Heart Disease

Muhammad Faraz Akram

Introduction

- Dataset

  Heart Diseases

- Data source

  CDC Official Website

- Reason

  To find which Age, Gender, Race is more at risk of getting Heart Disease

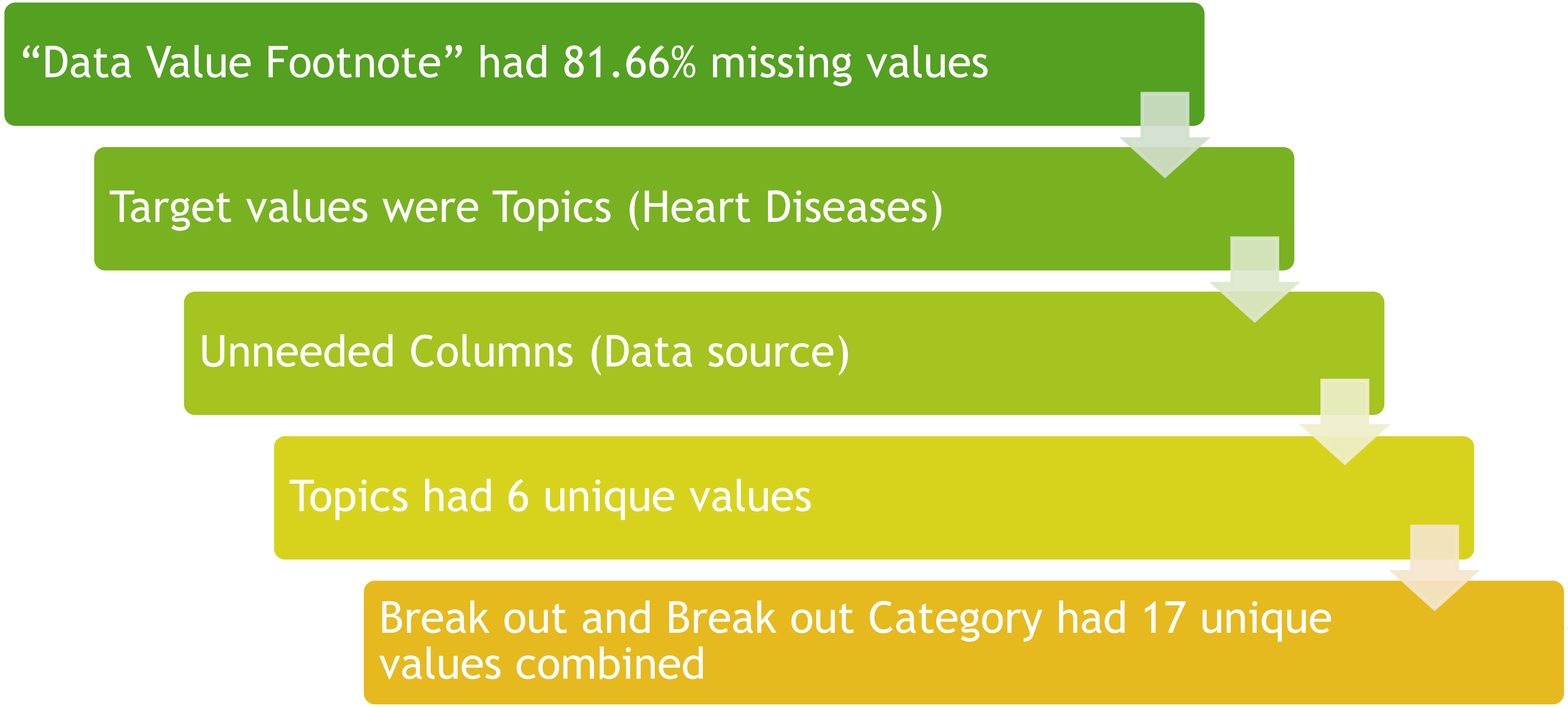
- Expectation

  To run classification models on the Dataset at hand

Cleaning of the Dataset

- Missing Values
- Target Values
- Unneeded Columns
- Plan of Action
- Unique Values
- Fixing Outliers

Cleaning Examples



“Data Value Footnote” had 81.66% missing values

Target values were Topics (Heart Diseases)

