Predicting Heart Disease

Shas [redacted]

Agenda

Introduction

Data Selection

Modeling and Methods Used

Interpretation of Analysis / Model Results

Conclusion

Introduction

- CDC reports that heart disease is one of the leading causes of death for people in the United States with one person dying from it every 34 seconds.
- From financial perspective, the average cost associated with this disease per year was about $229 billion between 2017 and 2019

Data Selection

Kaggle

Data Selection

- Resoure Link: Personal Key Indicators of Heart Disease | Kaggle
- Has 18 variables
- HeartDisease noted as Yes and No
- Collected by CDC in 2020 by telephonic survey
- Included 300 variables initially
  - Trimmed down to 17

Modeling and Methods Used

Visualizations

- Bar chart showing count of males and females having heart disease.
- Bar chart showing counts by races having heart disease.
- Bar chart showing counts by age group having heart disease.
- Bar chart showing count by general health having heart disease.

Data Preparation

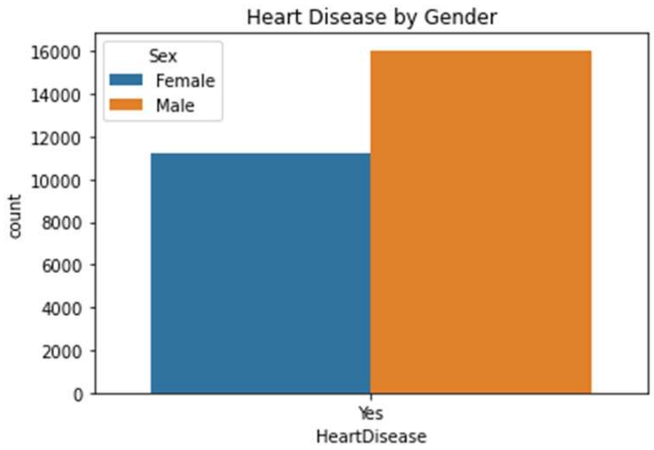
- Dummy variables were created for categorical variables
- Redundant variables removed after creation of dummy variables
- Checked for normalcy for race and gender categorical variables

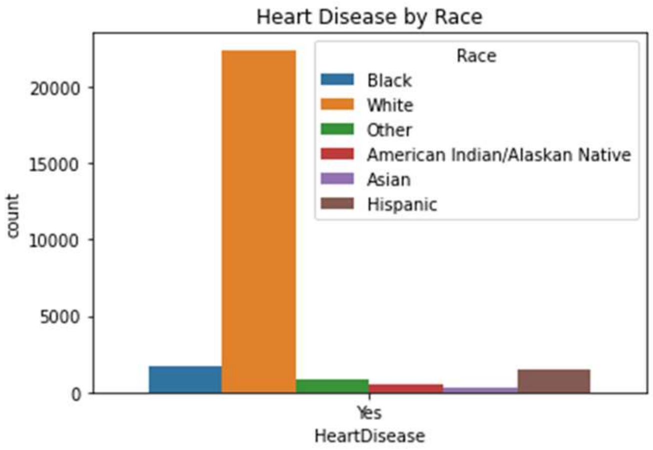
Modeling

- Target outcome is Yes or No / Binary
- Created two models:
  - Logistic Regression
  - Nearest Neighbor Algorithm

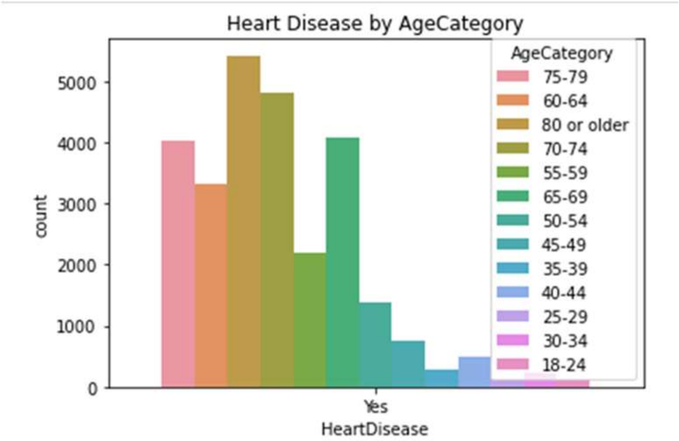
Interpretation of Analysis / Model Results

