



Heart Disease Dataset

Group 7

Stephan Dupoux, Allie Schneider & Francis Michael Villamater



Contents

Dataset Overview

Project Objective

Dataset Variables

Exploratory Data
Analysis

Hyper Parameter Tuning

Model Performance Reports

Proposed solution

Process



Overview

- Data is sourced from the CDC as part of the Behavioral Risk Factor Surveillance System and was cleaned by a user on Kaggle.
- Heart Disease is one of the leading causes of death amongst most races in the US.
- About half of all Americans have at least 1 of 3 key risk factors for heart disease: high blood pressure, high cholesterol, and smoking.
- Other key indicators: diabetic status, obesity (high BMI), not enough physical activity, or excessive alcohol consumption
- 18 Attributes and 319,795 instances.





Project Objective

Leverage machine learning methods to detect “patterns” from the data that can predict a patient’s condition (presence or absence of heart disease). Adjust weights/undersampling/oversampling and apply classifier models (logistic regression, random forest, Naive Bayes)

Dataset Variables



- HeartDisease (binary)
- BMI (continuous) - Body Mass Index
- Smoking (binary) - have you smoked ≥ 100 cigarettes in your life?
- AlcoholDrinking (binary) - Yes if male and >14 drinks/week or female and >7 drinks/week
- Stroke (binary) - Ever had a stroke
- PhysicalHealth (Continuous) - including physical illness and injury, how many days of the past 30 was your physical health not good?
- MentalHealth (Continuous) - how many days of the past 30 was your mental health not good?
- DiffWalking (binary) - Do you have serious difficulty walking or climbing stairs?
- Sex (binary) - male or female
- AgeCategory (category, 13 levels) - Age category

Dataset Variables (cont'd)

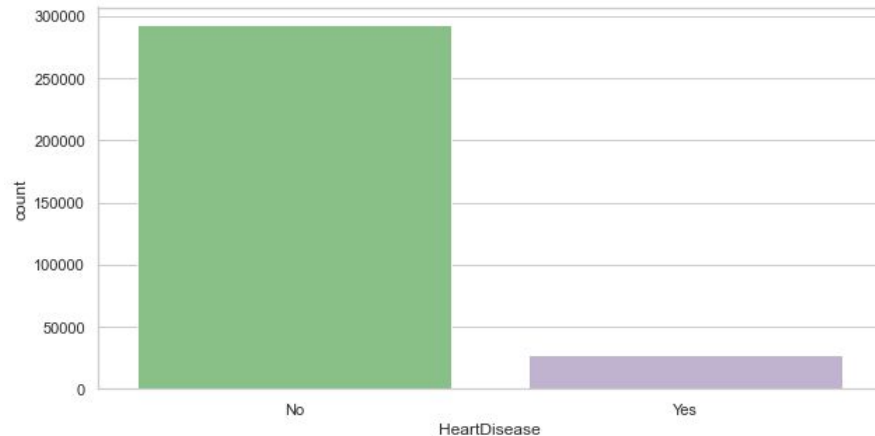
- Race (categorical, 6 levels) - Race/ethnicity
- Diabetic (categorical, 4 levels) - have they ever been diagnosed with diabetes?
- PhysicalActivity (binary) - Adults who reported engaging in physical activity/exercise over the last 30 days outside of their regular job
- GenHealth (categorical, 5 levels) - How would you rate their overall health?
- SleepTime (continuous) - how much sleep do they get in a 24 hour period?
- Asthma (binary) - have they been diagnosed with asthma?
- KidneyDisease (binary) - excluding kidney stones, bladder infection, or incontinence, have they ever been diagnosed with kidney disease?
- SkinCancer (binary) - have they ever been diagnosed with skin cancer?



