PREDICTING HEART DISEASE

Bailey Kui, Christy Hui, Shirley Tang

Statistics 101C | Professor Almohalwas | Lecture 2

OVERVIEW

1

Introduction

Overview of project context and dataset

2

Exploratory Data Analysis

Categorical and numerical variable exploration

3

Data Cleaning

Recoding and imputing data

4

Models

Analysis of attempted models

5

Discussion & Conclusion

Discussion, suggestions, and concluding remarks

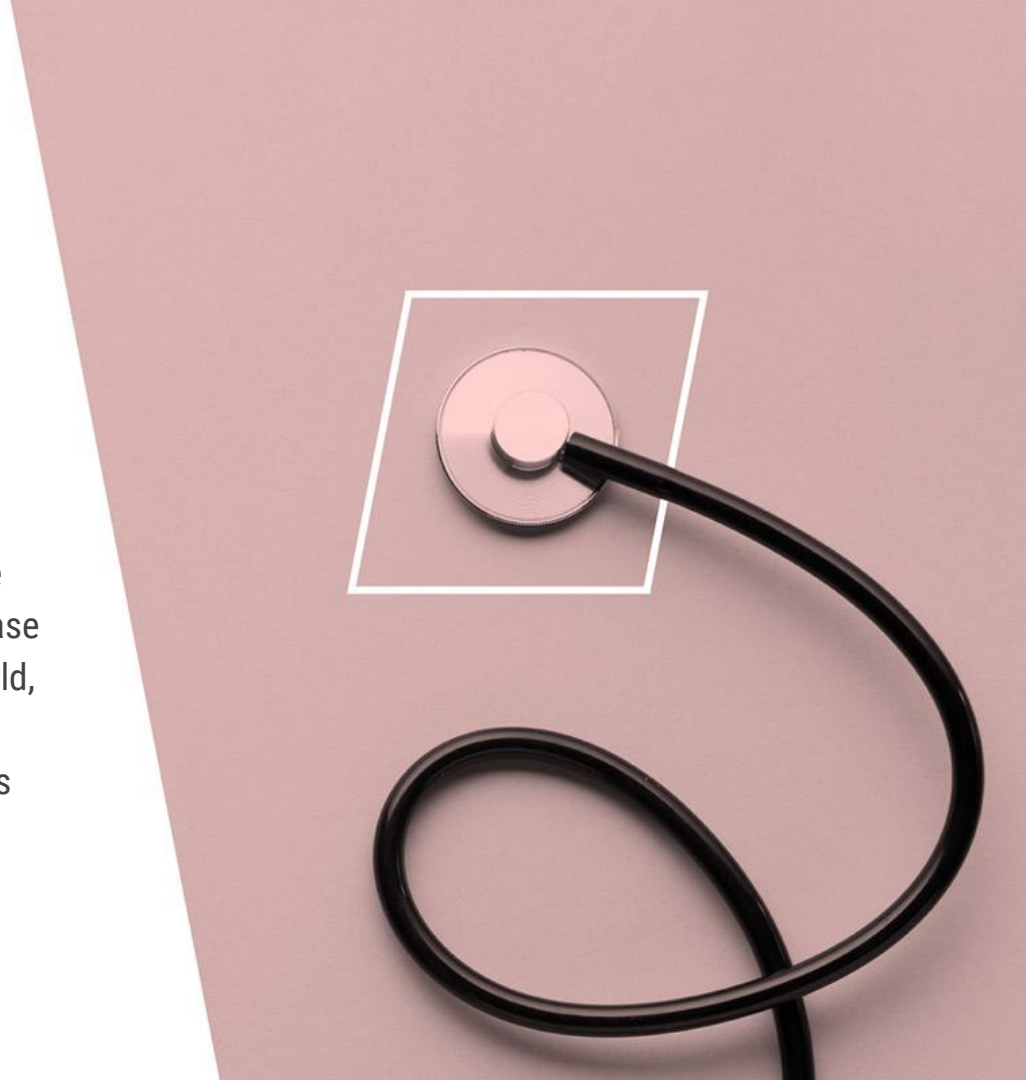
1

INTRODUCTION

What are we doing? Why are we interested?

