Phebe Carlson

Heart Disease Indicators Analysis

Problem Statement

In the United States, heart disease is the leading cause of death. The Centers for Disease Control and Prevention (CDC) conducts an annual health-related phone survey — the Behavioral Risk Factor Surveillance System (BRFSS) — to collect data about around 400,000 U.S. residents regarding their chronic health conditions, use of preventative services, and health-related risk behaviors.[2] The survey is often used in health care decision-making at the state level and is one of the CDC's most useful monitoring tools. Is it possible to predict an individual's heart disease prevalence by using the self-reported responses of the 2020 BRFSS?

Background

The causes of heart disease are complex and interconnected; they can be congenital (from birth) or a result of lifestyle factors and comorbidities accrued during a lifetime. In the United States, 659,000 people die each year from heart disease — that's 1 in 4 deaths.[1] Furthermore, someone has a heart attack every 40 seconds; with 1 in 5 heart attacks being silent — the damage is done, but the person is not aware of it — making preventative measures crucial. Costing $383 billion per year, heart disease also weighs heavily on the US economy.[1] Despite our best prophylactic measures, heart disease is still the leading cause of death in the US. With advances in machine learning, we are poised to make identifying and mitigating heart disease less costly and more accessible with predictive modeling of disease indicators, demographics, lifestyle factors. Moving into predictive medicine powered by machine learning and AI insights is not just good healthcare for patients but would save on the cost of treating a range of diseases once they reach maturity.

By using BRFSS features related to Heart Disease, I created a model to classify an instance —or patient — as having heart disease or not. After feature engineering and model tuning, my logistic regression model with lasso regularization and cross validation was able to predict Heart Disease with 77.15% accuracy.

Data Wrangling

This data retrieved from Kaggle was composed of a subset of features related to general health and heart disease chosen from the 2020 Annual CDC Behavioral Risk Factor Surveillance System (BRFSS). The BRFSS is conducted by mobile phone and landline survey data of about 400,000 adults related to their health. The dataset was previously processed and reduced to features likely useful as heart disease indicators and rows with null data were dropped.

I read the csv data in as a data frame called df. The df originally had 319,795 rows and 18 features. The target feature was Heart Disease — whether respondents had coronary heart disease (CHD). Less than 9% of the original data was in the positive Heart Disease class. A large pitfall of this type of data is model accuracy and precision being dictated by the a single class making up 90% of the original data. Undersampling was a necessary requirement to analyze the data more fully. To properly explore the data trends and interpret the model accuracies, I used random under sampling to resolve the high skew to 50% positive Heart Disease and 50% negative Heart Disease. I used RandomUnderSampler to resample the data from 319,795 rows to 54,746. There were no null or obviously incorrect data to handle. Some outliers were removed and features engineered. I utilized LocalOutlierFactor, an unsupervised outlier algorithm using nearest neighbors because with so many skewed features, identifying outliers was difficult. LocalOutlierFactor removed a further 5,214 rows, leaving the df with

38,582 rows for modeling. This resulted in higher accuracy, though in tandem, lowered explainability.

The dataset has 18 features are categorized as follows:
Lifestyle factors — PhysicalActivity, Smoking, AlcoholDrinking
Demographic — Sex, Race, AgeCategory
Health indicators — BMI, SleepTime, PhysicalHealth, MentalHealth
Diseases —Diabetic, Stroke, DiffWalking (difficulty walking), Asthma, KidneyDisease, SkinCancer

