Anthony Marcolongo

D209

01/05/2021

Medical Data Analysis with Random Forest

The goal of this project was to see if it is possible to use the random forest method to determine if a patient is likely to be readmitted to the hospital. The dataset in question is made up of 10,000 rows and 50 columns. The dataset was cleaned and prepared for use with the random forest method. Unlike other models, random forest can select features, so the data does not need to be normalized prior to use. The data needs to be split into a training and testing set, though. Once the data is split, the random forest model is applied to each set. The error is then plotted as a function of ntrees. This allows for selecting the number of trees that minimizes the error. For this project ntrees = 100 is the ideal value. The random forest model is again applied to the testing and training data, this time with ntrees fixed at 100. The accuracy of the models is then computed, and the most important factors are determined.

The most important factors for determining hospital readmittance are age and income. These are not surprising results. The older a person is, the more likely they are to suffer from medical complications. Since this project is considering people who have already been hospitalized once, it makes sense that older patients would be more likely to be readmitted. Higher income leading to a higher probability of readmittance, might seem like a surprising result on the surface due to people with more money eating better and being able to afford better care, but they are also much more likely to have health insurance. A person without health insurance might need to go to a hospital, but they cannot afford to and so they just do not go.

After creating the random forest models, the next step was to check the accuracy of both models. Both models are determined to be 60% effective in predicting if a patient will need to be readmitted to the hospital. When you consider all the factors that go into someone needing to be readmitted, this is an impressive result. A hospital being able to correctly identify 60% of their readmittance patients in advance would be extremely helpful. This would allow the hospital to target programs specifically to patients that are very high risk for readmittance. The hospital could create out-patient programs for older people to help detect problems before those patients need to be readmitted. The hospitals could also work with insurance companies and try to come up with ideas to discourage patients from seeking readmittance unless they truly needed it.