How may we improve existing predictive metrics for cardiovascular health outcomes using machine learning?
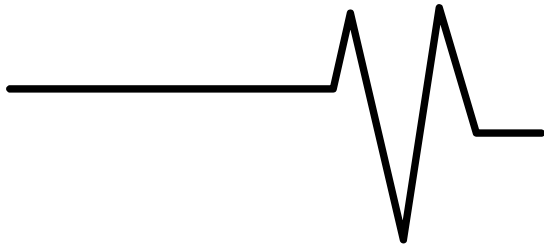
Noreen Mayat

Existing Metrics

Fig 1. Score demonstrating the variables used in the current ASCVD Risk Estimator Calculator using the ACC/AHA PCE to compute ASCVD risk scores for a 55 year–old white male [1].

Atherosclerotic Cardiovascular Disease Risk Calculator

Current Variables:

- Sex
- Age
- Race (Black/White)
- Total Cholesterol
- HDL–Cholesterol
- Systolic Blood Pressure
- Blood Pressure Medication (Yes/No)
- Diagnosed Diabetes (Yes/No)
- Smoker (Yes/No)
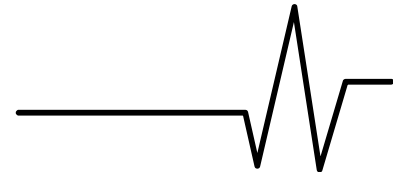
Social Determinants of Health

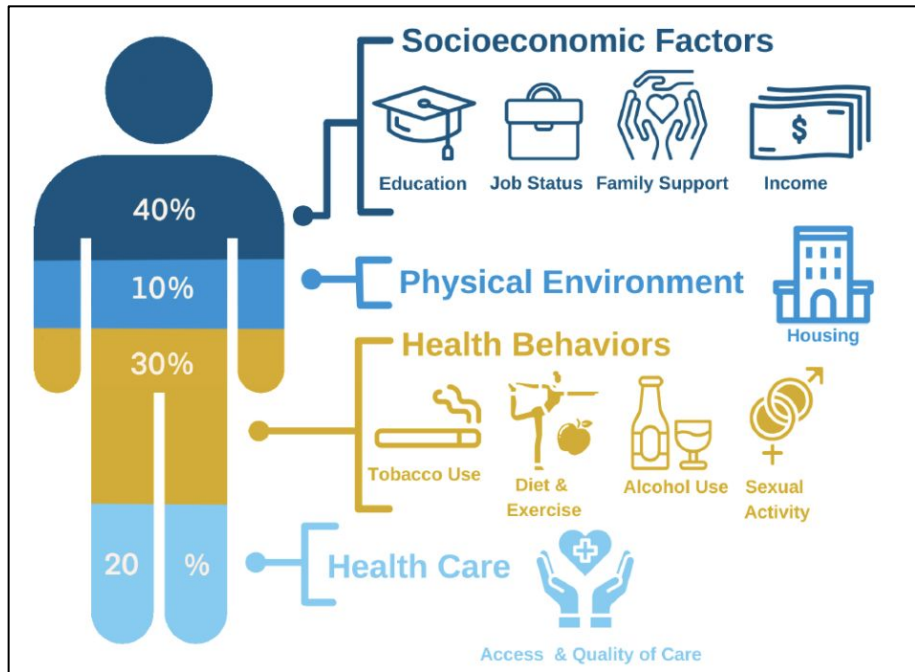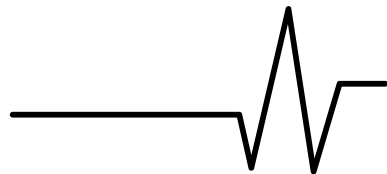Fig 2. Visualization of various social determinants of health and how they can contribute to overall health outcomes, developed by UCLA Health [2].
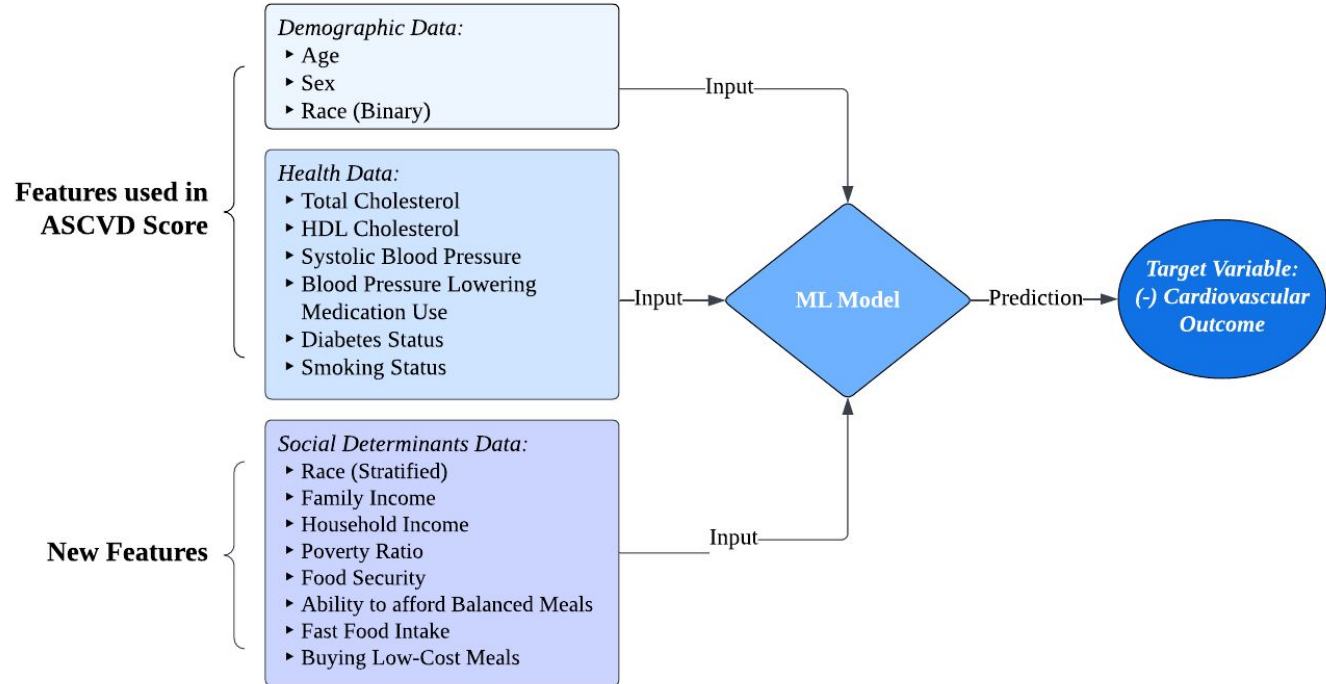
Social Determinants of Health

