ACME Learning Co.

CASE STUDY :

Insurance risk management

Vincent Roy

1. **Problem definition**
2. **Methodology**
3. **Data transformation and analysis**
4. **Model selection and evaluation**
5. **Conclusions**

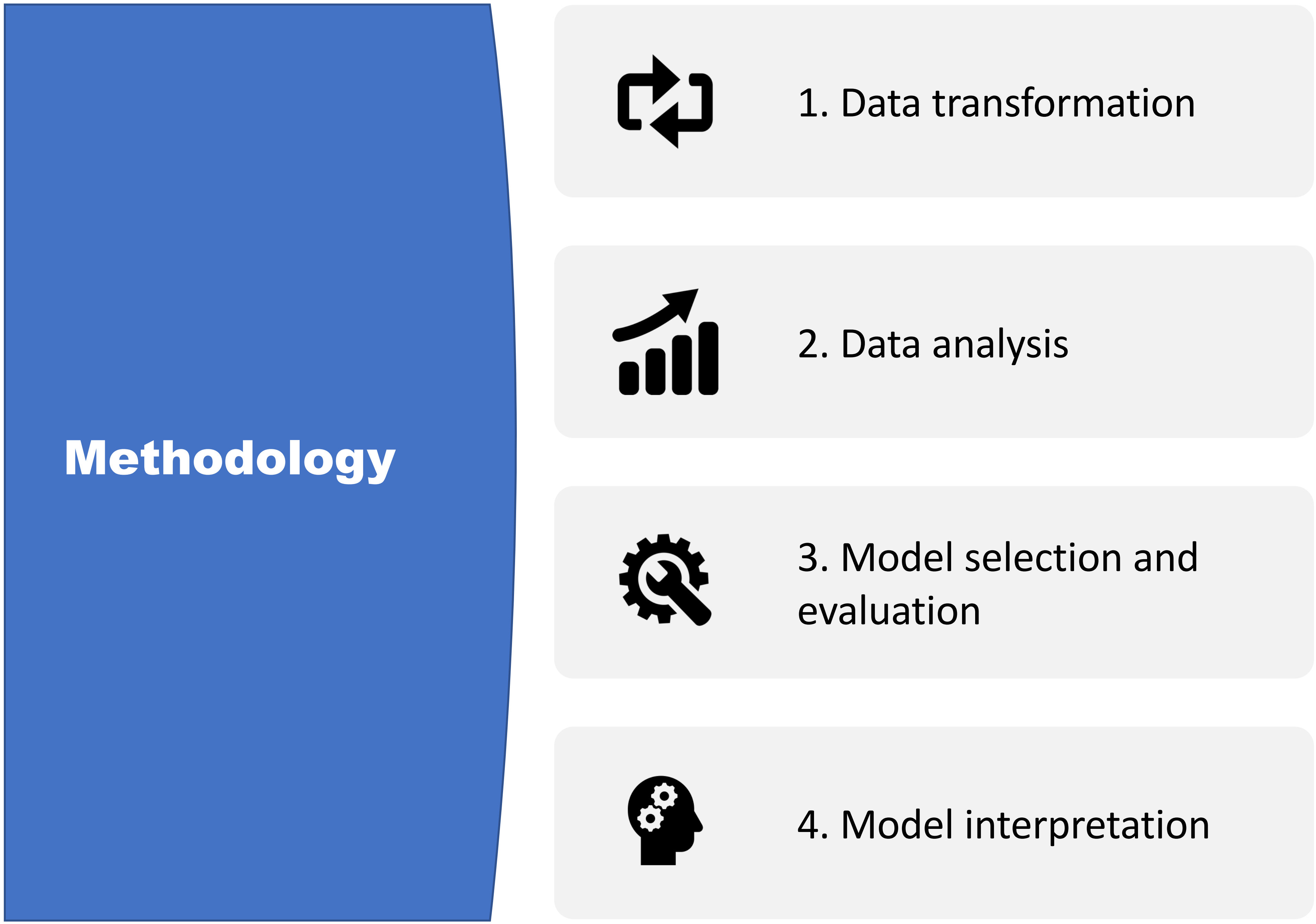
Problem definition

The ACME Learning Co. has been given a mandate by NoCoverage Insurance to identify ways to help it's client lower their insurance coverage risks.

