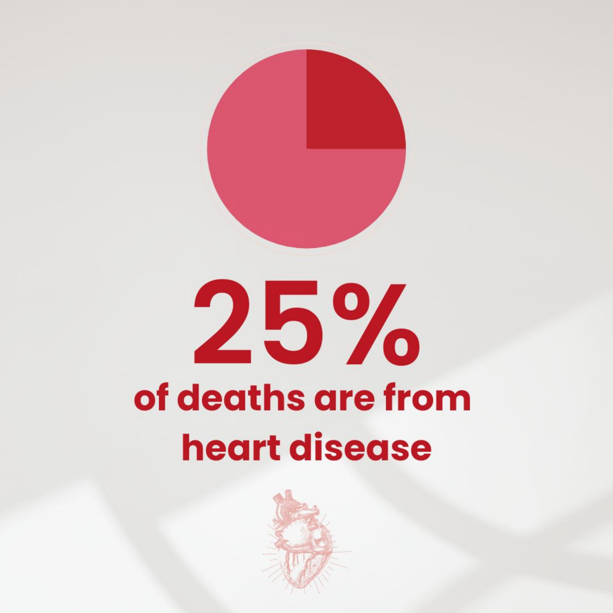
Heart Disease Indicators

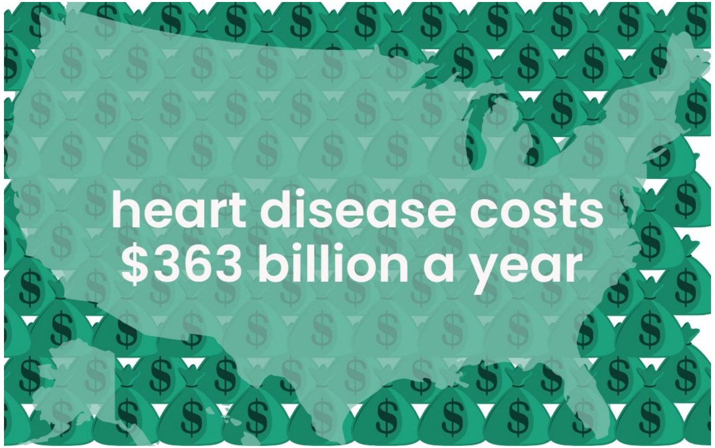
By
Phebe Carlson

Agenda

1. Introduction —
   Problem statement, solution, goals
2. Key findings
3. Data description and preprocessing
4. Exploratory data analysis
5. Modeling —
   Feature selection and evaluation
6. Takeaways and future research

The Problem:
Identifying risk factors to reduce heart disease prevalence





The causes and risk factors are complex and interconnected.

The Solution:
targeted early intervention using predictions

Classify, identify, and model heart disease indicators to use for prediction

Can use predictions to improve early intervention techniques

Our goal is to answer...

- Can we use this subset of the 2020 CDC BRFSS to correctly classify a respondent's heart disease status?
- Which factors have a significant influence on the likelihood of heart disease?
- What more needs to be done to be able to predict heart disease from data like this?

Key findings

Most accurate model —

Logistic regression with lasso regularization with **77.15% accuracy**

Mean Age Only —

68.52% accuracy test set

Most influential indicators — Mean Age and Physical Health

Potential improvements to methodology

The Data

Subset of the 2020 Annual CDC Behavioral Risk Factor Surveillance System (BRFSS) survey.

The survey is taken by mobile and landline phone calls.