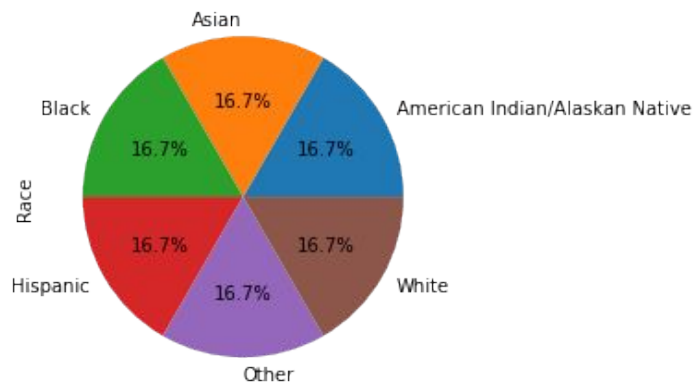
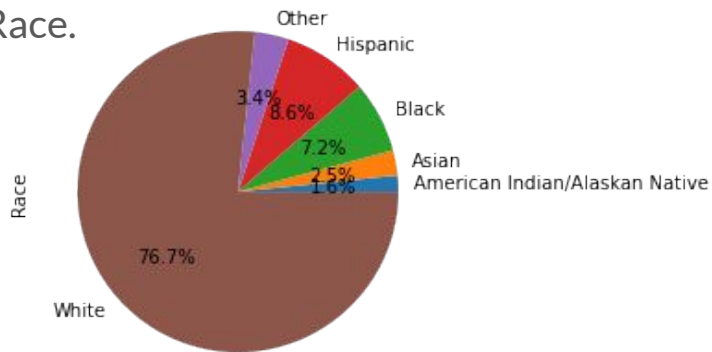
Exploratory Data Analysis

Heart Disease and Race

The dataset is highly imbalanced with most instances classified as “No” for Heart Disease (292422 negative, 27373 positive)



Overwhelming population of “White” in Race.





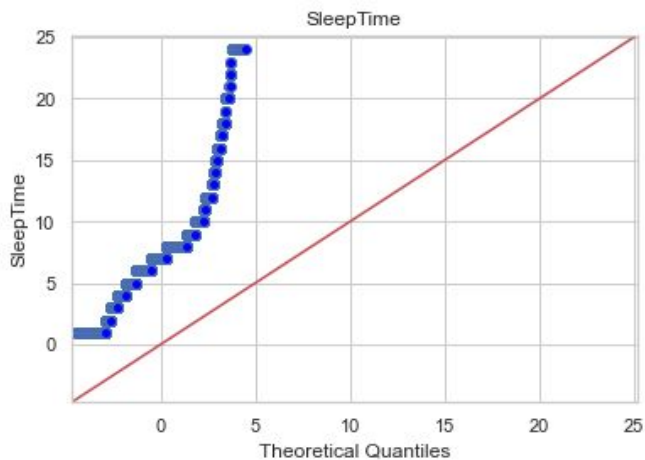
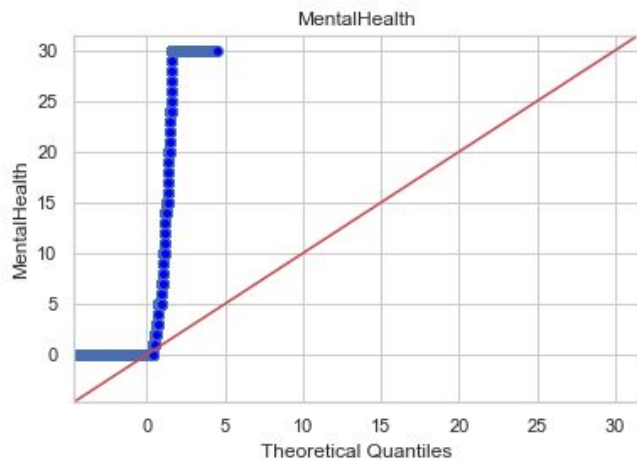
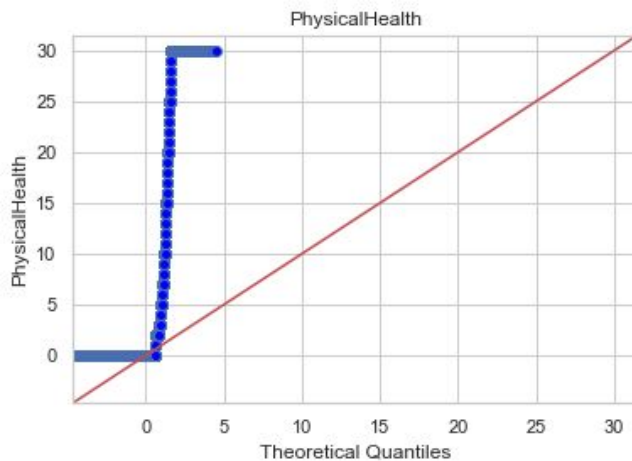
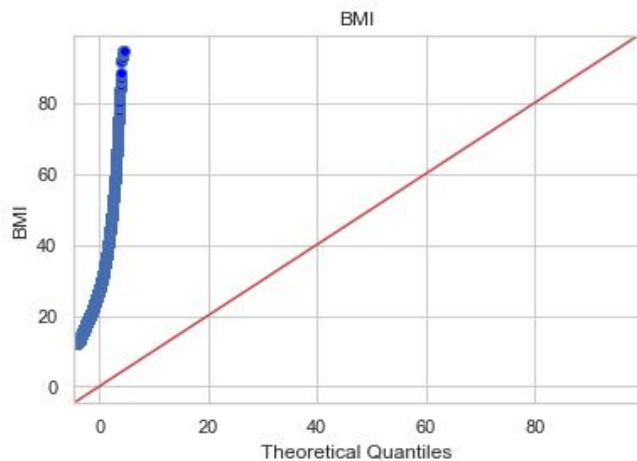
Handling Imbalanced Dataset:

- 292,422 Negative for Heart Disease
- 27,373 Positive for Heart Disease
- Resample on Race

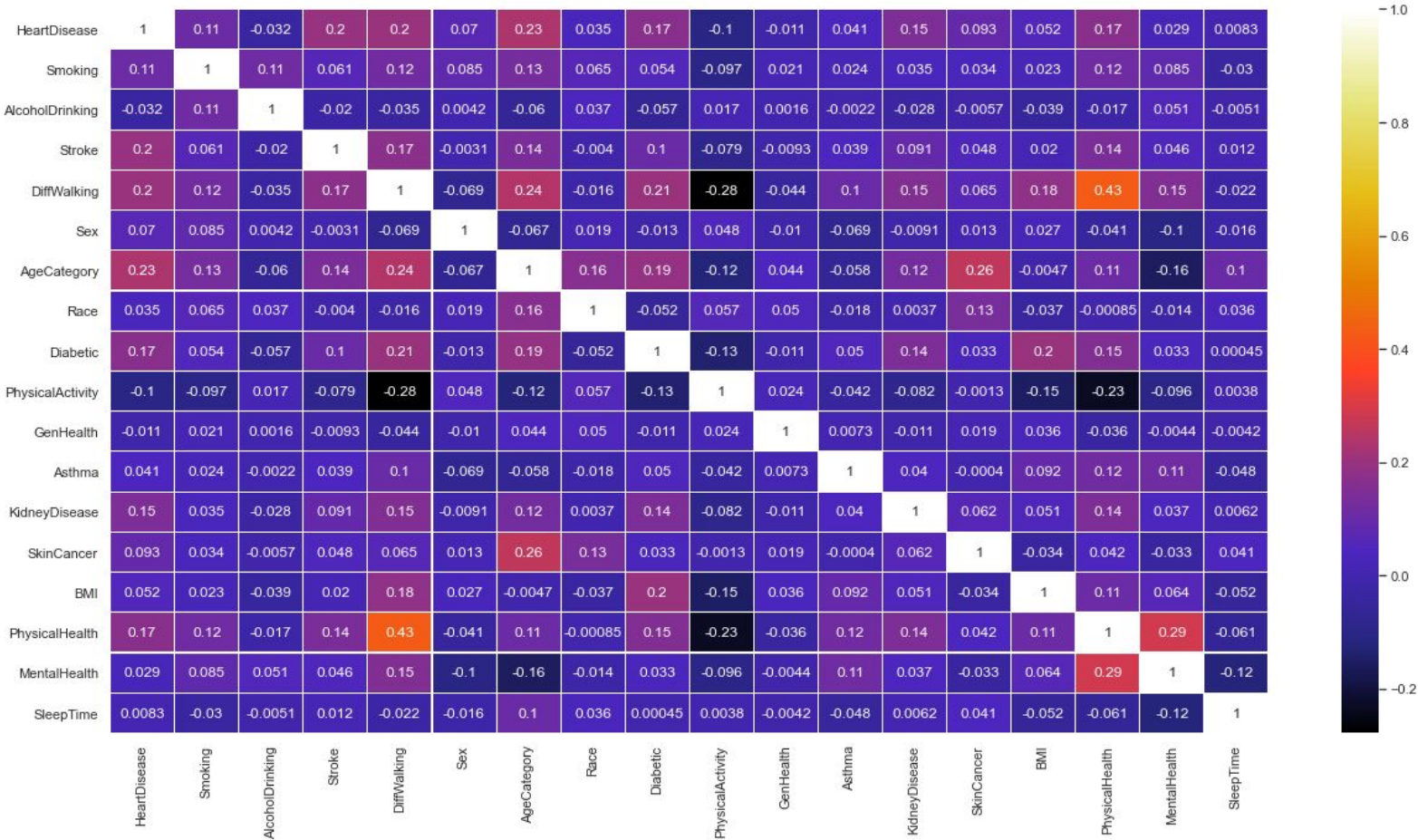
Continuous Variables

	BMI	PhysicalHealth	MentalHealth	SleepTime
count	319795.00	319795.00	319795.00	319795.00
mean	28.33	3.37	3.90	7.10
std	6.36	7.95	7.96	1.44
min	12.02	0.00	0.00	1.00
25%	24.03	0.00	0.00	6.00
50%	27.34	0.00	0.00	7.00
75%	31.42	2.00	3.00	8.00
max	94.85	30.00	30.00	24.00

Checking for Normality



Correlation Heat Map



Modeling



Models Chosen

- Logistic Regression
 - Simple to run and a good first step and helps select future models
 - Fastest to converge and not computationally intensive to train
- Random Forest
 - Chosen for its proficiency for handling high dimensional datasets
 - Algorithm will always converge with the datapoints
- Naive Bayes