INTRODUCTION

Heart Disease is the leading cause of death in the US (1 in 4 deaths). An early diagnosis of the disease is paramount in preventing deaths around the world, and if we are able to use data to reliably predict if someone has heart disease and treat it before it is too late, we could save millions.



Project Map

Explore

Get a sense of the “Heart Disease” data set.

Clean

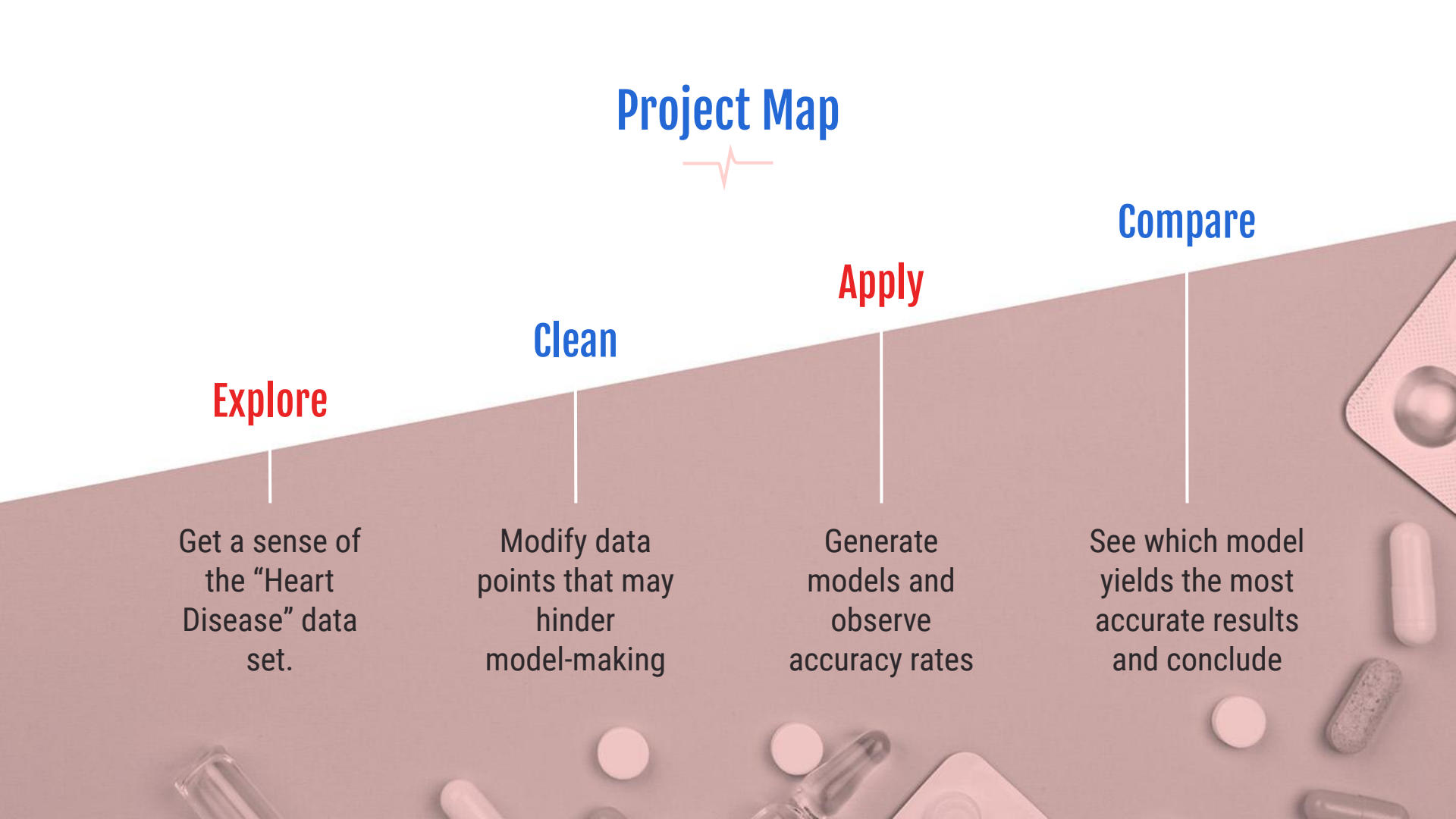
Modify data points that may hinder model-making

Apply

Generate models and observe accuracy rates

Compare

See which model yields the most accurate results and conclude



2

EXPLORATORY DATA ANALYSIS

What does our data look like?