Target variable: heart disease

Data source: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

Data wrangling and preprocessing

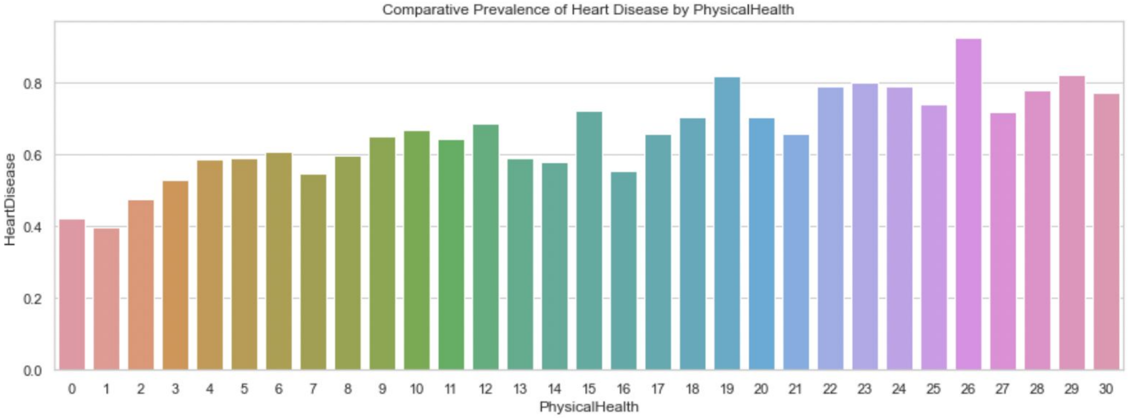
Originally had 319795 rows, 18 features

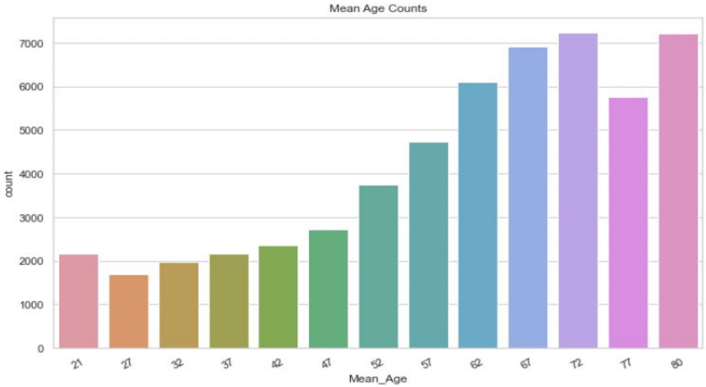
After preprocessing and feature engineering: 38582 rows, 47 features

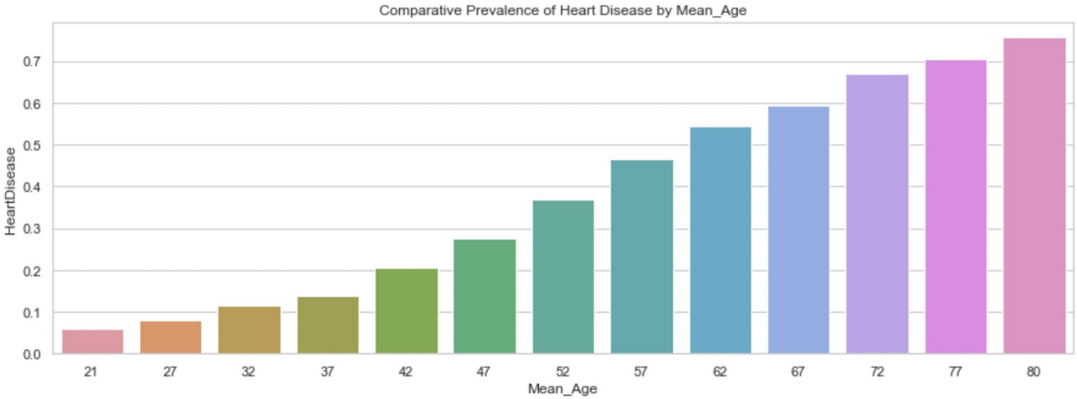
Manipulation steps –

- Undersampled – less than 9% of original data had heart disease
- Binned ordinal categorical variables
- Re-binned AgeCategory into numeric Mean_Age
- Encoded features
- Scaled numeric features
- Outliers