 - Assumes no relation between the variables, and performs well with categorical variables.



Hyperparameter Tuning

- Logistic Regression C values: [0.05, 0.5, 0.9]
- Random Forest Max Leafs: [5, 10, 15]
- Random Forest Number of Estimators : [500, 1000, 2000]
- Naive Bayes alpha values: [0.05, 0.5, 0.9]



Results and Conclusion

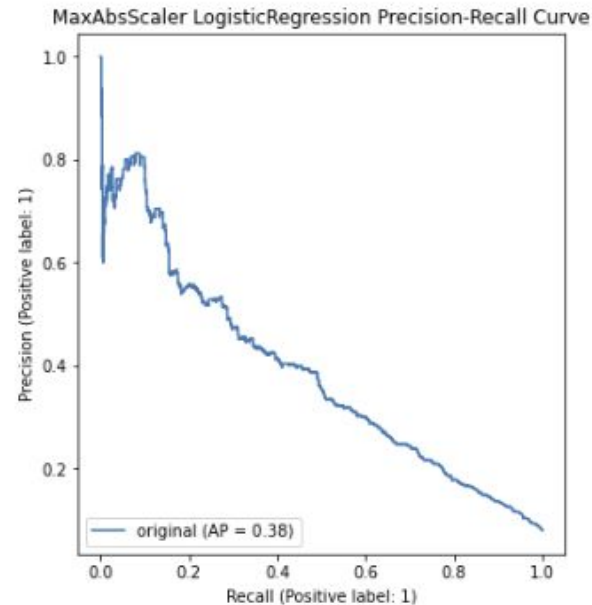
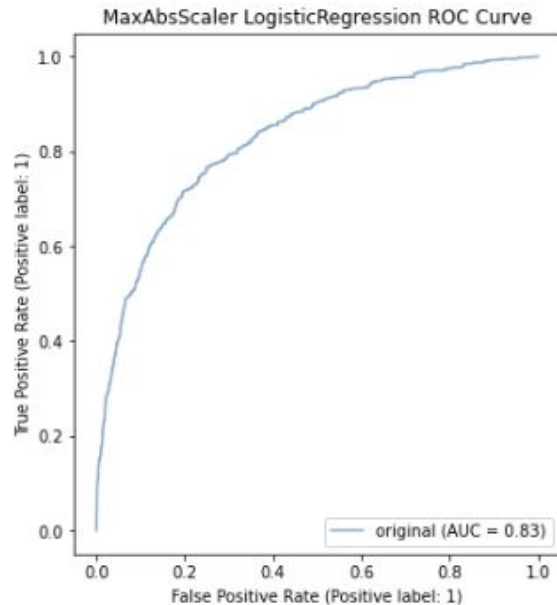


ML Models: Logistic Regression Algorithm

Scaler	MaxAbsScaler	PCA Dimensionality Reduction	Hyperparameter Tuning (C = 0.05)	Hyperparameter Tuning (C = 0.5)	Hyperparameter Tuning (C = 0.9)
Precision	0.5797	0.5678	0.5696	0.5786	0.5653
Recall	0.1712	0.1650	0.1674	0.1734	0.1690
F1	0.2643	0.2557	0.2587	0.2668	0.2603
False Positive Rate	0.0107	0.0107	0.0108	0.0108	0.0112
False Negative Rate	0.0715	0.0717	0.0714	0.0709	0.0717