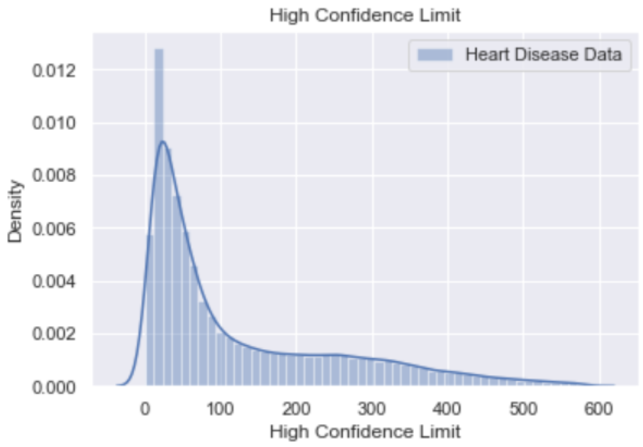
Unneeded Columns (Data source)

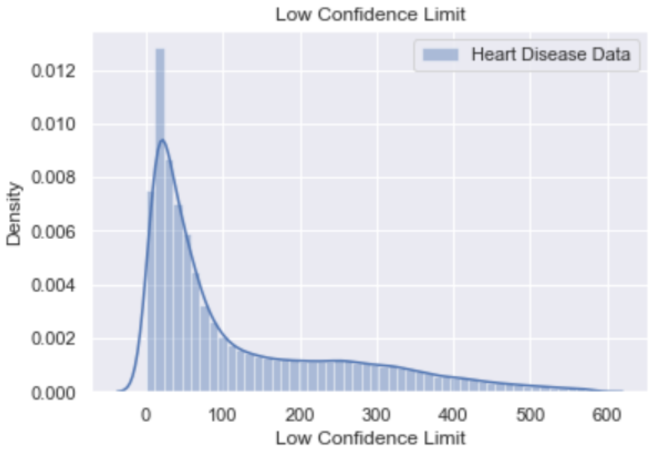
Topics had 6 unique values

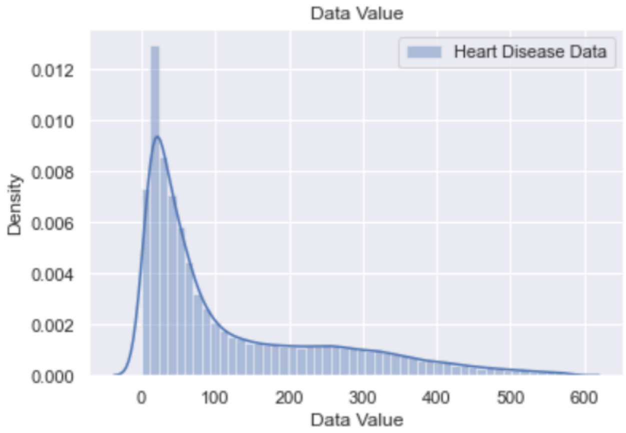
Break out and Break out Category had 17 unique values combined

