With the emergence of big data, a variety of information is being accumulated to hopefully one day be turned into meaningful insight. For this analysis, data taken from the first 24 hours of intensive care will be evaluated and models will be created to help predict patient survival based on the information gathered. It is known that African American women face disparities within healthcare realm, so this analysis will focus on this subset of these patients. Hopefully with this information, Cleveland Clinic will learn how to better serve their African American women population. By having predictions of the survival rate for each patient, Cleveland Clinic will be able to make more informed decisions about which patients to monitor more closely and help increase their rate of survival.

The original dataset had a lot of missing values. To handle missing categorical values, all NA’s were transformed to the string “Unknown”. For the quantitative variables, k-nearest neighbor method was employed to impute predicted values to replace the ones that were missing. But to begin, columns that had more than 30% missing were removed and then observations with more than 30% missing were removed.

After the handling of the missing data, about 131 variables remained. To allow an efficient analysis, the number of variables were further reduced to a more manageable size. Variables were removed if they served no purpose, for example readmission\_status was removed because it had the same value “1” for every observation. Variables were also removed if they caused multicollinearity within the dataset. Variance inflation factor (VIF) was used to help pinpoint the variables that showed signs of that. For example, the variables “d1\_diasbp\_invasive\_max”, “d1\_diasbp\_noninvasive\_max” and “d1\_diasbp\_max” showed multicollinearity between each other. After looking more at the data, they all three had very similar values so the first two were removed.

Most of the variables remaining align closely to the health of the individual patient. For example, the variable BMI, a high BMI can be an indicator of high body fatness. The variable d1\_temp\_max is the patient’s highest body temperature taken within the first 24 hours. A high temperature is a sign that something out of the ordinary is going on in the patient’s body, often due to an illness.

For the analysis, two data mining techniques were employed to help predict survival rate for patients, Logistic Regression and Artificial Neural Networks (ANN). These two models were chosen because they both work well for categorical response variables, for this case “hospital\_death” which states whether a patient is deceased or not. Since all the predictor variables for ANN has to be quantitative, all the categorical variables were transformed to binary formats. Both of the techniques were fit on all the remaining variables. Cross validation was used in the inside layer to select the best model and, due to timing, a validation set was used to validate the best model on the outside layer.

The result of the data mining showed that logistic regression fit on all the remaining variables held to be the best model. 98% of its validated classification matched the actual results. Since the original data set had 90% + of the data resulting in no deaths, the Area Under the Precision-Recall curve method was also used to assess the model. Shown in figure 1, this resulted in an AUC of 0.3454853. Because a validation set, and Cross-Validation was used, this model is not overfitted to this specific dataset. So, when new data is used from the patient information gathered from the Cleveland Clinic, this model will still be useful.

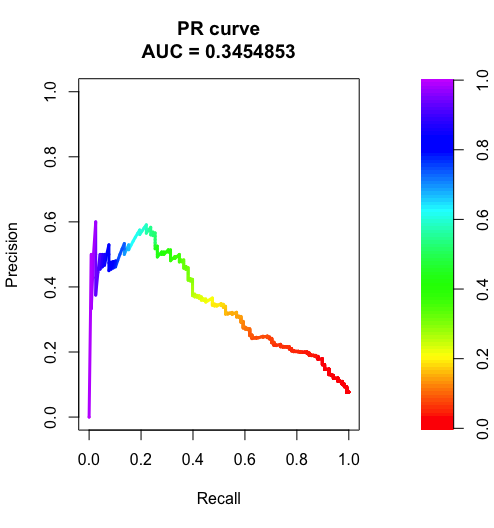


Figure 1