RECOGNITION OF HEART DISEASE

Part II

ABSTRACT

When selecting a hospital or a doctor, patients usually focus on the specialization and prestige of the institutions and staff. While reputation is important, the present work highlights the significance of implementing machine learning tools for the recognition of diseases.

Elianne Mora
Statistical Learning

Introduction

This project is a continuation of the previous classification problem, “Recognition of Heart Disease”, where we analyzed the ‘Heart Disease’ dataset from the clinical and noninvasive test results of 303 patients at the Cleveland Clinic in Cleveland, Ohio. Let us note that there have been no further manipulation of variables nor transformations from the previous portion of this project; no changes in data have been implemented.

In this project we are interested in the predictive capabilities of different machine learning models for the variable *Target*, given all other records (see Index Table 1 for variable names and definitions). We will explore *Decision Tree* models and different tuning of its parameters as well as deep learning methods. All models explored in the present work, will be compared to the LDA previously selected in *Part I*.

The present work implements a cost analysis focused on the patient rather than the institution. The institutions will profit regardless of the accuracy of the diagnosis, which is not of interest at this moment. However, development of algorithms for the detection of heart disease would lower the liabilities of the hospital (court cases, attorneys’ fees, settlement payments, etc.) and increase its prestige and funding potential.

Decision Trees

A decision tree is a flowchart-like structure in which each internal node represents a test on a feature, each leaf node represents a class label, and the branches represent the conjunctions of the features leading to those labels. Decision tree is a type of supervised learning algorithm, which we’ll use to predict heart disease. First, we fit a simple decision tree where we have again a training (60%) and test (40%) set. There are no correlations between the quantitative variables, so we keep them all. As it can be observed, this tree only used 8 variables. At each node we can see the criterion that lead to each following node and final decision. The root node is our variable *Chest Pain*, taking the factor/leaf “Typical Angina”. Then if the patient has typical anginal chest pain, we move to the left node towards the variable *Ca*; otherwise, we move to the right of the tree towards *Oldpeak*, and so on until we reach the final desired classification of our target variable.

