Minh Tran, Alex Hertadi,
Jack Kiishkila, and Olivia Plamann
Stat 451
Final Report

Model to Predict Heart Disease

I. Introduction

Heart disease is one of the leading causes of death for most people in the United States, and about half of Americans have at least one of three key risk factors for heart disease. Because of the prevalence of heart disease, we decided to build a model to predict heart disease based on some of the most common and influential predictors. This model could predict whether an individual has heart disease so they can receive the necessary treatment to reduce the burden of their condition.

II. Analysis

1. Data Preparation

We collected our source through kaggle.com. The data set has 18 variables, around 319,800 rows, and no missing values. For the 11 binary variables, we replaced their binary values with 0's and 1's. There were three multi-categorical variables: age, general health, and race. The Age variable was binned by 5-year bins. As the age variable is ordinal, we de-binned them using an average of each bin's bottom and top ages to retain its characteristics. For the Race variable, as it is nominal, we used one-hot encoding to handle it. The General Health variable has both ordinal and nominal characteristics and hence, is more complicated than the Age and Race. Scaling its values from 1 to 5 does not reflect well the relationship among five categorical values. For this reason, we used the weight of evidence (WOE) method to encode values of General Health.

After encoding all variables, we found no significant correlations (Fig 1. Correlation). Therefore, from the statistical point of view, we could keep all variables when modeling.

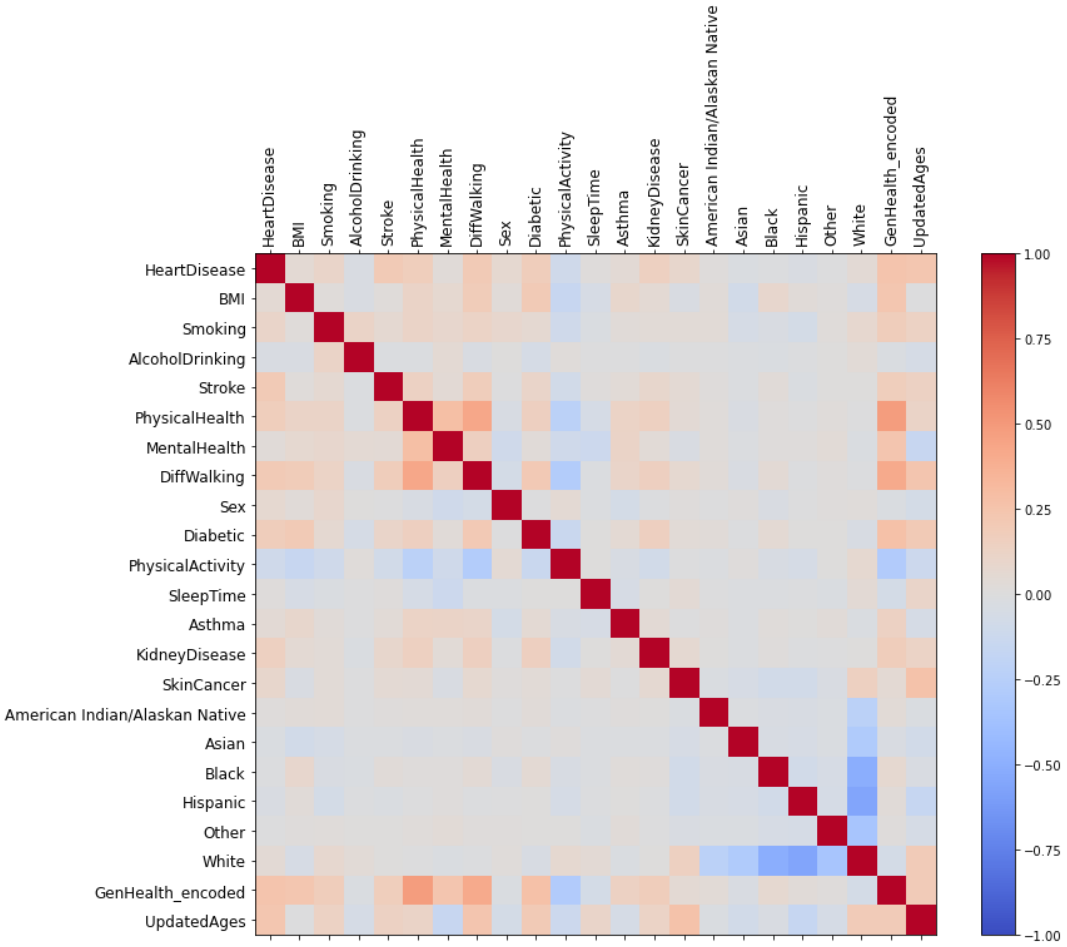


Fig 1. Correlation

2. Modeling Process

Using encoded variables and splitting them into 80% training, 10% validation, and 10% testing sets with stratify = y, we then tried multiple models and the grid search method to find the models and parameters that performed the best. Since the label variable and majority of the feature variables are categorical/binary variables, we tried classification models only.