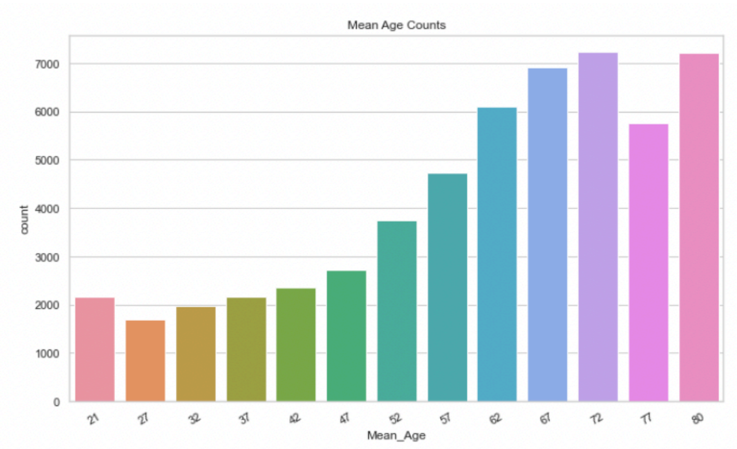
Or alternatively —
Numeric —
Continuous — BMI — calculated by BMI=weight in kg/(heigh in m^2)
Ordinal — SleepTime, PhysicalHealth, MentalHealth, AgeCategory
Categorical —
Over 2 categories — GenHealth, Race, Diabetic
2 categories or Y/N — Sex, Smoking, AlcoholDrinking, PhysicalActivity, Stroke, DiffWalking (difficulty walking), Asthma, KidneyDisease, SkinCancer

I reduced the dimensionality and in-group variation by mapping numeric data into bins. Created ranges for ordinal numeric columns. BMI bin ranges originated from my background knowledge in healthcare. Multiple encoding and scaling techniques were tested. I used Pandas get_dummies to OneHot encode columns with over 2 unique items and LabelEncoder for the columns with 2 or less. After some experimentation, higher accuracy models were achieved with LabelEncoding Physical Health. After exploring StandardScaler, RobustScaler was used, as it is better at handling outliers than other scalers and produced slightly higher accuracy. After preprocessing and wrangling, the shape of my dataset was 38,582 rows with 47 columns.

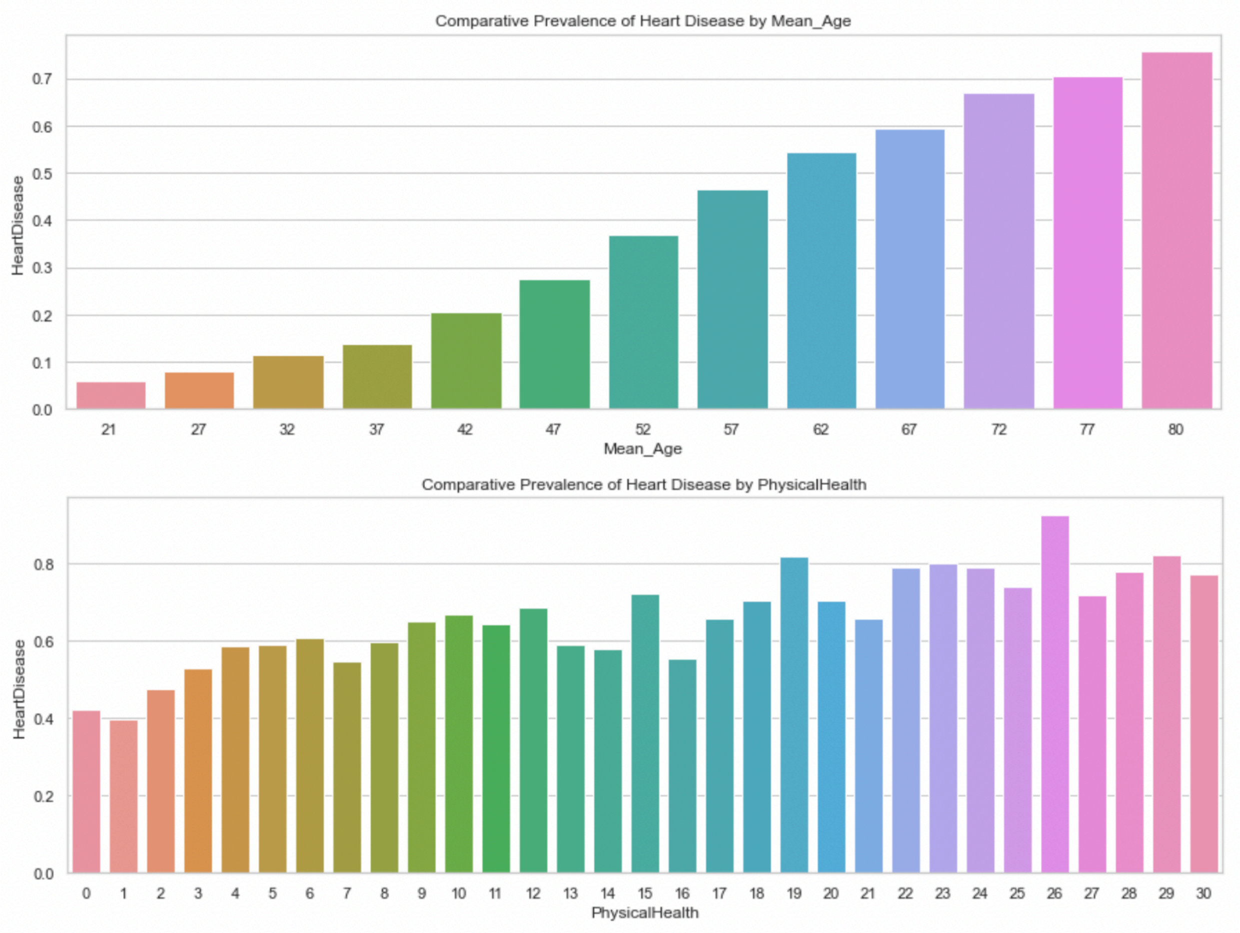
Exploratory Data Analysis & Initial Findings

