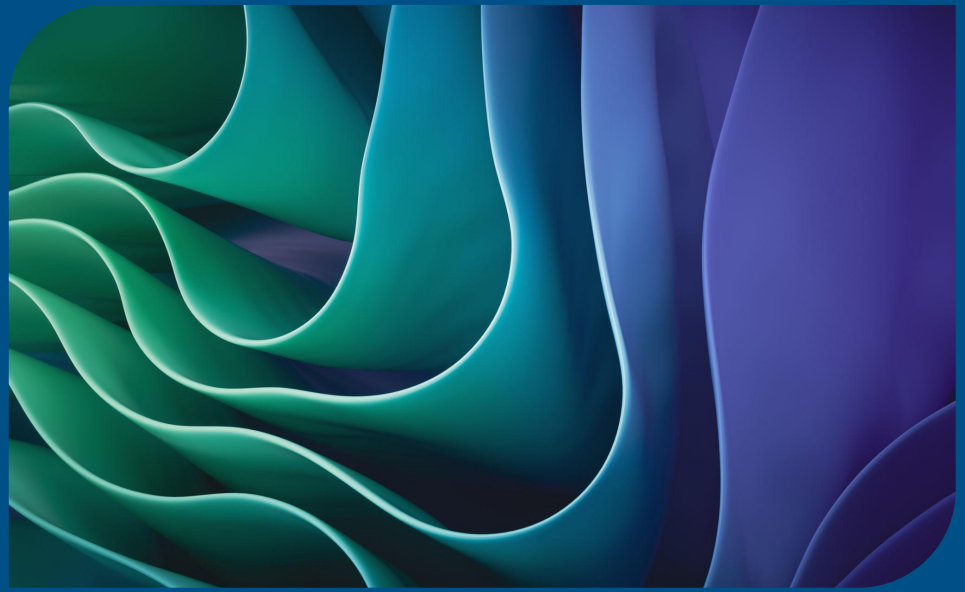
Coronary Artery Disease Prediction Using BRFSS Data

Fares Alahdab, AJ Strauman-Scott, Brandon Cunningham



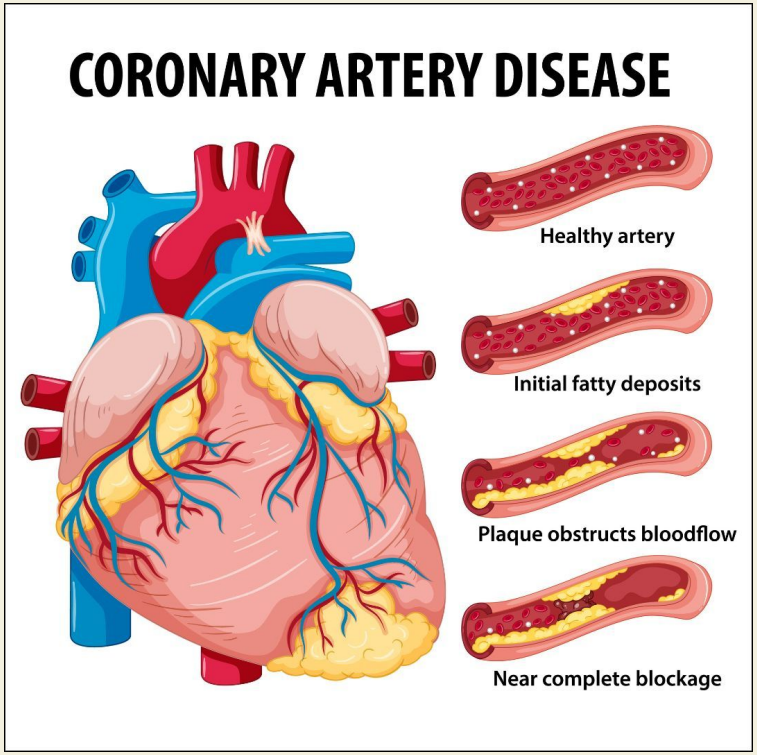
What is BRFSS?