BACKGROUND

19

Total predictor variables

12 Categorical

7 Numerical

4220

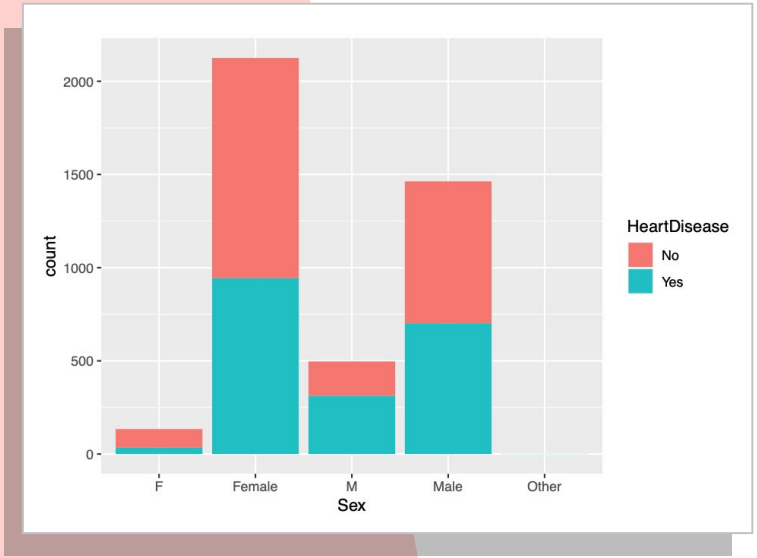
Total training observations

1808

Total testing observations

Our goal was to predict if a patient has heart disease (Yes or No) through comparative analysis of different classifiers.

EXPLORATORY DATA ANALYSIS – CATEGORICAL

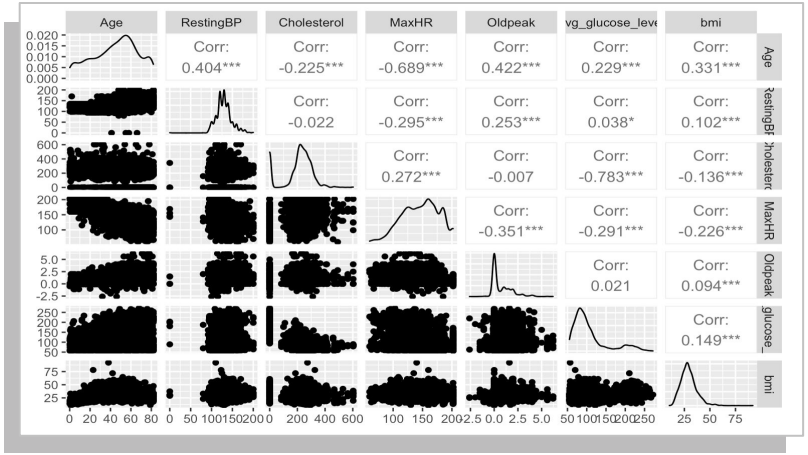


- Stacked bar charts of each categorical variable based on Heart Disease
- This helped us determine whether we needed to clean and recode variables and their levels (e.g. *Sex*)
- Or, if recoding could potentially help with the classification