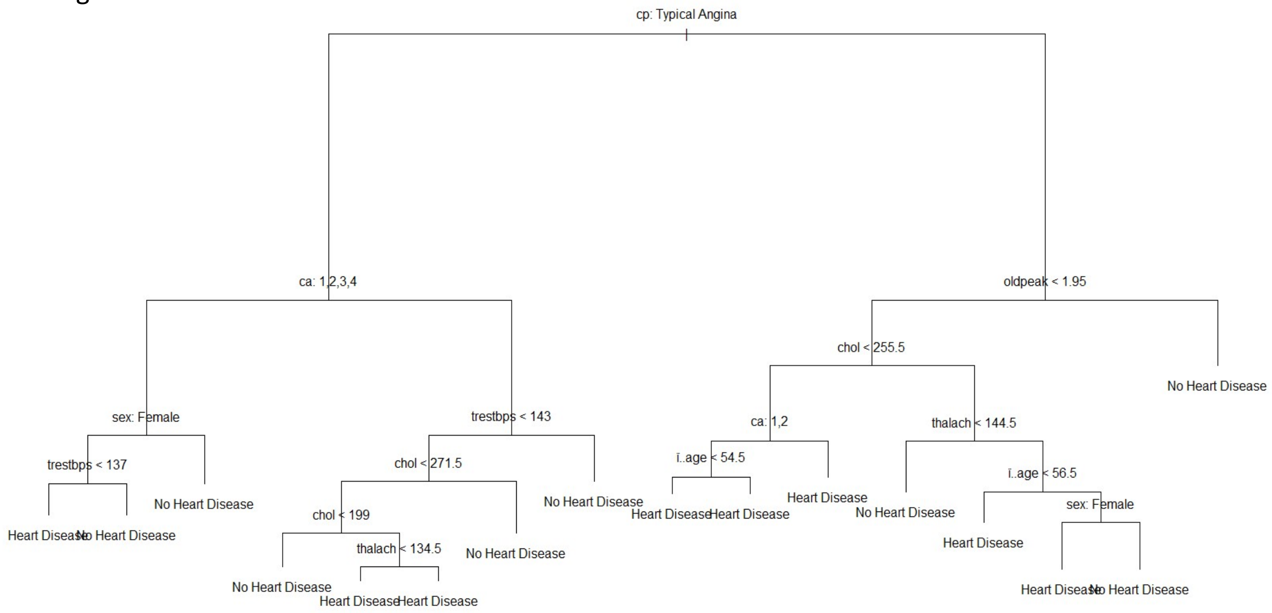


Figure 1. Simple decision tree

The simple tree gives an accuracy rate of 0.7190, so we will introduce cross-validation to optimize this tree and reduce the variance. Pruning the tree down to a size of 6 yields a simpler tree with a 0.7438 accuracy rate. The pruned tree has yielded a simpler tree with higher accuracy when implemented on the testing set; however, the misclassification error given by the trained models show that the latter is much higher.

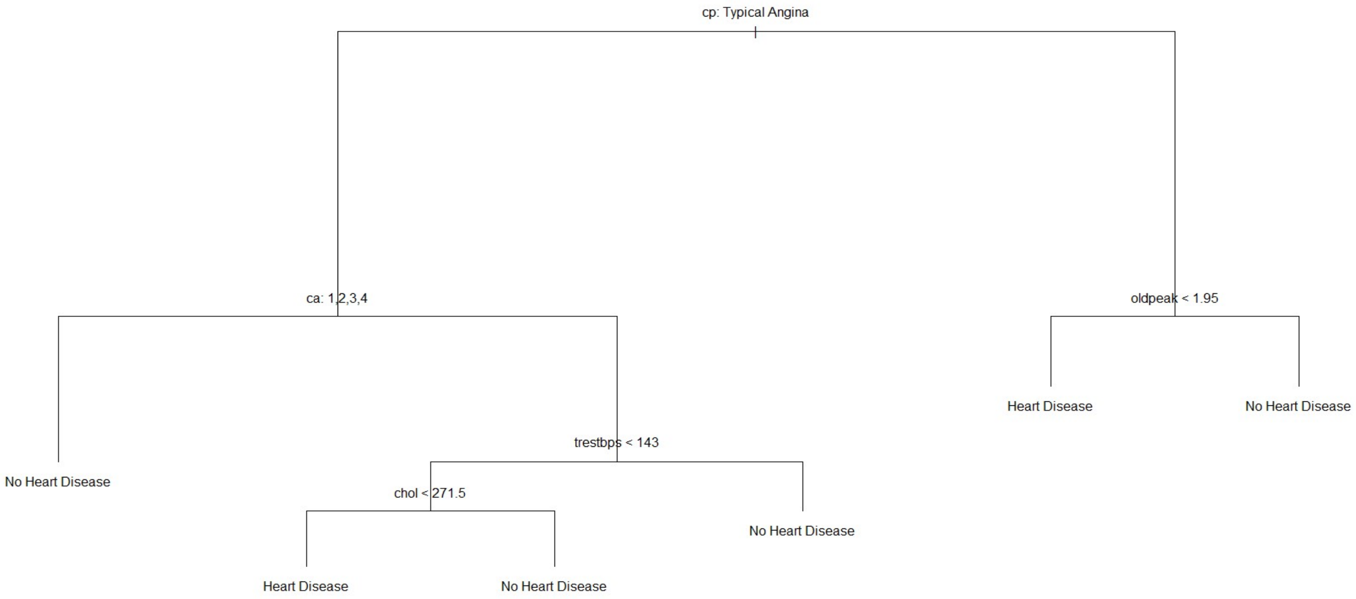
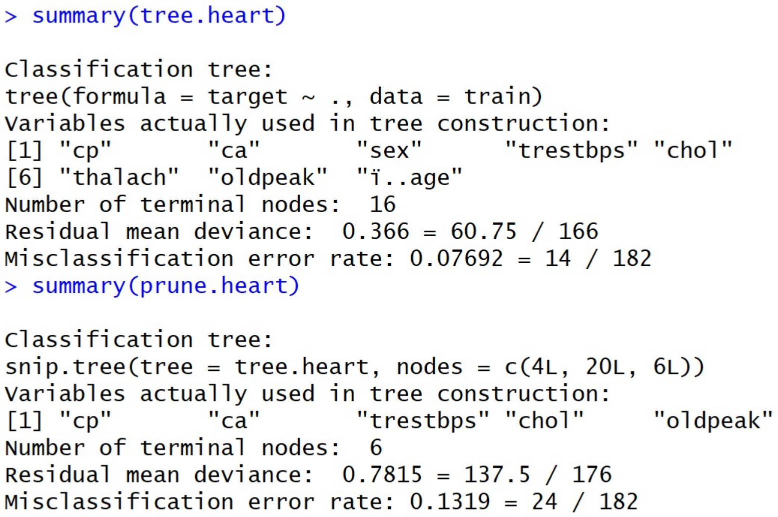


