Heart Disease Prediction

APPLYING SUPERVISED LEARNING TECHNIQUES TO HEALTH CARE

Introduction

- According to the CDC[1], About 610,000 people die of heart disease in the United States every year–that's 1 in every 4 deaths.
- The test to detect heart disease, the coronary angiography, costs more than 500$ to run[2]. Thus it is only run when doctors have good reasons to suspect heart disease in a patient.
- In this project, we will use Machine Learning techniques to try and **predict the disease without conducting the test**, as well as trying to **find the most efficient way to select patients** for a coronary angiography.

The Data

- The data was collected by the Cleveland Clinic Foundation and is available on the UCI Machine Learning Repository[3].
- It consists of 303 observations with 14 features each which makes it a rather small dataset.
- It was gathered in Cleveland in 1988 and might not reflect techniques used nowadays.
- Most of the observations come from medical tests results.

The Data

- Features description :

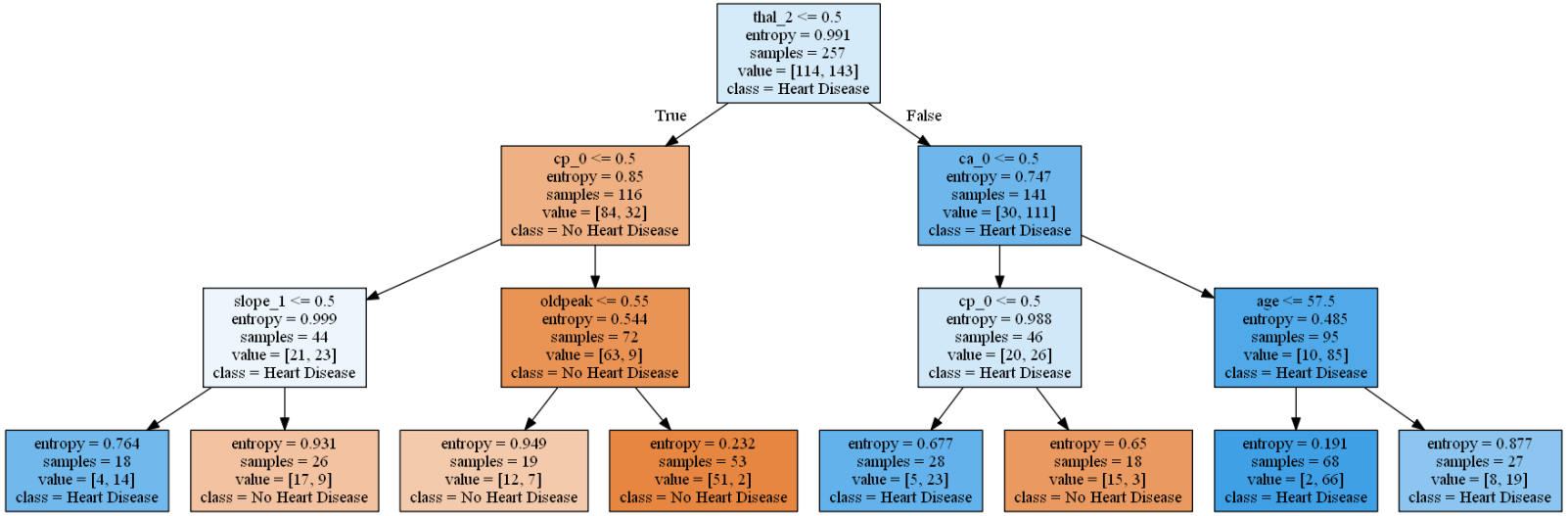
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

cp : Chest Pain Type
trestbps : Resting Blood Pressure
chol : Serum cholesterol
fbs : Fasting Blood Sugar
restecg : Resting electrocardiographic results
thalach : Maximum heart rate during exercise
exang : Exercise induced angina
oldpeak : ST depression induced by exercise
slope : Slope of the peak exercise ST segment
ca : Number of major vessels colored by fluoroscopy
thal : Thallium heart scan results

target : The predicted feature, the result of a coronary angiography

Exploration and first model

- After making sure our data does not have any class imbalance, we run a weak leaner to get an idea of the important features.
- It seems that the results of the thallium heart scan as well as chest pain absence and number of major vessels colored by fluoroscopy are important features.