New Variables:

- Stratified Race Variable
- Income
- Poverty Ratio
- Food Security
- Balanced Meals
- Fast-Food Intake
- Low-Cost Meals

Methodology

Features used in ASCVD Score

Demographic Data:
▸ Age
▸ Sex
▸ Race (Binary)

Health Data:
▸ Total Cholesterol
▸ HDL Cholesterol
▸ Systolic Blood Pressure
▸ Blood Pressure Lowering Medication Use
▸ Diabetes Status
▸ Smoking Status

New Features

Social Determinants Data:
▸ Race (Stratified)
▸ Family Income
▸ Household Income
▸ Poverty Ratio
▸ Food Security
▸ Ability to afford Balanced Meals
▸ Fast Food Intake
▸ Buying Low-Cost Meals

—Input—

—Input→

Input

ML Model

—Prediction→

Target Variable: (-) Cardiovascular Outcome

Methodology

1. **Data Pre-Processing:**
   a. Clean, filter, impute and organize NHANES data
   b. Cardiovascular event defined as: heart attack, stroke, heart failure, coronary heart disease
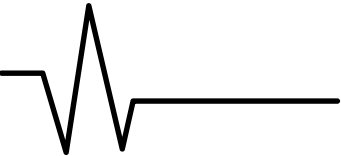   c. Split data (75%-25%)
2. **Exploratory Visuals**
3. **Model Training:**
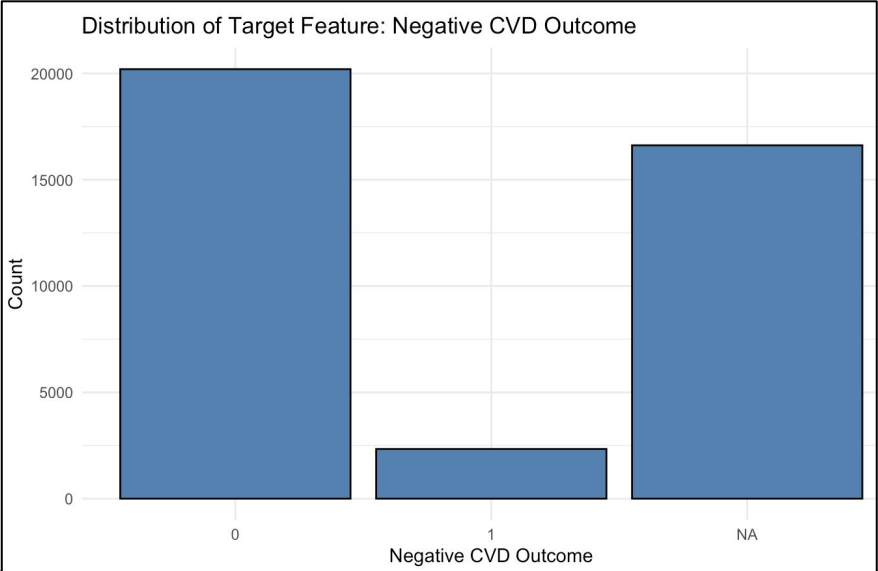   a. Logistic Regression (Existing PCE. vs. SDOH)
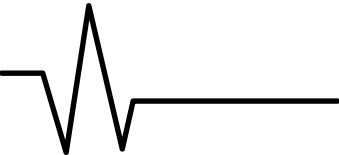4. **Model Tuning (hyper-parameters)**
5. **Testing & Evaluation**

Data Analysis

Distribution of Target Feature: Negative CVD Outcome

	No CVD Event (0)	Yes CVD Event (1)	NA
Overall	20,200 (51.59%)	2,337 (5.97%)	16,619 (42.44%)
Training Set	15,149 (51.59%)	1,726 (5.88%)	12,492 (42.54%)
Testing Set	5,051 (51.60%)	611 (6.24%)	4,127 (42.16%)

