Big Homework Neural FCA Report

Ayub Makhamaev

19.12.2023

1 Heart Disease.

Description of data imbalance choose the metric

First dataset that we use is "Heart Disease"dataset from `https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset/data`. It has 14 columns and 1025 rows, but after dropping duplicates it has only 302 rows.

Each row represents a person and contains general and medical information about him. And the target variable is the rate of heart disease of this person (0- patient is healthy).

Description of the dataset

column	description	unique values
age	Age in years	41
sex	Male/Female	2
cp	Chest pain type	4
trestbps	Resting blood pressure	49
chol	Serum cholestoral in mg/dl	152
fbs	fasting blood sugar > 120 mg/dl	2
restecg	resting electrocardiographic results	3
thalach	maximum heart rate achieved	91
exang	exercise-induced angina (True/ False)	2
oldpeak	ST depression induced by exercise relative to rest	40
slope	the slope of the peak exercise ST segment	3
ca	number of major vessels (0-3) colored by fluoroscopy	5
thal	0=normal; 1=fixed defect; 2=reversible defect	4
target	heart disease presence	2

We plot the corellation heatmap of features and plots of some moderately corellated features. SOME PLOTS OF MODERATELY CORELLATING FEATURES

Metrics For disease prediction recall is more important than accuracy, because is person gets false-negative classification it is very dangerous for him. So we choose "f1 score"as main metric, but we will also consider recall.

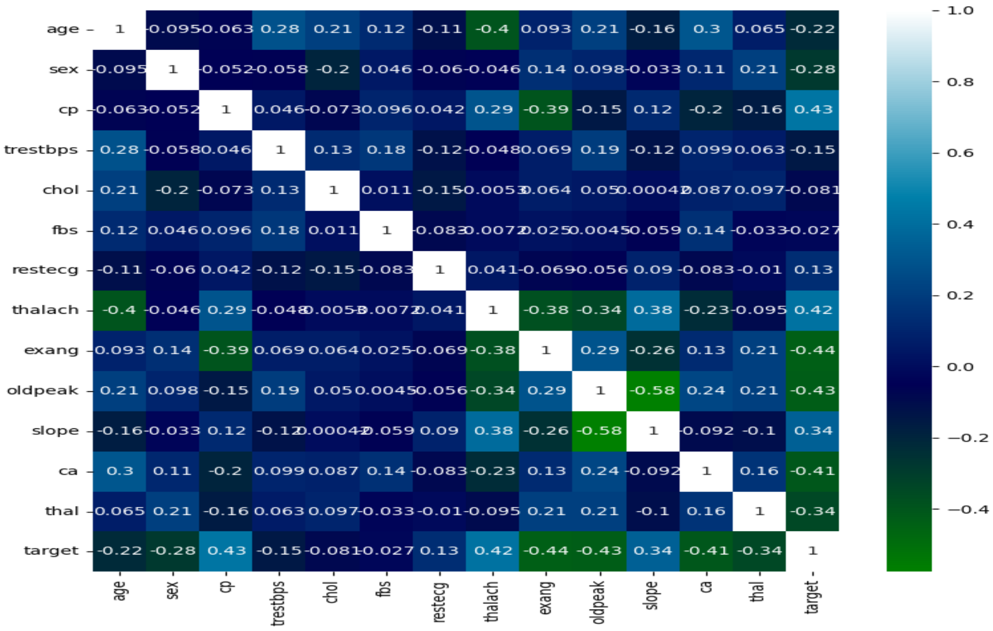


Рис. 1: Corellation heatmap

Standard ML baseline First of all let us build baseline with usual ML methods . We will use XGBoost classifier.