Problem definition - objectives

More specifically NoCoverage Insurance wishes to know :

- What are the risk-factors associated with the presence of a heart disease ;
- Provide a way to easily allow insurance brokers to predict the risk of future client having a heart disease based on lab tests.

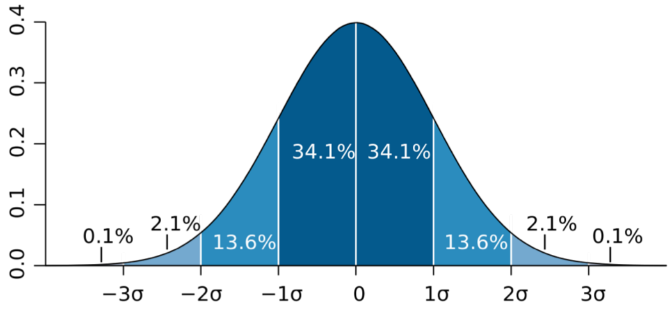


Methodology

1. Data transformation
2. Data analysis
3. Model selection and evaluation
4. Model interpretation

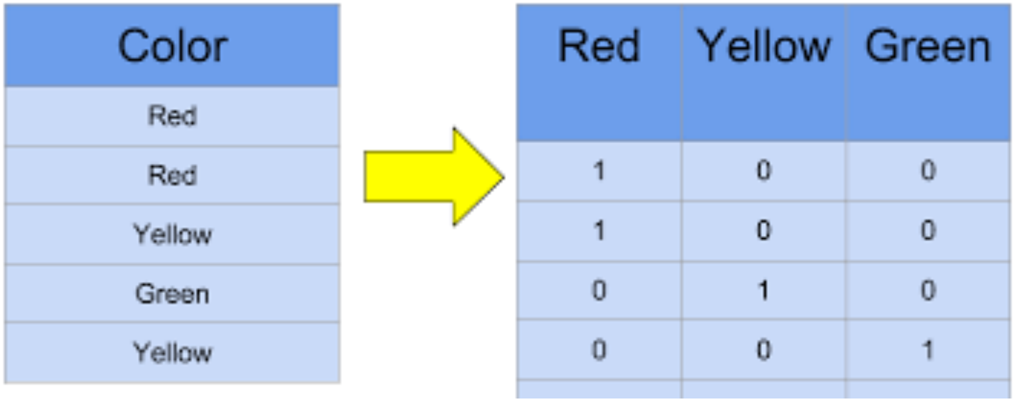
Data transformation

- ACME Learning Co. was provided a database of 303 insured clients containing lab results for 13 medical parameters ;
- The data contained both quantitative and qualitative data ;
- Categorical data was transformed to a binary values, one for each category ;
- Quantative data was standardized so that the mean is at 0 and that one standard deviation is one unit in value.



Data transformation

Categories to binary transformation



Color
Red
Red
Yellow
Green
Yellow

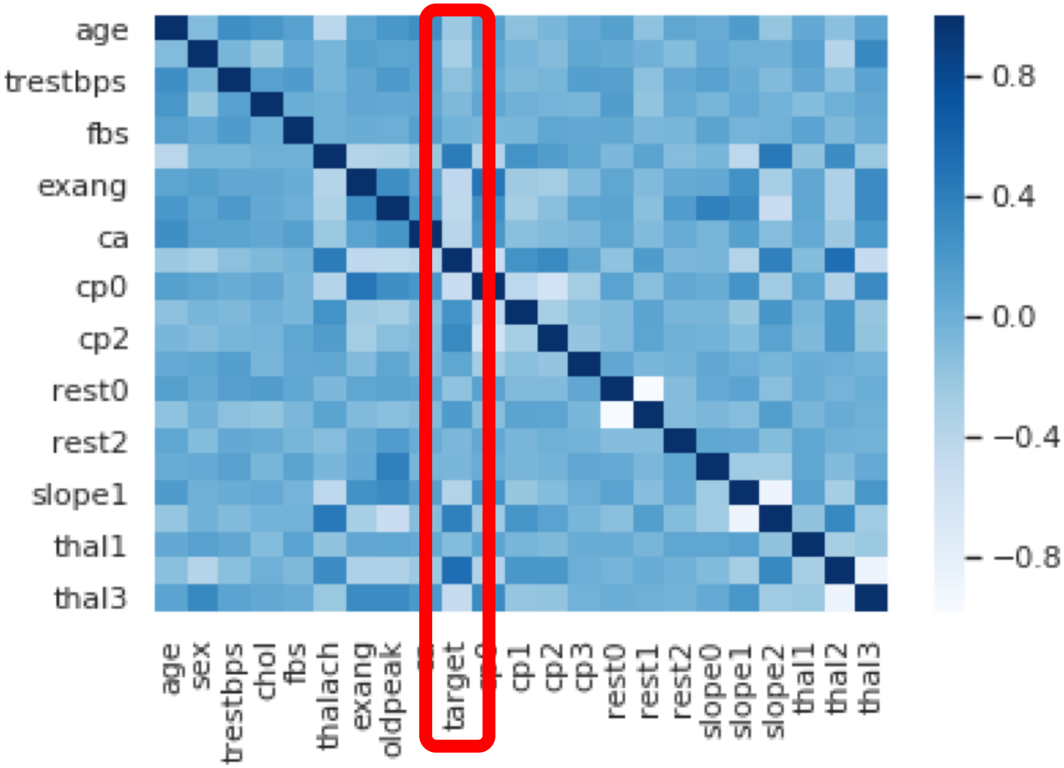
Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

