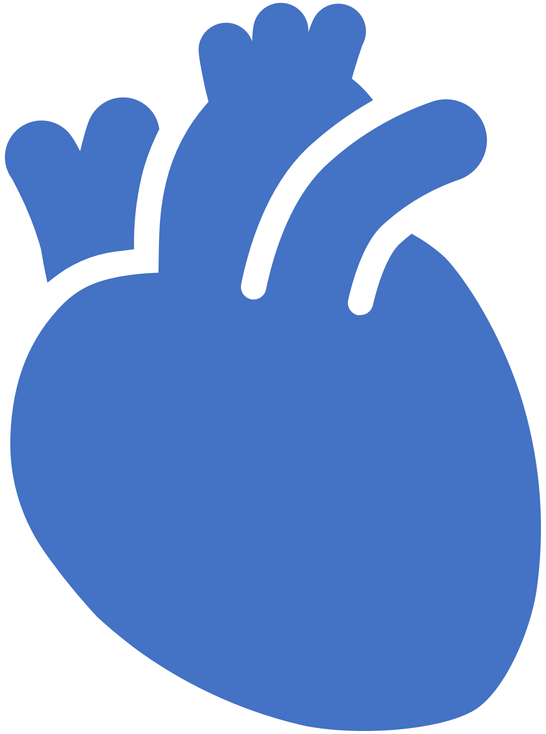


DETECTING HEART DISEASE

A DATA INTELLIGENCE FUSION

By: Roland Tetteh

Problem Statement

- Globally it’s estimated that 1 in 13 people are living with a heart or circulatory disease.
- In the United States, one person dies every 33 seconds from heart disease.
- According to the CDC about 5% of adults, age 20 and older have Coronary Artery Disease.
- In the United States, someone has a heart attack every 40 seconds.

Proposed Data Science Solution

In healthcare, identifying and preventing the factors that have the greatest impact on heart disease is very important.

Machine learning algorithms can detect "patterns" in data, using these risk factors, to predict a patient's condition or propensity to develop a heart disease.

This project seeks to apply machine learning techniques on a 2022 annual CDC survey data, of 400k+ adults related to their health status, to predict an individual's propensity to a heart attack.

Potential Impact

Early detection and risk assessment of heart disease.

Personalized risk assessment allowing for targeted interventions and prevention strategies.

Potentially be used in drug discovery.

Data Quality

- The dataset comes from the CDC, which conducts annual telephone surveys to collect data on the health status of U.S. residents.
- The original dataset of nearly 300 variables was reduced by the author to 40 most relevant variables.
- There were 157 duplicate rows out of the 445,132 rows and missing values in 38 out of the 40 columns.
- Some of the attributes in the data are State, Sex, MentalHealthDays, and HadDiabetes.
- The dataset is very imbalanced.