Data Analysis

Distribution of ASCVD Score by CVD Events

Data Analysis

Average ASCVD Scores by Race (Stratified)
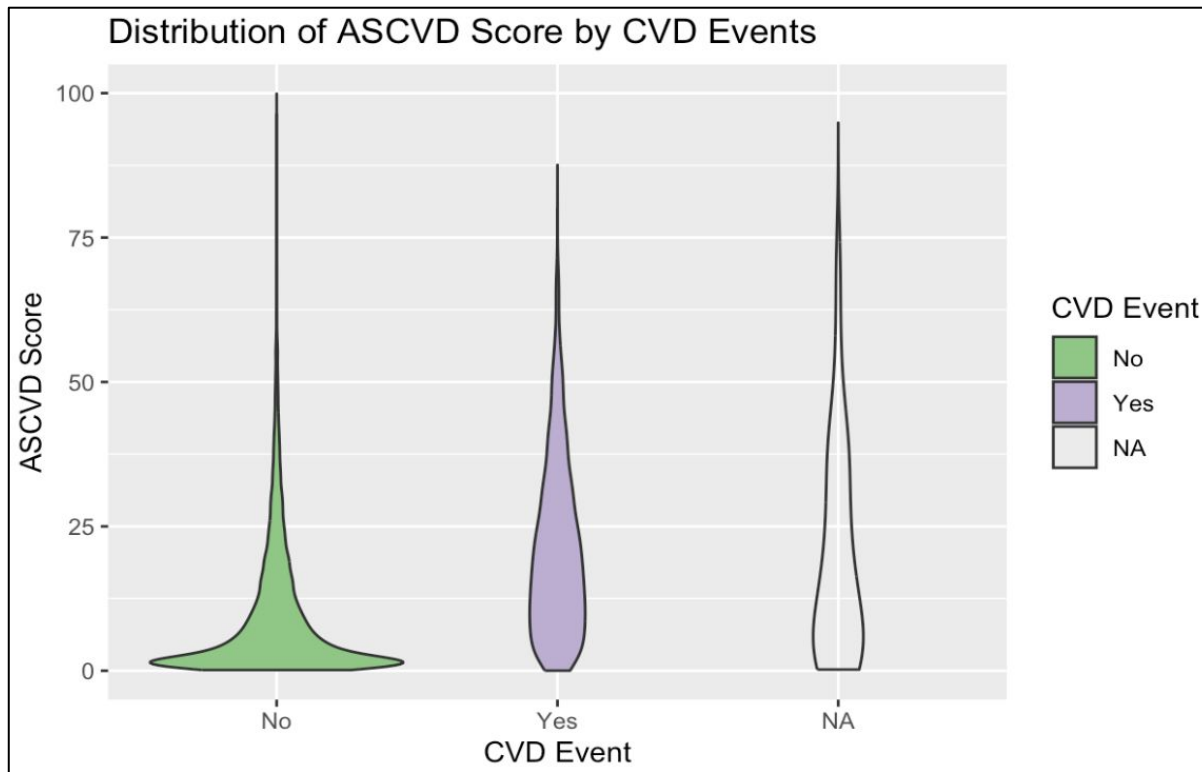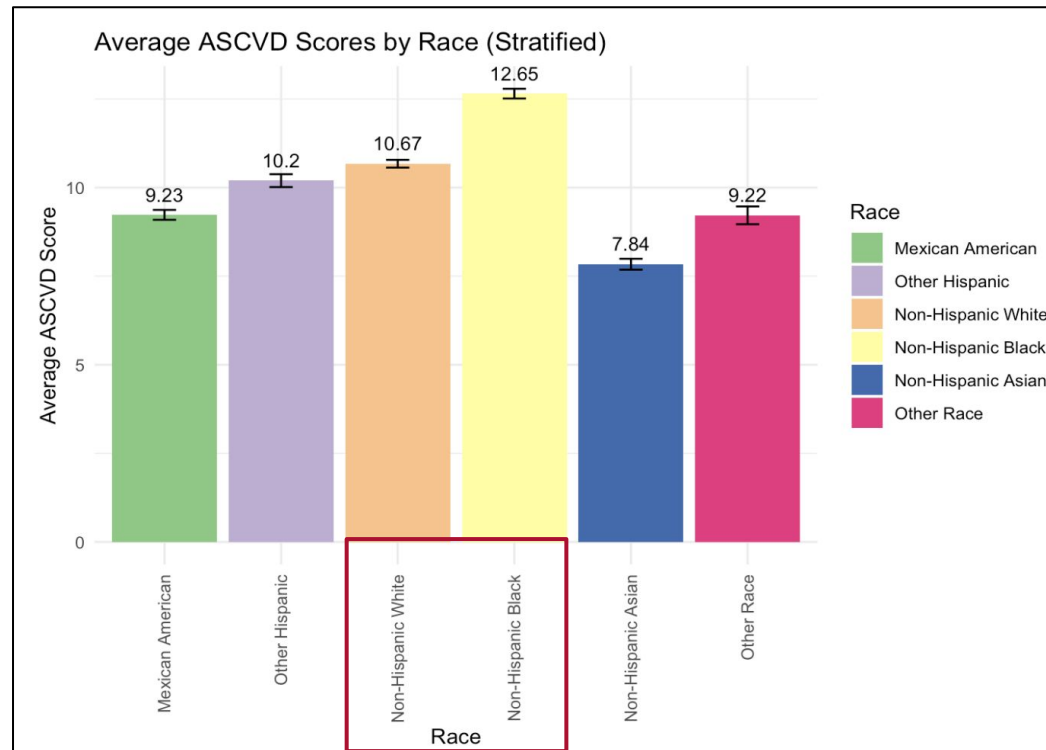
Results

PCE Variables

term	estimate	std.error	statistic	p.value
(Intercept)	-3.25796	0.09387	-34.707	< 2e-16*
Age	1.23491	0.04499	27.446	< 2e-16*
Sex	-0.14773	0.06785	-2.177	0.02945*
Race1	0.27233	0.07478	3.642	2.71e-04*
HDLChol	-0.24941	0.03636	-6.859	6.92e-12*
TotalChol	-0.26274	0.03421	-7.681	1.58e-14*
AvgSysBP	0.08139	0.03173	2.565	0.010319*
BPMed	-0.01581	0.07084	-0.223	0.823434
Diabetes	0.13513	0.07102	1.903	0.057081
Smoking	0.31486	0.06961	4.523	6.09e-06*
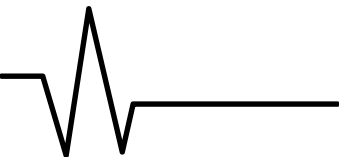
Results

PCE Variables

PCE	Accuracy	F1	Sensitivity	Specificity
Training Set	0.9006623	0.2580316	0.3218830	0.9334824
Testing Set	0.9054665	0.2901235	0.3868313	0.9327277