The top 5 most correlated columns with the target feature, Heart Disease, were Mean Age, Difficulty Walking, Diabetic Yes, Diabetic No, and GenHealth Excellent.
Most diagnosis columns are skewed heavily towards No, similar to the disparity in proportion of positive Heart Disease responses. All feature distributions were plotted using a variation of countplots, bar charts, and histograms, both standalone and stratified by Heart Disease. Not only does the survey have a higher proportion of older respondents, but those respondents have a higher incidence of Heart Disease with age. Many of the features have heavy skew and kurtosis that is difficult to assess.

In-Depth Analysis



One feature - Mean Age - appears to account for over half the variability in Heart Disease. There are more respondents to the CDC BRFSS survey as age increases and there is higher prevalence of heart disease as age increases. Heart disease prevalence also appears to increase with more days of reportedly low physical health.

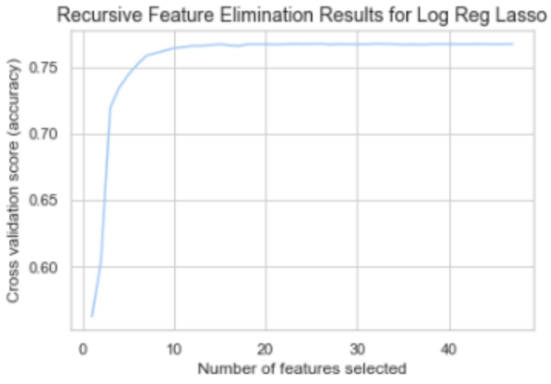


Obesity III BMI is the maximum category for BMI at a value of 40. Many respondents having BMI's well above this number. Numeric data appears to have many values lying outside the 1.5 IQR range, though they may not be outliers because these data are bound between values (e.g. 30 days in a month). Investigating how to split these variables into normal and abnormal features could be useful in predictions. A higher incidence of heart disease is correlated with increasing age in reality but it is also possible that bias like response bias and bias involving the age at which one seeks treatment are likely to be involved.

