Predicting the maximum heart rate of patients with presence and absence of heart disease

Agenda

- ***Description and importance of project***
- ***Description of project methodology***
- ***Explanation of multiple regression analysis and results***
- ***Explanation of decision tree algorithm and results***
- ***Comparison of results***
- ***Implication for future research***

Project Description

- ***Heart attack or stroke as major causes of many deaths in United States!***
- ***Cardiovascular disease being the number one killer in America !***
- ***The things that are known to increase the risk of heart attack are called risk factors***
- ***There are two types of risk factors: <u>risk factors that can be influenced-</u> raised blood cholesterol, raised blood pressure, stress, smoking, diabetes and <u>risk factors that can not be changed</u> – age, gender, previous history of heart disease, family history of heart disease***
- ***Purpose of project is to differentiate patients with heart disease from healthy based on certain patient characteristics***

An-Initial Look at the Data

- ***What type of variables are in the dataset?***
  -Combination of real, ordered, binary and nominal

Number of Classes: 2

Class	*Nr. Observations*
1-healthy	*150 (55.56%)*
2-with disease	*120 (44.44%)*

- **The total number of variables in the dataset – 14, which have been extracted from a larger set of 75**
- **The training data set consists of 179 randomly selected observations, testing of 91 observations**

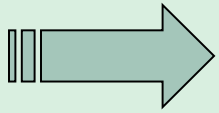
Project Methodology

Statistical Methods Applied:

- Decision Tree as Technique for Prediction

- Multiple Regression Analysis as Feature Selection Technique

Age
Sex
Chest Pain Type
Resting Blood Pressure
Serum Cholesterol
Fasting Blood Sugar
Resting Electrocardiographic Results
Maximum Heart Rate
Exercise induced Angina
Old peak=ST depression
Slope of peak exercise ST segment
Number of major vessels
Thal
Class of patient



MODEL SELECTION METHODS

- ***Backward Elimination Method***
- ***Forward Selection Method***
- ***Step-wise Selection Method***

Multiple Linear Regression Analysis

- ***Is gender a good predictor of the maximum heart rate?***

  The fact of being selected by only one selection method- Forward Selection and yet having p-value greater than 0.05 allows us to say that gender does not contribute to prediction of the maximum heart rate

- ***Selection of the best model predicting the maximum heart rate***

  Backward Elimination and Step-wise Selection methods turned out to have the same attributes as predictors of the maximum heart rate:

SELECTED PATIENT'S ATTRIBUTES

Age	p-value 0.0001
Serum cholesterol	p-value 0.0695
Blood sugar	p-value 0.0164
Ex. Ind. Angina	p-value 0.0022
Slope of Peak Ex.	p-value 0.0011
Class-heartdse	p-value 0.0047

Multiple Linear Regression Analysis

- **The best model predicting *Maximum Heart Rate***

y = 221.41 – 0.8082 age + 0.0095 cholstrl – 2.4828 bloodsugar – 11.299 indangina – 11.153 slopepeak – 6.2644 heartdse

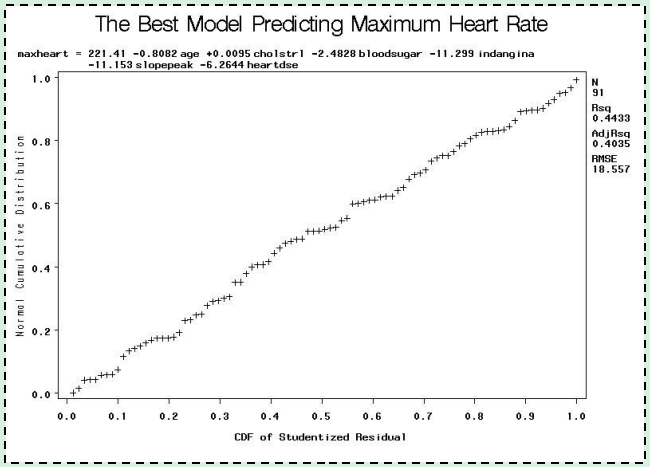
- Maximum heart rate increases as serum cholesterol increases

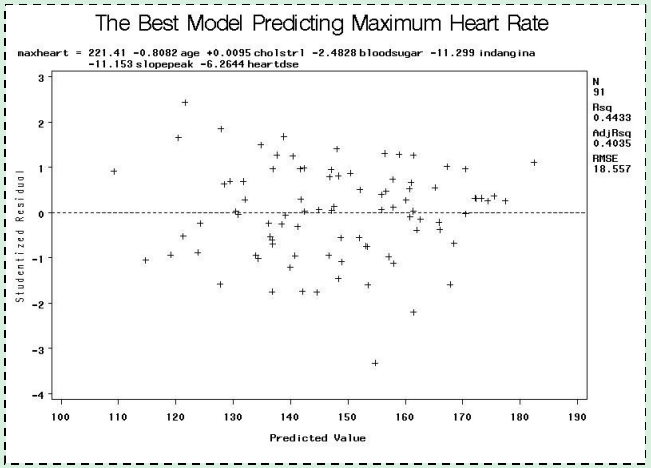
- Maximum heart rate decreases as age, blood sugar, exercise induced angina and slope of peak exercise increases

- Patients with heart disease have lower maximum heart rate

Regression: Results

- **Regression Analysis conducted on 13 independent variables and one dependent –maximum heart rate shows the liner association between maximum heart rate and all other variables**

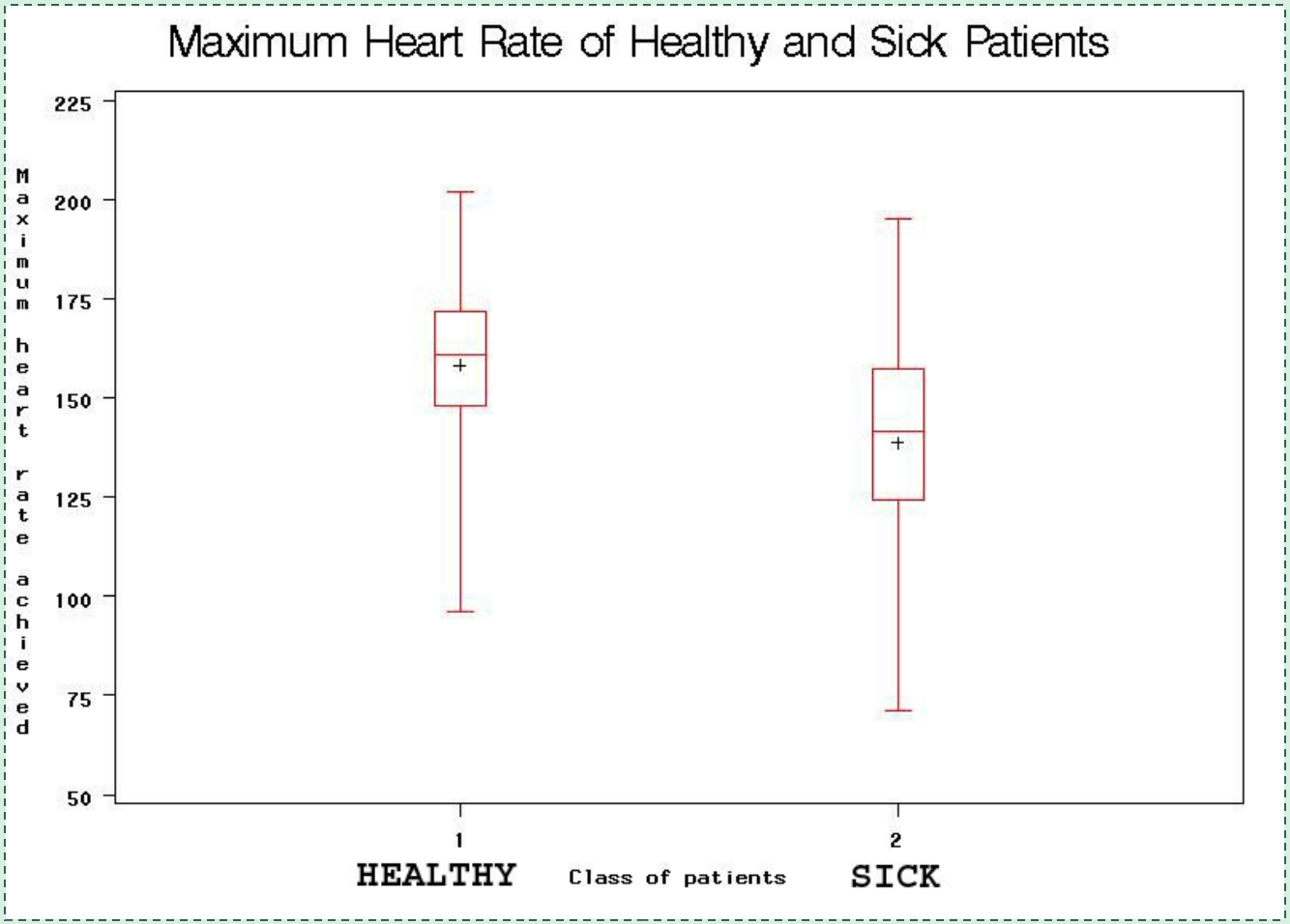




- **Residuals appear to be randomly spread showing that equal variance assumption has been satisfied**
- **R^2 shows that 44.33% of variation in the maximum heart rate about its mean has been explained by the model**

Maximum Heart Rate:

Healthy Patients versus Patients with Heart Disease



Conclusion of Regression Analysis

- **The maximum heart rate shows to be higher for healthy patients**
- **Patient characteristics predicting the maximum heart rate can be divided as:**

  - Risk factors that can be influenced: serum cholesterol, blood sugar and exercise induced angina

  - Risk factors that can not be changed: age and class of patient
- **Gender does not appear to have an influence on the maximum heart rate**
- **Blood pressure selected by Forward Selection Method, however with non-significant prediction of the maximum heart rate**

Decision Tree As Technique for Classification or Prediction

- **Performs exploratory and confirmatory segmentation and classification analysis**
- **Can be defined as a collection of decision rules that predict or classify future observations**

- **Decision tree was constructed based on 66% of the original data set – training sample**
- **The best decision tree model in terms of accuracy was tested on 34% of the original data set –testing sample**

Decision Tree Definition

- Decision tree consists of leafs showing the object class, and branches
- The search begins from root until the object class is reached
- Decision tree model predicted the maximum heart rate of patients based on the decision rules
- After running several models, the one with the best accuracy was based on the following growing method and stopping rules:

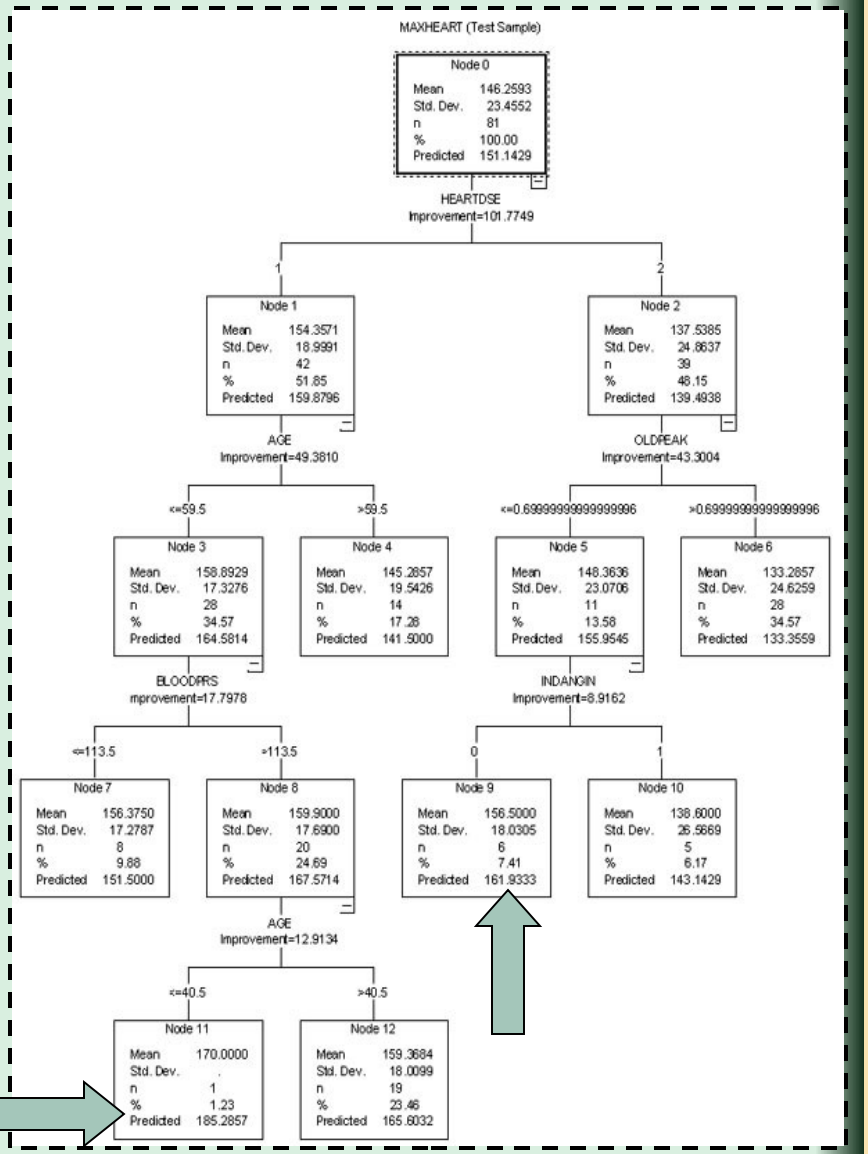
  - Growing method: C&RT
  - Impurity Measure : Gini
  - Stopping rules: Levels below root : 4

  Minimum number of Cases for Parent node : 15

  Minimum number of Cases for Child node : 5

Decision Tree Results

- First split is made on class of patients: healthy or sick. This is not surprising, since as previously discovered patients with heart disease have lower maximum heart rate
- Next, the two nodes representing class of patients are split on different predictors. For healthy patients splitting is on age, whereas for sick splitting is on oldpeak-depression induced by exercise relative to rest
- For healthy patients and older than 59 years the split is based on the blood pressure, whereas for sick patients with oldpeak being greater than 0.69 the split is based on exercise induced angina
- Last split for healthy patients with blood pressure greater than 113.5 is based again on age by classifying them into two age groups greater and less than 40 years old



Decision Tree Conclusion

- **Classification tree was build on the chosen patient characteristics and shows the ones that are most important for determination of the maximum heart rate**

  - Risk factors that can be influenced : blood pressure, exercise induced angina and oldpeak – depression induced by exercise relative to rest

  - Risk factors that can not be changed: age and class of patient (healthy or with heart disease)

- **The maximum heart rate for healthy patients can be calculated according to a new formula from "The Journal of the American College of Cardiology" :**

  208 – (0.7)*age

  According to the formula for healthy patients age less than 40 with blood pressure greater than 113 the maximum heart rate is 180. Predicted by decision tree is 185, which in terms of value was the highest maximum rate predicted by decision tree algorithm

Evaluation of Decision Tree

- **How well the grown tree does at predicting the maximum heart rate in terms of examining the risk summary?**

  Total variance = within-node (error) variance + between-node (explained) variance

  within-node variance = 416.242 Risk Estimate

  SE of Risk Estimate = 67.1026

  total variance (Risk Estimate for tree with only one node) = 543.353

  Proportion of variance due to error is 416.242 / 543.353 and equals to 0.7661

  Proportion of variance explained by model is:

  100% - 76.61% = 23.39 %

- **The ability of the model to capture the variation of the maximum heart rate is less than optimal**

Regression vs. Decision Tree

- **When comparing regression results to decision tree results, I came up with similar attributes and variation explained by models (44.33% and 23.39%) respectively**

  - Risk Factors common between two analysis were: age, exercise induced angina and class of patients

  - Risk Factors different between two analysis were: blood sugar versus blood pressure, slope of the peak exercise versus oldpeak – exercise relative to rest and serum cholesterol being selected by regression, however not selected as an important attribute by the decision tree algorithm

- **In conclusion, the multiple regression model and the decision tree model predicted that healthy patients have higher maximum heart rate. However, neither model provided strong power in terms of finding patient characteristics that contribute to the prediction of the maximum heart rate as the variance explained by both models was less than optimal**

Implications for Future Research

- **Two prediction methods applied on the same data yield different models for predicting the maximum heart rate**

- **For the future, one possible rout would be to construct the decision tree model first whose structure can be used to determine the factors and interactions that should be considered for regression analysis. Perhaps, my decision would be to go with Forward Selection Method as it might give better accuracy of prediction by using all of the information contained in predictors that were excluded**

- **Conducting the analysis of this data by applying Incomplete Principal Component Regression Analysis and deleting variables of weak relationship to the maximum heart rate. The idea behind this is to create a new set of uncorrelated variables and show their relation to the original variables.**

References

- “Answer Tree 3.0 User’s Guide” by SPSS Inc. 2001
- “Maximum Heart Rate” in “New York Times” 2001 www.bodyresults.com/E2maxheartrate.asp
- “SAS System for Regression” by Rudolf J. Freund and Ramon C. Littell, SAS Institute Inc., Third Edition, 2000
- “SPSS 12.0 Statistical Procedures Companion” by Marija J. Norusis, Prentice Hall, Inc., 2003