We get the confusion matrix:

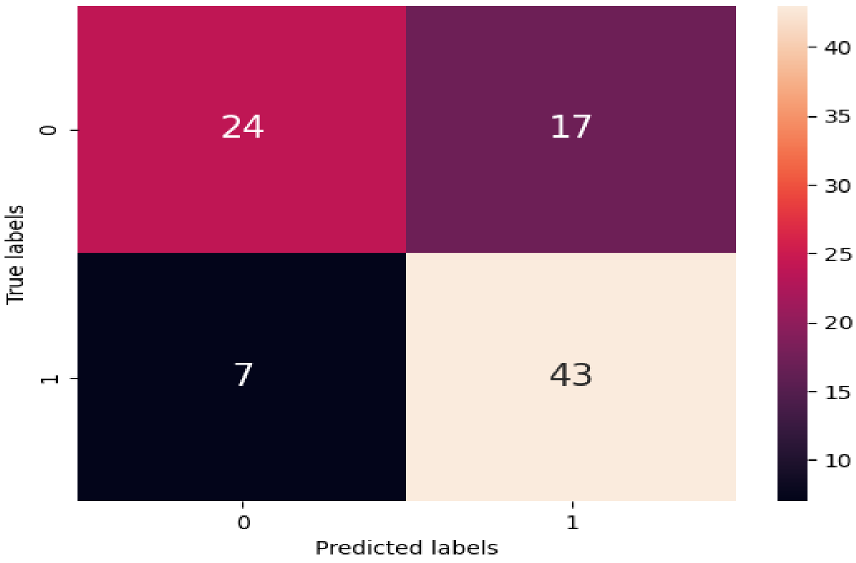


Рис. 2: Confusion matrix

Some scores for XGBoost Classifier

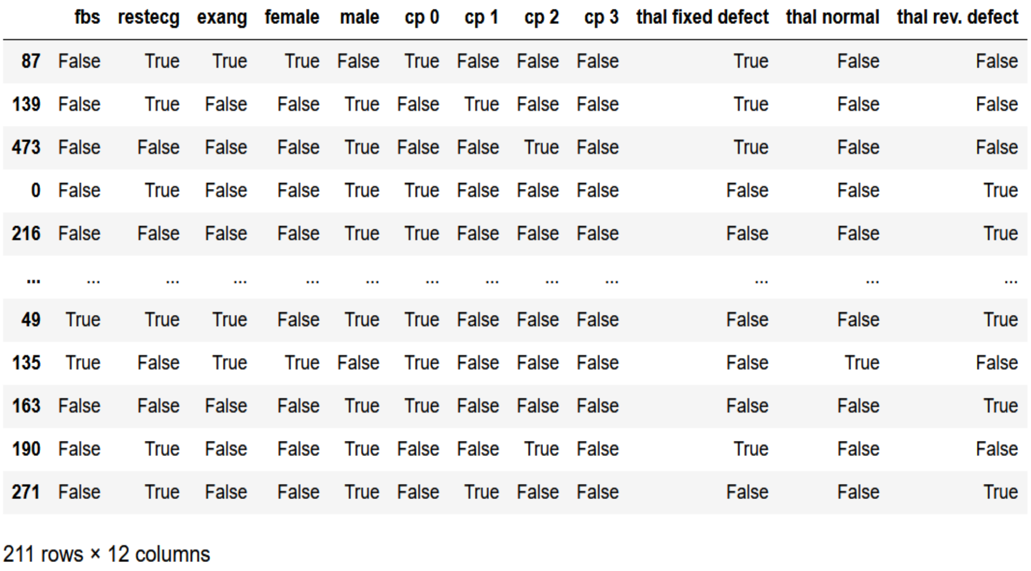
	f1-score	recall	roc-auc
XGBoost	0.782	0.86	0.723

1.1 Neural FCA

Baseline 1. Lets build first baseline based only on categorical features. We rename categories of categorical columns with their real values (for example for sex 0-"female 1-"male") and one-hot encode them.

Categorical columns: ['sex', 'cp', 'fbs', 'restecg', 'exang', 'thal'].