SDOH Variables

SDOH	Accuracy	F1	Sensitivity	Specificity
Training Set	0.9053731	0.3014113	0.3695921	0.9366961
Testing Set	0.9071106	0.3024691	0.4032922	0.9335929

Results

PCE Variables

ROC Curve

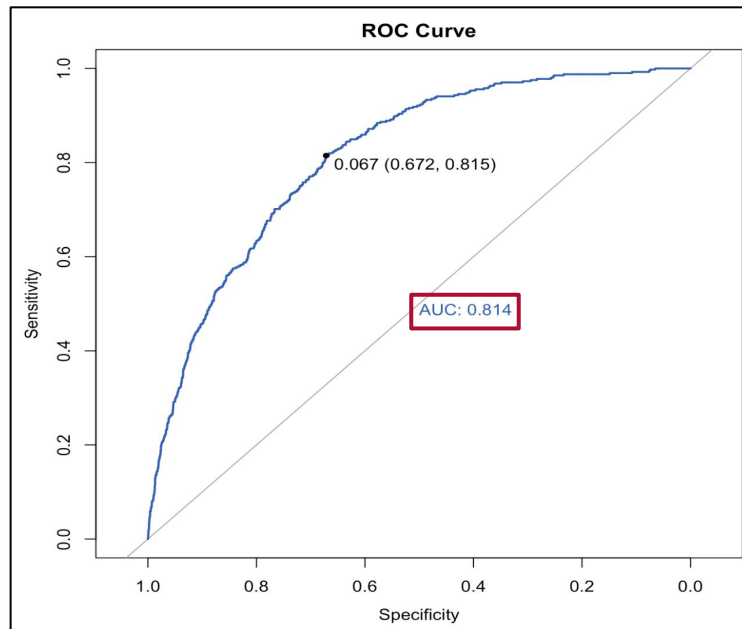0.066 (0.696, 0.812)

AUC: 0.831

SDOH Variables

ROC Curve

0.067 (0.672, 0.815)

AUC: 0.814
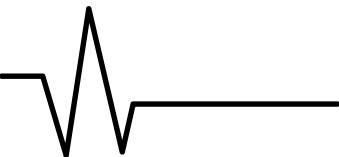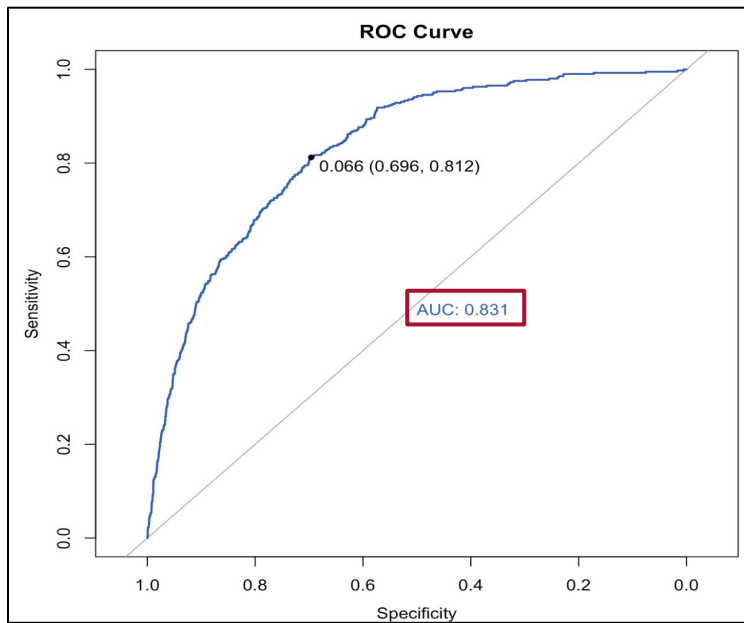
Conclusion and Next Steps

- Potentially use SMOTE to oversample negative outcomes.
- Add/Remove SDOH variables and re-run our models to observe how they impact model performances.
- Build and hyper-parameter tune 3rd machine learning model using only PCE variables and a stratified race variable to observe accuracy/performance.
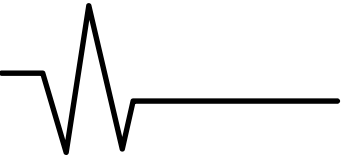
References

- *Fig. 1* "Ascvd risk estimator," in *MDedge Federal Practitioner*, vol. 31, no. 5, 2014.
- *Fig 2.* "Social determinants of health," in *UCLA Health: Sustainability*, 2018.