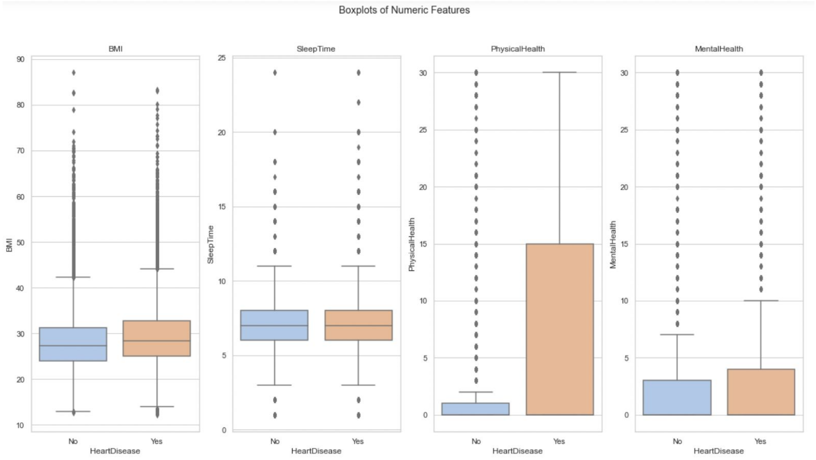
Exploratory Data Analysis

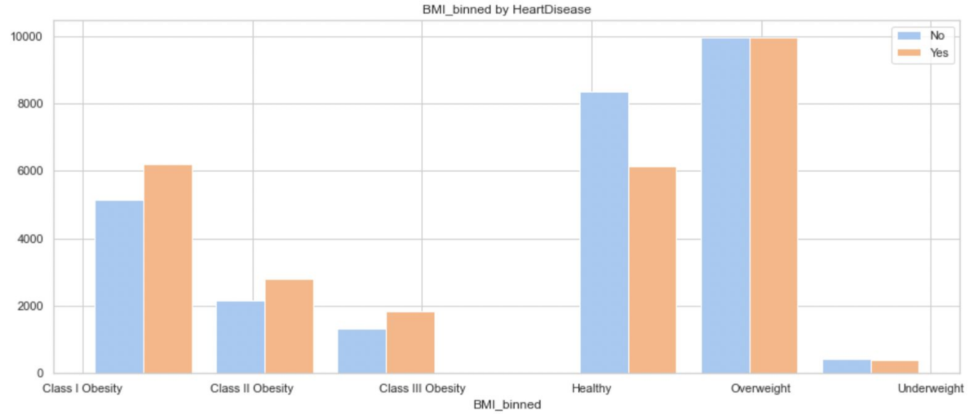






Exploratory Data Analysis: continued





Modeling

Models —

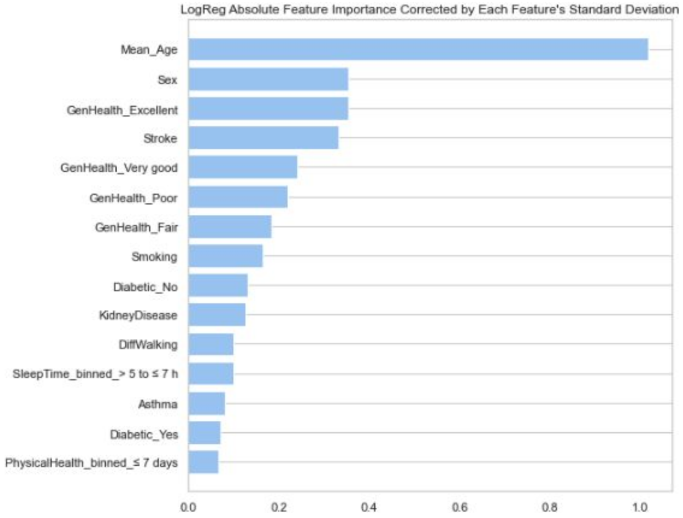
- Logistic Regression
- LinearSVC
- RandomForest
- XGBoost

