Final project

Due April 25, 2024

The final project requires that you build a predictive model based on real data – your own or the provided National Impatient data– and a paper-style short report (2-3 of pages long) describing the problem, the approach(es) taken, and the results. Below is a *guideline* structure for the report. You should use the section breakdown into intro, methods, results, conclusions/discussion but don't have to necessarily include every element listed below within those sections. And you may want to include elements not listed below. Use your judgement.

Introduction

1. Describe the problem explaining in particular why prediction is of primary interest (inference could also be of interest but there has to be a good reason for wanting to predict a particular outcome)
2. Describe the data (e.g. data source, data collection, outcome of interest, available features, sample size, missing data, etc.)

Methods

1. Describe any data pre-processing steps (e.g. cleaning, recoding, variable transformation, dealing with missing data, selection of features to be included in your models, etc)
2. Briefly describe the Machine learning methods you will be using and why they are appropriate for your data (e.g. given the sample size and dimensionality of your training data, are you more concerned about bias or variance?) You should try and compare at least 3 distinct appropriate methods.
3. Describe how you are splitting the data into testing and training and/or resampling strategy used for comparing methods, tuning parameters, and/or model/feature selection.
4. If applicable, describe any model/feature selection used.
5. If applicable, describe any tuning parameters and how you will be tuning them.
6. Describe what performance metric(s) you will be using and why.

Results

1. Present key summaries (table and/or plots, but plots preferred when both available) of your data (e.g. class frequencies if a classification problem)
2. Report training, validation/cross-validation, and test errors. Present cross-validation plots for tuning parameters if available. Report variable importance (e.g. p-values, model coefficients, Random forest and boosting variable importance).

Conclusions/discussion

Discuss whether and why the prediction model(s) developed achieved sufficient high accuracy to be usefully deployed to predict new observations.

#Additional notes for those using the NIS data The data provided consists of a random subset of 200,000 patients from 2012 from the National Impatient Sample (NIS) data collected by the Healthcare Cost and Utilization Project (HCUP). You can find information on the HCUP database at https://www.hcup-us.ahrq.gov (https://www.hcup-us.ahrq.gov). You can choose to develop a model to predict death during hospitalization also known as impatient mortality (variable DIED in the dataset) or hospital length of stay (variable LOS in the dataset). For extra credit, you can also choose to predict both. The dataset has a relatively large number of variables. In the provided data dictionary I preselected variables (highlighted) which are both availabe (not all variables in the dictionary are available for 2012) which might be relevant for predicting impatient mortality and/or hospital length of stay. Based on their description and additional info from the HCUP site you should choose which variables among the preselected ones you will consider as features/predictors. You don't have to use them all. There maybe variables that are redundant (capture pretty much the same info others already capture), variables that are too complex (e.g. categorical with way too many levels), or that based on your judgment are unlikely to be important. Be aware that the data is real and has not been pre-processed in any way and you will have to do some data cleaning. For example, you should carefully check the variables you consider as possible predictors for correctness of type (e.g. many numeric variables will be read in as factor variables when you use `read.csv`), outliers, missing observations, nonsensical values, etc.