Final Project Proposal

Statistical Computing

E. Kelley

Fall 2022

Introduction

I will be analyzing a Kaggle data set, Length of Hospital Stay (LHS) (https://www.kaggle.com/datasets/aayushchou/hospital-length-of-stay-dataset-microsoft). My motivation for selecting this data set was driven by a few aspects. This was the first time I have explored Kaggle, and I was surprised by how many of the datasets included relatively few variables. I wanted to find a data set rich in predictors, so I could perform model feature selection. Additionally, I reasoned that a "medical" data set might present challenges similar to what I might encounter in the "wild," and, consequently, would help me gain practical experience. This data set has 29 variables and 100,000 observations. Given the large number of observations, I will take a subsample of 1000 observations for ease of computation. The focus of this analysis will be prediction of `lengthofstay` (days spent in hospital) using a mix of categorical (n=13) and numeric (n=10) variables. `lengthofstay` ranges from one to 14 days (Figure 1). The categorical variables appear to be a collection of various diseases (eg, asthma) and risk factors (eg, malnutrition) (Table 1). I expect to drop `fibrosisandother` as my sample only contains two observations that are positive instances. There is a fairly even split across gender. The numeric variables are largely composed of blood metrics, but BMI, pulse, and respiration are also present. Some dates are included: vdate (visit date) and discharge (discharge date). I may consider month of visit date in the modeling, as well. Reviewing histograms of the numeric predictors, I see most look approximately normally distributed with the exceptions of `bloodureanitro` and `neutrophils`(Figure 2). There are some weak correlations present in the set of continuous predictors (Figure 3), specifically between `hematocrit` and `respiration`, between `lengthofstay` and both `bloodureanitro` and `respiration`. I anticipate those metrics being important for prediction of `lengthofstay`. I also see `lenghtofstay` and `neutrophils` are negatively correlated. Based on the histograms, summary table, and correlation analysis, I don't see any "red flags" that need to be addressed with further data cleaning. Additionally, this data set, provided by Microsoft, was assigned a "use-ability" score of 10 on Kaggle. While I'm not sure how much blind faith to put in that score, it does give me a bit more confidence in my assessment that this data set is ready for analysis.

Analysis

Starting with a linear regression will provide a nice comparator for some of the more complex model types, and will provide an opportunity for model inference. In addition to linear regression, I plan to implement lasso for feature selection, and decision trees for prediction. If time permits, I will include GAMs as well since there are several continuous predictors. This will all be performed in a cross-validation framework. I have purposefully aimed to include techniques that are well suited for data sets with large `p`. I often work with data that has large numbers of predictors, as it is produced using Next Generation Sequencing (but not all). A typical data set might involve 10-100s thousands of predictors along with some lower dimensional subject metadata (often categorical).

Goals

I will consider the project successful, if I can implement what I have put forth in the 'Analysis' section. Through this project, I hope to grow a deeper understanding of decision trees, particularly with respect to model over-fitting and confidently choosing the optimal complexity parameter. I feel well-equipped to apply linear models and lasso. Having an opportunity for a deeper review of linear models, decision trees, lasso (and maybe GAMs) all across the same data set will make for a fairly comprehensive project. My hope is this might also be reflective of a 'real' data science project which requires application of multiple techniques to settle on the optimal model. I expect this project will encourage me to develop some additional R markdown knowledge as an added benefit.

