Stroke Prediction with Decision Trees

Thomas Bourton

October 13, 2019

1 Problem Description

Predictive model: We use a predictive model to classify patients who have stroke. We evaluate the performance of the model and suggest which features may be useful in stroke prediction.

Based on the `train_2v.csv` from the Kaggle stroke dataset: www.kaggle.com/asaumya/healthcare-dataset-stroke-data

2 Preliminary Analysis and Visualisation

The training data contains data regarding stroke patients. The feature variables are `gender, age, hypertension, heart_disease, ever_married, work_type, residence_type, avg_glucose_level, bmi, smoking_status` while the target variable is `stroke`. The training dataset contains 43400 entries of which approximately 1.8% have stroke.

Our aim is to determine which features are most important in determining whether a given person has stroke or not. Hence it is useful to separate the data into those patients with stroke and those without. We can make some preliminary plots. For example, one might expect that e.g. a patient who has never suffered from heart disease is less likely to have stroke and vice versa. We can plot this (see Figure 1) and it somewhat confirms our initial estimate. On the other hand, for

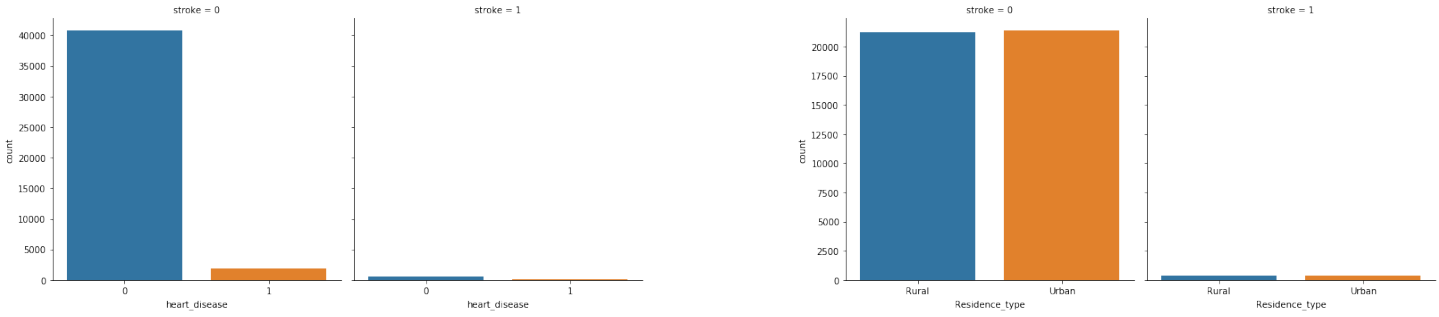

Figure 1: Left: *Number of patients with heart disease with/without stroke.* Right: *Residence types having not with/without stroke.*

example one would also expect that the area in which a patient lives (rural or urban) appears to have little direct importance in a medical condition such as stroke, see Figure 1. We can again plot this and be fairly confident that this initial assumption is somewhat correct. We can also plot the age distribution for stroke/no-stroke. We can see that the proportion of patients with stroke is skewed heavily towards the ages 60+. There is similar patterns for average glucose level, a larger proportion of stroke patients have a high average glucose level. See Figure 2.

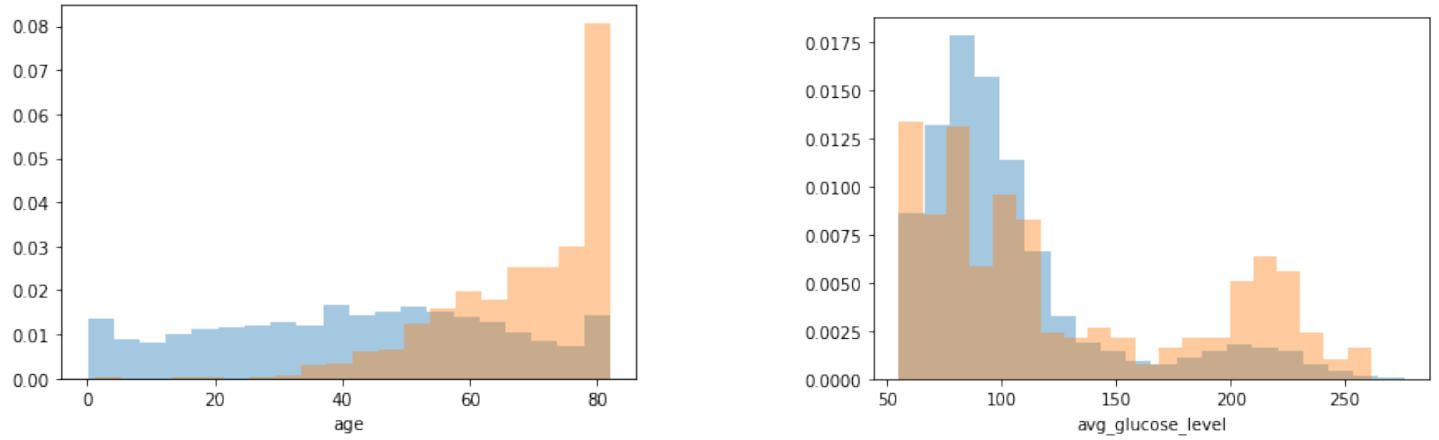

Figure 2: Left: *Distribution of ages without stroke (blue) / with stroke (orange).* Right: *Average glucose level of patients without stroke (blue) / with stroke (orange).*

3 Data Clean-up and Wrangling

Having performed some basic analysis on our data we now would like to build our model. However before doing this we have to prepare the data. In particular we have two main concerns. The first is that both the training and test data have missing data for `bmi` and `smoking_status`. Approximately 3% of `bmi` data and 30% of `smoking_status` data is missing from both sets. For `bmi` data we will simply replace the missing data with the average for each set. On the other hand a much larger portion of `smoking_status` data is missing, also this label is discrete hence replacing such a large proportion with an average is potentially troublesome. We will pursue two approaches, the first being to replace this data with the mode (which is never smoked) the second is to completely drop the `smoking_status` data from the dataset.