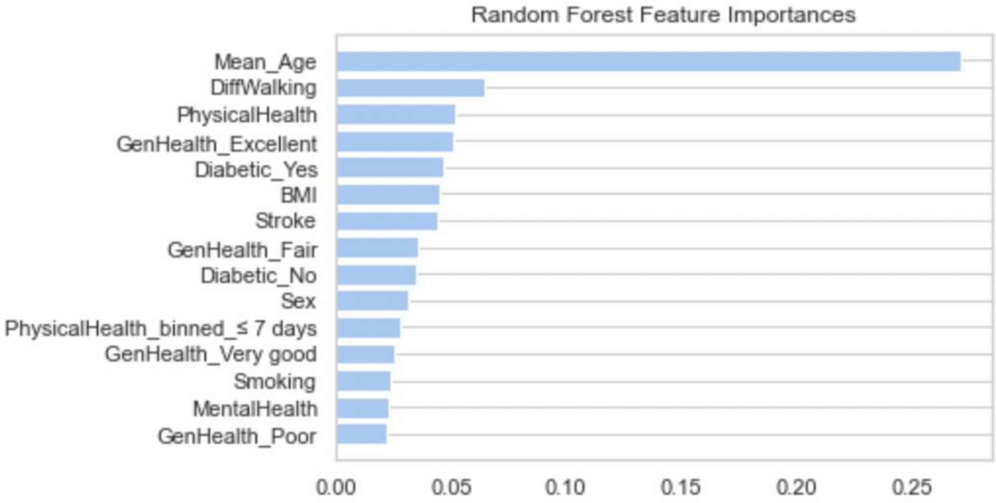
Hypertuning done using GridSearchCV and RandomizedSearchCV