Figure 2. Pruned decision tree.



```
> summary(tree.heart)

Classification tree:
tree(formula = target ~ ., data = train)
Variables actually used in tree construction:
[1] "cp"       "ca"       "sex"      "trestbps" "chol"
[6] "thalach"  "oldpeak"  "ï..age"
Number of terminal nodes:  16
Residual mean deviance:  0.366 = 60.75 / 166
Misclassification error rate: 0.07692 = 14 / 182
> summary(prune.heart)

Classification tree:
snip.tree(tree = tree.heart, nodes = c(4L, 20L, 6L))
Variables actually used in tree construction:
[1] "cp"       "ca"       "trestbps" "chol"     "oldpeak"
Number of terminal nodes:  6
Residual mean deviance:  0.7815 = 137.5 / 176
Misclassification error rate: 0.1319 = 24 / 182
```

Figure 3. Results of the simple decision tree and the pruned tree.

Before we move on to other models and tuning, let us further define and shape the initial challenge to better understand the impact of the classification problem in the real world. We define the associated costs a patient may incur given a diagnosis. In the presence of a true negative – the patient is healthy and is correctly diagnosed as such – the patient incurs no further costs associated with the diagnosis as he/she needs no treatment. In the presence of a true positive – the patient is diagnosed with heart disease and is in fact sick – the patient will incur further medical expenses for medications, consults, exams, surgery, etc. Now, if the patients are given a false positive – diagnosis shows heart disease, but patient is healthy – the patient will incur medical costs, possibly develop complications from unnecessary treatments and drugs, until the diagnosis is assessed and "fixed". Lastly, the consequences of a false negative would lead to death and the possibility of the relatives suing the hospital, initially incurring costs for funerary services, attorneys' fees, etc. Hence, the problem now highlights the importance of selecting a clinic at par with the latest technologies for the detection of diseases.

For simplicity, let us define the costs as follows, where they are expressed in dollars spent monthly given the diagnosis:

	No Heart Disease	Heart Disease
Predicted No Heart Disease	$0	$500
Predicted Heart Disease	$150	$200

Figure 4. Patient's cost matrix

From the conclusions drawn from the first part of this project, the LDA approach was selected. Let us implement once again the LDA approach adding the relative costs of the classification problem. If a patient was to attend a hospital, where the LDA model is the preferred method for diagnosis of heart disease, he/she is likely to further incur expenses in the amount of $141.32/month.

Random Forest

Random forest (RF) trees outperform the previous tree model when it comes to predictive power and misclassification errors. For the current section and subsequent, the associated costs will be used as comparative measure between models. Let us first fit a RF that grows 100 trees and samples 6 random variables at each split, using cross-validation as the basis for the pruning (5-fold), and a cutoff of 0.7. The image below showcases the process selection of the random forest given the specified parameters over observed points (black line), and the predictions. As the trees grow the observed error remains between 0.2-0.3, while the green line gives much lower values throughout the process. The associated patient cost under this model is $152.48/month, much higher than with LDA which means the error rate of a false diagnosis is even higher.

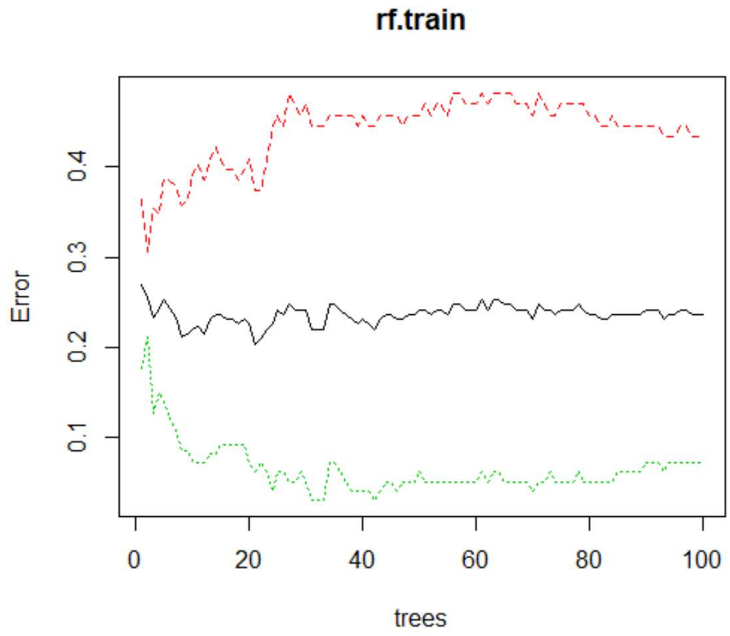


Figure 5. Random forest process with five-fold cross-validation, ntree=100, mtry=5, cutoff=(0.7,0.3).

Refitting the RF model with the cutoff (0.6, 0.4) and a threshold of 0.5 used in LDA, we obtain a cost of $147.52, much lower than the previous RF results but still higher than the LDA results. Note that further growing the tree does not yield better results, so this parameter remains fixed at 100 and we focus on optimization of all other parameters.

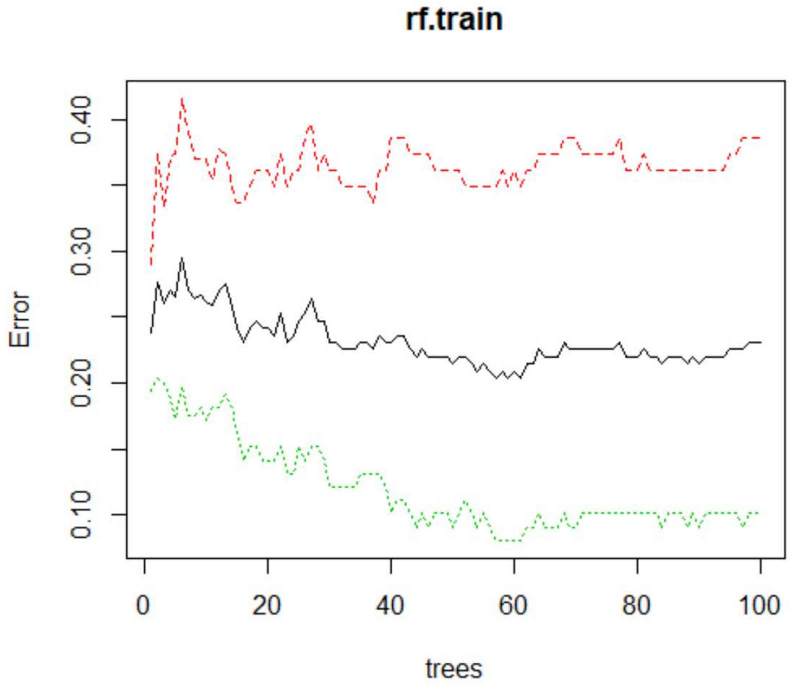


Figure 6. Random forest process with five-fold cross-validation, ntree=100, mtry=5, cutoff=(0.6,0.4).