The second main concern with the data contains is that it contains many categorical variables. The decision tree algorithm cannot deal with categorised data, hence we need to turn the features `ever_married, work_type, residence_type, smoking_status` into numerical variables. We do this using the dummy variables methods in the pandas library.

4 Building Decision Tree Model

We build the model with the `smoking_status` data replaced by the mode. We then build the model using the feature variables with target being `stroke`. We are going to use the `tree` library from the `sklearn` package. In order to build the model we further split the training data into a trainset and a testset. We chose a 80/20 split. We train the model on the trainset, this builds the predictive model. By using this model on the testset and comparing with the known data from the testset we can determine several analytical metrics on the model. The model with the smoking data replaced by mode performed with accuracy of 96.2%.

We also repeated the same model build but with the smoking data removed. In which case the model was 96.1% accurate at predicting stroke on the testset.

5 Analysis of the Model and Conclusions

Our main goal is to determine the importance of the features on determining stroke. Hence we can plot the feature importance metrics, see Figure 3. The models both confirm our early expectations that for example `age` and `bmi` are quite important in predicting stroke, whereas `residence_type` is fairly unimportant. A surprising output of the model is that `heart_disease` appears to be very unimportant in determining stroke, where the preliminary analysis suggested to expect otherwise. We

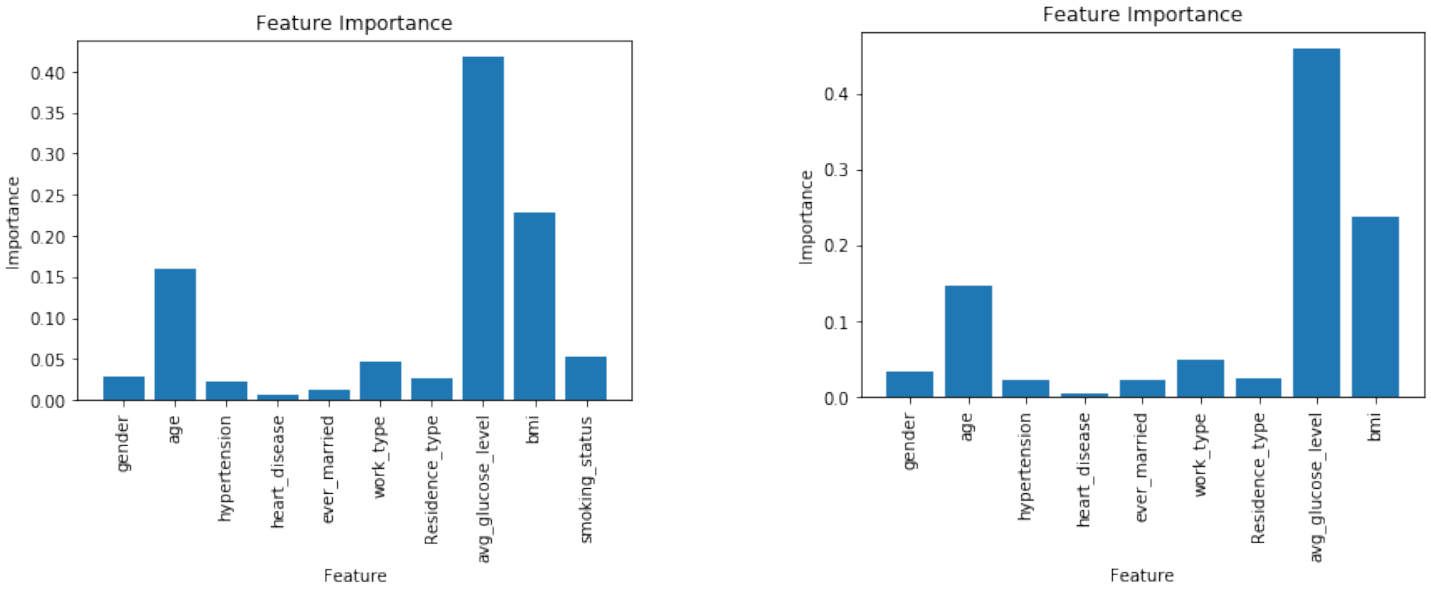

Figure 3: Left: *Feature importance for model with smoking data replaced by the mode.* Right: *Feature importance with smoking data removed.*

can also see that our model matches the approximate data visualisations in each case (see Jupyer notebook for more plots). When applied to the test data with unknown stroke data the model with smoking replaced by the mode predicted that $350/18261 \sim 0.019$ fraction of patients have stroke, this is to be compared with $783/43400 \sim 0.018$ in the known training data.

One potential way that the model could be improved is by dealing more systematically with the smoking status data. Instead of simply replacing each missing entry with the mode we could have placed instead a more uniform distribution of data which could have been derived from the training data. This is important as one would naively think that smoking should have a larger impact on determining stroke than our model predicted.