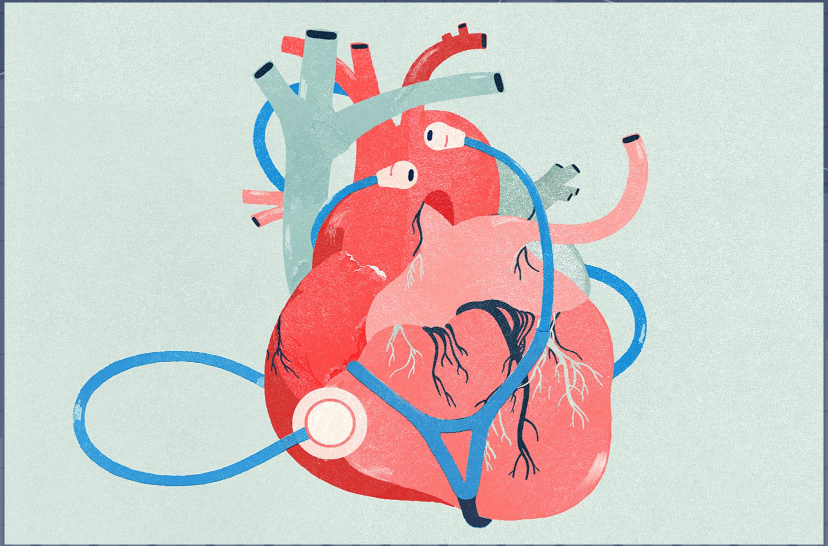
PREDICTING HEART DISEASE

By Tristan Dewing, Vivian Luk, Karina Santoso, and Brandon Wang (Lecture 1)



"The problem with heart disease is that the first symptom is often fatal." — Michael Phelps

THE PROBLEM

Heart disease is the #1 leading cause of death in America.
To prevent as many of these deaths as possible,
we have to detect it early.

That's where statistical modeling comes in.

THE DATA

4200 training observations

Each observation represents a single patient tested for heart disease.

We trained our models on these 4200 observations and then tested them against 1808 additional observations.

20 variables (not including Ob.)

Each patient was screened for demographic information and various risk factors, including age, sex, occupation type, chest pain, blood pressure, cholesterol, and smoking status.

The final target variable was the diagnosis of heart disease. (“Yes”/”No”)

GOAL

- **Train a classification model that can accurately predict whether or not a patient has heart disease based off of demographic information and various risk factors.**

METHODOLOGY

In order to build a model that could accurately predict whether or not a patient has heart disease, we went through a four-step process.

Here's how we did it.

2

OUR PROCESS

Data Cleaning

Step 1

Model Evaluation

Step 3

Step 2

Model Training

Step 4

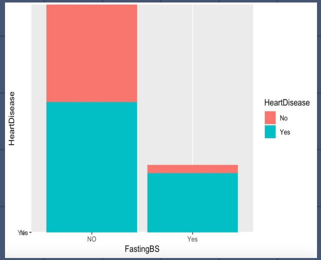
Model Comparison

DATA CLEANING: HANDLING NA'S

- With **2524** NA's in the training data and **1148** NA's in the test data, we had to either drop or impute them to ensure our models could successfully run
- We tried **dropping** all columns with NA's (only 4 out of 20 predictors had NA's) as well as **imputing** all NA's for numerical predictors with the mean or median and all NA's for categorical predictors with the mode
- **Ultimately, we imputed all NA's with the non-NA value that was closest to them in their respective column**

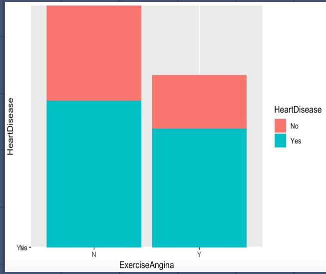
DATA CLEANING: FEATURE SELECTION

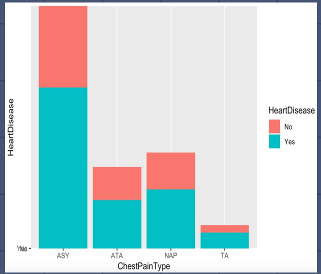
- We tried training our models using both **"full" models** in which we used all 20 predictors as well as **"reduced" models** where we used a subset of the best predictors to make the model simpler
- Barplots and density plots showed that categorical predictors such as **Fasting Blood Sugar**, **Exercise Angina**, and **Chest Pain** Type are generally better at separating the categories of the response variable than numerical predictors
- **Ultimately, the models scored the highest when we used ALL predictors from the dataset**