Table 1 below shows the accuracy scores of each model with the best parameters we run on the validation data set after fitting it to the training set. In summary, Logistic Regression with C = 100 and max_tier = 5,000 produced the highest accuracy of 0.917.

We tested Decision Tree with max_depth = [1, 3, None], and it gave a pretty close accuracy to Logistic Regression, but its best parameter for max_depth was 1, which was not valuable in classifying features. Gradient Boosting also produced a good accuracy when we tested it on the four most important factors (from experts' view): Smoking, Diabetic, Age, and Sex. However, it fell if we ran this model on all variables due to the too large data. K-NN model faced the same

problem as Gradient Boosting, and when testing with 3 and 4 neighbors, k-NN's accuracy was worse than Gradient Boosting's. We also tested Random Forest with n_estimators = [5, 10, 100]. Its best estimator parameter of 100 gave an accuracy of only 0.905. Other stacking models introduced worse accuracies than running each model independently. From this finding, we selected the logistic regression model as the best model.

Rank	Models	Accuracy Score (Validation Data)
1	Logistic Regression	0.917
2	Decision Tree Classifier	0.914
3	Gradient Boosting Classifier	0.914
4	k-NN Classifier	0.910
5	Random Forest Classifier	0.905
6	Stacking: 1st Layer Estimator = Logistic Regression Final Layer Estimator = Random Forest Classifier	0.873

Table 1. Models' Accuracy Summary

The coefficients obtained from the logistic regression model are as below. Stroke, Sex, Race, GenHealth, and KidneyDisease appear to be the most critical factors.

	Variables	Coeficients
0	BMI	0.008917
1	Smoking	0.348268
2	AlcoholDrinking	-0.240519
3	Stroke	1.044154
4	PhysicalHealth	0.002061
5	MentalHealth	0.004218
6	DiffWalking	0.238808
7	Sex	0.715309
8	Diabetic	0.469355
9	PhysicalActivity	0.020429
10	SleepTime	-0.028685
11	Asthma	0.285002
12	KidneyDisease	0.573590
13	SkinCancer	0.117036
14	American Indian/Alaskan Native	-0.609911
15	Asian	-1.054690
16	Black	-0.906037
17	Hispanic	-0.797958
18	Other	-0.592686
19	White	-0.632704
20	GenHealth_encoded	0.629555
21	UpdatedAges	0.055546

3. Model Evaluation

We used the model's ROC curve and the false-negative rate to evaluate the model result. Minimizing the false-negative rate is critical because a false-negative result means missing out on

a positive diagnosis of heart disease. This ignored diagnosis would leave the individual to forgo care that could help to reduce the burden of their condition.

Using the validation data set, we started with a model that used only the first two columns of the data (BMI and Smoking) and got a false-negative rate of 1.0 (Fig 2). We added two more variables (Stroke and AlcoholDrinking) and improved the false-negative rate to 0.989 (Fig 3). Finally, we used all variables (Fig 4) and got the best ROC with the lowest false-negative of 0.885 among the three tries. However, this false-negative rate is still high, and in future models we hope to decrease this.

