + payercode + medicalspecialty + numlabprocedures + numprocedures + nummedications + numberoutpatient + numberemergency + numberinpatient + diagnosis + numberdiagnoses + diabetesMed***

Upon initial creation of this logistic regression model, it was tuned to have a decreased logLoss value. This value is what will be measured upon entry into the Kaggle competition, therefore logLoss was the single most important evaluation technique used. The model produced a logLoss of 0.641143, which by competition standards is quite high and far exceeds the logLoss of the sample submission and most other competitors. It is also the most important because we are reporting predictions based upon probabilities of classification. The loss function enables the inclusion of the price paid for inaccuracy. This particular logreg model passed this initial standard test.

The ROC function applied to the logreg model shows three different evaluation values under the same umbella. The ROC measurement. The ROC value is a measurement of area under the curve. An area of 0.5 represents a model that is as good as random. The ROC value for the logreg model is 0.6798, well above random and seemingly pretty decent. AUROC sensitivity and AUROC specificity are metrics associated with area under the ROC curve. Sensitivity is the true positive rate or recall. It represents the percentage of instances from the first positive class, “Yes”, that were actually predicted correctly. This metric at 0.7506 is very high, this model is quite good at predicting positive values for “Yes”. However, there was a lower success rate in specificity at 0.4935, which is essentially a measurement of negative instance that were predicted correctly. This model is much better at the former.

This leads to the more formalized measurement of percentage correct classifies out of all instances. The accuracy metric as applied to the logreg model was 0.631942. Very much significant and validates this model as being more than random guesses. The kappa metric is even more validation that the logreg model is useful, the kappa metric is a measure of accuracy considering that the possibility of correctness could happen by chance. The kappa metric on the logreg model sits at 0.252354.

The Kolmogorov-Smirnov chart is another metric that validates the usefulness of the logreg model. This statistic is a measure of the distance between the empirical distributions (Blue) and cumulative distributions (Red). The visualization to the right displays the difference between these two curves, which is the calculation for the KS- Test. The difference in our model is about 0.25. Not great, but good enough as to not discredit the logreg model.

The d-statistic shows the difference in the mean probability of readmittance and no readmittance (both factor levels). This difference indeed shows that mean probability of readmittance is about 0.10 greater than the latter. The d-statistic does not validate the accuracy of the model, but rather provides valuable information about each prediction as compared to the alternative “negative” prediction.

Overall, the evaluation metrics applied to the logreg model for predicting readmittance, did not show extraordinarily accurate results. However, the logreg model did prove to be a feasible model that predict readmittance at a significantly greater rate than a random prediction might.

**Part (C)**

**Summarize Model Performances**

The Random Forrest Model provided the most accurate predictions on readmittance. This training technique also required the most computing power and time. The hyperparameters were optimized using the train function in the caret library. There are significant advantages in creating easily adjustable and simple models, but the Random Forrest model ultimately proved to be the most useful.

**Part (D)**

Model’s have been submitted to Kaggle as well as a Notebook that shows how to create proper column and factor level naming conventions required for certain model types.