Feature Engineering

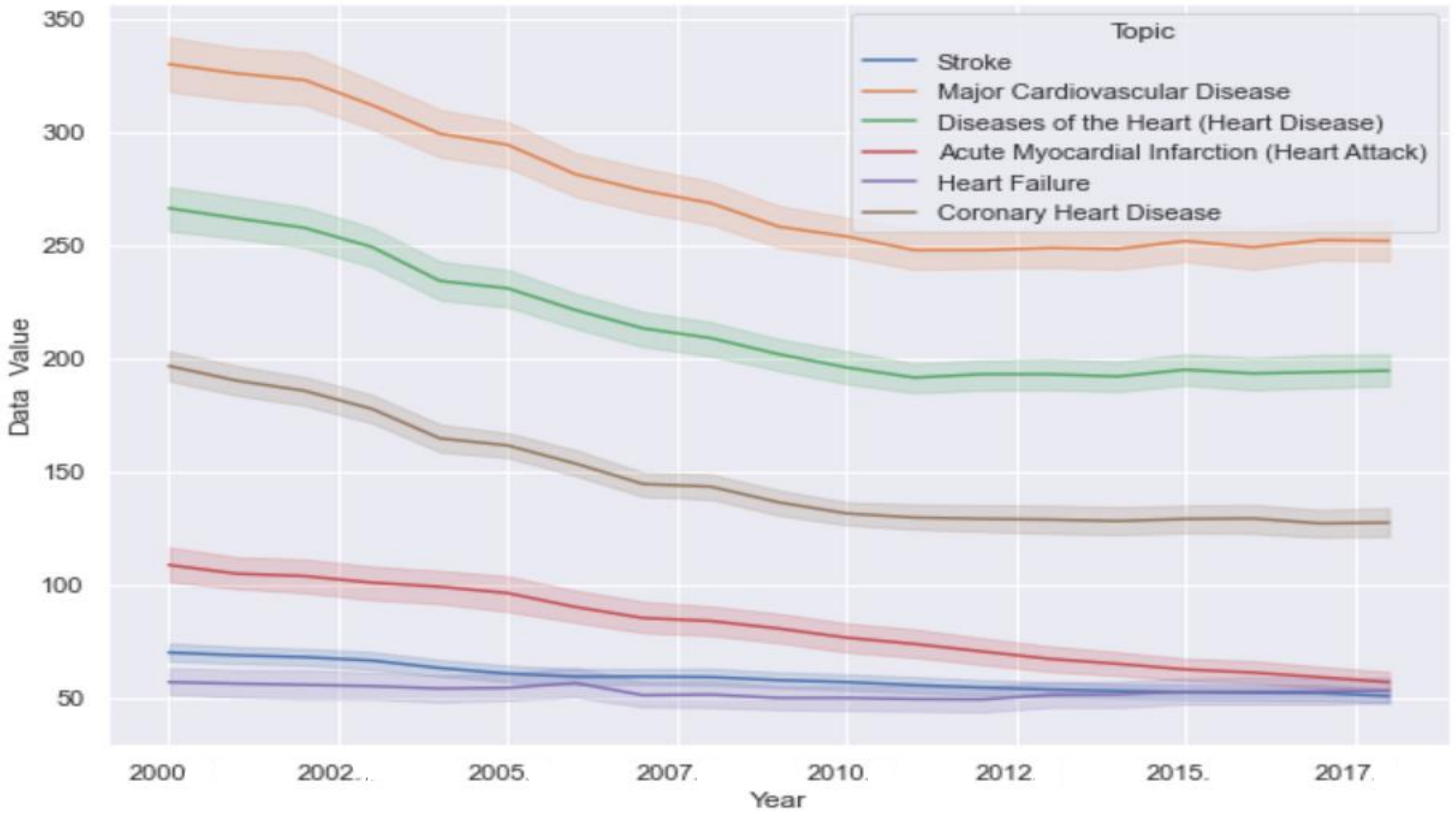
Diseases	Dummying up Target Variable (6 Types of Heart Conditions)
Categories	Dummying up Categories (Age, Race, Gender,..)
Binary Data	Converting Columns into Binary Data