Feature engineering

- Let's create a couple of features based on our observations and intuition.
- First the absence or presence of chest pain is a new binary feature.
- Then we add a couple of intuitive things like an old age feature (age >= 57.5) as well as high blood pressure and high cholesterol.
- Finally, it seems like combining the 'sex' and 'slope' features gives a couple of very decisive features (females with a slope of 2 all have heart disease for example).
- After getting dummies, we end up with 40 features.

Grid Search

- To select our best model, we will run a grid search across several supervised learning models and several parameters for each.
- The models are :

  K Nearest Neighbors Classifier, Random Forest Classifier, Logistic Regression, Ridge Classifier, Support Vector Classifier and Gradient Boosting Classifier.
- We run the grid on train splits with test splits consisting of 20% of the total observations.
- Since we are working with life threatening medical conditions, we want to avoid false negatives as much as possible. We will thus focus on maximizing the recall score of our models.

Grid Search results

- On average, Random Forest models seem to be performing better in our case.
- Their average recall score is close to SVC but they also offer notably better Precision, which can be important time- and/or cost-wise(not sending healthy patients to take an expensive test).

Method	Accuracy	Precision	Recall	Auroc	False negative	False positive
boosting	0.858227	0.854936	0.869176	0.936878	4.055556	4.592593
forest	0.888889	0.865391	0.926523	0.966398	2.277778	4.500000
knn	0.633197	0.620032	0.750000	0.670766	7.750000	14.625000
logregr	0.744080	0.650772	0.697133	0.802330	9.388889	6.222222
ridge	0.867638	0.853578	0.891278	0.500000	3.370370	4.703704
svc	0.683060	0.675951	0.924731	0.500000	2.333333	17.000000

Model Validation

- We will select this Random Forest Classifier and check for overfitting through cross validation :

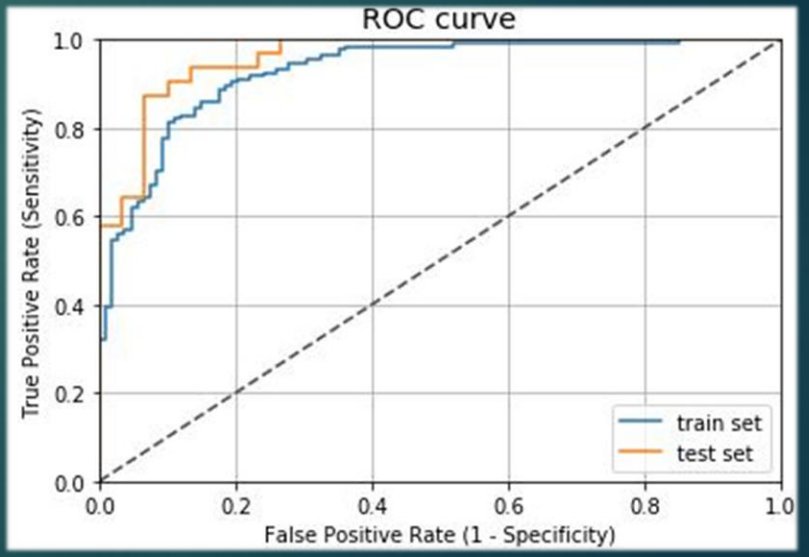
```
Entrée [54]: best = ensemble.RandomForestClassifier(n_estimators=1000, criterion='entropy', max_depth=2)

             np.mean(cross_val_score(best, X, y, cv=10, scoring='recall'))

Out[54]: 0.9080882352941178
```

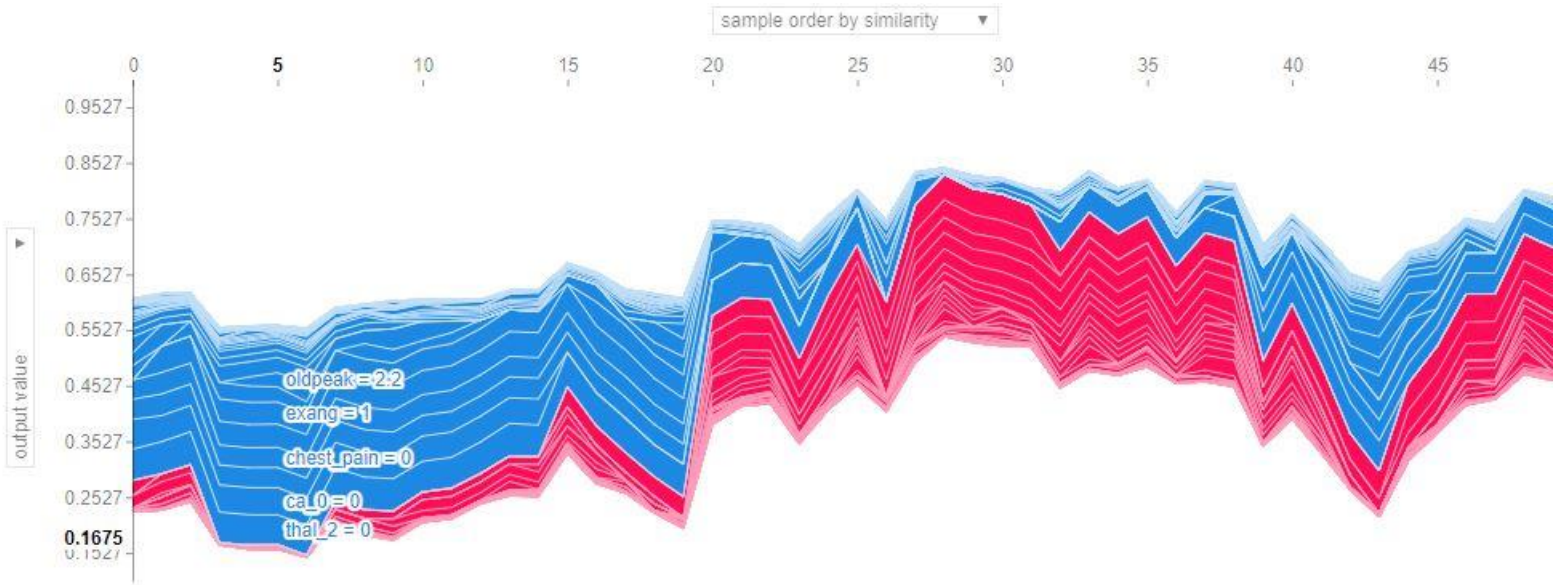
- We also check the ROC curve on train and test splits.

