Introduction to Profile Areas Datascience Project Week 3

by Eva Aßmann, Paul Vogler

1 Scientific Background

A chain of hospitals has retained McKinsey to help creating the next generation of healthcare for its patients. The company founded a Center of Data Science Excellence with the purpose of providing proactive health care for its patients. Within the scope of the Data Science Excellence initiative, the client wanted to carry out a study around strokes. A Stroke is a disease that affects the arteries running to and within the brain. A stroke occurs when a blood vessel that carries oxygen and nutrients to the brain is either blocked by a clot, bursts or ruptures. When that happens, part of the brain is not provided the blood and oxygen it needs anymore and dies.

2 Goal

This project aimed at developing a model predicting the probability of a stroke happening to the hospital chains' patients to help doctors take proactive health measures for these patients.

3 Data and Preprocessing

Over the last few years, the client has captured several health, demographic and lifestyle details about its patients. The resulting McKinsey Stroke Data Set contained 12 attributes for 62.0001 patients (Tab.1). The data set came split into a training data set containing 43400 patients and a test set containing data on 18602 patients. The test data set does not provide the target value 'stroke' for the patients.

Variable	Definition	Data Type
id	Patient ID	int
gender	Gender of a patient (male, female, other)	string
age	Age of a patient in years	double
hypertension	0 - no hypertension, 1 - suffering from hypertension	int
heart_disease	0 - no heart disease, 1 - suffering from heart_disease	int
ever_married	Yes/No	string
work_type	Type of occupation (Govt job, Never worked, Private, Self-employed, children)	string
Residence_type	Area type of residence (Urban/Rural)	string
avg_glucose_level	Average Glucose level (measured after meal)	double
bmi	Body mass index	double
smoking_status	patient's smoking status (formerly smoked, never smoked, smokes)	string
stroke	0 - no stroke, 1 - suffered stroke	int

Table 1: Health, demographic and lifestyle attributes on each patient in the train and test data set.

The training data contained the label variable id, categorical feature variables (gender, hypertension, heart_disease, ever_married, working_type, Residence_type, smoking_status, stroke) and numeric continuous feature variables (age, avb_glucose_level, bmi)).
Looking at the value distributions showed that the training data set was highly imbalanced, since the number of observations belonging to patients who already had experienced a stroke was significantly lower than those belonging to patients who had not suffered from one yet (Fig1). The histogram plots for the continuous feature variables bmi and avg_glucose_level matched a normal distribution, while the age variable showed equally distributed values.

In the outlier detection for the continuous variables, no outliers could be observed.
There were bmi values missing for 1462 patients and no smoking_status was recorded for 13292 patients. Neither the bmi nor smoking_status was recorded for 426 patients. Since the overlapping missing feature values represented less than 1% of the training data, no further measures were taken. The missing bmi values represented merely around 3% of the training data, thus the respective subjects were dropped. The missing smoking_status was replaced with 'no info' as a dummy value.
For the pairwise feature correlation analysis, all categorical feature variables stored as string values were binary-encoded, i.e., gender, ever_married, work_type, Residence_type, smoking_status). Performing correlation analysis using the Pearson correlation score showed a strong negative relationship between the feature variables ever_married and age (score=-0.69) and a weak negative relationship between ever_married and bmi (Fig.2).

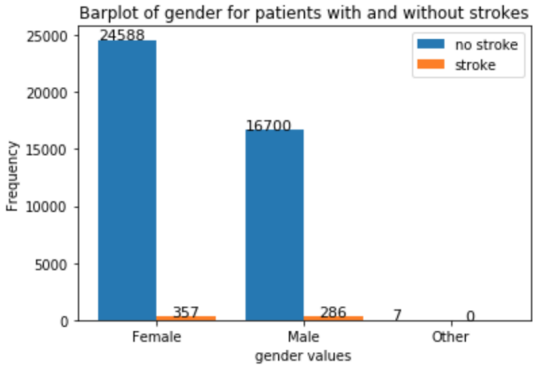

Figure 1: Frequency for patients' gender categories grouped by stroke status.