The following quantitative variables were transformed :

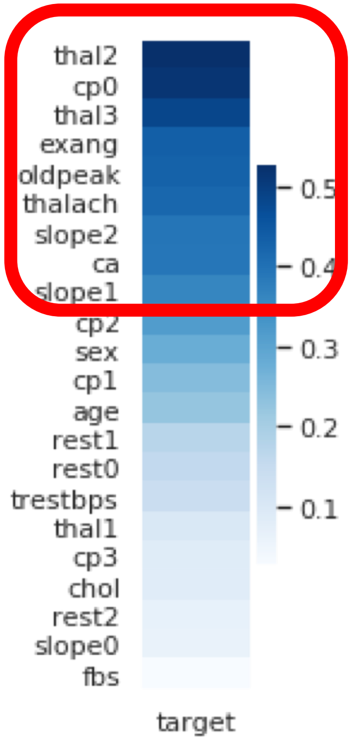
- cp: chest pain type – 4 categ.
- restecg: resting electrocardiographic results – 3 categ.
- slope: the slope of the peak exercise ST segment - 3 categ.
- thal – 3 categ.

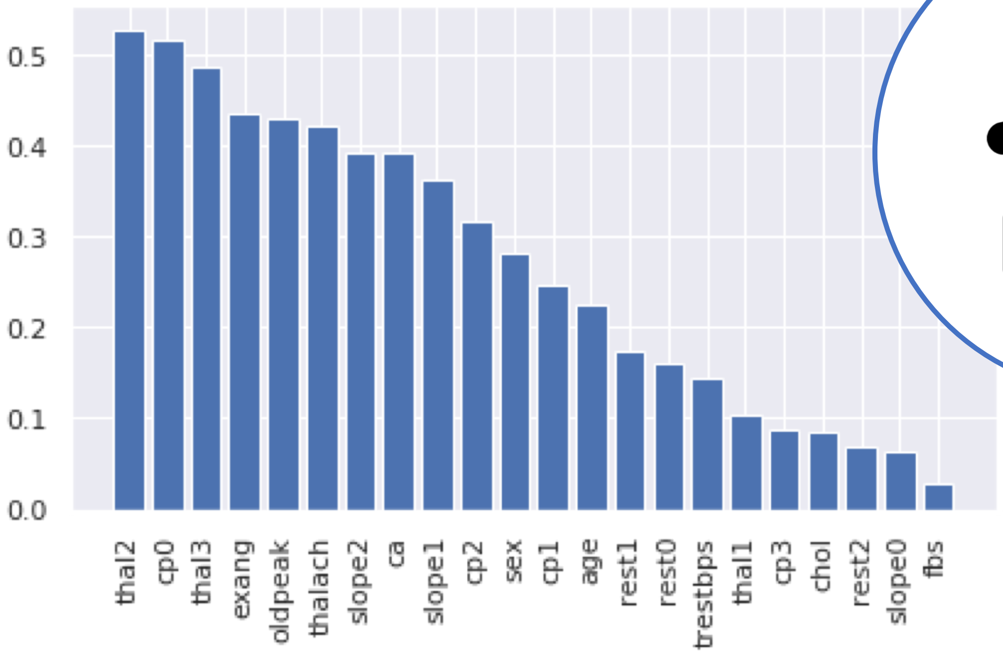
Data analysis

- 165 out of 303 candidates have a heart disease ;
- Correlation matrix indicates the variables in the data set that are highly related to the risk of heart disease :



