RESEARCH

Introduction to the profile areas of data sciences: project 1

Eva Aßmann* and Paul Vogler

*Correspondence:
eva.assmann@fu-berlin.de
Institute of Mathematics and Informatics, Berlin, Takustr. 9, Berlin, Germany
Full list of author information is available at the end of the article

Abstract

Goal of the project: Develop two prediction models to predict the probability of a patient having coronary artery disease.

Main results of the project: The neural network and the gradient boost classifiers were accurate at predicting the target groups, with an accuracy of about 95%.

Personal key learnings: We applied a neural network and a gradient boost tree for the first time.

Estimated working hours: 8

Project evaluation: 2

Number of words: 1386 Words

Suggestions for improvements of the project: Too little information about what the meaning behind our plot should be.

1 Scientific Background

Probability analysis aiming at the diagnosis of common diseases is a major part of medicine science. A large amount of clinical patient data is required to derive accurate prediction models that are reliable and universally applicable. Since obtaining data from these large numbers of patients is difficult oftentimes alternative approaches are needed. One solution lies in statistical prediction models, that try to categorize future patients based on previous patients data. The Cleveland database was collected in conjunction with the development of a logistic regression algorithm, to predict coronary artery disease (CAD). CAD describes the narrowing or blockage of arteries, typically caused by atherosclerosis, the deposit of cholesterol and plaques on the inner wall of arteries. This hinders the blood flow to the heart, sometimes causing a chest pain called angina. If the blood flow is cut completely, a heart attack can occur [1].

2 Goal

The goal of this project was to develop two different prediction models derived from the relatively small clinical and test data of the Cleveland Data Set (303 patients) that could accurately estimate probabilities of CAD.

3 Data and Preprocessing

3.1 Data

For 303 patients referred for coronary angiography at the Cleveland Clinic historical data was collected, invasive and noninvasive clinical tests were run plus angiography

was performed. All tests were analyzed and the results recorded independently and without knowledge of other test data meaning no work-up bias. From the collected patient data in the test group 13 variables were derived that were entered into a computerized database [1].

The data was structured in a table with 303 entries and 14 attribute columns. The first 13 attributes contained the clinical and test results, the additional attribute was the target variable describing the CAD diagnosis of each patient. The attributes contained the following information:
Variables containing categorical values:

attribute	description	values
sex	Sex	1: male, 2: female
cp	Chest pain type	1: typical anginal, 2: atypical, anginal, 3: non-anginal, 4: asymptomatic
fps	Fasting blood sugar	0: fbs <= 120mg/dl, 1: fbs >120mg/dl
restecg	ECG result at rest	0 = normal, 1: ST-T wave abnormality, 2: probable or definite left ventricular hypertrophy
exang	Exercise induced angina	0: no, 1: yes
slope	Slope of the peak exercise ST segment	1: upsloping, 2: flat, 3: downsloping
thal	Exercise thallium scintigraphic defects	3: normal, 6: fixed defect, 7: reversible defect
num	Diagnosis of heart disease	0: <50% vessel diameter narrowing, 1-4: >50% diameter narrowing

Variables containing continuous values:

attribute	description	values
age	Age	age in years, range 29-77
trestbps	Resting blood pressure	recorded in mm Hg, range 94-200 mmHg
chol	Serum cholesterol	recorded in mg/dl, range 126-564 mg/dl
thalach	Maximum heart rate achieved during exercise	range 71-202 bpm
oldpeak	ST depression induced by exercise relative to rest	range 0-6.2 mm
ca	Number of major vessel colored by fluoroscopy	values 0-3

3.2 Preprocessing

There were nan entries found in the column storing information on the number of major vessels colored by fluoroscopy (ca) and for thallium scintigraphy (thal). Since the nan values comprised around 1.9% of the data set (6 entries), the respective subjects were deleted instead of performing imputation or chosing dummy values. The target attribute (num) encoded the magnitude of artery diameter narrowing in an angiogram with 5 values (0,1,2,3,4). Since this project aimed at simply distinguishing between the presence or absence of CAD the target attribute was redefined as a boolean variable. All values ¿=1 (1,2,3,4) were changed into 1 to encode the presence of CAD, zero values encoding the absence of CAD remained unchanged. All variables that were read in as floating point numbers but only contained integer values were redefined to integer values. Looking at the correlation between each attribute and the CAD status most attributes showed moderate to strong correlation with the target attribute ‘num’.

For each attribute the value distribution of patients with CAD and without CAD were compared to see their relation to CAD and how they can be used in a prediction model. Except for fbs, restecg and trestbps all the other variables showed visibly significant differences between the CAD and non-CAD groups. The most interesting ones were Serum Cholesterol (chol), thallium scintigraphy (thal) and the number

	fbs	chol	trestbps	restecg	age	sex	slope	cp	exang	thalach	oldpeak	ca	thal
num	.003	.08	.153	.166	.277	.278	.333	.409	.421	.424	.424	.463	.527

Table 1: Correlations of features with the target.

of blood vessels coloured by fluoroscopy (ca). While the serum cholesterol level for patients with CAD was normally distributed, it was observed that patients without CAD more often have a serum cholesterol value between 200 mg/dl and 250 mg/dl. People who have higher cholesterol in their blood are more likely to suffer from atherosclerosis which causes CAD. Thallium Scintigraphy is an imaging method for measuring and assessing a patient's blood flow by comparing the status at rest with the post-exercise image [2]. The exercise thallium scintigraphic defect for patients without CAD mainly matched a normal degree. Most patients with CAD showed a reversible defect. The number of vessels coloured by fluoroscopy is mainly 0 for patients without CAD, while diseased patients values are more or less equally distributed. With this method only blood vessels containing calcium are fluoroscopically marked.