“The Behavioral Risk Factor Surveillance System (BRFSS) is the nation’s premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world. “ - CDC

Coronary Artery Disease (CAD)

A type of heart disease where the arteries of the heart cannot deliver enough oxygen-rich blood to the heart.

Heart disease is the most common cause of death for Americans, and over 20 million adults in the US have CAD making it the most prevalent heart disease in America.



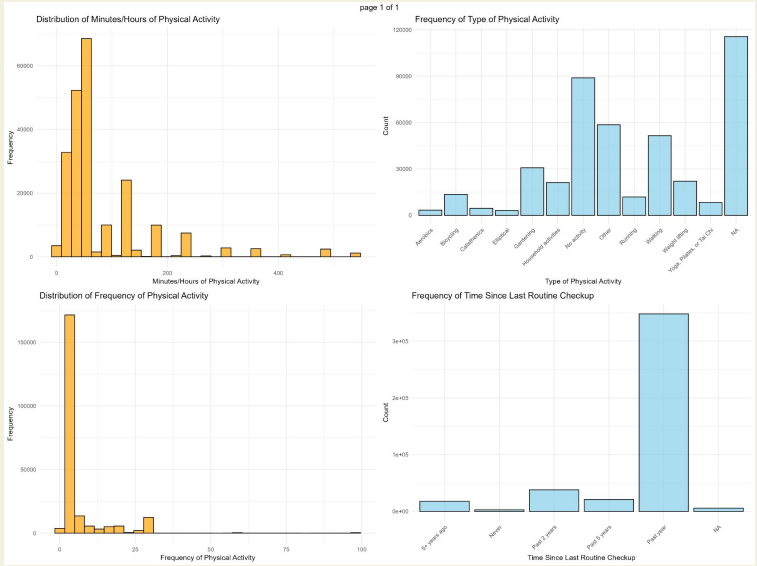
What factors contribute most to the risk of developing CAD?

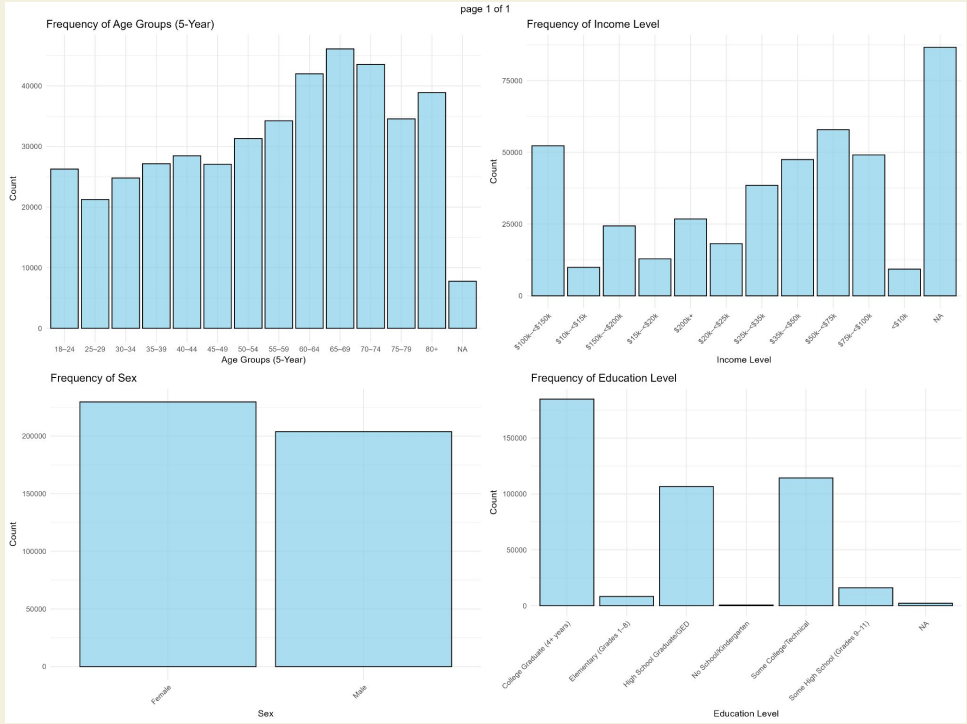
Some of the factors we will investigate include:

- Comorbidity Presence
  - Hypertension
  - Diabetes
  - High cholesterol
- Covid-19 Exposure
- Demographic and Clinical Factors
  - Age
  - Sex
- Lifestyle Factors
  - Physical activity
  - Alcohol consumption
- Physical Health and Functional Status
  - Self reported health status
  - Frequency/type of Exercise
- Psychosocial and Mental Health Factors
  - Perceived low health
- Social Determinants of Health
  - Education Level
  - Income
  - Healthcare Access
- Strength Training

Data Exploration and Imputation

Relationships between these explanatory factors were explored. KNN imputation fill in missing values for continuous variables, and mode imputation was used for categorical variables.

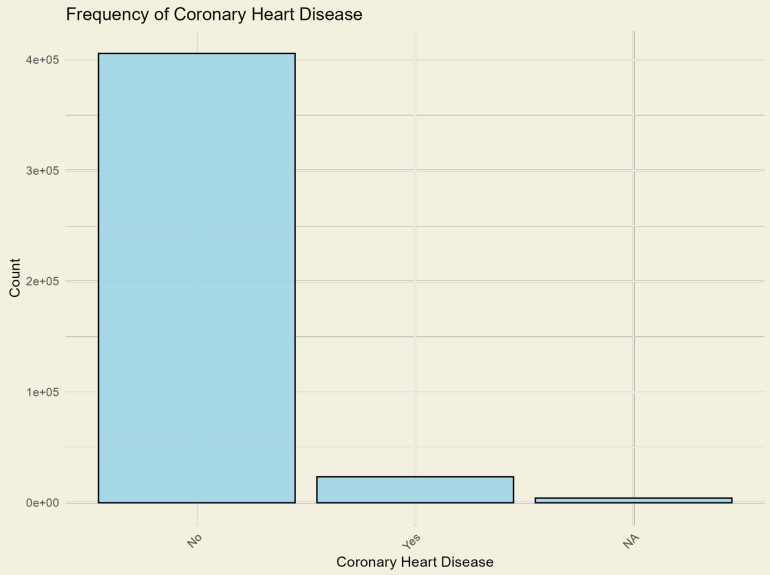




Model Metrics

All models were compared using the following metrics:

- Area Under Curve (AUC)
  - Measures a model's ability to distinguish between classes across all classification thresholds, making it a robust metric for evaluating overall performance, especially in imbalanced datasets.
- Balanced Accuracy
  - Given the significant imbalance between the frequencies of "No CAD" and "Yes CAD," balanced accuracy provides a fairer assessment by averaging the percentage of correctly predicted "No" and "Yes" cases, ensuring equal consideration for both classes.
- Akaike Information Criterion (AIC)
  - AIC balances model fit and complexity, lower values indicate a better trade-off between the model's accuracy and its simplicity.



Model 1: Logistic regression using GLM and all features included

AUC: 0.854
Balanced Accuracy: 51.2%
AIC: 1043297

The high AUC and low Balanced accuracy indicates that with a different threshold value the model could be a good predictor, but in its current state it is not.

```
Generalized Linear Model

300364 samples
    27 predictor
     2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 270327, 270327,
Resampling results:

  Accuracy   Kappa
  0.9455694  0.06386221
```