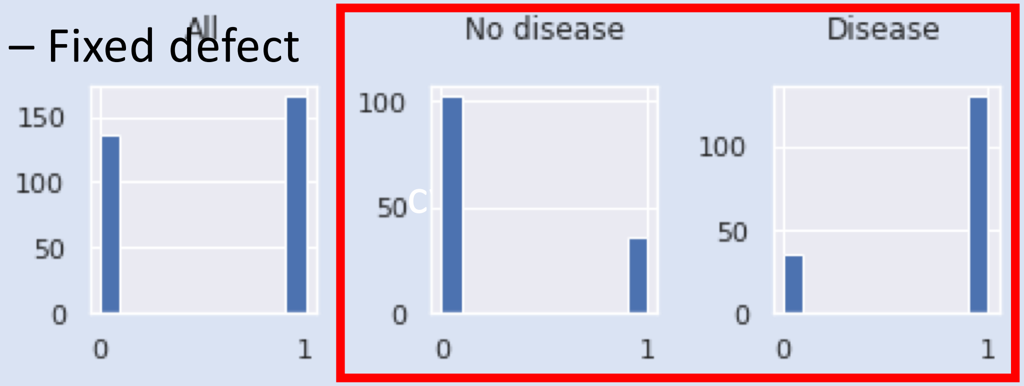
Data analysis



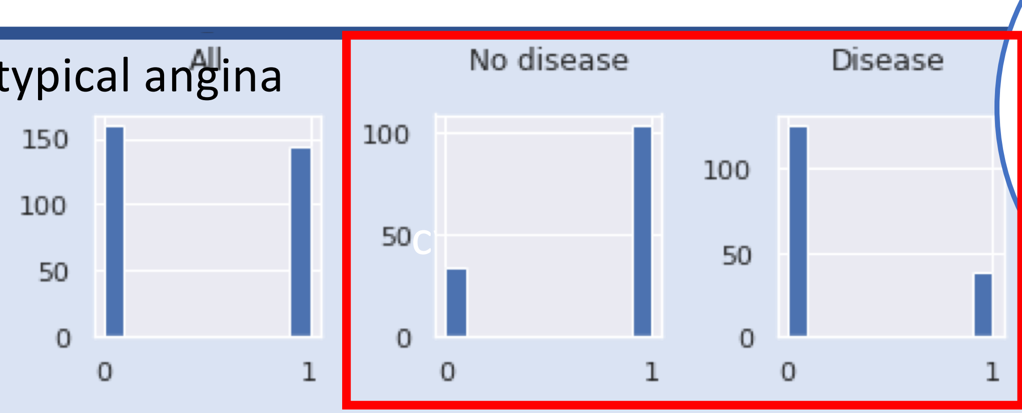


Data analysis

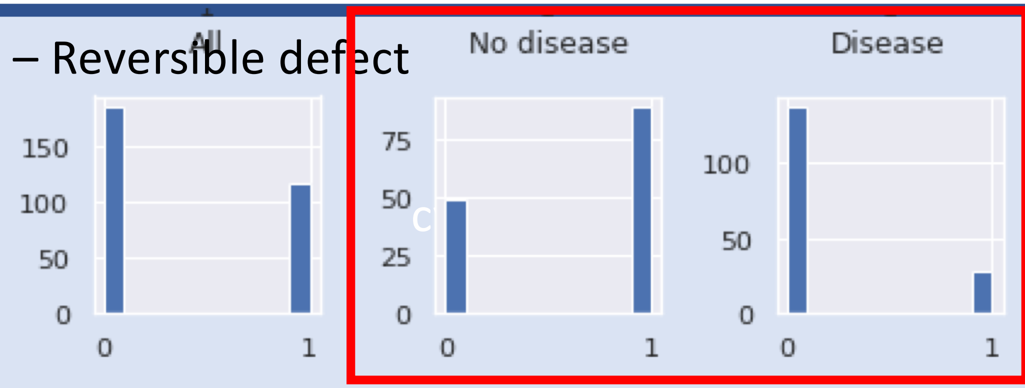
THAL2 – Fixed defect



CPO - typical angina

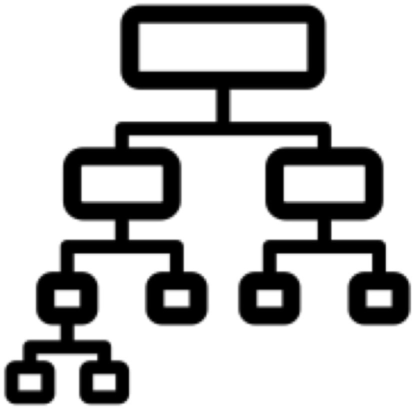


THAL3 – Reversible defect



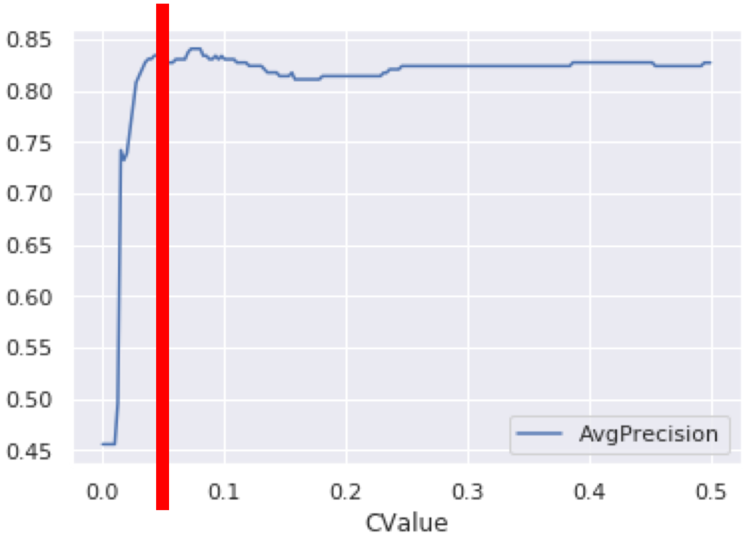
Model selection and evaluation

- Two different types of models were created from the transformed data :
  - Logistic Regression model with regularization ;
  - Decision Tree with bagging.
- These models are suited for classificatior
- Nocoverage is interested in high **recall** as it relates to a **low** false negative rate.



Logistic Regression model

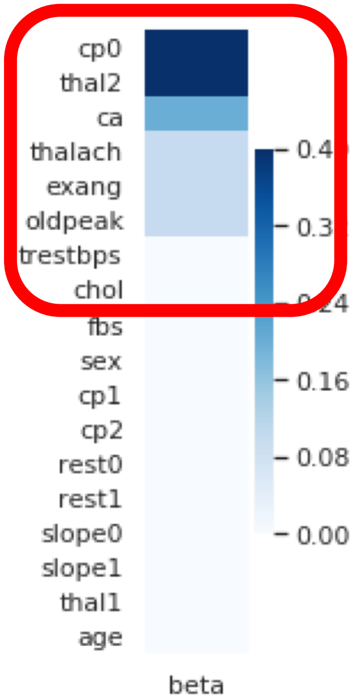
- Logistic model was created with 10 fold cross-validation and l1 norm regularization to eliminate unnecessary parameters ;
- Optimized regularization parameter without compromising model recall.