Data Preprocessing

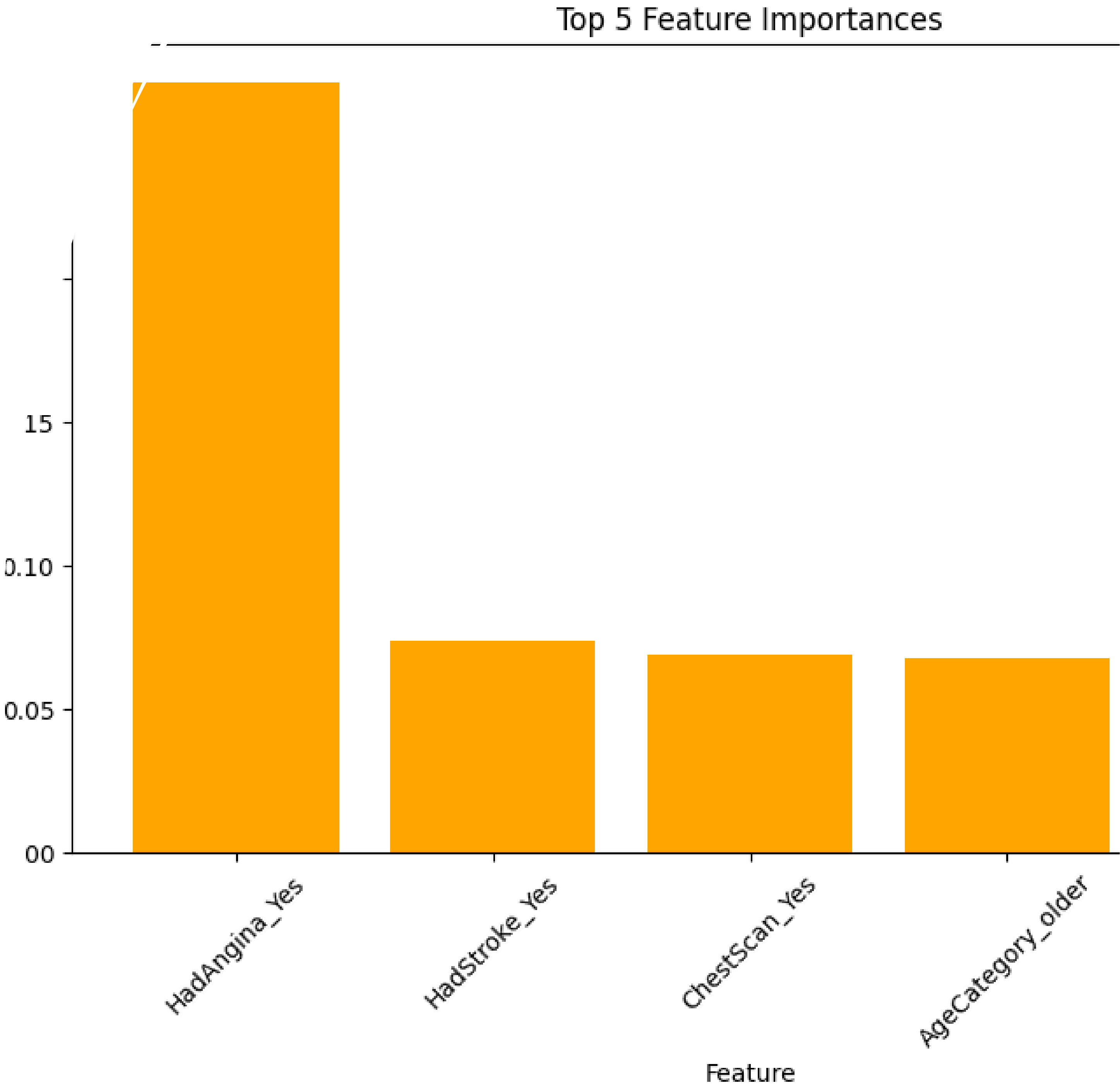
Categorical to Numerical Conversion.

Detecting and removing multicollinearity via the Variance Inflation Factors.

Upsampling and Downsampling, to deal with class imbalance.

Important Findings

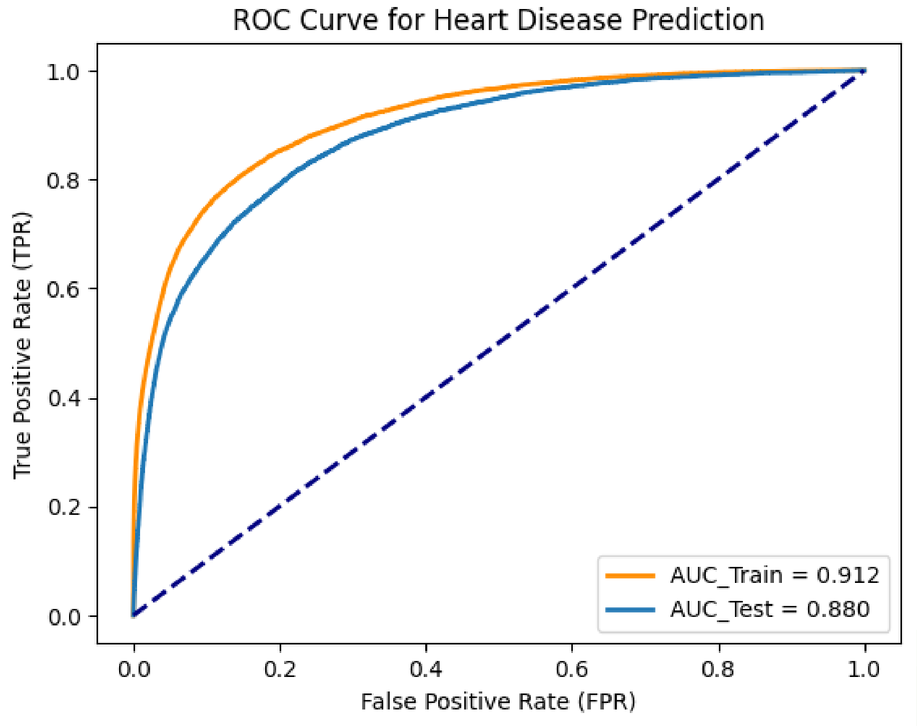
The graph depicts a Random Forest Classifier’s most important features in predicting the target variable.