Model 2: GLM using stepwise feature selection

AUC: 0.854
Balanced Accuracy: 51.2%
AIC: 99269

With a similar AUC and Balanced Accuracy, the logistic model after feature selection preforms just as well as the full model, but with fewer features reducing the risk of overfitting, however with a higher AIC score than the full model this tradeoff does not appear to be worth it.

```
Call:
glm(formula = CVDCRHD4 ~ X_AGEG5YR + GENHLTH + TOLDHI3 + SEXVAR +
    BPHIGH6 + CVDSTRK3 + PERSDOC3 + SMOKE100 + DIABETE4 + EMPLOY1 +
    CHECKUP1 + COVIDPO1 + PHYSHLTH + ALCDAY4 + MARITAL + PRIMINS1 +
    MEDCOST1 + EDUCA + MENTHLTH + EXRACT22 + INCOME3 + STRENGTH +
    BMI, family = binomial, data = train_data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.1026291  0.1408918 -14.924  < 2e-16 ***
X_AGEG5YR    0.2482304  0.0045993  53.971  < 2e-16 ***
GENHLTH      0.4953251  0.0109786  45.117  < 2e-16 ***
TOLDHI3     -0.6265312  0.0186279 -33.634  < 2e-16 ***
SEXVAR      -0.6695136  0.0181259 -36.937  < 2e-16 ***
BPHIGH6     -0.3022255  0.0103189 -29.289  < 2e-16 ***
CVDSTRK3    -0.7157803  0.0261641 -27.357  < 2e-16 ***
PERSDOC3     0.3105684  0.0148407  20.927  < 2e-16 ***
SMOKE100    -0.3143702  0.0178468 -17.615  < 2e-16 ***
DIABETE4    -0.1381668  0.0095157 -14.520  < 2e-16 ***
EMPLOY1      0.0493522  0.0041467  11.902  < 2e-16 ***
CHECKUP1    -0.1745829  0.0178524  -9.779  < 2e-16 ***
COVIDPO1    -0.1385395  0.0177904  -7.787 6.85e-15 ***
PHYSHLTH     0.0058928  0.0009446   6.238 4.43e-10 ***
ALCDAY4     -0.0090216  0.0013781  -6.546 5.90e-11 ***
MARITAL     -0.0372561  0.0068146  -5.467 4.58e-08 ***
PRIMINS1    -0.0043779  0.0008050  -5.439 5.37e-08 ***
MEDCOST1    -0.1547790  0.0358919  -4.312 1.62e-05 ***
EDUCA        0.0382272  0.0093169   4.103 4.08e-05 ***
MENTHLTH     0.0036792  0.0010708   3.436  0.00059 ***
EXRACT22    -0.0005501  0.0002322  -2.369  0.01783 *
INCOME3     -0.0103270  0.0048947  -2.110  0.03487 *
STRENGTH     0.0023810  0.0011719   2.032  0.04218 *
BMI          0.0023708  0.0014176   1.672  0.09444 .
```

Model 3: XGBoost with Oversampling

AUC: 0.859
Balanced Accuracy: 78.7%
AIC: 448160

For this and all future models, we apply oversampling to increase the representation of the minority class (Yes CAD) beyond its proportion in the original dataset, enabling the model to better learn and accurately predict these cases. XGBoost as a model will be able to find more complex non-linear relationships than the logistic regression models used before, so if any exist it should perform significantly better.`

```
eXtreme Gradient Boosting

568138 samples
    27 predictor
     2 classes: 'class0', 'class1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 511325, 511324,
Resampling results:

  ROC        Sens       Spec
  0.8661721  0.7390423  0.8431895
```

Model 4: Ridge With Over-Sampling

AUC: 0.854
Balanced Accuracy: 78.0%
AIC: 380372

The similar performance of this ridge model compared to the earlier XGBoost model but with a much lower AIC suggests that the relationship between the features and CAD presence is linear.

```
glmnet

568138 samples
    27 predictor
     2 classes: 'class0', 'class1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 511324, 511324, 511324,
Resampling results across tuning parameters:

  lambda  ROC        Sens       Spec
  0.001   0.8537735  0.7415698  0.8133834
  0.112   0.8502092  0.7352228  0.8133939
  0.223   0.8475341  0.7311886  0.8125104
  0.334   0.8457583  0.7291679  0.8110037
  0.445   0.8444671  0.7277070  0.8106270
  0.556   0.8434774  0.7267917  0.8091907
  0.667   0.8426718  0.7259891  0.8076559
  0.778   0.8419996  0.7257075  0.8074236
  0.889   0.8414490  0.7252780  0.8068850
  1.000   0.8409684  0.7249401  0.8060718
```

Model 5: LASSO With Over-Sampling

AUC: 0.854
Balanced Accuracy: 78.0%
AIC: 411734

The Lasso model performs similarly to the ridge and XGBoost model before it, but has a higher AIC value than the ridge model meaning it has a worse trade-off between accuracy and simplicity.