The variable importance given by this RF is as follows:

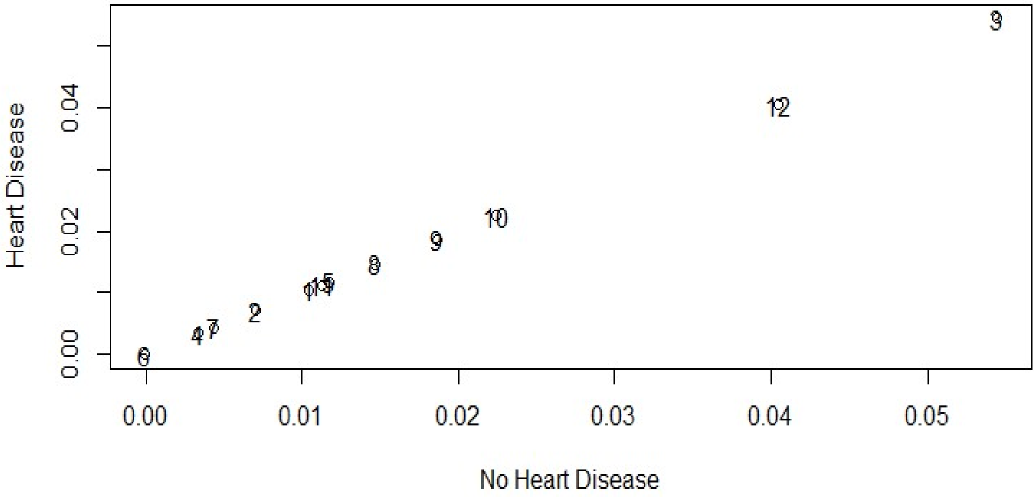


Figure 7. Variable importance plot for the classification problem.

The most important variable is given by label 3, which corresponds to chest pain (Cp), followed by the number of major vessels colored by fluoroscopy (Ca), the slope results from the stress test (Oldpeak), exercise induced angina results (Exang), the person's record maximum heart rate achieved during the stress test (thalach). However, we're unable to establish the factors and values of these variables at this time.

Gradient Boosting

Gradient boosting allows us to further improve our tree models. While bagging creates multiple copies of the training dataset using bootstrap and then combining all trees to create a single model, boosting grows the trees sequentially using information from the previous trees, which does not involve bootstrapping. To slow the sequential partition down, we will set the *shrinkage* parameter to 0.01, so that more diverse trees are created to minimize residuals.

We implement a gradient boosting model with 100 trees to be grown as in the previous sections. Out of the 12 available predictors, 11 have non-zero influence, however the model does not identify which ones. Setting the threshold to 0.5, this model yields an expected cost of $133.88/month. A significant improvement from both LDA and RF alone. This model offers an optimization advantage, so we build a more intricate model where each parameter (number of trees, shrinkage, the number of variables sampled at each split, etc.), while keeping the same threshold. The results are not better compared to the previous methodology as the cost has nearly doubled, which is an indication that rather than focusing on the growth of the trees, we should focus on the tuning of other parameters. From the graph below, we can see that the significant variables under this model remain the same for the first four –

although the order has certainly changed – and age has replaced Ca. Here we finally identify fasting blood sugar (*fbs)* as the 12th variable that had a zero influence on the model.