Linear Discriminant Analysis (LDA) was performed on the processed data set. LDA is a classifier that tries to find a linear combination of variables that can separate two or more classes from each other. The resulting plot showed that the labelled data could approximately be separated into two groups of data points (Figure 1). This LDA result indicated that a model trained on the data could have a good performance.

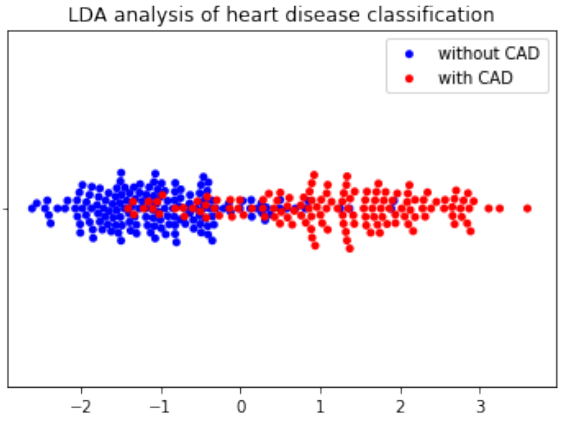

Figure 1: LDA analysis plot for the cleaned data set.

Neural networks and linear models can only interpret continuous numerical values, so each categorical value feature needs to be encoded with a numerical value. Here the enum-like attributes cp, thal, slope and restecg were encoded to be continuous using dummy variables. In addition to non-continuous values a model can be affected by outliers. To detect and analyse potential outliers Principal Component Analysis (PCA) was applied to the processed data. PCA tries to find a transformation of possibly correlated variables to produce a linearly uncorrelated set. The resulting plot showed two or three outliers. We did not drop any of the outliers, because we had no confirmation that they would negatively affect the models.

4 Methods

For evaluation purposes, the processed data was split into 80% training set and 20% test using the sklearn.model_selection library. Two basic classification solutions were implemented as prediction models for CAD. First a 3-layers Neural Network (NN) was trained. The network was trained in 5000 cycles using pytorch. Next a decision tree model using gradient boost was trained and tested on the same train and sets. The model was trained in 1000 iterations using the lightgbm library.

5 Results

The learning curve of the training process shows the average difference between the actual target value and the predicted one. Looking at the learning curve for the 3-layers NN the average difference decreased with each iteration step (Figure 2). Evaluated on the same test set, both the trained NN and the trained decision tree model yielded a 95% accuracy (Figures 3 and 4).

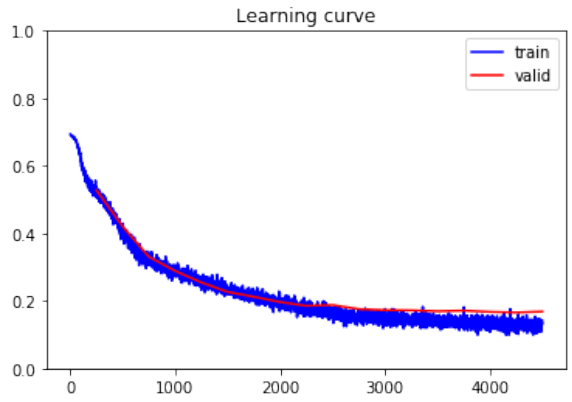

Figure 2: Learning curve for the training of the 3-layers NN.

6 Discussion

The herein described project was a good example for the typical tasks of a Data Scientist. Structured Data from an external study needed to be cleaned up and some attributes needed to be redefined. In preparation for following probabilistic analyses and modeling, attributes needed to be examined regarding their general applicability in diagnosis prediction. Furthermore it needed to be checked if the attributes allowed an accurate and valid prediction model between non-CAD and CAD patients. The data had to match the requisites of the applied machine learning methods and had to be preprocessed accordingly. Since Data Science rather deals with the preparation steps paving the way for the actual machine learning part of a project, developing and tuning a prediction model does not really fall into the area of typical Data Science tasks.

References

1. R Detrano, W.S. A Janosi: International application of a new probability algorithm for the diagnosis of coronary artery disease. The American Journal of Cardiology **64**, 304–310
2. McKillop, J.H.: Thallium 201 scintigraphy. The Western journal of medicine **1331**, 26–43 (1980)

Accuracy = 95.00%

	T	F
P	32	2
N	1	25

Figure 3: Accuracy for the trained 3-layers NN.

Accuracy = 95.00%

	T	F
P	31	3
N	0	26

Figure 4: Accuracy for the trained decision tree.