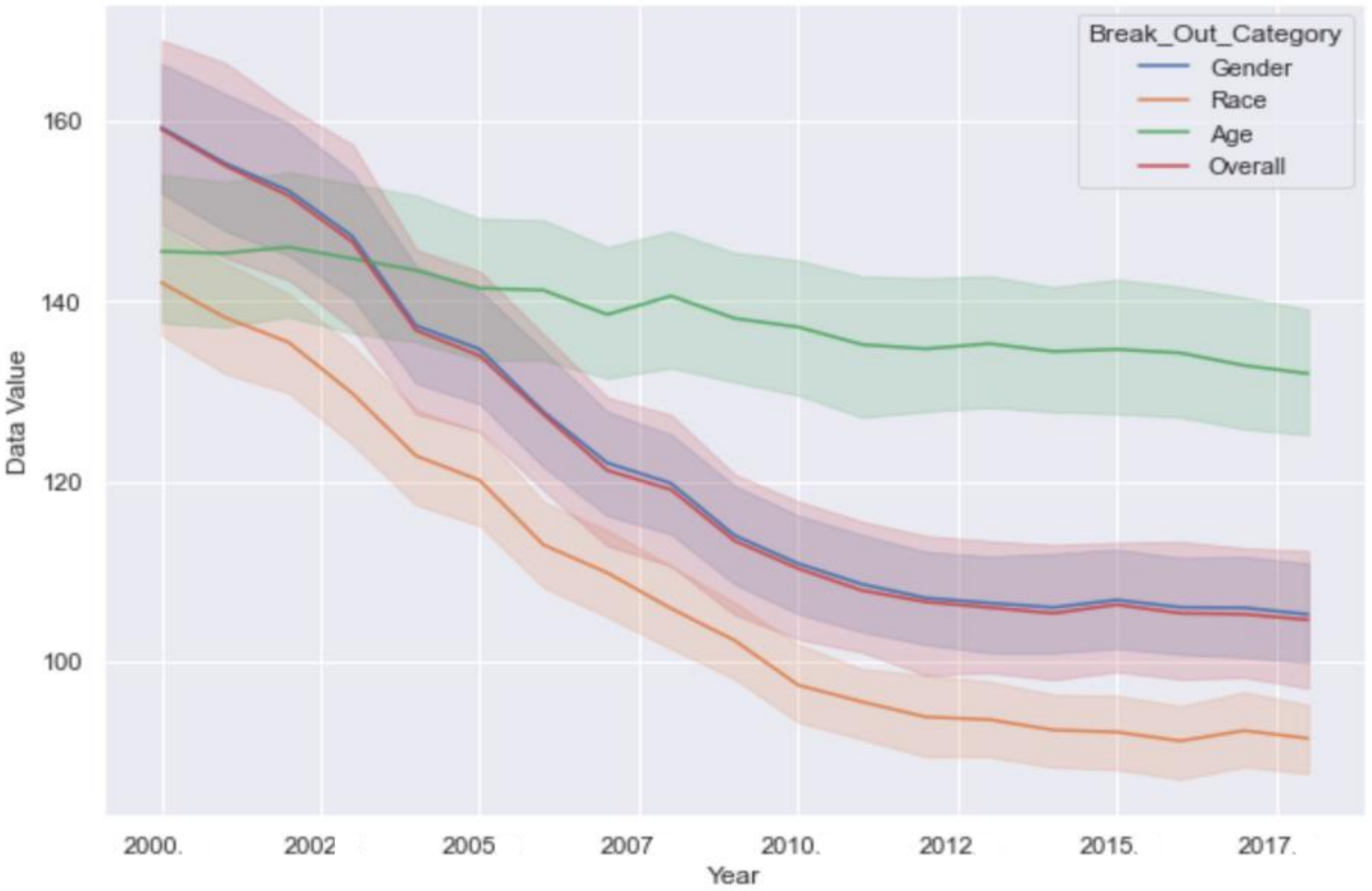




Distribution



Lineplot for Data Value and Years



Lineplot for Confidence Limits

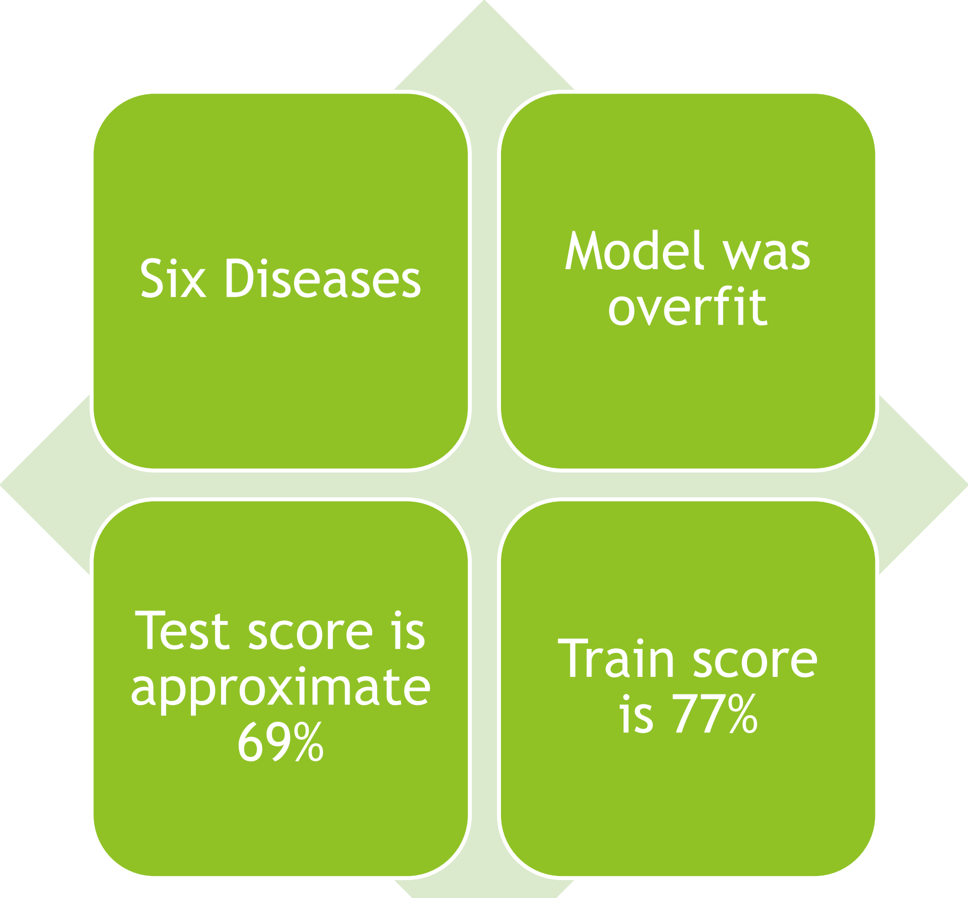
Types of Heart Condition

- Stroke
- Cardiovascular Disease
- Heart Attack
- Heart Failure
- Coronary Heart Disease
- Other Heart Diseases