ML Models: Logistic Regression Algorithm





ML Models: Logistic Regression Algorithm

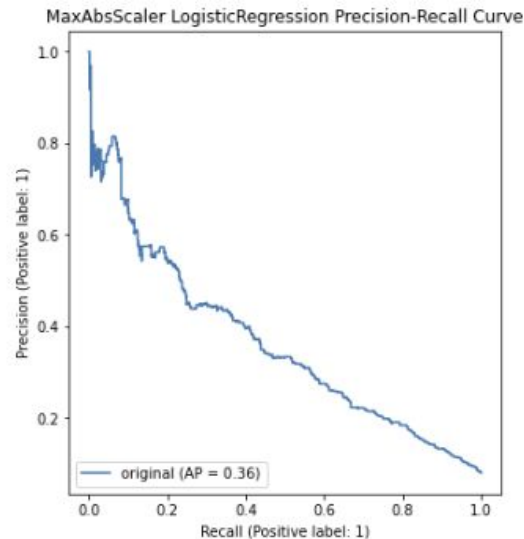
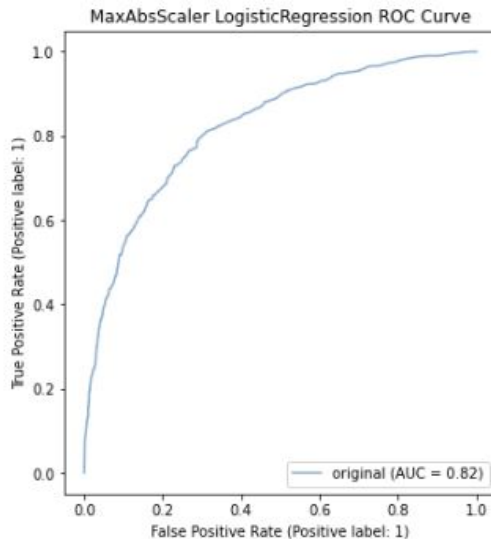
Log Transformation

Scaler	MaxAbsScaler	PCA Dimensionality Reduction	Hyperparam eter Tuning (C = 0.05)	Hyperparameter Tuning (C = 0.5)	Hyperparameter Tuning (C = 0.9)
Precision	0.5689	0.5637	0.5720	0.5657	0.5749
Recall	0.1556	0.1586	0.1551	0.1575	0.1562
F1	0.2443	0.2475	0.2441	0.2465	0.2457
False Positive Rate	0.0101	0.0105	0.0099	0.0103	0.0098
False Negative Rate	0.0729	0.0723	0.0722	0.0721	0.0722



ML Models: Logistic Regression Algorithm

Log Transformation





ML Models: Logistic Regression Algorithm

Feature Scaling

Scaler	MaxAbsScaler	PCA Dimensionality Reduction	Hyperparameter Tuning (C = 0.05)	Hyperparameter Tuning (C = 0.5)	Hyperparameter Tuning (C = 0.9)
Threshold	0.5787	0.4881	0.5470	0.5760	0.5823