Tables and Figures

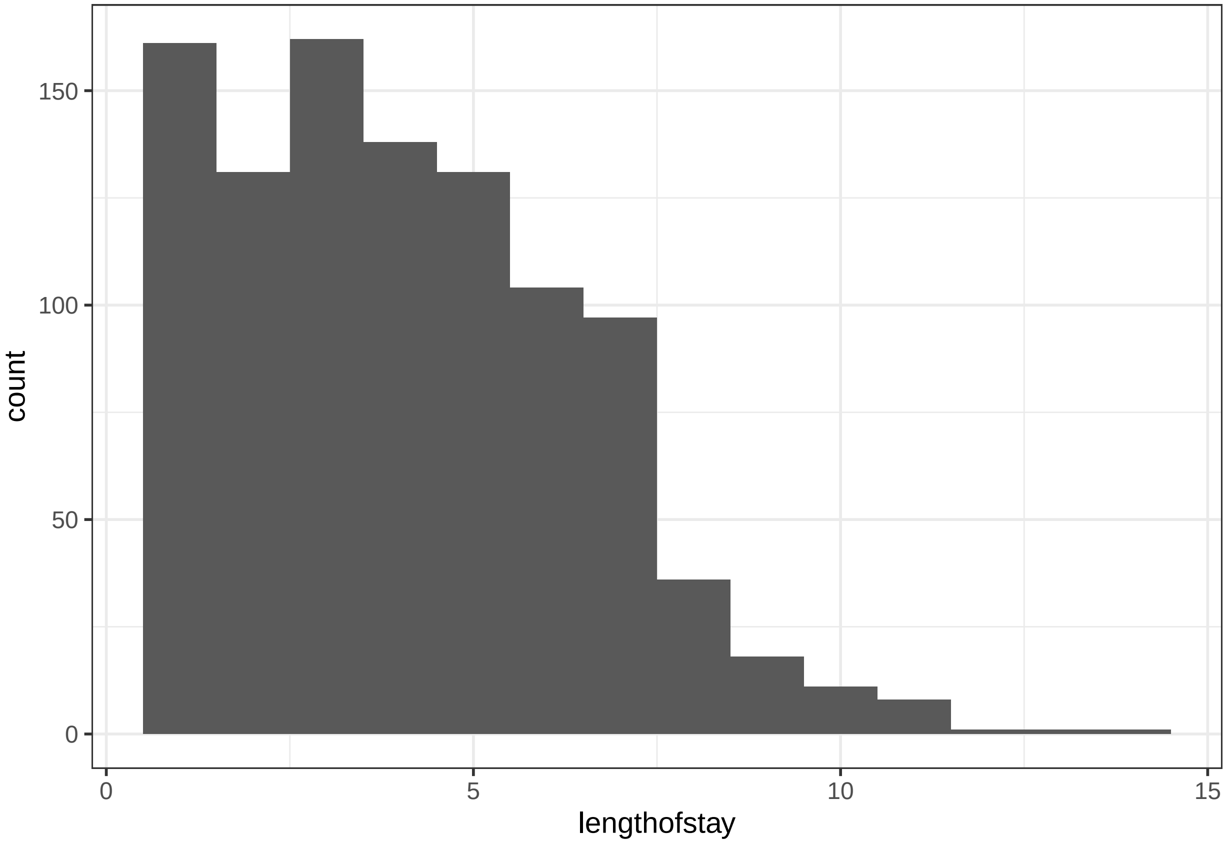

Figure 1. Histogram of LHS

Table 1. Summary of categorical predictors

name	0	1
asthma	969	31
depress	954	46
dialysisrenalendstage	961	39
fibrosisandother	993	7
hemo	915	85
irondef	905	95
malnutrition	952	48
pneum	964	36
psychologicaldisordermajor	746	254
psychother	946	54
substancedependence	923	77