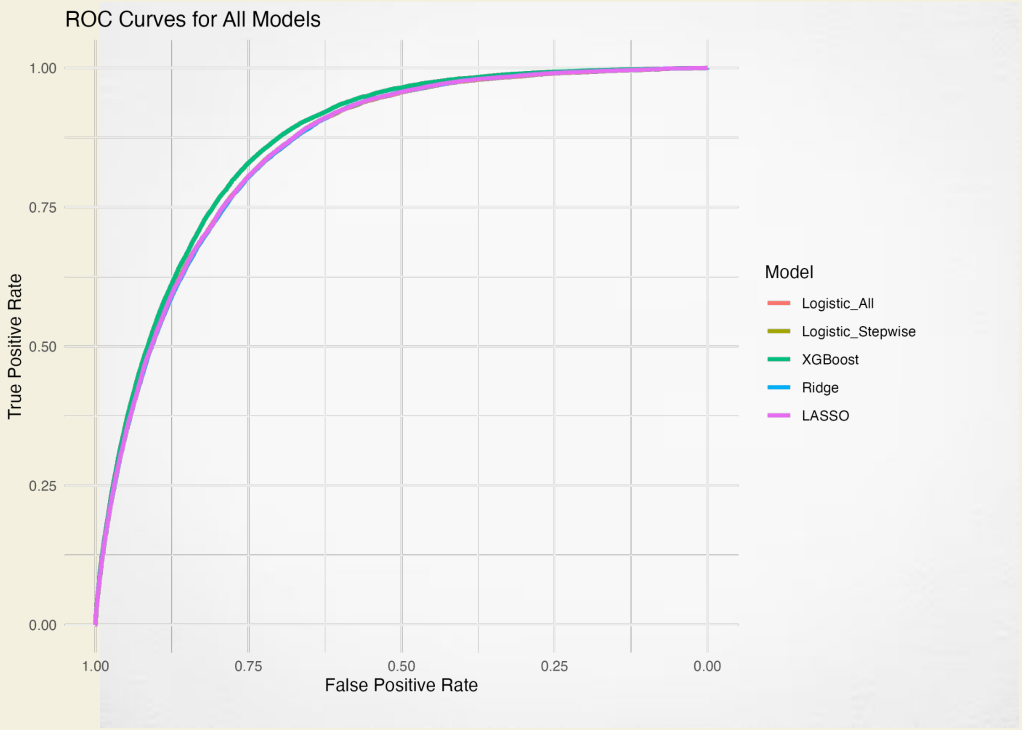
```
568138 samples
    27 predictor
     2 classes: 'class0', 'class1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 511324, 511324, 511325,
Resampling results across tuning parameters:

  lambda  ROC        Sens       Spec
  0.001   0.8544977  0.7432596  0.8125631
  0.112   0.8248141  0.6844710  0.8200332
  0.223   0.7497227  0.6284494  0.7419113
  0.334   0.5000000  0.1000000  0.9000000
  0.445   0.5000000  0.1000000  0.9000000
  0.556   0.5000000  0.1000000  0.9000000
  0.667   0.5000000  0.1000000  0.9000000
  0.778   0.5000000  0.1000000  0.9000000
  0.889   0.5000000  0.1000000  0.9000000
  1.000   0.5000000  0.1000000  0.9000000
```