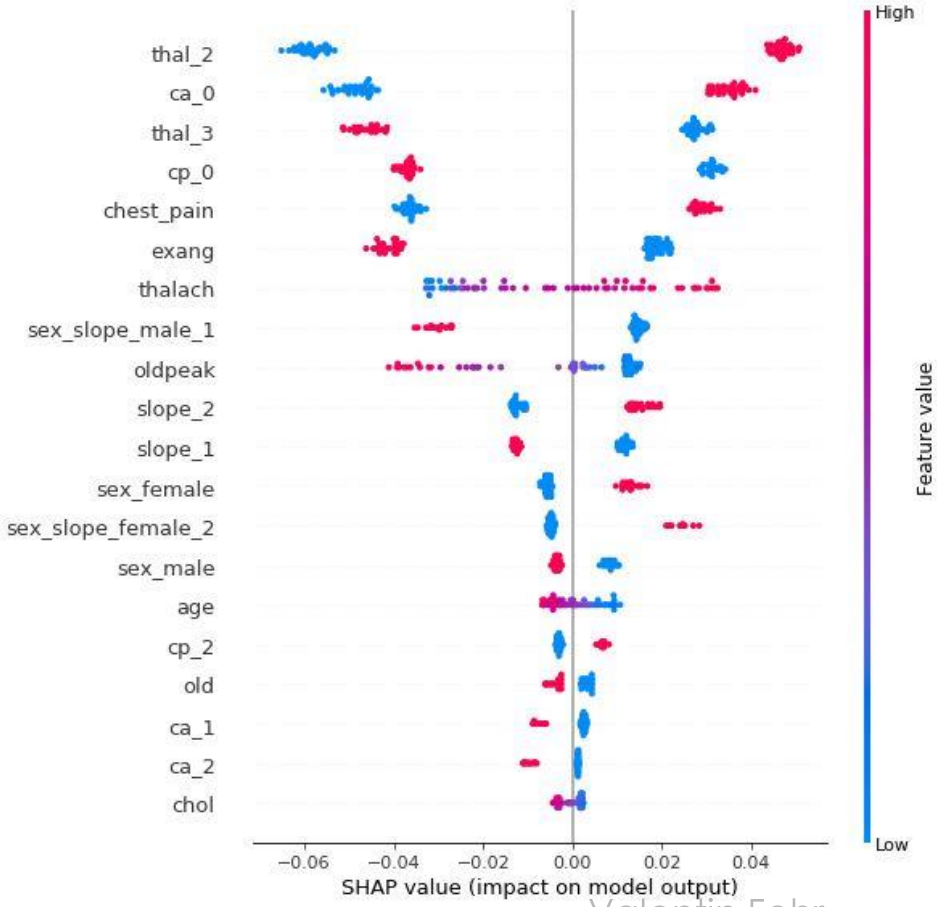
- Our model seems to do great in terms of prediction.



Explanatory Power

- Using the SHAP method we look at the explanatory power of our model and each of its features





Feature Evaluation

- The cost of each group of test is as follows :
  - Age, Sex, Chest Pain, Blood Pressure : Free
  - Cholesterol, fbs, restecg : 27.97 CAD
  - Exang, oldpeak, slope (effort test) : 87.30 CAD
  - Ca (vessels fluorescence) : 100.90 CAD
  - Thal, thalach (thallium heart scan) : 102.90 CAD

- Based on this, we can run our model with only a certain set of features based on what tests have been done.
- Explanatory analysis suggested thallium scan was the most important and we will check that now.

Feature Evaluation

- We now have 5 new sets of features :
  1. Only free observations
  2. Free observations + Cholesterol, fbs, restecg results
  3. Free observations + Effort test results
  4. Free observations + Vessels fluorescence results
  5. Free observations + Thallium heart scan

- We run them to cross validation again and look at the mean recall score for each.

Set of features n°	10 fold cross validation mean Recall score
1	0.762868
2	0.762868
3	0.836029
4	0.818382
5	0.913971

Conclusions

- Our model can **accurately predict when a patient is at risk of heart disease** so it can help health professionals decide when to send a patient for a coronary angiography.
- Moreover, we showed that the best test to predict heart disease was the thallium heart scan. Even if it the most expensive, it is by far **the most efficient at predicting heart disease and would save time and money** if conducted right away.

- Finally, our model has one main issue which is the low number of observations. In the future, it would be helpful to try to add more data by tracking more patients. Even with more data though, the process would stay similar.

References

1. https://www.cdc.gov/heartdisease/facts.htm
2. https://www.mdsave.com/procedures/ct-angiography-coronary-angiography/d786ffc9
3. https://archive.ics.uci.edu/ml/datasets/Heart+Disease