Final Project

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 8,000 observations for ease of computation. I had originally proposed a subsample of 1,000 observations, but after I began modeling, realized there were too few samples in some factor categories which were causing errors. The focus of this analysis will be prediction of `lengthofstay` (days spent in hospital) using a mix of categorical (n=12) and numeric (n=10) variables. `lengthofstay` ranges from one to 17 days (Figure 1).

Materials

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 contains few 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.

*After starting analysis, I realized that numerical data should be centered and scaled prior to modeling, so I implemented that using the `skimr` package. It has a nice feature for summarizing data in a data frame, including variable types, and even plots little histograms within a column of the data frame. I was going to include it here as a figure but the unicode characters were creating a problem in pdf document knitting.

Methods

caret for cross validation

After a bit of research, it looks as though using the `caret` package for cross-validation of models will be a stream-lined solution and address the problem of model-specific function arguments. Looking into `caretEnsemble`, it appears well-suited for cross validating a list of models across the same splits of training data for direct comparison of models. You must make sure to provide the indices to `trainControl` to ensure the same data is used for each model k-fold. Additionally, `caretEnsemble` provides functionality to create ensemble models, but that is outside the scope of this project. As a starting point, I will use `method = cv` instead of `repeatedcv`, but if possible I will look into the latter. It seems as though `repeatedcv` is a bootstrapping method for getting estimates of spread for cv error across folds. Looking into the available methods listed in the package documentation, all of my selected models are available (Table 2).
`rfe()` is a potential option for feature selection for linear models. I am uncertain what is the benefit over lasso, if any. As I understand it, they both aim to increase model bias. Additionally, I feel some confusion around the proper use of `set.seed()` within R markdown and also between model implementations. I've seen recommendations of calling `set.seed()` prior to running every model. That seems unnecessary, but perhaps there is some esoteric reason behind that line of thinking.

Selection of models

A critical aspect of the project that I overlooked in my project proposal was the approximate distribution of my response variable. Once I started looking at model results and thinking about interpretation, I realized my error. `lengthofdays` is count data, and thus violates the assumption of normally distributed data for a linear model. Given my error, I researched an appropriate model that would be available in `caret` and came up with poisson glm as a good replacement. A negative binomial model would likely also be good to try since the mean and variance are different for `lengthofdays`. Decision trees are still an appropriate model, since they are non-parametric. So, I will move forward with applying poisson regression, negative binomial regression and decision trees for this project. Before I realized the problem with modeling a count data response variable, I had applied linear models and lasso regression, and was surprised by the RMSE for both being relatively low (even across the test data set). Since the lm and lasso seemed to perform better than expected, out of curiosity, I will keep them in the modeling, but will focus my discussion mostly on the glms and decision tree.

Results

RMSE for 10-fold cross validation

For 10-fold cross validation, lasso produced the lowest RMSE, followed closely by negative binomial and linear models (Figure 4 cv rmse). Poisson had the highest RMSE (2.53), and the decision tree was second highest. The 0.95 confidence interval for Poisson RMSE was much larger than all other models (~2.2-2.9).

The negative binomial, poisson, linear and lasso models all performed closely when measured as RMSE on the unseen test data of 4000 observations (Table 2). The rpart decision was less accurate.

Tracing errors in `caretList`

Initially, I ran into a warning coming from the `caretList` function: "There were missing values in resampled performance measures." I found a related question on stackoverflow, and the best answer was that the decision tree in rpart couldn't find a split, so it averages the outcome and uses that as the predictor. One drawback of using `caretList` is that errors can be more challenging to trace since it is unclear from which model method they originate. I ended up just searching the web for the error message and/or commenting out code chunks until I could locate the source. I wondered if the errors are captured in the returned list, and browsed through it but didn't see any list element that looked like an error/warning message.

Negative Binomial GLM model: tuning of the link function

The model results for negative binomial list the 'identity' as the `bestTune` link function for the glm. I think this may be a clue as to why the linear model performed better than expected, in fact, comparably to the negative binomial model. Essentially, what I think is happening is all three link functions are being evaluated (sqrt, identity, log), and identity is the best. I don't have intuition for why this might be the case. The 'link' function is listed as a tuning parameter in the `caret` manual; `caret` wraps the `MASS` `glm.nb` method. The help documentation for `MASS::glm.nb` has a default of `link=log`. This supports the theory that `caret` is optimizing the link function and `identity` just so happens to be the best. I reviewed the `caret` output, and this theory was correct (Table 3).

Rpart

The optimal complexity parameter determined for the tree was 0.028. This resulted in two terminal nodes with the only internal node `facidE`, facility id E. This result has limited utility for inference, aside from the conclusion that facility E generally keeps patients for fewer days (Figure 6). Some things to look into are expanding the grid of complexity parameters to optimize (I used the default) and the other, less likely problem, `minsplit` which is the parameter that defines number of observations required to create a new node (the default is 20). I think `minsplit` is less likely to be the issue since I have a large enough sample size that I would anticipate at least 20 observations falling into new nodes.

Discussion

Model selection via cross validation

The negative binomial model performed the best when applied to the unseen test data, but the linear model, Poisson, and lasso were all very close (Table 4).

Model predictions

In order to understand the usefulness of the final model, it's important to think about the observed error in the context of our observed responses. The first quartile is 2 days spent in the hospital and the third quartile is 6 days. I'm unclear on how the average error applies across our model; are lower counts more impacted by an RMSE of ~2 days vs. higher counts it is relatively less? The context of how the model will be applied is also important. If we are looking for a general idea of factors leading to longer hospital stay (ie, inference), then I think this is "good enough," but if we're looking to predict length of stay for an individual and implement some kind of patient-specific mitigation, then I think we have more work to do.

Model inference

Since the negative binomial model performed the best on the test data, I will focus on it for model inference. Generally, there would be a log link function to take into account, but since the best link function identified by cross validation was the identity function, the response will not be transformed. The most important variables contributing to increase in `lengthofstay` are `hemo1`, `facid` (C/D/E), types of psychological problems: `psychologicaldisordermajor1`, `substancedependence1`, `psychother1`; `dialysisrenalendstage1` (Table 5). For our most influential variable, `hemo1`, a one unit increase will result in a 1.4 unit increase in `lengthofstay`. The relative importance of variables can be seen in the importance plot output by `caret` (Figure 7).

Conclusions

Overall, I really liked using `caret` for cross validation since it's quite flexible and efficient. A potential drawback is that it could be easier to misuse a method without realizing it. I spent a fair bit of time just familiarizing myself with the model methods and tuning parameter optimization, along with learning the output structure in order to troubleshoot and extract results, and I still feel like I'm only scratching the

surface. Regardless, I think it is worth the time spent learning it since it removes the burden of learning several different packages, especially if you use the `caretList` function from `caretEnsemble`.

Citations

caretEnsemble

Avoid Mistakes in Machine Learning Models with Skewed Count Data