Adaptive Boosting Classifier

Using the resampled training data, scaling the data and applying GridsearchCV over a range of learning rates.

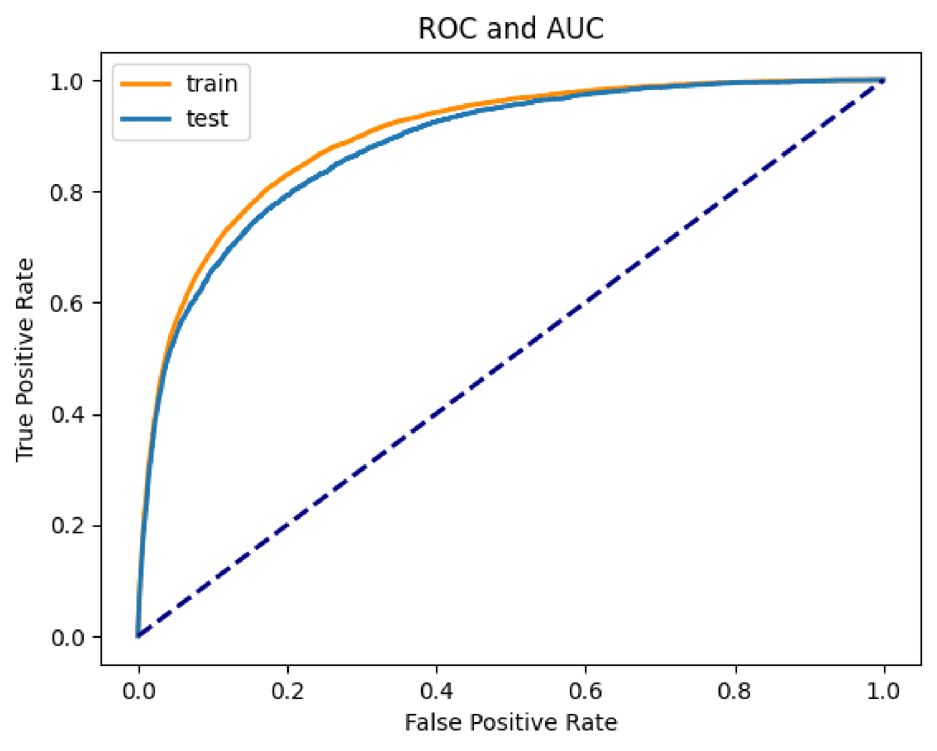
- The base estimator - Voting classifier of a logistic regression and Random Forest Classifier.
- For a FPR of 0.2%, a TPR of 0.79% can be achieved with a threshold of 0.49.