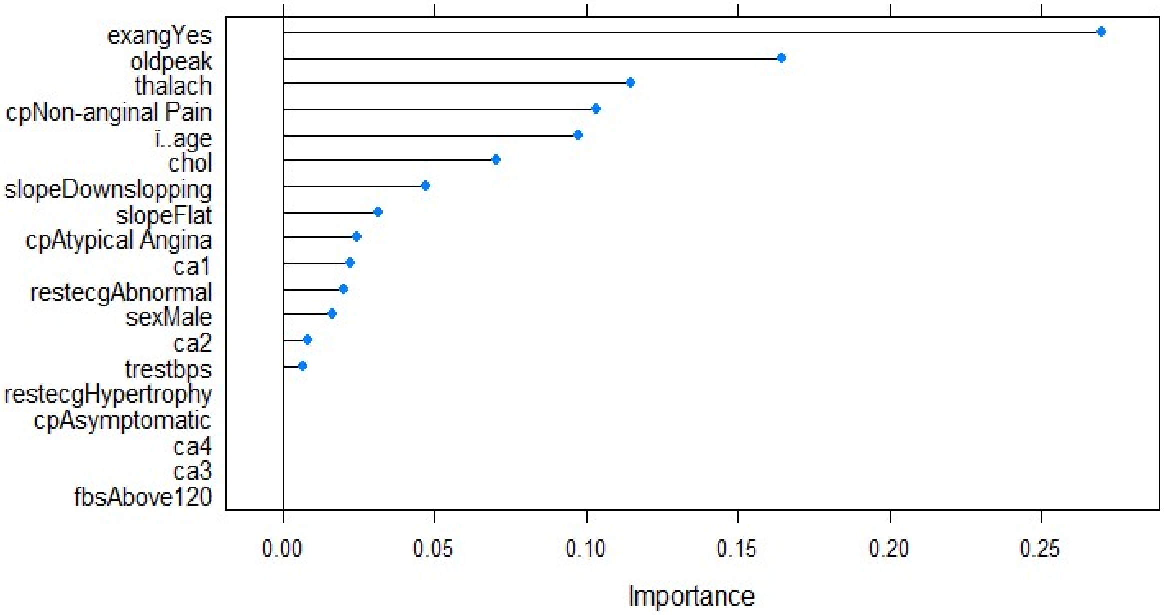


Figure 8. Importance of variables under the tuned boosting model.

Deep Learning

Deep learning is a subset of machine learning that uses multi-layered artificial neural networks to deliver high accuracy in diverse tasks, such as object detection, speech recognition and language translation, among many others. Neural networks are a set of algorithms designed to recognize patterns. The patterns they recognize are numerical, contained in vectors, into which all real-world data – be it images, sound, or text, must be translated.

Neural Network

NN models require large amounts of data and performance is better when there are many variables involved, however, we will fit a simple NN model with just one hidden layer, using the same grid parameters used in boosting, a maximum number of networks to be expanded up to 300 in 100 iterations. From the plotted results below we can observe that the optimal weight is 0.001 with 2 levels (grid) as it yields the lowest errors, which consequently yields the lowest cost found slightly above $135 – much better than LDA and RF, but not better than the simple boosting model that yielded $133.88.