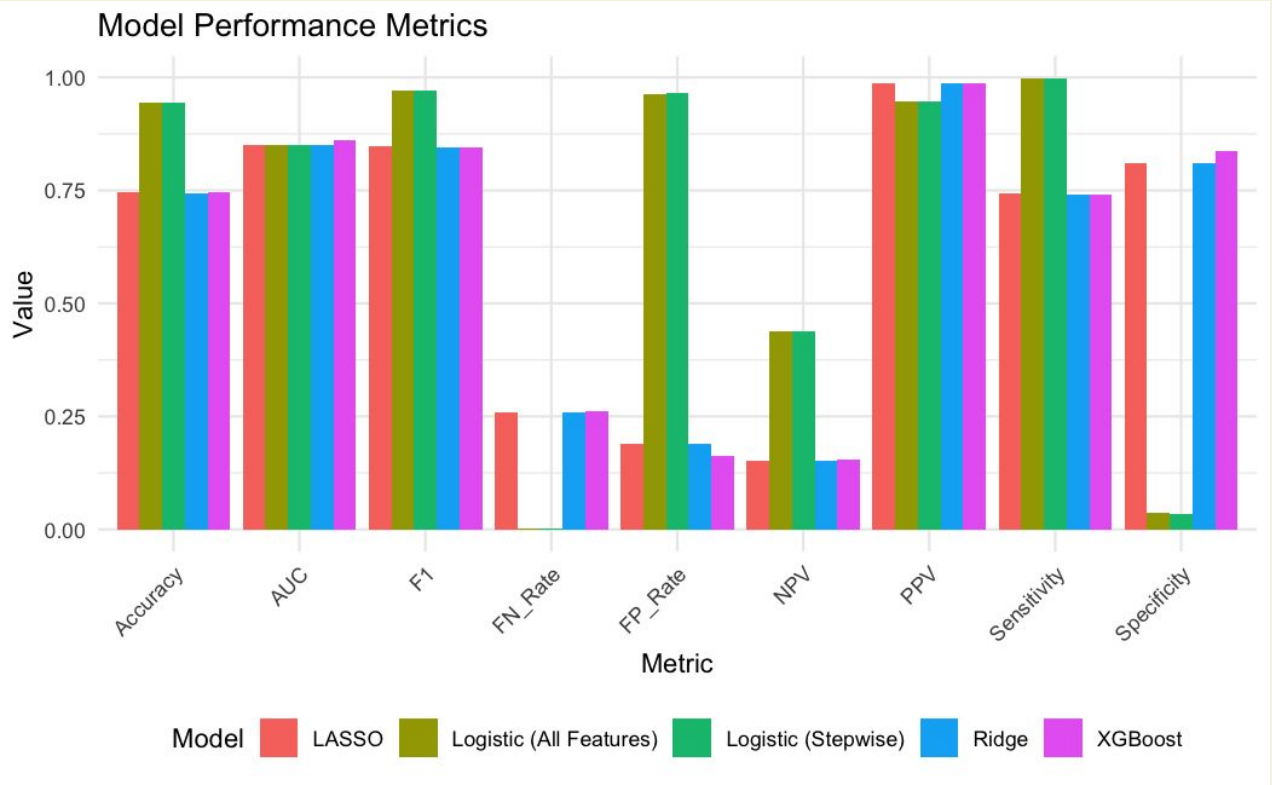
ROC Plot

The ROC (Receiver Operating Characteristic) curve is a tool used to evaluate the trade-off between true positive and false positive rates for a binary classification model across different thresholds. From the ROC Curve, we observe that all models display nearly identical sensitivity, as shown by their similar slopes. However, the XGBoost model exhibits slightly better specificity, maintaining a lower false positive rate compared to the other four models.