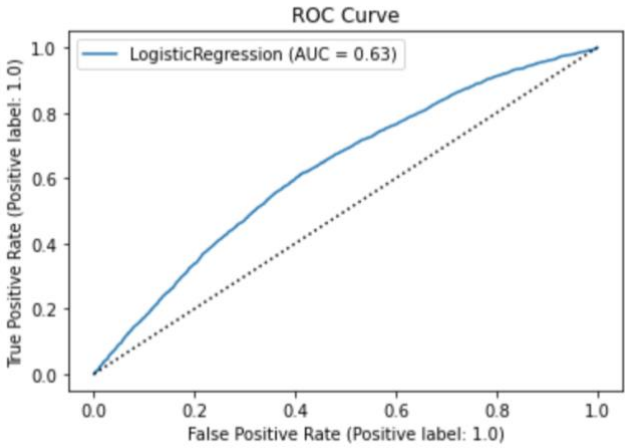


Fig 2. ROC Curve with Two Features

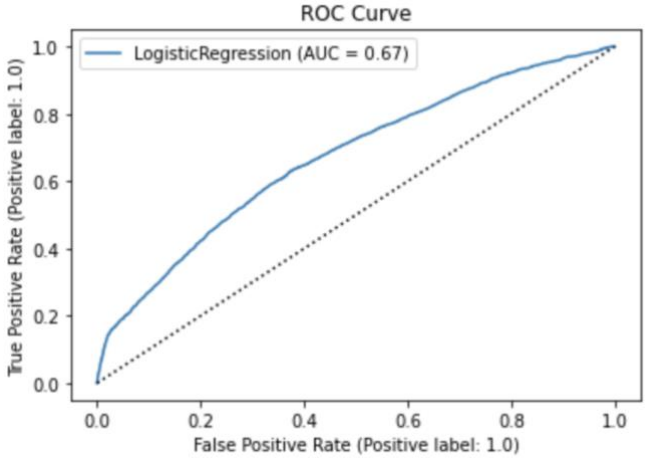
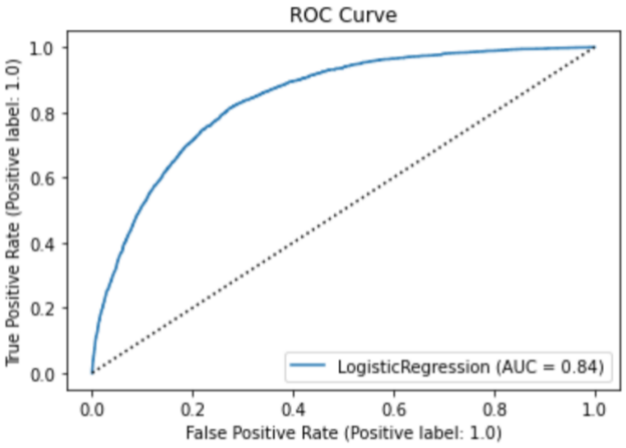


Fig 3. ROC Curve with Four Features

Fig 4. ROC Curve with All Features



After training the model on the training data, we achieved a score of 0.915 on the testing data set.

```
model = linear_model.LogisticRegression(max_iter = 5000, C = 100)
model.fit(X_train, y_train)
print(f'Validation data score: {model.score(X_valid, y_valid):.3}')
print(f'Test data score      : {model.score(X_test, y_test):.3}')
```

```
Validation data score: 0.916
Test data score      : 0.915
```

The confusion matrix associated with our chosen model on the testing data set is as follows:

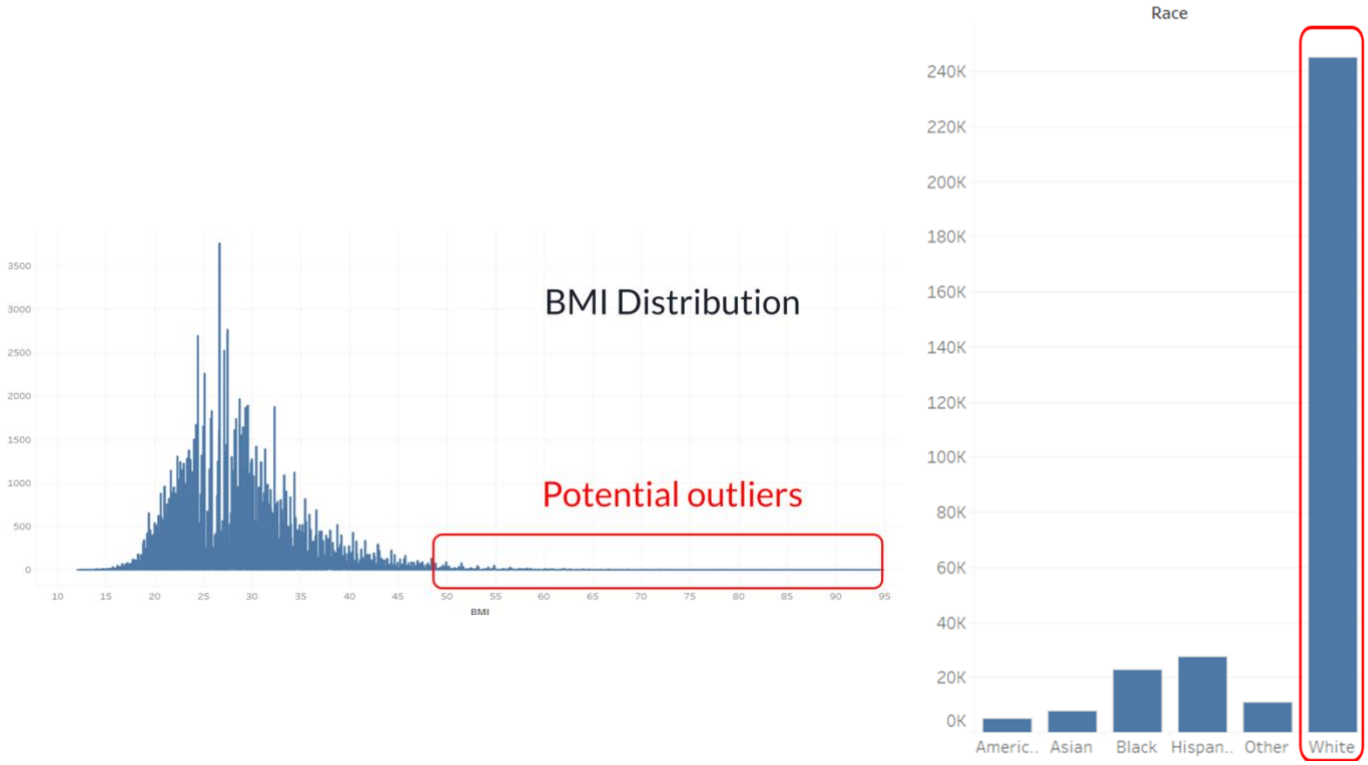
```
df:
        0    1
0   28974  268
1    2446  292

TN: 28974, FP: 268, FN: 2446, TP: 292

TPR=0.107, FPR=0.00916, FNR=0.893
```

While the overall score is relatively high due to the imbalance of individuals with heart disease in our data, our false-negative rate is very high, with only 10.7% of the positive cases correctly predicted. This high rate is due to the dataset's non-equal positive and negative heart diseases diagnosis samples. We had about 9% of individuals with heart disease in this dataset. It might help improve our model if it was balanced in the dataset.

III. Conclusion

All in all, there are certain areas of improvement and other methods we could apply to improve the accuracy of our model. For example, we could do some over and under sampling

techniques to try and even out the distribution of individuals with and without heart disease in the dataset. We should also do this with the consideration of other features that are skewed in the data as well like BMI and Race.



As pictured above on the left, BMI is a feature in the dataset with potential outliers that may be impacting our model and could improve its accuracy and reduce its False Negative Rate. Race is another variable we should absolutely consider trying to even out in our future improvements of this model. The number of individuals who are white in the data is much larger than any other race as pictured above on the right. This can be extremely problematic in our model, and Race feature that certainly should be more balanced in our future work.

Though we implemented some forms of feature selection, a more rigorous approach to feature selection may be used for logistic regression in the future. To reduce data skewness, we might consider increasing the percentage of heart disease individuals and collect more data from other races other than White. With the current dataset proportion of 9% positive diagnosis, it could be an option to select a random subset of the negative diagnosis samples such that the ratio between the two groups is more balanced. We will also experiment with other models to reduce the False Negative Rates further while not compromising the overall testing accuracy.

IV. Contributions

Minh

- Converted General Health variable using WOE
- Experimented with Gradient Boosting, Random Forest, and Stacking models
- Drew Correlation chart and visualizations for presentation
- Edited overall final report

Alex

- Wrote up Model Evaluation and Conclusion for the final report
- Coded model diagnostics with the validation and test data score, as well as the confusion matrix of the test data.
- Extracted coefficient values from the resulting logistic regression

Jack

- Wrote up Data Preparation and Modeling Process for the final report
- Experimented with feature engineering of OneHotEncoding and mutating binary variables
- Split the data into training, test, and validation data
- Helped with the grid search code to find the model with best accuracy

Olivia

- Found the data source
- Mutated the Age Group variable in a for-loop.
- Experimented with Logistic Regression, k-NN, and Decision Tree
- Outputted different ROC curves, confusion matrices, and true and false-negative rates