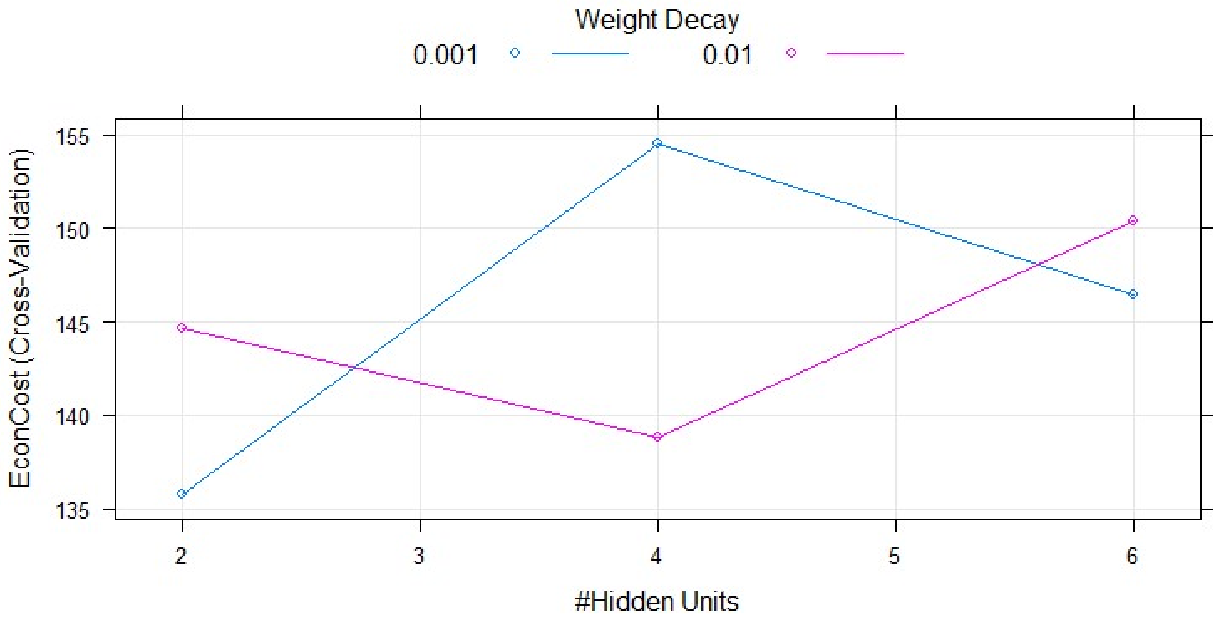
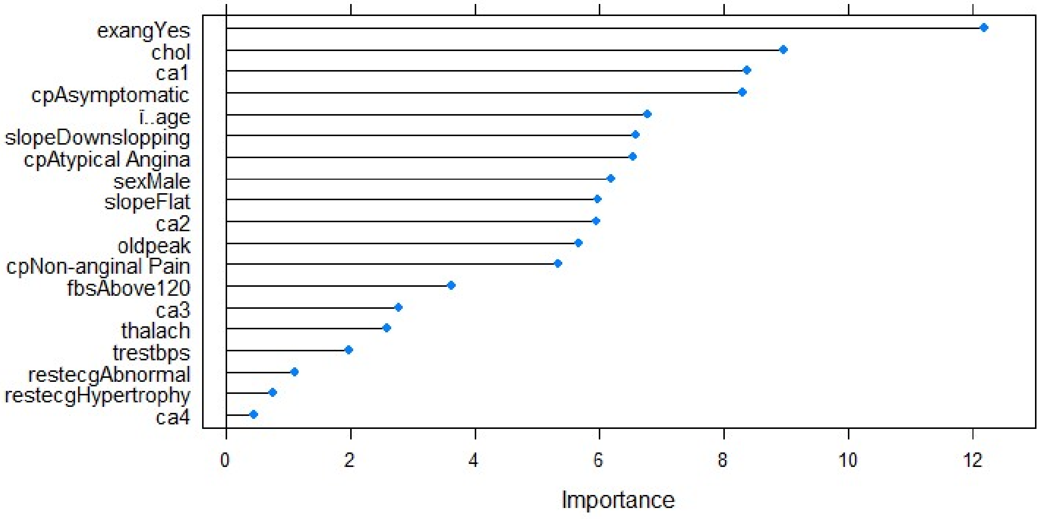


Figure 9. One-layer neural network results with tuned parameters

Once again, we can observe that most of the significant variables are repeatedly found in each model. This model does not include the variable *Thalach* as it has been replaced by Ca level 1 (only one vessel has been lighted through fluoroscopy), and Fbs is not included at some level of significance, which the gradient boosting method did not.