Log Transformation and Feature Scaling

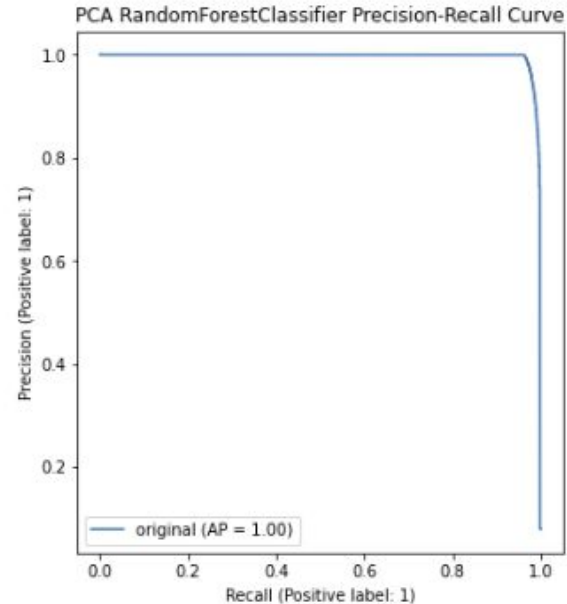
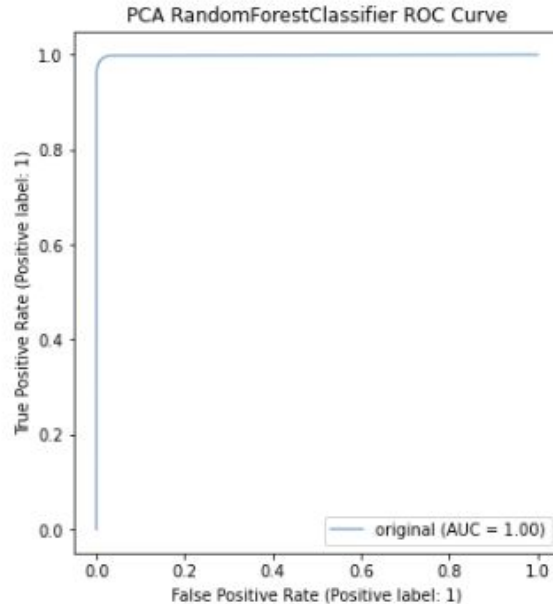
Scaler	MaxAbsScaler	PCA Dimensionality Reduction	Hyperparameter Tuning (C = 0.05)	Hyperparameter Tuning (C = 0.5)	Hyperparameter Tuning (C = 0.9)
Threshold	0.5801	0.5790	0.5216	0.5244	0.5801



ML Models: Random Forest Classifier Algorithm

Scaler	MaxAbsScaler	PCA Dimensionality Reduction	Hyperparameter Tuning (n_estimators = 500 / max_leaf_nodes = 5)	Hyperparameter Tuning (n_estimators = 1000 / max_leaf_nodes = 10)	Hyperparameter Tuning (n_estimators = 2000 / max_leaf_nodes = 15)
Precision	0.9930	0.9964	1.0	0.9330	0.9521
Recall	0.9645	0.9643	0.0	0.0104	0.0612
F1	0.9785	0.9801	0.0	0.0207	0.1150
False Positive Rate	0.0005	0.0002	0.0	6.4271	0.0002
False Negative Rate	0.0030	0.0030	0.0858	0.0846	0.0805

ML Models: Random Forest Classifier Algorithm



ML Models: Random Forest Classifier Algorithm

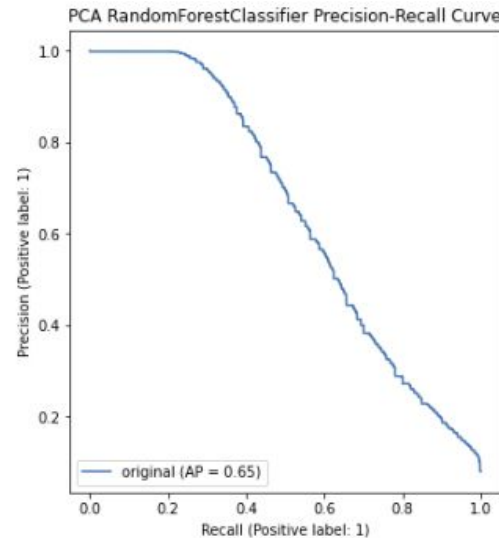
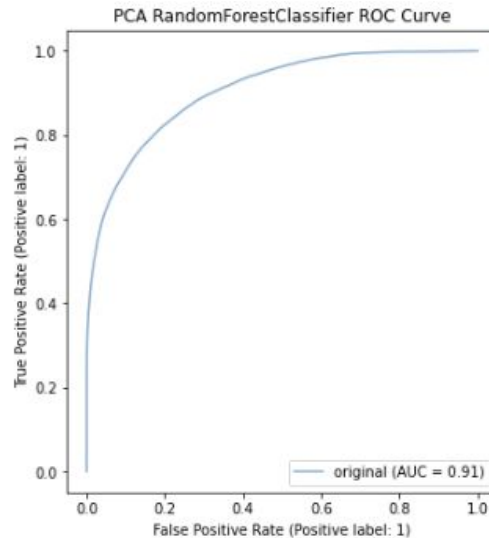
Log Transformation