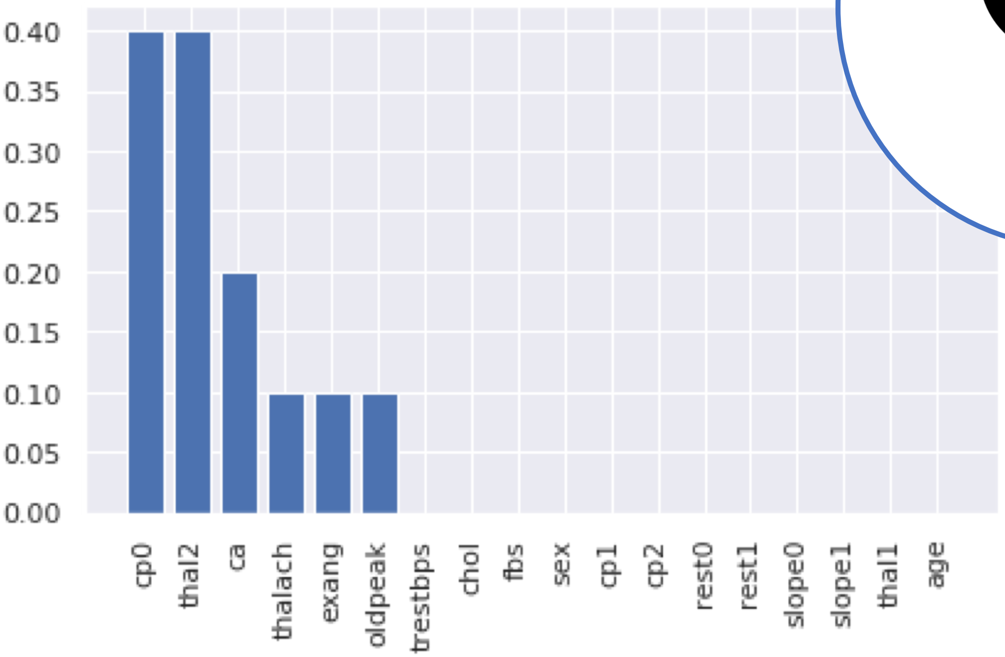


Reg value = 0.05
Recall = 0.86
Accuracy = 0.85

Logistic Regression model

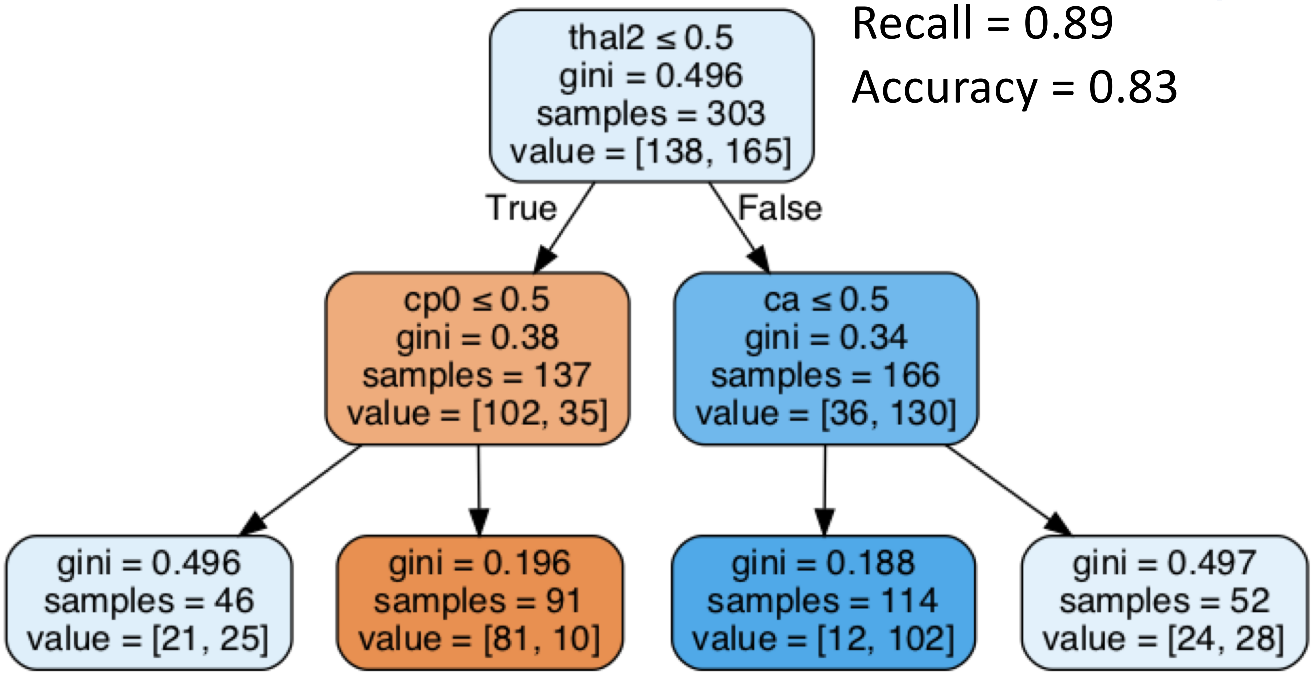
- Logistic with regularization identified the following predominant parameters. That is the high risk factors associated with a heart condition





Decision Tree

- First a decision tree was created with optimized depth = 2 based on recall ;
- Cross validation was used to get the recall and accuracy.