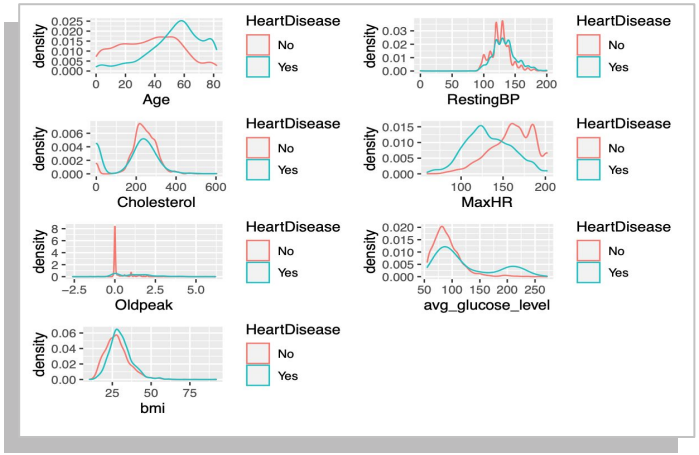
EXPLORATORY DATA ANALYSIS – NUMERICAL

Patterns in density plots

Some variables showed a good pattern in their density plots (e.g. *MaxHR*), which suggested that they may be useful in predicting heart disease. It was also good to check if any of the numerical predictors were highly correlated, to avoid multicollinearity.



Matrix plot



Density plots

3

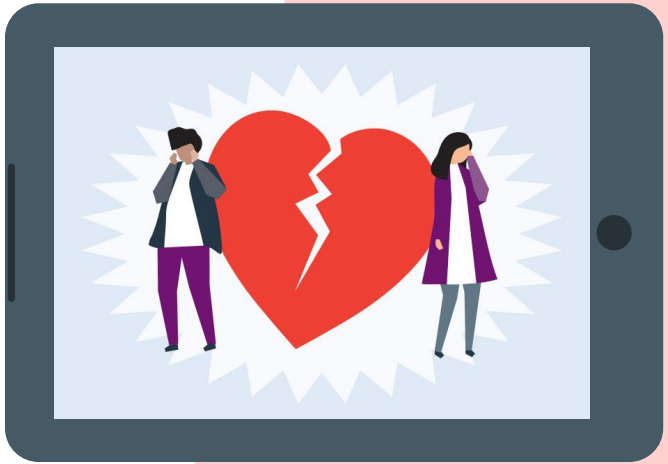
DATA CLEANING

How did we deal with hurtful data observations?

THE DATA VALUE PROBLEM – MISSING VALUES

Most of the time, data is incomplete. People may choose to refrain to answer or they may not have an answer, leading to some missing values in our data.

We fill in those missing values with the average of existing observations (if the data column is numeric) or with values that appear the most (if the data column is categorical).



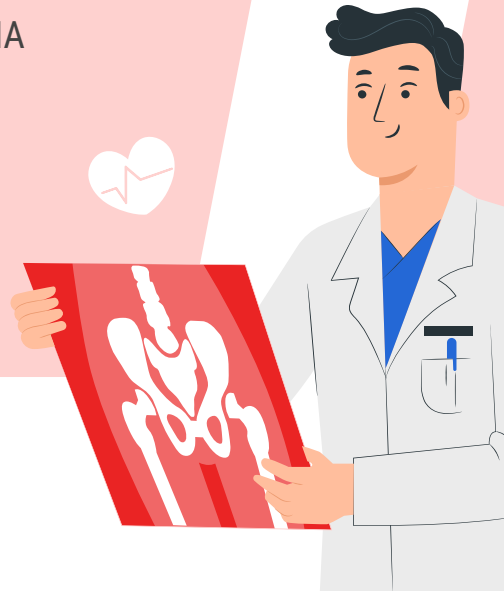
THE DATA VALUE PROBLEM – INCONSISTENT RESULTS

- Furthermore, data can be inconsistent.
- For example, in this data set, under *Sex*, we see multiple answers: “Female,” “Male,” “F,” “M,” and “Other”, where “F” is the same thing as “Female” and “M” is the same thing as “Male”.
- We want to make data observations consistent and easier to read so the computer can interpret them as the same.
- Thus, we should change these variables to be more consistent.

DEALING WITH INCOMPLETE RESULTS - MISSING VALUES

We dealt with missing values by imputing them.