"FastingBS"

"ExerciseAngina"





"ChestPainType"

DATA CLEANING: LABEL ENCODING

- Some models we trained only accept numerical predictors and thus do not accept categorical variables as strings
- This requires them to be **encoded** as integers so they can be treated as numerical predictors
- As a result, **we encoded ALL 12 categorical variables as integers** so that our models could properly train and run

Sex (**text**)
Female
Male
Other

⇨

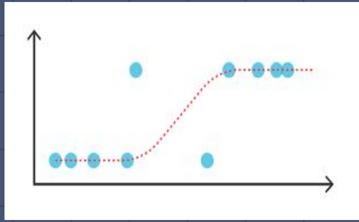
Sex (**numerical**)
1
2
3

MODEL TRAINING

- Once we cleaned our data, we trained an assortment of supervised **classification models** on the data, including:

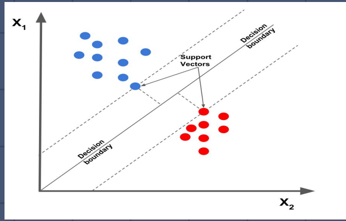
Logistic Regression

Calculates class probabilities of a binary response variable using the logistic function



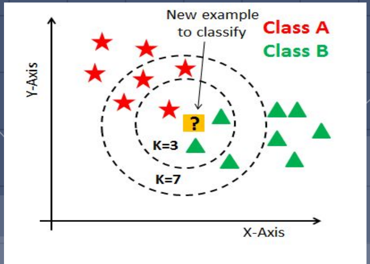
SVM

Finds the decision boundary that best separates classes of the response variable



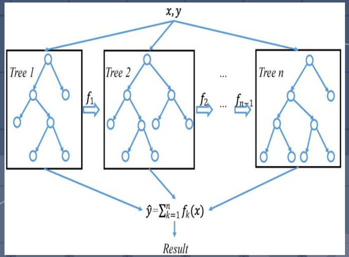
KNN

Classifies observations based on distance from other observations with known classes



XGBoost

Builds gradient boosted trees one at a time, allowing for new trees to use results of old trees to classify observations



MODEL EVALUATION

- To evaluate our models, we created **confusion matrices** and computed **correct classification rates** for our training data
- We used **k-fold cross validation** to determine the best hyperparameters and used different methods for handling NA values to optimize the performance of each model
- Models that achieved at least 80% training accuracy were submitted to Kaggle

	No	Yes
No	✓	✗
Yes	✗	✓

MODEL COMPARISON

- In choosing our final model, we ranked each type of model by their **best Kaggle accuracy score** and their **simplicity/interpretability**. Overall score would be based on averaging the rankings of accuracy and simplicity, **with accuracy taking precedence**

- We tried other types of models, but for now we will discuss our 4 most successful models: logistic regression, support vector machine (SVM), K-nearest neighbors (KNN), and XGBoost

Curious which model performs the best? Let's find out!
We will be using this chart to make compare the models!

	Logistic	SVM	KNN	XGBoost
Accuracy	?	?	?	?
Simplicity	?	?	?	?
Overall	?	?	?	?

RESULTS AND DISCUSSION

Here are the results of our modeling process!

The outcome may surprise you....

3

LOGISTIC REGRESSION

- Training Classification Rate: **0.8135071**
- Testing Classification Rate (Kaggle):
  - Public: **0.81422**
  - Private: **0.79373**
  - Overall: **0.808073**
- Logistic regression was the first model we used, which was successful even with its simplicity. However, we wanted to see first if other models with more hyperparameters could perform better before declaring this as the winner.