Scaler	MaxAbsScaler	PCA Dimensionality Reduction	Hyperparameter Tuning (n_estimators = 500 / max_leaf_nodes = 5)	Hyperparameter Tuning (n_estimators = 1000 / max_leaf_nodes = 10)	Hyperparameter Tuning (n_estimators = 2000 / max_leaf_nodes = 15)
Precision	0.8261	0.8339	1.0	0.8413	0.8924
Recall	0.4036	0.4064	0.0	0.0139	0.0372
F1	0.5423	0.5465	0.0	0.0273	0.0715
False Positive Rate	0.0072	0.0069	0.0	0.0002	0.0003
False Negative Rate	0.0511	0.0510	0.0856	0.0838	0.0826



ML Models: Random Forest Classifier Algorithm

Log Transformation



ML Models: Random Forest Classifier Algorithm

Feature Scaling

Scaler	MaxAbsScaler	PCA Dimensionality Reduction	Hyperparameter Tuning (n_estimators = 500 / max_leaf_nodes = 5)	Hyperparameter Tuning (n_estimators = 1000 / max_leaf_nodes = 10)	Hyperparameter Tuning (n_estimators = 2000 / max_leaf_nodes = 15)
Threshold	0.0400	0.04	0.04	0.04	0.04

Log Transformation and Feature Scaling

Scaler	MaxAbsScaler	PCA Dimensionality Reduction	Hyperparameter Tuning (n_estimators = 500 / max_leaf_nodes = 5)	Hyperparameter Tuning (n_estimators = 1000 / max_leaf_nodes = 10)	Hyperparameter Tuning (n_estimators = 2000 / max_leaf_nodes = 15)
Threshold	0.05	0.0499	0.0500	0.0499	0.0499

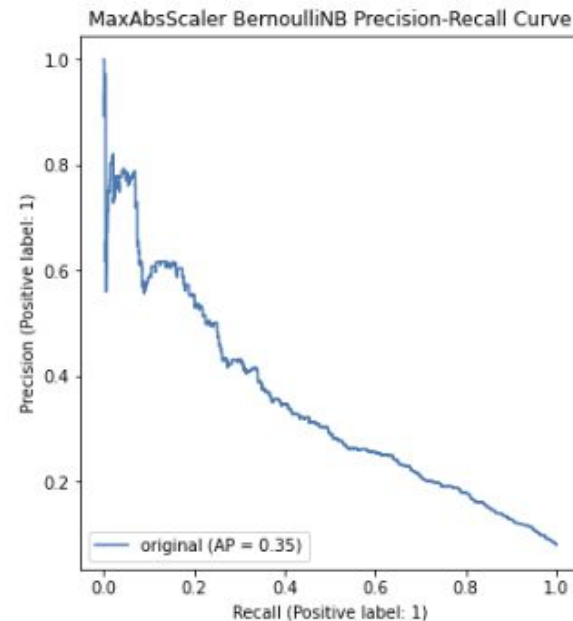
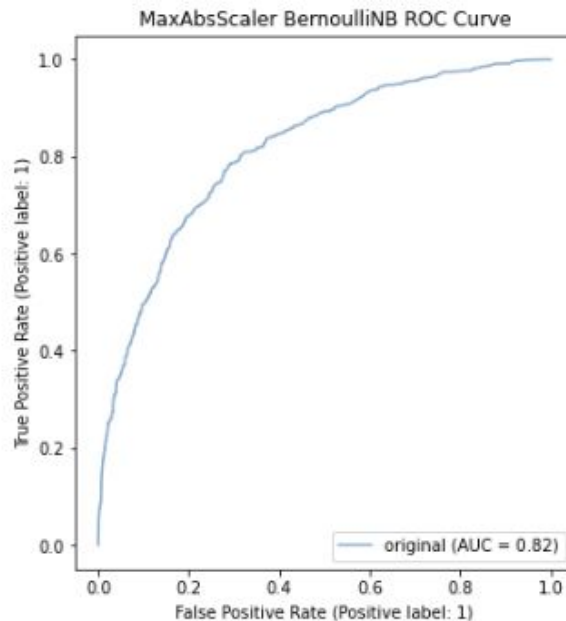