Model Selection

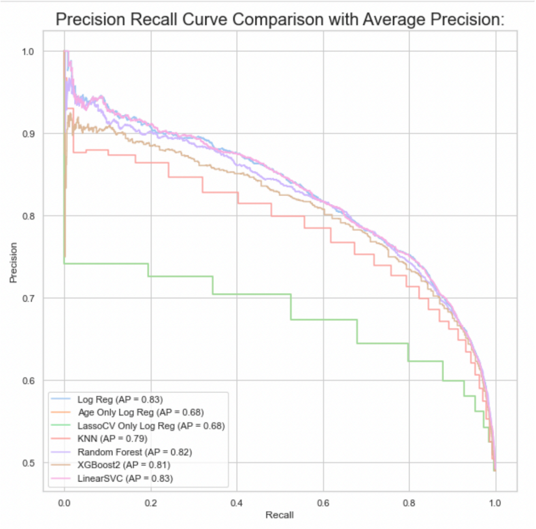
A number of models were implemented to see which worked best; the chosen models were Logistic Regression, Random Forest, LinearSVC, and XGBoost. I used recursive feature elimination with cross validation (RFECV) to choose features for Logistic Regression and XGBoost, and also gathered feature importances from Logistic Regression and Random Forest model coefficients. When assessing the models, I used the best set of features for each model.

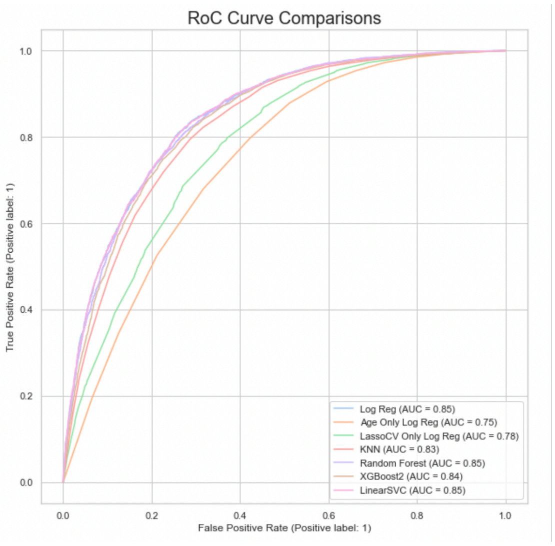
The most influential features were Mean Age and Physical Health. These two features alone achieved 71.16% accuracy on logistic



regression of the test data, while Mean Age alone achieved 68.52% accuracy.

RandomForest and XGBoost appeared to overfit on training data compared to test data and were more accurate at predicting the 0 class than the 1 Heart Disease class. Using selected features lowered the Logistic Regression Lasso model's accuracy. The model didn't appear overfit on training data. In the end, my logistic regression model with lasso regularization achieved an accuracy of 77.15%.


Precision Recall Curve Comparison with Average Precision:


RoC Curve Comparisons

Takeaways and Future Research

If I had more time on this project, I would try different types of binning to see if any other predictive features exist — e.g. splitting BMI column into normal BMI and abnormal BMI for very high values. I would encode ordinal numeric data in a way that best preserves the hierarchy order, which may not have happened with LabelEncoder.

I would like to try Heart Disease predictions with the full CDC BRFSS survey that includes geographic and socioeconomic factors. These other geographic and socioeconomic factors exhibit great influence over the comparative health of Americans.

While the Behavioral Risk Factor Surveillance System (BRFSS) from is impressive and well-enacted, I would be interested to look into the experimental design and whether implicit bias is affecting the type of data being collected and therefore the representativeness of the sample.

Heart disease and many other diseases are public health and economic weights. If we are able to utilize the CDC BRFSS survey information to predict disease prevalence, it is one more tool toward proactively combating the leading cause of death in the United States.

References:

1. Heart Disease Facts. Centers for Disease Control and Prevention. (2022). Retrieved 28 June 2022, from https://www.cdc.gov/heartdisease/facts.htm.
("Heart Disease Facts", 2022)

2. Kamil Pytlak. (2022 February). Personal Key Indicators of Heart Disease. Retrieved May 15 2022, from https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease.