Predictive features for Stroke

Daniel Jonathan Leirer

August 12, 2019

# Overview

Using Stroke as the outcome I was able to achieve a sensitivity and specificity of 0.797 and 0.734 respectively in the test set. The models I generated are not clinically useful, but they are robust and rely on features that seem plausible. The most consistent features were Age, Hypertension and Heart Disease. BMI, Glucose and Smoking were also highlighted.

This analysis is broken down into four sections. 1.) Basic data exploration. 2.) Processing data just enough to build predictive models. 3.) “Spot checking” of diverse algorithms to compare their performances. 4.) Analysis of results to guide future direction.

This is very much intended as the start of an iterative process, rather than an endpoint. There are a lot of things I was unable to do. A to-do list would include significantly more processing of features, especially Glucose and BMI. Feature engineering (thinking of diabetes), Recursive feature elimination, alternatives to dealing with the imbalanced data (rose and smote), focusing on model performance in high-risk groups, and proper tuning of Random Forrest or Boosted Trees. Finally, caret ensemble to create a blend of models.

# Data Exploration and Processing

I first identified missing values in BMI and Smoking status. I felt both of them would be important. For the missing Smoking data, I added a new “unknown” category. This essentially just adds a dummy variable, and I felt there was a good chance the missing data was structured. I decided to drop rows with missing BMI data. This is not an ideal solution and I briefly considered several alternatives. 1) Imputing BMI. 2) Dropping BMI rather than missing BMI columns. 3) Turning BMI into a categorical variable and adding an “unknown” variable.

The other major point for me is Glucose. The distribution for glucose clearly shows two populations, the larger one being in the normal range below ~140 and the other one above, which indicates Diabetes. I feel this is important, and I have several ideas for dealing with this. However, since my strategy was to use a range of predictive models I felt it would be more fruitful to press on and prioritise based on the results.

Finally, I dropped the “other” category in gender.

# Predictive Modelling: Method and Approach

My predictive modelling approach was to use several algorithms that are likely to be uncorrelated. This is to make sure my initial results are robust, I can quickly identify the most promising direction, and to have the option of stacking models at a later stage. I choose the following 6 models, in part because they are well known and tested, and because they run relatively quickly.

1. Generalised Linear Models with Regularisation (glmnet)
2. K-nearest Neighbour (KNN)
3. Naive Bayes (NB)
4. Random Forrest (RF)
5. Support Vector Machine with Linear Kernel (svmLinear)
6. Support Vector Machine with Radial Kernel (svmRadial)

I used the caret package, to split the data 70-30 into train and test sets, and used 3\*5 repeated K-fold cross-validation.

I intentionally left the hyperparameter tuning settings as the default. Caret defaults tend to be reasonable, and I would prefer to tune one or two promising models once everything else is in place.

I used downsampling, which is not ideal, but the dataset is a reasonable size, and it solves any potential issues I have with class imbalances and reduces computational time. Besides, the test data still has its original class proportions, so if the results are biased it would be immediately obvious.

Finally, I set the model selection metric to ROC. In hindsight, F1 score would have probably been better.

# Analysis of results

I hypothesised that a Tree-based model would be especially well suited for this problem, and was slightly surprised that the rf model somewhat underperformed. All models except KNN achieved and AUC of around ~0.85 (See plot 1). I extracted the Glmnet and RF models to make use of their internal feature importance rankings, and they both showed that Age unsurprisingly was the most important feature. I choose to focus on the glmnet model in more detail, which achieved a sensitivity of 0.81 and a specificity of 0.74. This was comparable to the results in the test data which were 0.797 and 0.734. Finally, I used the predicted probability for stroke from the glmnet model to explore my features further. I noticed that almost all individuals with heart disease and hypertension are predicted to have a stroke. I think the most sensible direction to go in next would be to focus on high-risk individuals and build separate models for different age categories.