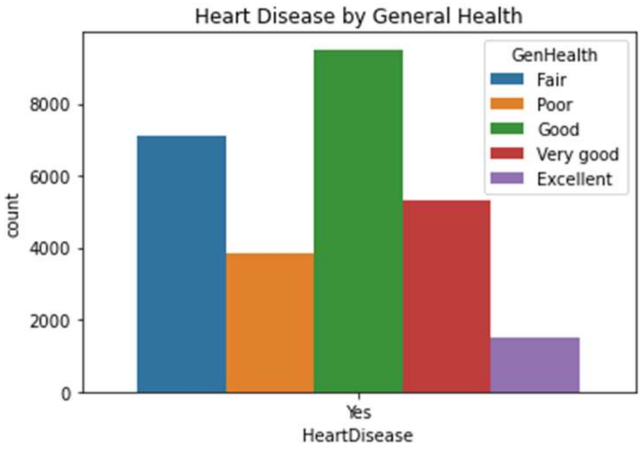
Visualizations





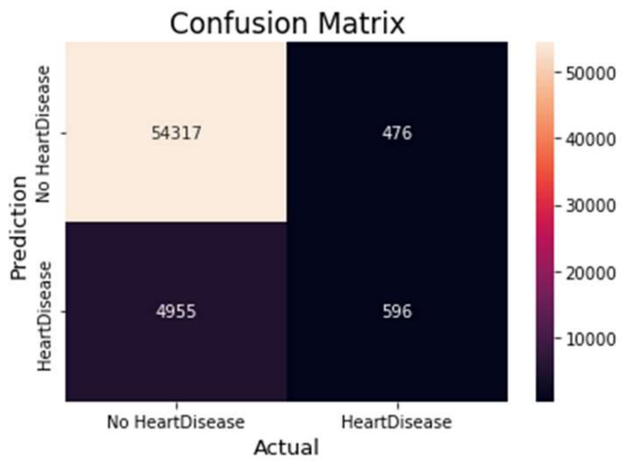
Visualizations Continued

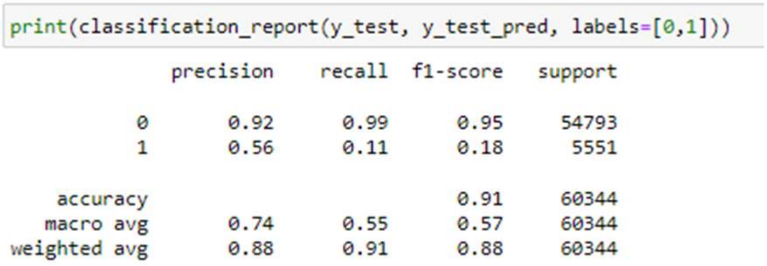




Model Result Interpretation

- Logistic Regression

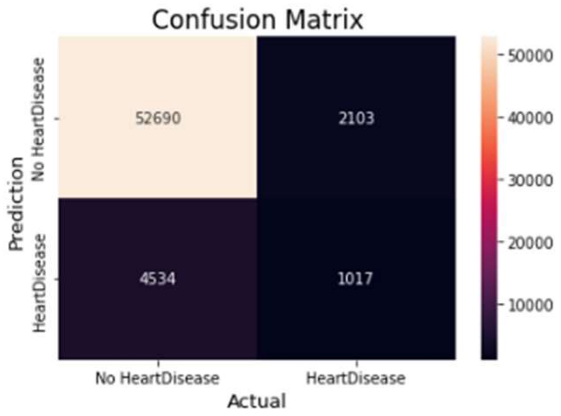


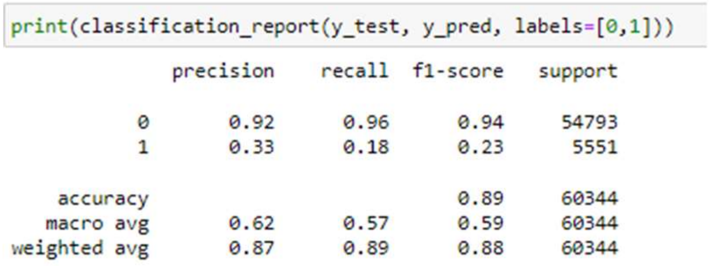


```
print(classification_report(y_test, y_test_pred, labels=[0,1]))
              precision    recall  f1-score   support

           0       0.92      0.99      0.95     54793
           1       0.56      0.11      0.18      5551

    accuracy                           0.91     60344
   macro avg       0.74      0.55      0.57     60344
weighted avg       0.88      0.91      0.88     60344
```

Model Result Interpretation - Continued

- Nearest Neighbor

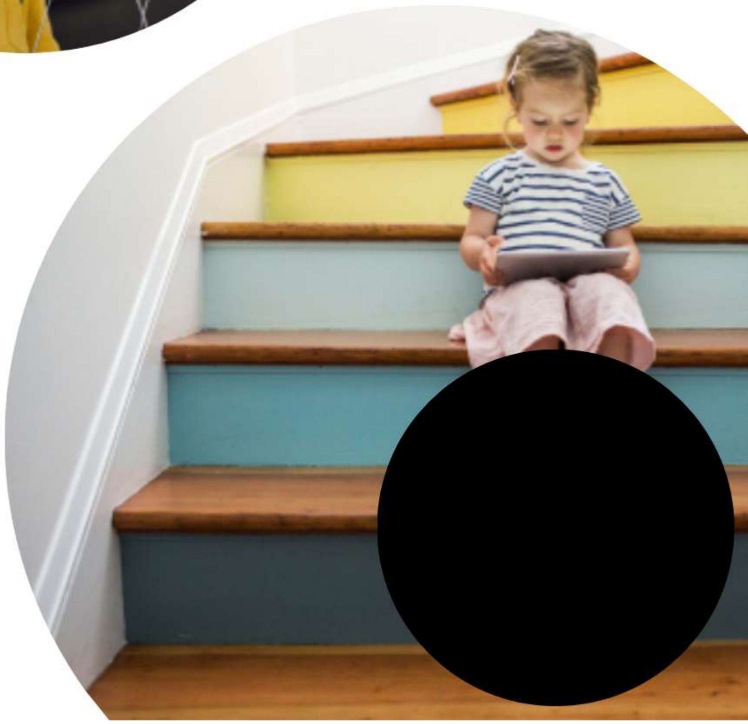




```
print(classification_report(y_test, y_pred, labels=[0,1]))
              precision    recall  f1-score   support

           0       0.92      0.96      0.94     54793
           1       0.33      0.18      0.23      5551

    accuracy                           0.89     60344
   macro avg       0.62      0.57      0.59     60344
weighted avg       0.87      0.89      0.88     60344
```

Conclusion

- Since the maximum number of people in the data set having heart disease were in good general health, a prediction model becomes important.
- Recommend Logistic Regression - Higher accuracy together with highest precision scores for both predicting “heart disease” and “no heart disease” between the two models created.
- Only slight overfitting observed.



Thank you

Shashi Bhushan