Modeling

One Vs Rest Classification	Multiclassification Model used for more than two classes. (6 diseases)
Random Forest Classification	Using it to predict One Disease at a time (Binary Data)
Extra Tree Classification	Using it to predict One Disease at a time (Binary Data)

One Vs Rest Classification



Six Diseases

Model was overfit

Test score is approximate 69%

Train score is 77%

Baseline Accuracy = 63.2%

Disease	Train Score	Test Scores	Overfit
Stroke	86 %	81 %	Yes
Cardiovascular Disease	95 %	93 %	No
Coronary Heart Disease	95 %	94 %	No
Heart Attack	89 %	88 %	No
Heart Failure	92 %	91 %	No
Other Heart Disease	93 %	90 %	No

Baseline Accuracy = 63.2%

Classification Models for Diseases
Random Forest Classification

Classification Models for Diseases Extra Tree Classification

Diseases	Train Score	Test Score	Overfit
Stroke	84 %	80 %	Yes
Cardiovascular Disease	94 %	93 %	No
Coronary Heart Disease	94 %	93 %	No
Heart Attack	89 %	88 %	No
Heart Failure	90.5 %	90.2 %	No
Other Heart Diseases	91 %	89 %	No

Baseline Accuracy = 63.2%

Confusion Matrix

In this Confusion Matrix, we can better explain the outcome

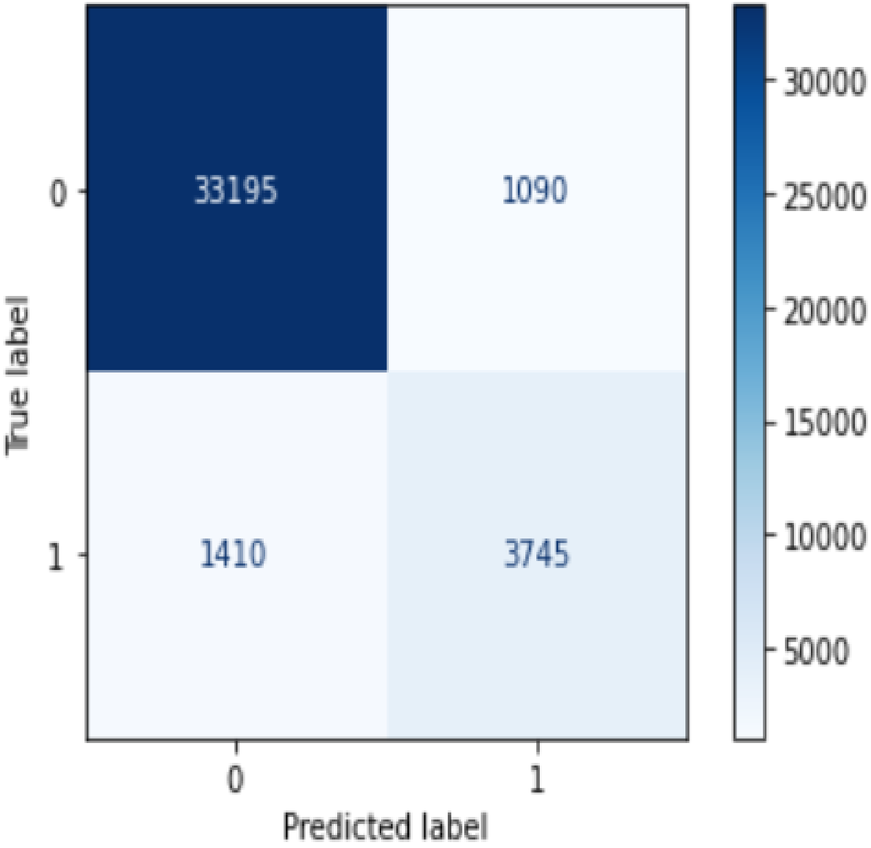
True negatives means that we correctly predicted the people who don't have Heart Disease

True positives means that we correctly predicted the people who have Heart Disease

False positives means that we incorrectly predicted the people who have Heart Disease(who don't have Heart Disease)

False negatives means that we incorrectly predicted the people who don't have Heart Disease(who have Heart Disease)

Confusion Matrix for Cardiovascular Disease



- Better able to predict True Negatives
- Will be helpful in specifying the people with no Heart Conditions
- Will be able to make better Healthcare plans for the members

Conclusion

Values related to Stoke 62%

One Vs Rest Classification was Overfit

Models worked well

Success in True Negatives

Thank You for Listening