Feature importances

Most influential features from LassoCV: Physical Health and Mean_Age

LogReg with Lasso regularization and Random Forest Feature importances



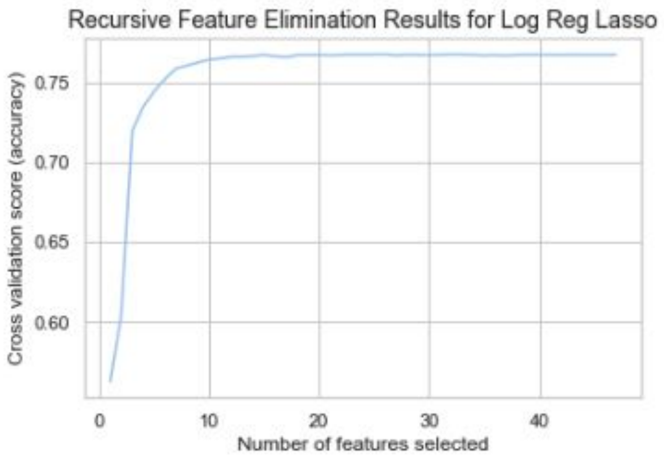


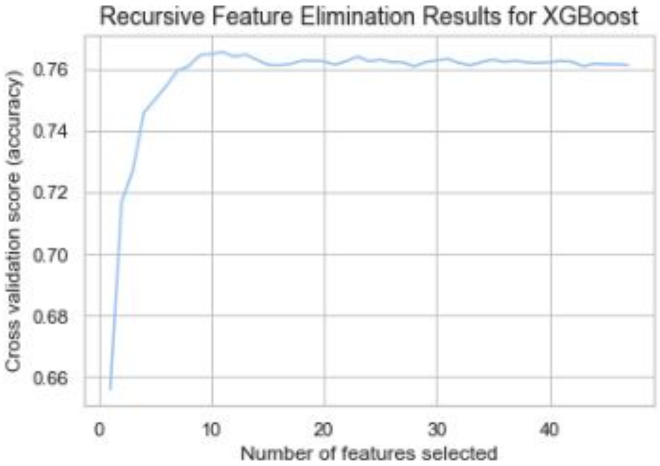
Feature selection

Hyperparameter tuning — GridSearchCV, RandomizedSearchCV

Sklearn RFECV for cross-validating recursive feature elimination

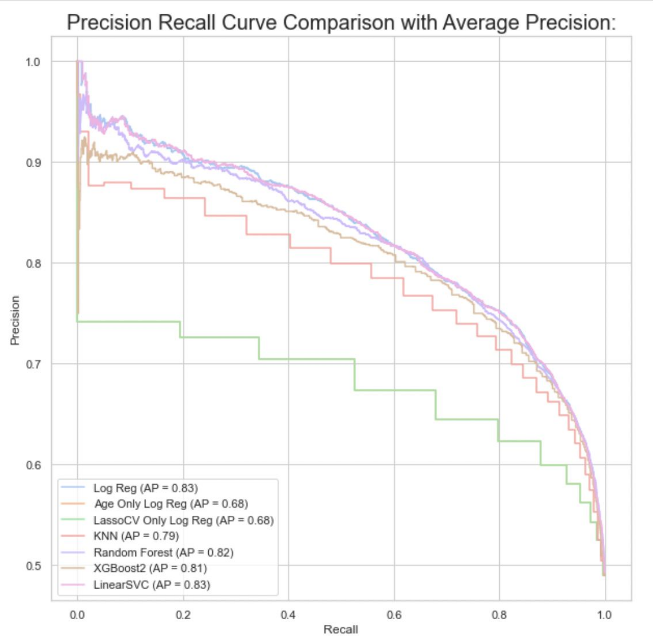
RFECV Log Reg Lasso – optimal number of features 26

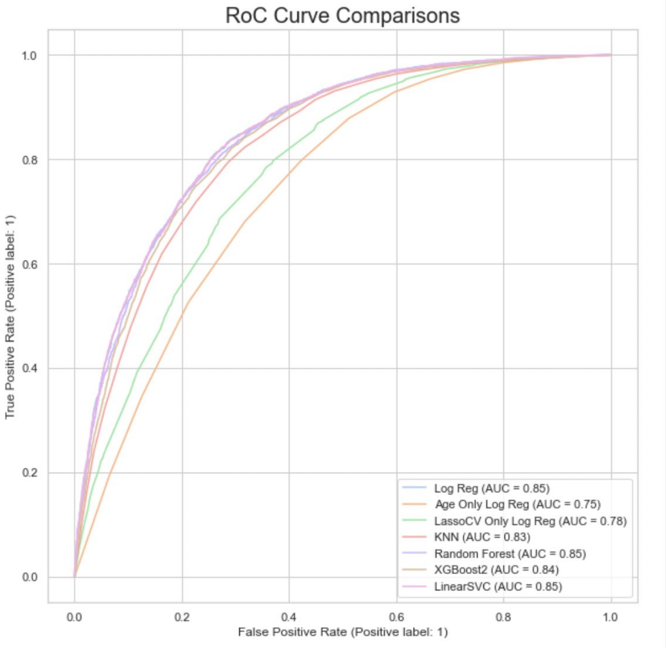




Evaluation with precision-recall and ROC curves

Models – Logistic Regression, KNN, LinearSVC, Random Forest, XGBoost





Notable Modeling Details

Most accurate model — 77.15% accuracy

Logistic regression with lasso regularization with all features

Mean Age Only —

68.52% accuracy test set

Most influential indicators —

Mean Age and Physical Health

Takeaways & Future Research

With more time –

- Create BMI column of normal and abnormal
- Try different encoding for ordinal columns
  - Reassess after this
- Investigate representative nature of CDC BRFSS responses and any possible improvements to methodology
- Work with full CDC BRFSS dataset

Questions?