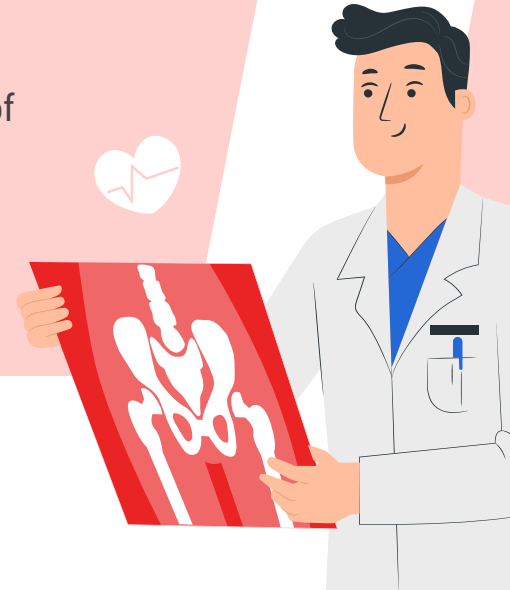
- There were 2524 NA values in the training data and 1148 NA values in the testing data
- Only categorical variables had NA values
  - Variables with NAs include:
    - *Ever Married, Work Type, Residence Type*



DEALING WITH INCONSISTENT RESULTS – NONSENSICAL VALUES

We changed values that do not make sense.

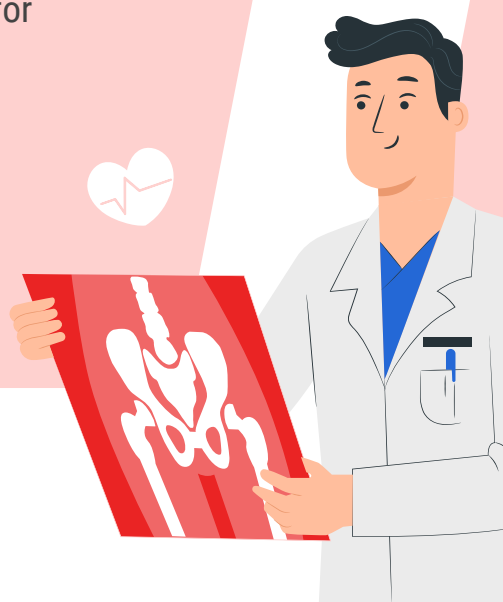
- *RestingBP* and *Cholesterol* had values of 0
  - Altered the values of 0s to the respective averages of said columns
- We later found that these values seemed to be intentional



DEALING WITH INCONSISTENT RESULTS – REPEATED LEVELS

We changed values that meant the same thing.

- The Sex variable had 5 unique levels when it was intended for 3: “Female,” “Male,” and “Other.”
  - Altered the values of “F” and “M” to “Female” and “Male,” respectively
- The Smoking variable had NAs when it had an “Unknown” value
  - Altered the NAs to “Unknown”



4

MODELS

What kind of models did we apply on our data?

HOW DO WE CHOOSE OUR PREDICTORS?

- We determined which predictors were significant based on the summary of the full **Logistic Regression**.

- We also used **regsubsets** (forwards & backwards stepwise regression) with the cleaned data and found that the only variable choice difference between the GLM and regsubsets using the cleaned data was:

  ➔ GLM has *SexMale* as a significant predictor
  ➔ Regsubsets has *ever_marriedYes* as a significant predictor

Using non-cleaned data:

- *Cholesterol, FastingBSYes, MaxHR, Oldpeak, avg_glucose_level, strokeYes* were significant.

Using cleaned data:

- *SexMale, ChestPainTypeNAP, Cholesterol, FastingBSYes, MaxHR, ExerciseAnginaY, Oldpeak, ST_SlopeUp, avg_glucose_level, strokeYes* were significant.

Significant predictors from Logistic Regression models

OVERVIEW OF RANDOM FOREST

Basic Idea

Construct a multitude of decision trees and average them, trying to reduce variance and prevent overfitting. Random Forest usually works best when we use a random subset of the predictors. In particular, the square root of the total number of predictors works best.