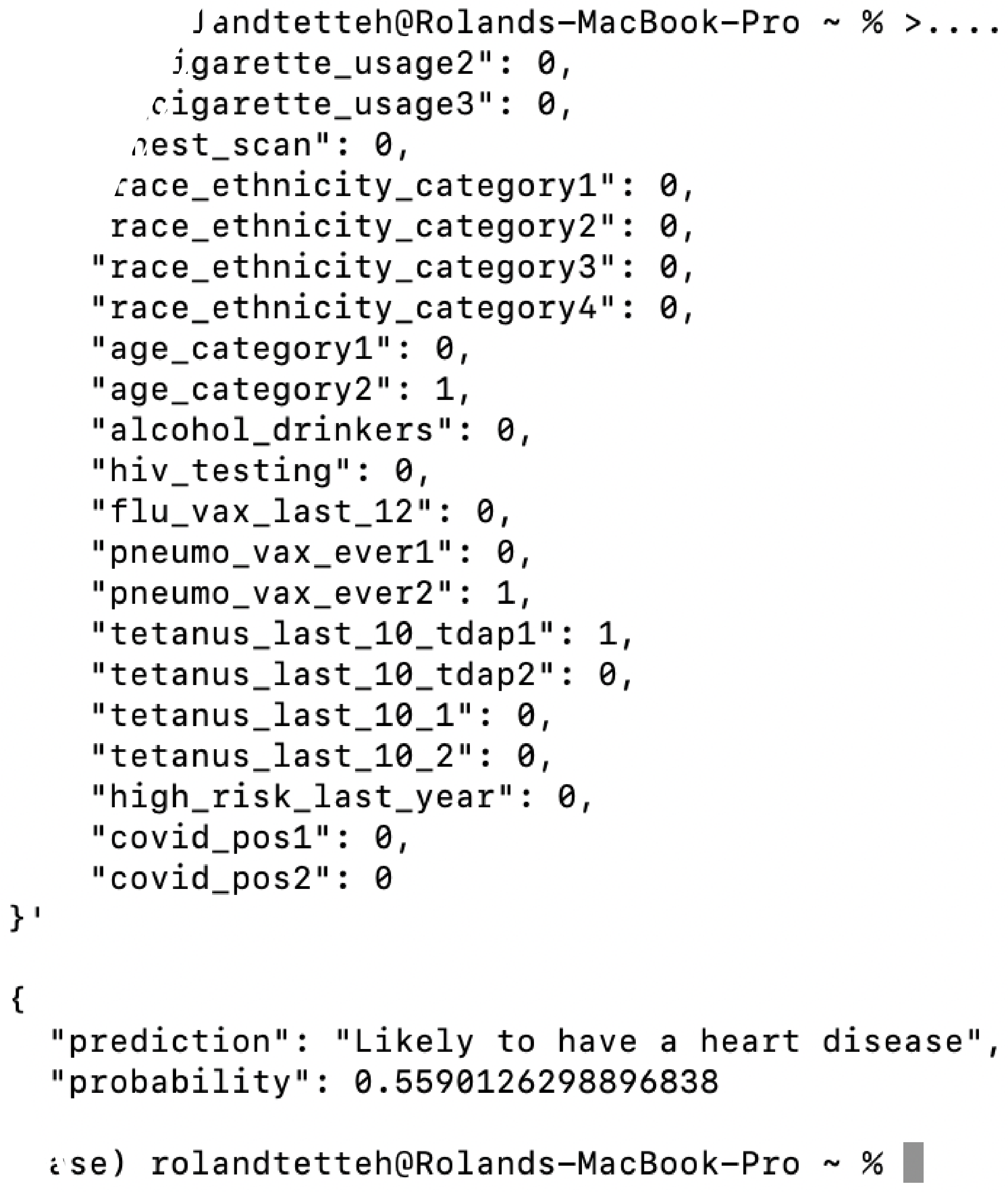
Neural Network Model

Using the resampled training data, scaling the data and applying Gridsearch on the Adam on SGD optimizers over a range of dropout rates.

- Batch Normalization was applied
- Allowing for a FPR of 0.2%, a TPR of 0.79% can be achieved with a threshold of 0.44.
- Train AUC – 0.897
- Test AUC – 0.882



Product Design



```
         andtetteh@Rolands-MacBook-Pro ~ % >....
        igarette_usage2": 0,
       cigarette_usage3": 0,
     est_scan": 0,
    ace_ethnicity_category1": 0,
     race_ethnicity_category2": 0,
    "race_ethnicity_category3": 0,
    "race_ethnicity_category4": 0,
    "age_category1": 0,
    "age_category2": 1,
    "alcohol_drinkers": 0,
    "hiv_testing": 0,
    "flu_vax_last_12": 0,
    "pneumo_vax_ever1": 0,
    "pneumo_vax_ever2": 1,
    "tetanus_last_10_tdap1": 1,
    "tetanus_last_10_tdap2": 0,
    "tetanus_last_10_1": 0,
    "tetanus_last_10_2": 0,
    "high_risk_last_year": 0,
    "covid_pos1": 0,
    "covid_pos2": 0
}'

{
  "prediction": "Likely to have a heart disease",
  "probability": 0.5590126298896838

   se) rolandtetteh@Rolands-MacBook-Pro ~ %
```