After one-hot encoding and binarizing we get:

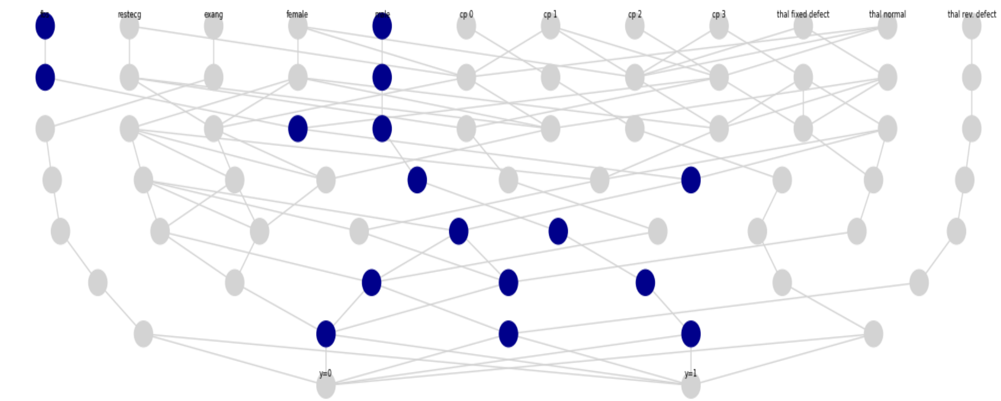


	fbs	restecg	exang	female	male	cp 0	cp 1	cp 2	cp 3	thal fixed defect	thal normal	thal rev. defect
87	False	True	True	True	False	True	False	False	False	True	False	False
139	False	True	False	False	True	False	True	False	False	True	False	False
473	False	False	False	False	True	False	False	True	False	True	False	False
0	False	True	False	False	True	True	False	False	False	False	False	True
216	False	False	False	False	True	True	False	False	False	False	False	True
...	...	...	...	...	...	...	...	...	...	...	...	...
49	True	True	True	False	True	True	False	False	False	False	False	True
135	True	False	True	True	False	True	False	False	False	False	True	False
163	False	False	False	False	True	True	False	False	False	False	False	True
190	False	True	False	False	True	False	False	True	False	True	False	False
271	False	True	False	False	True	False	True	False	False	False	False	True

211 rows × 12 columns

Now we build neural network. Choose best concepts based on "f1metric.

NN based on 13 best concepts from monotone concept lattice



*Blue neurons are the ones activated by description {'fbs', 'male'}

Рис. 3: Concept lattice

Fitted network with edge weights:

Prediction scores:

	f1-score	recall	roc-auc
1 baseline	0.779	0.822	0.770

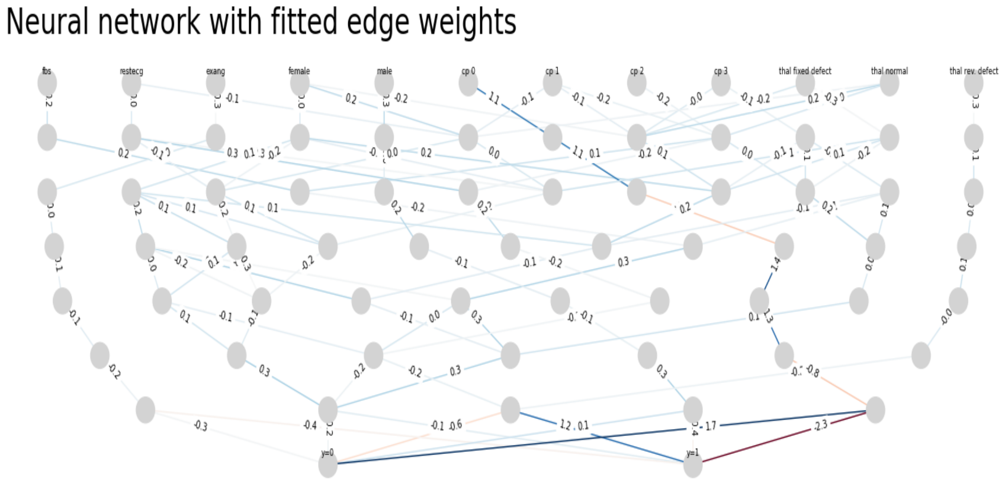


Рис. 4: Fitted network

Baseline 2. Now we use all the attributes and scale numerical features using different strategies

.