	No	Yes
No	1909	467
Yes	320	1524

```
Call:  glm(formula = as.factor(HeartDisease) ~ ., family = "binomial",
    data = h_train)

Coefficients:
     (Intercept)              Ob             Sex             Age   ChestPainType
      -3.199e-01      -5.801e-05       1.103e-01      -3.794e-04      -2.227e-01
       RestingBP     Cholesterol       FastingBS      RestingECG           MaxHR
       2.577e-03       2.128e-03       6.308e-01       5.824e-02      -1.974e-02
  ExerciseAngina         Oldpeak        ST_Slope    hypertension    ever_married
       1.013e+00       1.342e+00      -5.027e-01       2.512e-01       5.682e-02
       work_type  Residence_type avg_glucose_level            bmi  smoking_status
       2.642e-02       4.737e-02       2.492e-02      -3.036e-05      -2.731e-03
          stroke
      -2.463e+00

Degrees of Freedom: 4219 Total (i.e. Null);  4199 Residual
Null Deviance:       5837
Residual Deviance: 3627        AIC: 3669
```

K-NEAREST NEIGHBORS (KNN)

- Training Classification Rate: **0.8028436**
- Testing Classification Rate (Kaggle):
  - Public: **0.78893**
  - Private: **0.77716**
  - Overall: **0.785399**
- The next model we tried was a KNN model using all 7 numerical predictors in the dataset and hyperparameter k = 25. However, one limitation to this model is that it can only take into account numerical predictors, and none of the information our categorical predictors provide.

	No	Yes
No	1941	544
Yes	288	1447

SUPPORT VECTOR MACHINE (SVM)

- Training Classification Rate: **0.8421801**

- Testing Classification Rate (Kaggle):
  - Public: **0.80316**
  - Private: **0.79005**
  - Overall: **0.799227**

- While SVM performed relatively well, it had a tendency to overfit as shown by the higher training accuracy compared to the testing accuracy. The more we increased the value of the hyperparameter gamma, the more the model overfit the test data.

	No	Yes
No	2030	467
Yes	199	1524

XGBOOST

- Training Classification Rate: **0.9388626**

- Testing Classification Rate (Kaggle):
  - Public: **0.81343**
  - Private: **0.79373**
  - Overall: **0.79964**

	No	Yes
No	2103	132
Yes	126	1859

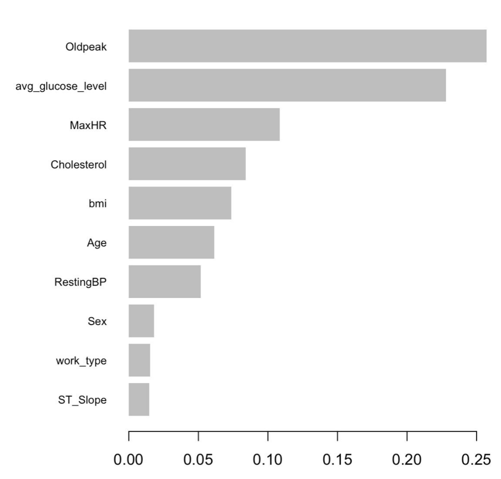
- After transforming all the predictors to be numeric, we also tried an XGBoost model with max depth = 1000, eta = 0.3, nthread = 2, nrounds = 25, and objective = "binary:logistic". Although this model had a high training classification rate, it did not classify as accurately with the test data.

XGBOOST WITH FEATURE SELECTION

- Training Classification Rate: **0.8547393**
- Testing Classification Rate (Kaggle):
  - Public: **0.79367**
  - Private: **0.78268**
  - Overall: **0.785977**

	No	Yes
No	2046	430
Yes	183	1561

- After analyzing the importance matrix, of the XGBoost model with all predictors, we did feature selection and made a model with the 7 most important predictors: **Old Peak, Average Glucose Level, Max Heart Rate, Cholesterol, Body Mass Index, Age, and Resting Blood Pressure**. Unfortunately, this did not result in a better classification rate.