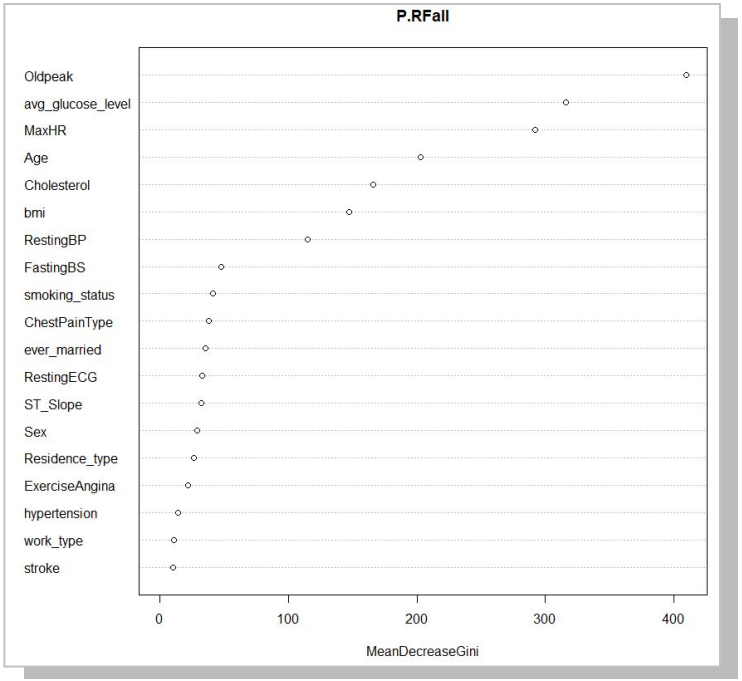
Application

We will apply a Random Forest algorithm, with a parameter of 4 random predictors. We choose 4 because the total number of predictors is 19. The square root of 19 is around 4.

RANDOM FOREST MODEL

We constructed a random forest based on cleaned data, selecting the best 4 predictors and using 100 trees.

This model gave us a public score of .81185.



Variable importance plot

OVERVIEW OF BEST MODEL – LOGISTIC REGRESSION

Summary

After initially fitting a full GLM model and exploring other methods, we found that Logistic Regression with 9 variables performed the best out of all of the models we tried, based on a public Kaggle score of .81264.

Approach

To optimize our model, we also tried Logistic Regression with ROCR, which gave us the same score of .81264.

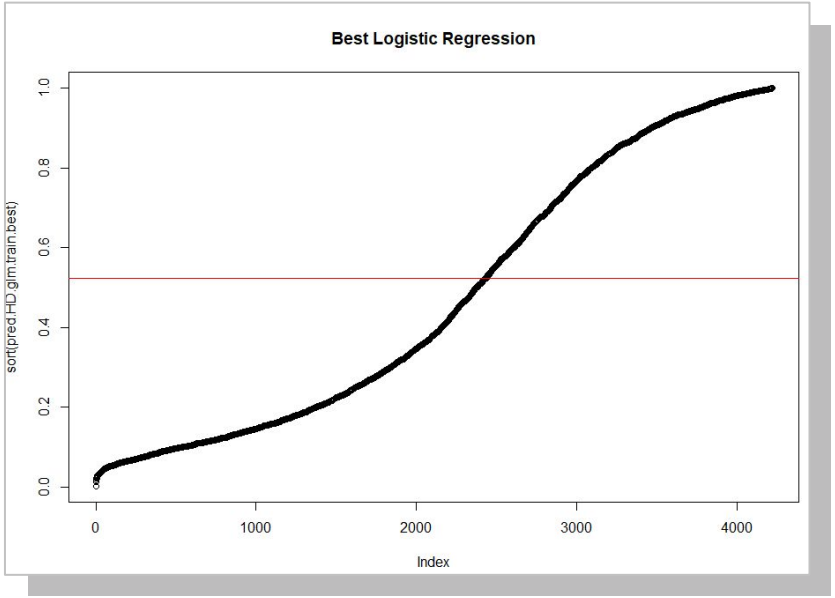
Another thing that we tried was to change the **tuning parameter** in classifying Yes vs. No.
A for-loop was used to find the best tuning parameter with the lowest error rate, and we tested values between .4xxx and .6xxx to classify Heart Disease.