ML Models: Naive Bayes Algorithm

Scaler	MaxAbsScaler	PCA Dimensionality Reduction	Hyperparameter Tuning (alpha = 0.05)	Hyperparameter Tuning (alpha = 0.5)	Hyperparameter Tuning (alpha = 0.9)
Precision	0.3217	0.3188	0.3198	0.3208	0.3189
Recall	0.4454	0.4393	0.4458	0.4514	0.4450
F1	0.3736	0.3695	0.3725	0.3750	0.3715
False Positive Rate	0.0806	0.0807	0.0812	0.0816	0.0818
False Negative Rate	0.0476	0.0482	0.0474	0.0468	0.0477



ML Models: Naive Bayes Algorithm





ML Models: Naive Bayes Algorithm

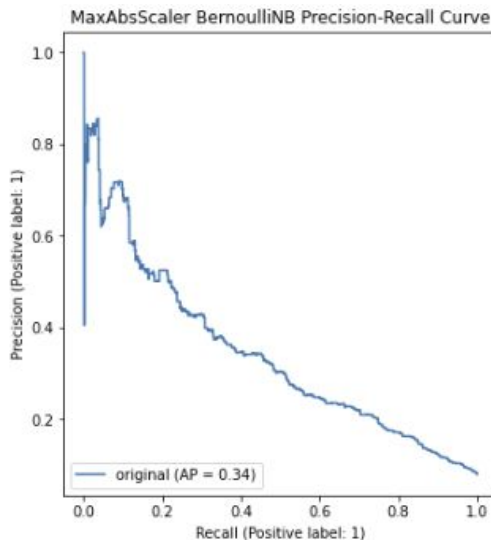
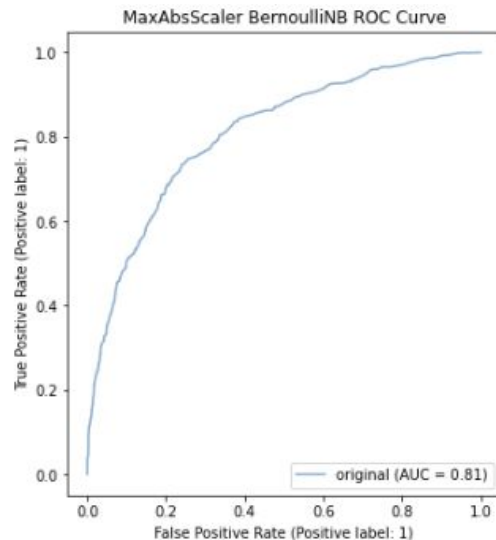
Log Transformation

Scaler	MaxAbsScaler	PCA Dimensionality Reduction	Hyperparam eter Tuning (alpha = 0.05)	Hyperparameter Tuning (alpha = 0.5)	Hyperparameter Tuning (alpha = 0.9)
Precision	0.3521	0.3566	0.3582	0.3553	0.3575
Recall	0.3826	0.3899	0.3866	0.3901	0.3894
F1	0.3667	0.3725	0.3718	0.3719	0.3728
False Positive Rate	0.0604	0.0605	0.0594	0.060	0.0602
False Negative Rate	0.0529	0.0524	0.0526	0.0523	0.0525



ML Models: Naive Bayes Algorithm

Log Transformation



ML Models: Naive Bayes Algorithm

Feature Scaling

Scaler	MaxAbsScaler	PCA Dimensionality Reduction	Hyperparameter Tuning (alpha = 0.05)	Hyperparameter Tuning (alpha = 0.5)	Hyperparameter Tuning (alpha = 0.9)
Threshold	1.6284	1.6872	1.6291	1.6274	1.6287

Log Transformation and Feature Scaling

Scaler	MaxAbsScaler	PCA Dimensionality Reduction	Hyperparameter Tuning (alpha = 0.05)	Hyperparameter Tuning (alpha = 0.5)	Hyperparameter Tuning (alpha = 0.9)
Threshold	1.9146	1.9136	1.9139	1.9146	1.9134



ANOMALY DETECTION

- Used algo called Isolation Forest on continuous data
- Wanted to see if there was any unexpected data points within the dataset
- Aims to see how well the algorithm performs with the prediction of anomalies.




ML Models: Isolation Forest

Scaler	None
Precision	0.1459
Recall	0.5436
F1	0.2301
False Positive Rate	0.2731
False Negative Rate	0.0392



Recommendations

- Health-centric data, more specific to the individual
 - HIPAA compliance will be an significant obstacle
- Identifying socio-economic status
 - People with greater income, will have access to wider resources to better their health
- Specifying location (zip code)
 - Certain cities/states may have better health/Medicaid programs than others.



Thank you