FINAL MODEL

- The model that produced the best results and was the simplest was actually our first **logistic regression model using all predictors**

- According to the summary of the model, the most significant predictors included **Chest Pain Type**, **Cholesterol**, **Fasting Blood Sugar**, **Max Heart Rate**, **Exercise Angina**, **Old Peak**, **ST Slope**, **Average Glucose Level**, and **Stroke**, though ALL predictors were needed in the model to achieve the best accuracy

	Logistic	SVM	KNN	XGBoost
Accuracy	1	3	4	2
Simplicity	1	3	2	4
Overall	1	3	4	2

```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.199e-01  7.056e-01  -0.453 0.650342
Ob               -5.801e-05  3.434e-05  -1.689 0.091173 .
Sex               1.103e-01  4.330e-02   2.546 0.010888 *
Age              -3.794e-04  3.475e-03  -0.109 0.913062
ChestPainType    -2.227e-01  7.200e-02  -3.093 0.001983 **
RestingBP         2.577e-03  2.529e-03   1.019 0.308128
Cholesterol       2.128e-03  6.349e-04   3.352 0.000804 ***
FastingBS         6.308e-01  1.790e-01   3.524 0.000425 ***
RestingECG        5.824e-02  8.956e-02   0.650 0.515514
MaxHR            -1.974e-02  1.974e-03 -10.002  < 2e-16 ***
ExerciseAngina    1.013e+00  1.409e-01   7.189 6.54e-13 ***
Oldpeak           1.342e+00  6.098e-02  22.002  < 2e-16 ***
ST_Slope         -5.027e-01  1.152e-01  -4.363 1.28e-05 ***
hypertension      2.512e-01  2.229e-01   1.127 0.259761
ever_married      5.682e-02  1.126e-01   0.505 0.613868
work_type         2.642e-02  3.970e-02   0.666 0.505662
Residence_type    4.737e-02  8.317e-02   0.570 0.568975
avg_glucose_level 2.492e-02  1.699e-03  14.669  < 2e-16 ***
bmi              -3.036e-05  5.896e-03  -0.005 0.995892
smoking_status   -2.731e-03  4.190e-02  -0.065 0.948031
stroke           -2.463e+00  2.693e-01  -9.144  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

LIMITATIONS AND CONCLUSIONS

Though our team ranked pretty high on the scoreboard, there will always be ways to improve our model and process.

Here are some issues possibly limiting our model's success.

4

LIMITATIONS

DATA IMPUTATION

Four of our predictor variables that were ultimately used in the final model had NA values for around 15% of its entries. As mentioned previously, we dealt with this issue by imputing all NAs with their closest values, but there are many different methods that could have been used to handle these entries.

VARIABLE SELECTION

From the density plots and bar charts we produced in our exploratory data analysis, we definitely saw that some of the predictor variables seemed more significant than others. However, our best model incorporated all predictor variables in it, and taking any predictors out decreased our accuracy score.

ASSUMPTIONS

One assumption of logistic regression is that there are no extreme outliers. In our data preparation process, we did not remove any outliers. Another assumption is that there is no multicollinearity between explanatory variables. Since we used all predictors in our final model, some correlation between our predictors may exist.

CONCLUSIONS

- After trying a few different approaches, our best performing model was a logistic regression model using all predictor variables in the dataset. The model also happened to be highly usable and interpretable due to its simplicity.
- Interestingly, trying to remove any predictor variables we thought to be less significant only decreased the accuracy of our model.
- As with most modeling approaches, there are some limitations to our process, but we were able to achieve a high accuracy rate with our final model.
- **Thank you for listening!**

REFERENCES

- Almohalwas, Akram. "Introduction to Statistical Models and Data Mining." Statistics 101C, Fall Quarter 2021, UCLA, Los Angeles. PowerPoint presentation.
- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2013.
- Centers for Disease Control and Prevention, National Center for Health Statistics. About Multiple Cause of Death, 1999–2019. CDC WONDER Online Database website. Atlanta, GA: Centers for Disease Control and Prevention; 2019.