BEST MODEL – LOGISTIC REGRESSION

Our best model consisted of 9 predictors:

- Sex (recoded)
- ChestPainType
- Cholesterol
- FastingBS
- MaxHR
- ExerciseAngina
- Oldpeak
- Avg_glucose_level
- stroke

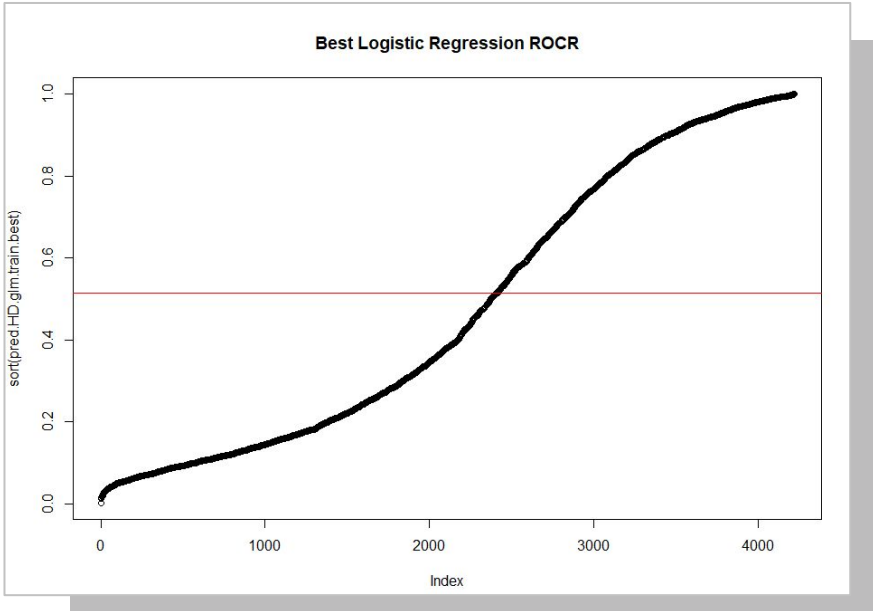
This model used the least amount of predictors and had a tuning parameter of >.524 for classifying “Yes”



BEST MODEL: LOGISTIC REGRESSION WITH ROCR

Our other best model (with ROCR) consisted of 11 predictors:

- Sex (recoded)
- Age
- ChestPainType
- Cholesterol
- FastingBS
- MaxHR
- ExerciseAngina
- Oldpeak
- ST_Slope
- Avg_glucose_level
- stroke



This model also had a tuning parameter of >.5134 for classifying “Yes”

COMPARISON OF BEST ACCURACY RATES

Model	Testing Accuracy
Logistic Regression	0.81264
Logistic Regression (ROCR)	0.81264
LDA	0.79189
QDA	0.79123
KNN	0.76126
Random Forest	0.81185

5

DISCUSSION & CONCLUSION

What questions should we ask? What do we conclude?

DISCUSSION

It would interesting to see if there was more data on the patients to see how significant the variables are.

How accurately can doctors predict heart disease compared to the accuracy of our modeling techniques?

What other techniques and approaches can we use to transform and improve our models?

SUGGESTIONS FOR FUTURE ANALYSIS

01 Categorize numerical variables

E.g: Splitting *Cholesterol* into different groups like low/normal/high

02 Try more variable transformations

Such as creating a polynomial model, using log transformations, etc.

03 Bring in external data

More biological/physiological data about the patient (e.g Ancestral history of Heart Disease, Patient ethnicity)

CONCLUSIONS

- Our modeling accuracy of ~81% is decent, but it can be improved upon
- Further research is needed to fully understand what causes Heart Disease and why
- Best overall predictors
  - *Sex, ChestPainType, Cholesterol, FastingBS, MaxHR, ExerciseAngina, Oldpeak, avg_glucose_level,* and *stroke*

REFERENCES

- https://www.cdc.gov/heartdisease/index.htm#:~:text=Heart%20disease%20is%20the%20leading,can%20lead%20to%20heart%20attack.
- Almohalwas' 101C Lecture
- R libraries used: class, leaps, MASS, randomForest, ROCR, corrplot, GGally

THANK YOU!

CREDITS: This presentation template was created by **Slidesgo**, incluiding icons by **Flaticon**, and infographics & images by **Freepik**

Please, keep this slide for attribution.