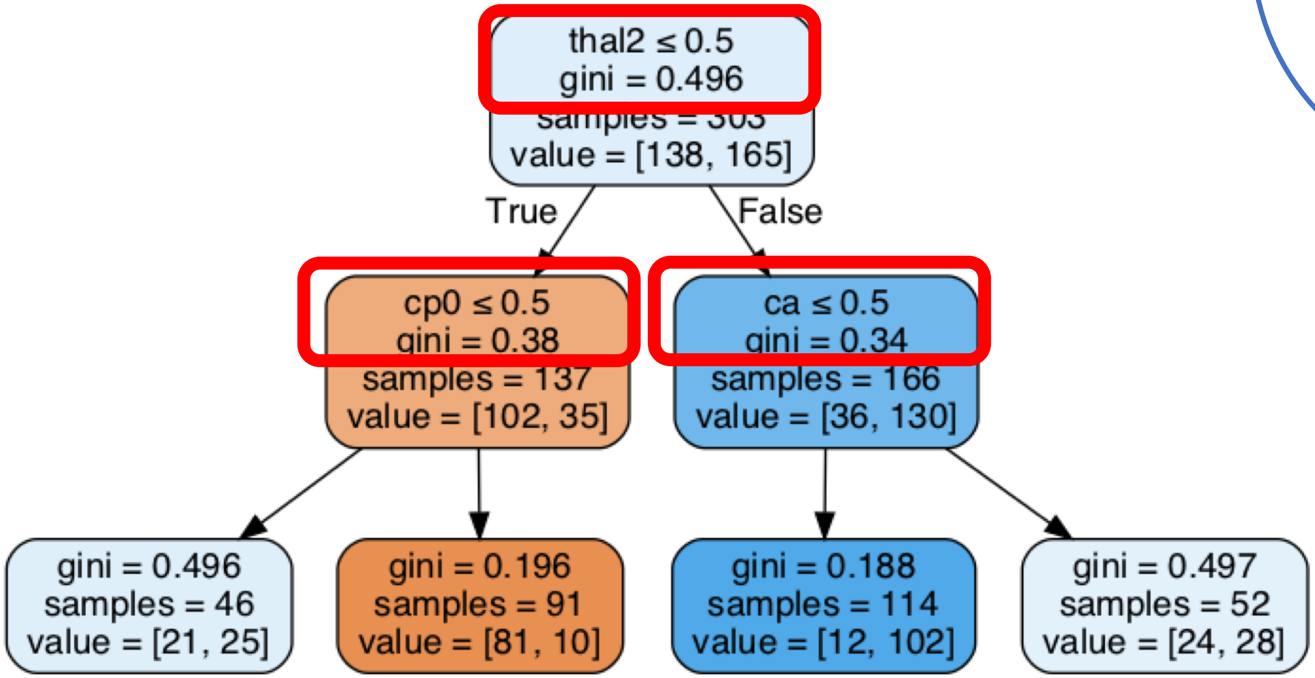


Recall = 0.89

Accuracy = 0.83

Decision Tree

- From the decision tree it can be seen that the following parameters are risk factors of heart disease :
  - thal2 : fixed defect
  - cp0 : typical angina
  - ca : small number of major vessels



Conclusions

- Both Logistic and Decision Tree models identified that the following parameters are highly associated with the risk of heart disease :
  - thal2 : fixed defect ;
  - cp0 : typical angina – chest pain ;
  - ca : small number of major vessels.
- This is in accordance with the raw analysis of the data ;
- It is possible to build a reliable model for the prediction of heart disease from the data ;
- With a decision tree graph an insurance broker could easily predict the risk of heart disease from three parameters ;
- A more advanced model could be built … but we need another mandate $$$.

Questions