Fitting a deep neural network model with equal number of iterations as the NN model and 3 hidden layers leads to an expected patient cost of $298.76. Which is by far the worst result, although not surprising given the dimensions of our dataset.

As a final step, we consider the combination of the LDA, gradient boosting, and NN models. As it can be observed below, the models are highly correlated. The ensemble yields a cost of $176.03 optimizing the threshold at 0.25.

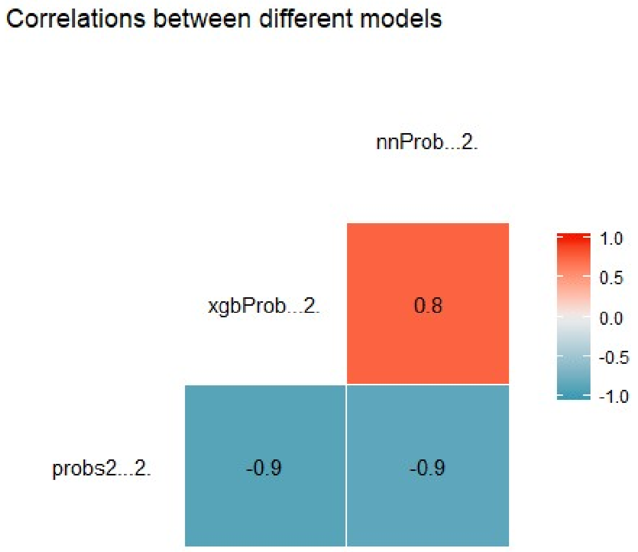


Figure 10. Correlation among the predicted probabilities of the Neural Network, Gradient Boosting, and LDA models.

Since the NN model alone did not yield improved results, we try a second ensemble where we only consider the best two models, LDA and gradient boosting model. This ensemble proves much useful and better with a predicted cost of $132.64/month at an optimized threshold of 0.35. This ensemble is slightly better that the gradient boosting model alone.

Conclusions

The deep and neural network models performed relatively poor compared to the LDA and tree approaches; however, this is not surprising given the rather “small” dimension of our data, which is not apt to be fed into such models. Although the best performing model on its own was the simple gradient boosting model, combining it with the LDA – whose predicting power (93% AUC) and error rate set it apart from all other models – yields much better results.

While technologies set apart hospitals and doctors, it is important to understand the methodologies and implementation of said technologies. As a patient searching for the best institution to be treated, one may consider the innovative approaches and research being conducted at these organizations. Therefore, this analysis was conducted from the perspective of the patients rather than the institutions, as the latter will benefit from any false diagnosis is inadvertently gives. While the cost analysis is subjective, the overall approach is relevant.

Bibliography

James Le, **Decision Trees in R** (2018) – https://www.datacamp.com/community/tutorials/decision-trees-R

Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar, **Introduction to Data Mining** (2018) – https://www-users.cs.umn.edu/~kumar001/dmbook/index.php

Tables

- Index table 1 – variables and definitions

ID	*Description*
Age	Patient's age
Sex	Patient's sex (0=female and 1=male)
Cp	The chest pain experienced (Value 0: typical angina, Value 1: atypical angina, Value 2: non-anginal pain, Value 3: asymptomatic)
Trestbps	Patient's resting blood pressure (mm Hg on admission to the hospital)
Chol	Patient's cholesterol levels in mg/dl
Fbs	Person's fasting blood sugar (if > 120 mg/dl, 1 = true; 0 = false)
restecg	Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' crite
Thalach	The person's maximum heart rate achieved during Stress Test (exercise)
Exang	Exercise induced angina (1=yes, 0=no)
Oldpeak	Stress test depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot)
Slope	the slope of the peak exercise ST segment (Value 0: upsloping, Value 1: flat, Value 2: downsloping)
Ca	The number of major vessels colored by fluoroscopy
Target	Diagnosis of heart disease (1=yes) or no heart disease (0=no).