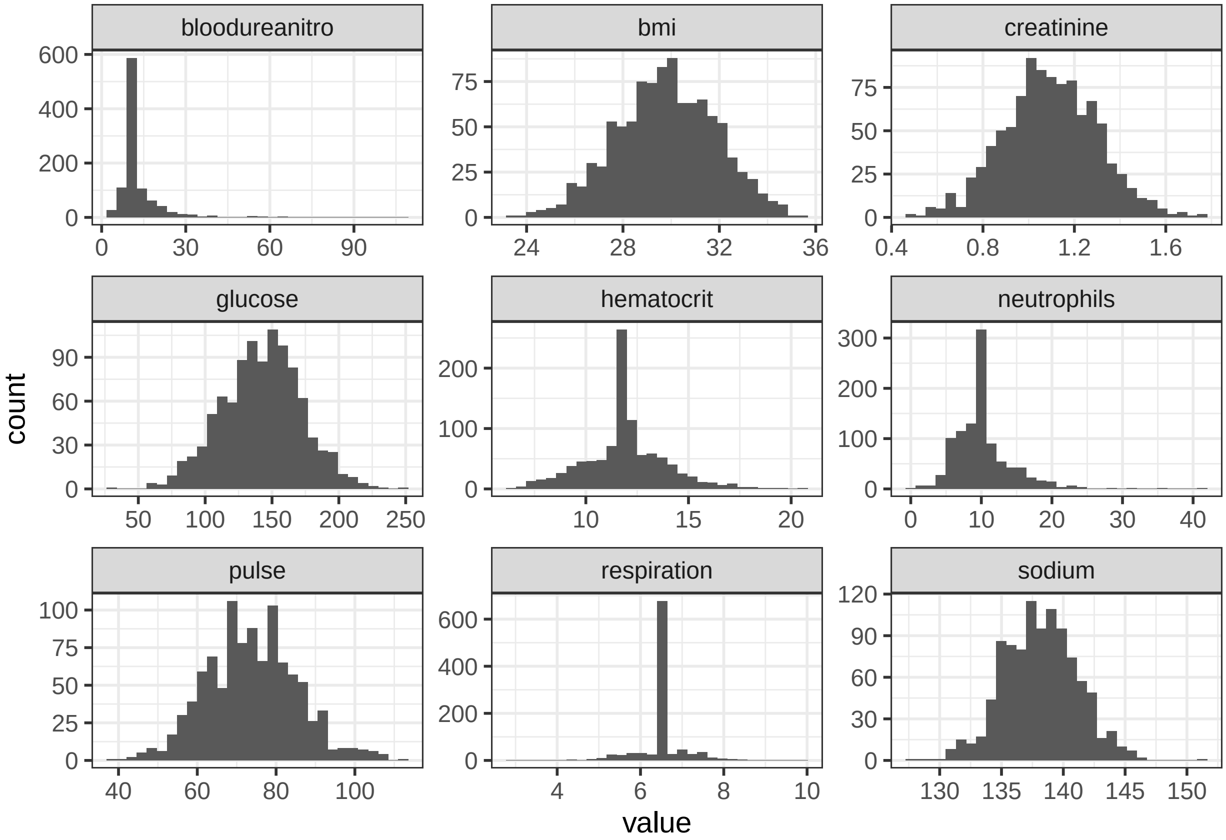

Figure 2. Histograms of continuous predictors

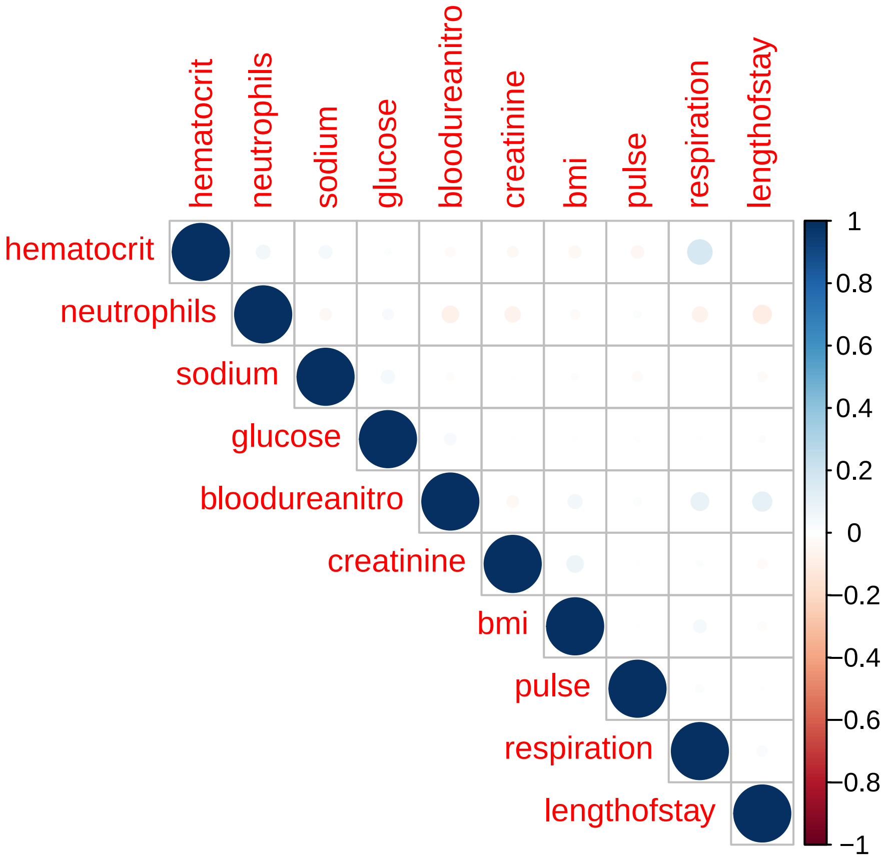

Figure 3. Correlation of continuous predictors