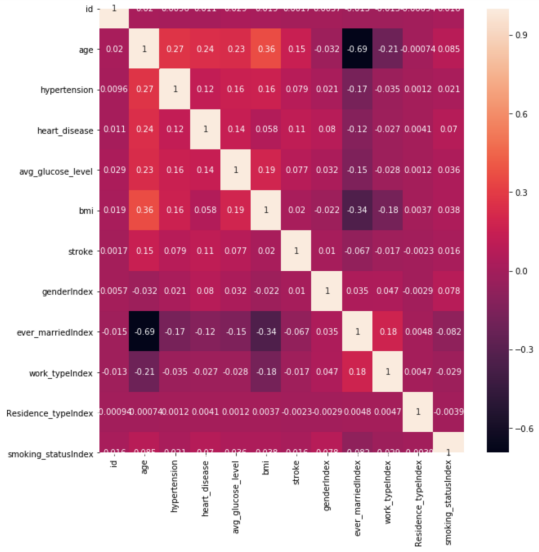

Figure 2: Pairwise correlation plot for all features of the processed training data set.

The processed training data was visualized using Chernoff faces. Chernoff faces display multivariate data in the shape of a human face making it easier to intuitively interpret data based on the resulting expression. The individual facial parts represent values of the variables by their shape, size, placement and orientation [1]. The facial parts were encoded by 18 variables, comprising the 11 available attributes from the training data (excluding id) and 7 constant variables which were manually set to 0.5. The height of the upper face, overlap with the lower face, half of vertical face size, width of the upper face and width of the lower face were encoded by gender, marriage status, age, work type and hypertension status, respectively. The length of the nose, vertical position of the mouth, mouth curvature and mouth width were encoded by residence type, heart disease status, smoking status and a constant value. Eye features like vertical position, separation, slant, eccentricity, size and position of the pupils were encoded by four constant values, average glucose level and another constant value. The vertical position of the eyebrows, their slant and size were encoded by a patients bmi, stroke status and a constant value.
Chernoff faces were created for 25 random subjects (Fig.3). The most notable variations between the displayed Chernoff faces were observed in the widths and overlap of the upper and lower face as well as in the eyebrow orientation. Eyebrow orientation was encoded by a patient's stroke status and translated into upwards directed (stroke=0) and downwards directed brows (stroke=1). Age, marriage status and gender defined the proportion of upper and lower face and the overlap position. A patient's work type and hypertension status influenced the width of both face halfs, causing three degrees of upper half width and two possible lower face widths.

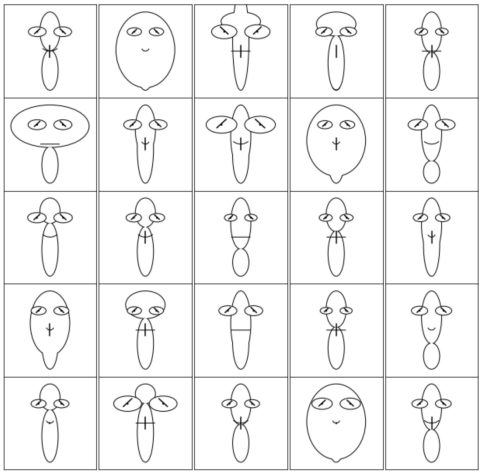

Figure 3: Chernoff faces for 25 random patients from the processed training data set. Facial features are encoded by attributes from the data set and constant values.

4 Methods

Two prediction models were trained to predict the risk of a stroke happening to a patient. For each prediction model the training data was randomly split into 70% training data and 30% test data, respectively. A decision tree classifier and a gradient boost classifier were trained and evaluated.

5 Results

Evaluated on an independent 30% test data sets, the trained decision tree classifier yielded 98.51% accuracy, the trained gradient boost classifier reached an accuracy of 98.42% (Fig.4). The decision tree model showed only a slightly better evaluation result than the gradient boost model. Both models are based on decision trees and yielded very good accuracies. In theory, a decision tree classifiers in combination with gradient boosting should be performing better since decision trees on their own are weak learners. Yet, with the data set at hand, the basic decision tree model seemed to be the better classifier, even if it showed just a small margin.

6 Discussion

This week's project was typical for a data scientist's work, because we got a larger external data set and the task to prepare it appropriately for machine learning. We had to understand the data, visualize various statistical properties, clean and reformat it for the use with the spark framework. Furthermore we had to evaluate how the processed data would work for different classification approaches.

```
accuracy of: 98.51%
accuracy of: 98.42%
```

Figure 4: Evaluation accuracy for trained decision tree classifier (top) and gradient boosting decision tree classifier (bottom).

References

[1] Herman Chernoff. The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, 68:361–368, 1973.