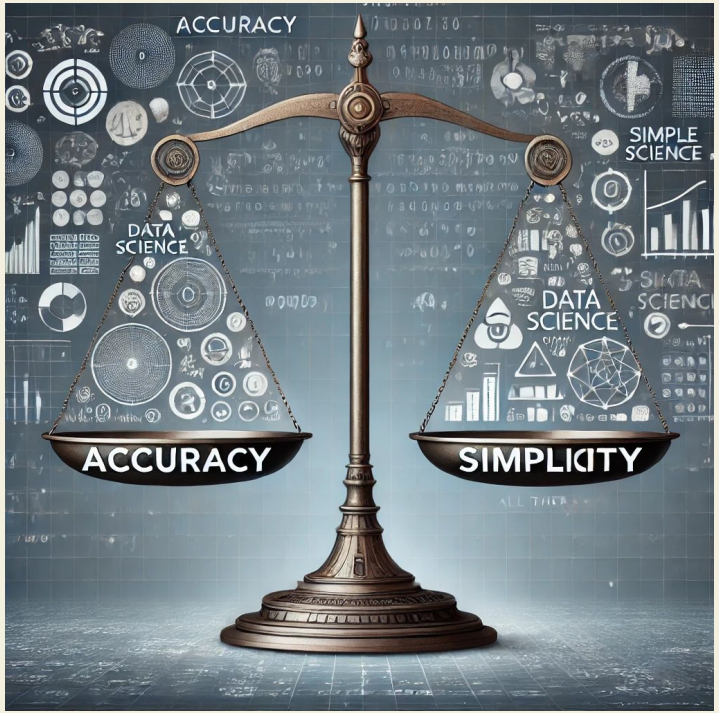
Comparing Metrics

- Both Logistic models perform best in Accuracy, F1 Score, and NPV, but struggle with extremely low Specificity, due to the imbalance nature of CAD.
- LASSO, XGBoost, and Ridge achieve high Specificity but lower Sensitivity, exhibiting a trade-off between these metrics, reflecting challenges in handling the class imbalance.



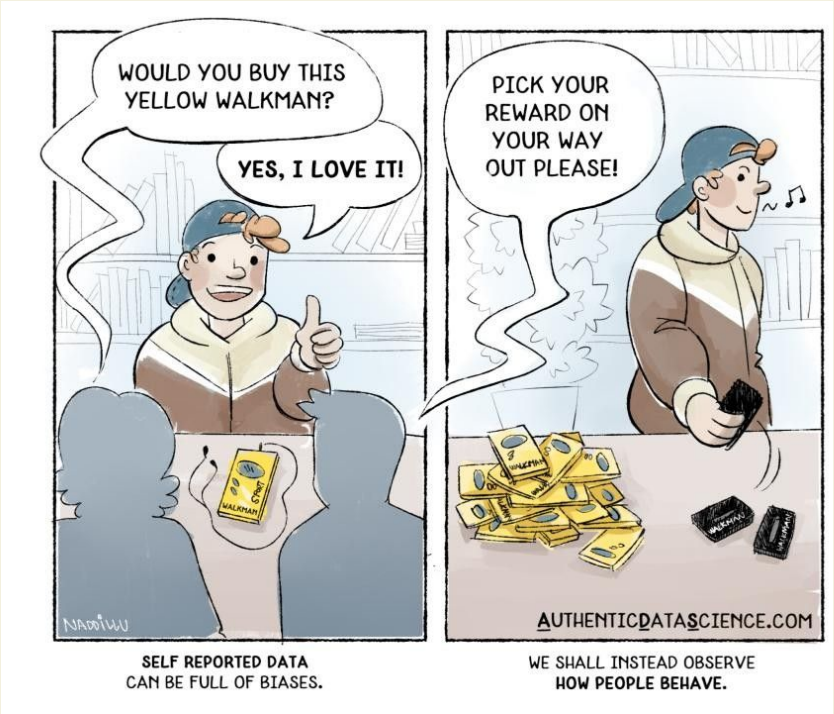
Selected Model: Ridge Regression

The Ridge model stands out as the best choice for CAD prediction due to its robust performance and superior efficiency. While it achieves high Balanced Accuracy and AUC scores comparable to those of the Lasso and XGBoost models, it also boasts the lowest AIC score among all models. This combination of strong predictive power and exceptional model parsimony makes Ridge the most reliable and effective option for this application.



Study Limitations

- Data is self-reported, so some answers (like exercise, alcohol use, or mental health) may not be completely accurate.
- The study only looks at one point in time, so it can't show cause-and-effect relationships.
- Missing data, even with imputing, might still affect the results, especially for features with lots of missing data.
- Model selection: Ridge regression balances performance and interpretability, but advanced models like XGBoost could add value in specific contexts.



Conclusion

This study demonstrates that integrating lifestyle factors, mental health measures, and social determinants of health significantly improves CAD risk prediction. Key findings reveal that, alongside traditional predictors like hypertension, diabetes, and high cholesterol, mental health indicators, like poor mental health days, and socioeconomic factors, like income level, and healthcare access, are critical contributors to CAD risk. These results underscore the role of psychosocial stress and health inequities in driving cardiovascular outcomes, highlighting the importance of addressing upstream determinants of health. Incorporating these non-traditional predictors into CAD risk models enhances their accuracy and provides a more comprehensive foundation for prevention and intervention efforts.