Assessing heart disease and heart attack risk using self-reported health, lifestyle, and demographic data

Cindy Zheng
Brown University
https://github.com/czheng27/data1030-midterm/tree/main

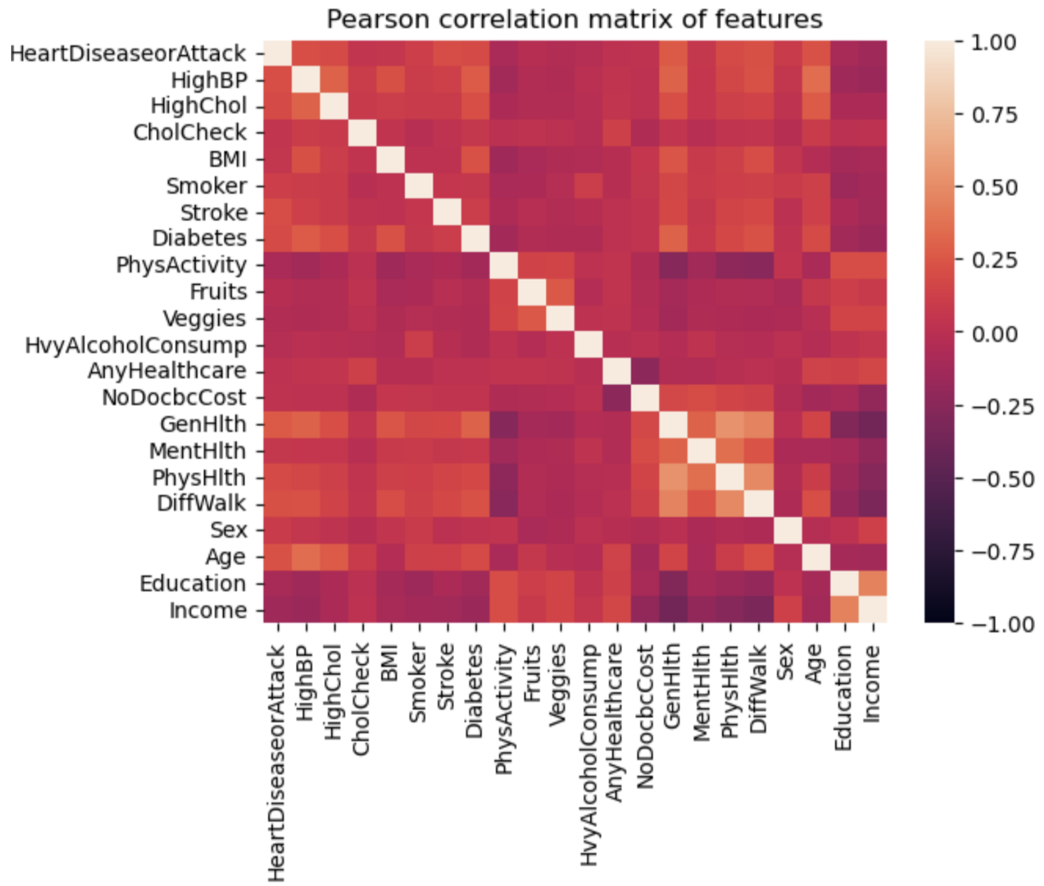
INTRODUCTION

Heart disease is the leading cause of death in the US [1]. Heart disease encapsulates numerous cardiovascular conditions, many of which can lead to heart attacks [1]. Risk factors for heart disease include age, sex, smoking, family history, an unhealthy diet, high blood pressure, and high cholesterol (2). Treatment includes medication and surgery, although lifestyle changes can prevent or significantly improve heart disease outcomes [3,4]. Annually, heart disease and stroke costs the US around $329.7 billion for diagnosis and care [5]. Thus, a predictive model to assess a patient's risk of developing heart disease or having a heart attack holds the potential to not only improve patients' health, but also cut costs for both the patient and the healthcare system. Existing risk calculators, such as the Atherosclerotic Cardiovascular Disease (ASCVD) risk estimator rely on tests (such as a patient's cholesterol panel) which a patient may struggle to receive without a healthcare provider, limiting the accessibility of the tool [6].

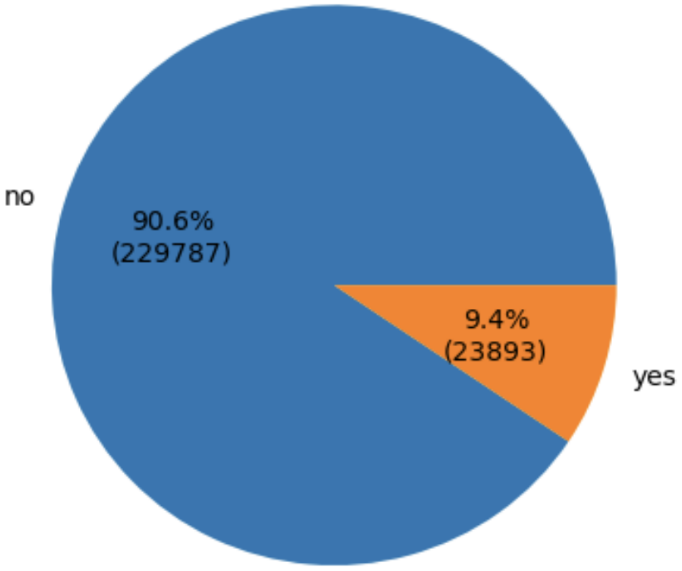
The models in this paper use a cleaned version of the 2015 Behavioral Risk Factor Surveillance System (BRFSS) provided by Alex Teboul on Kaggle [7]. The BRFSS is an annual nationwide telephone survey conducted by the CDC to assess the distribution of health-related risk factors, chronic conditions, and use of preventative services across the US [7]. A quasibinomial model trained by Dolezel et al. on the 2019 BRFSS assessing only heart attack for adults 35 and older using over 50 of the demographic, geographic, socioeconomic, and health related factors achieved an F1 score of 0.898 [8]. As this model engages with far more variables than available in our cleaned dataset and only focuses on heart attack risk, its impressive performance provides hope for the predictive power of the information within the BRFSS dataset and may emphasize the importance of various factors to heart disease and heart attacks.

EDA

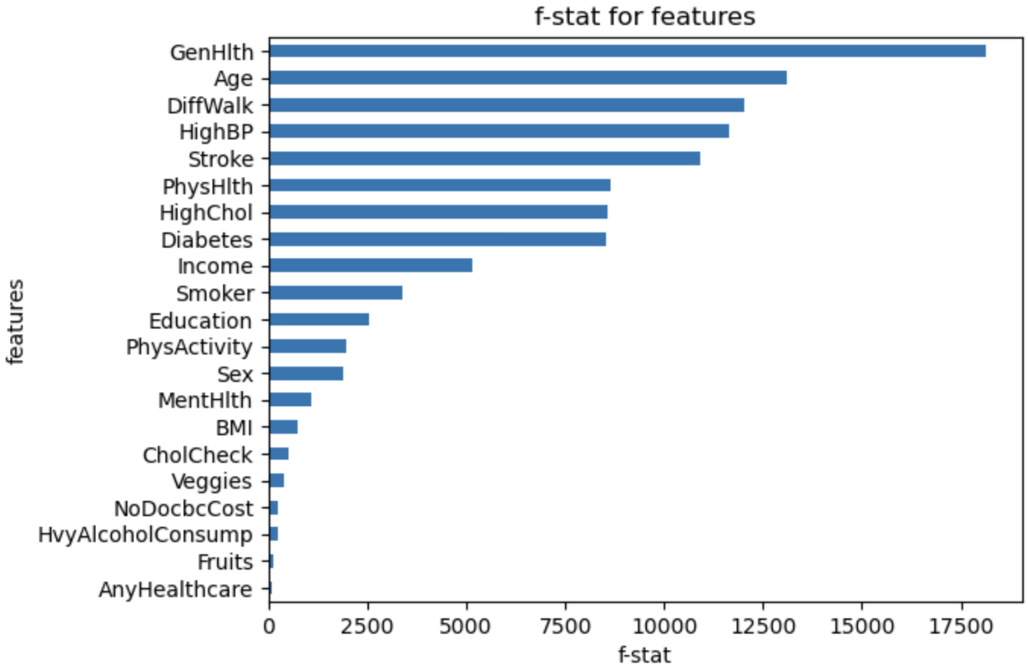
The dataset consists of 253,680 responses and 22 features. The dataset had already been cleaned such that there were no missing values in any of the features as a result. The target variable, "HeartDiseaseorAttack", encapsulates both causes where a patient has been told by a provider they have heart disease and/or a history of heart attacks: 0.0 represents "no" and 1.0 represents "yes". Thus, we are working with a classification problem. This target variable is highly unbalanced in this dataset with the majority of responses for the target variable being 0.0/"no".


Pearson correlation matrix of features
HeartDiseaseorAttack
HighBP
HighChol
CholCheck
BMI
Smoker
Stroke
Diabetes
PhysActivity
Fruits
Veggies
HvyAlcoholConsump
AnyHealthcare
NoDocbcCost
GenHlth
MentHlth
PhysHlth
DiffWalk
Sex
Age
Education
Income
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

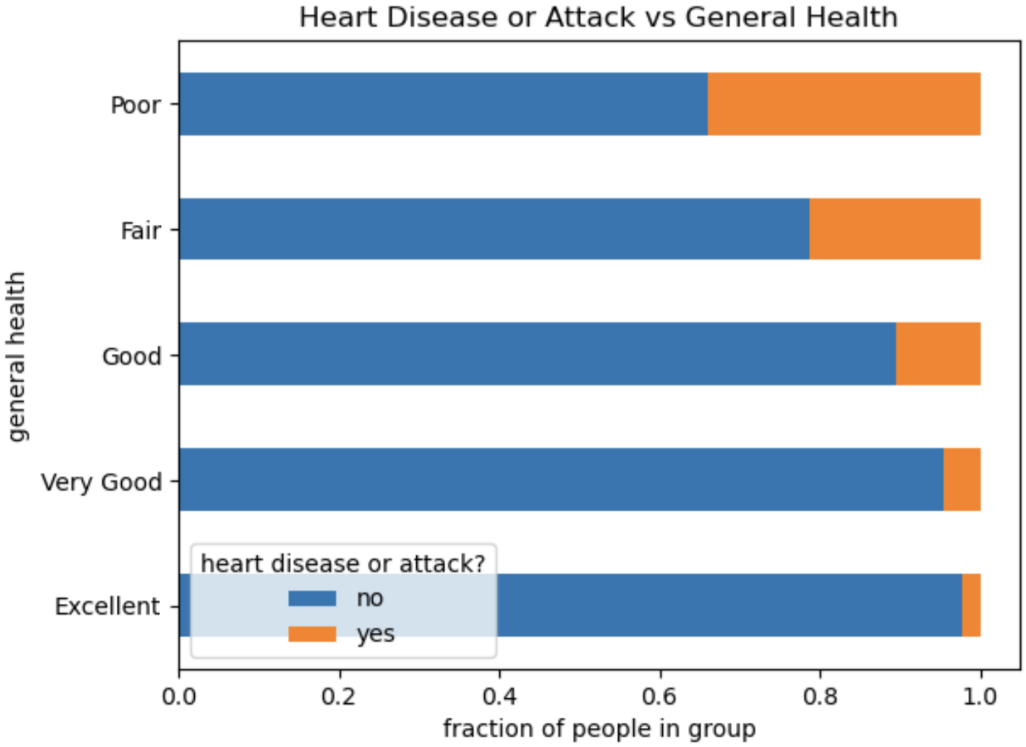
<Fig 1: Pearson Correlation Matrix of features. This correlation matrix shows that no features are significantly correlated with each other. Thus, no features were dropped.>


no
90.6%
(229787)
9.4%
(23893)
yes

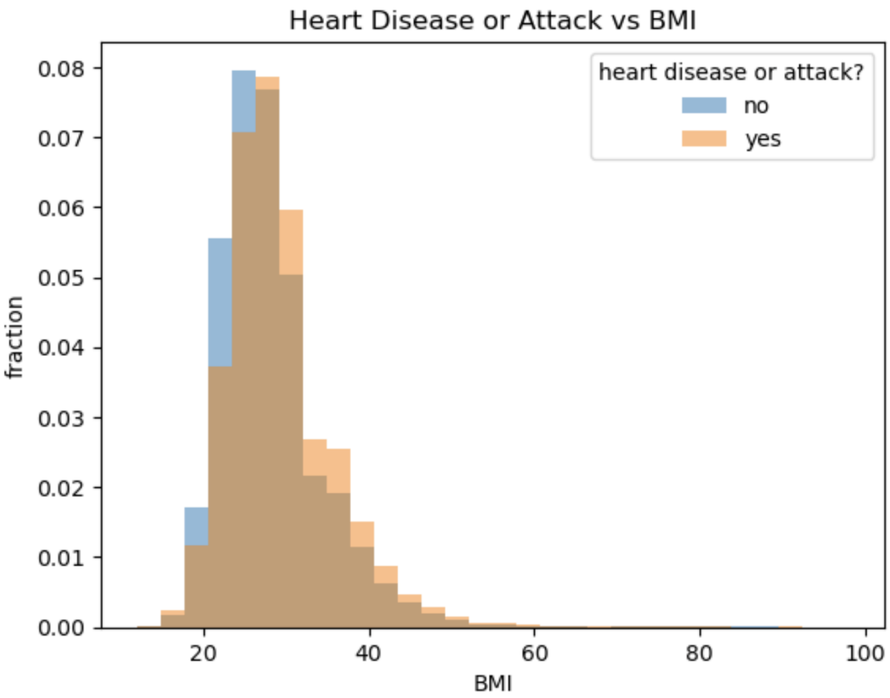
<Fig 2: Pie chart of counts for the target variable “HeartDiseaseorAttack”. This figure shows that the target variable is highly unbalanced, favoring 0.0/“no” outcomes>


f-stat for features
features
GenHlth
Age
DiffWalk
HighBP
Stroke
PhysHlth
HighChol
Diabetes
Income
Smoker
Education
PhysActivity
Sex
MentHlth
BMI
CholCheck
Veggies
NoDocbcCost
HvyAlcoholConsump
Fruits
AnyHealthcare
0
2500
5000
7500
10000
12500
15000
17500
f-stat

<Fig 3: F-stats for features graph: This graph shows the top features that impact the target variable. In this case, General Health, Age, and Difficulty Walking are the top 3 factors.>


Heart Disease or Attack vs General Health
general health
Poor
Fair
Good
Very Good
Excellent
heart disease or attack?
no
yes
0.0
0.2
0.4
0.6
0.8
1.0
fraction of people in group

<Fig 4: HeartDiseaseorAttack vs General Health: As one of the top F-stat features, General Health shows a close connection to the HeartDiseaseorAttack distribution. As general health improves, the proportion of individual with heart disease or a history of heart attack decreases. >



<Fig 5. HeartDiseaseorAttack vs BMI: This graph shows that the HeartDiseaseorAttack has little connection to BMI, a factor often connected to chronic conditions like Heart Disease>

METHODS

SPLITTING

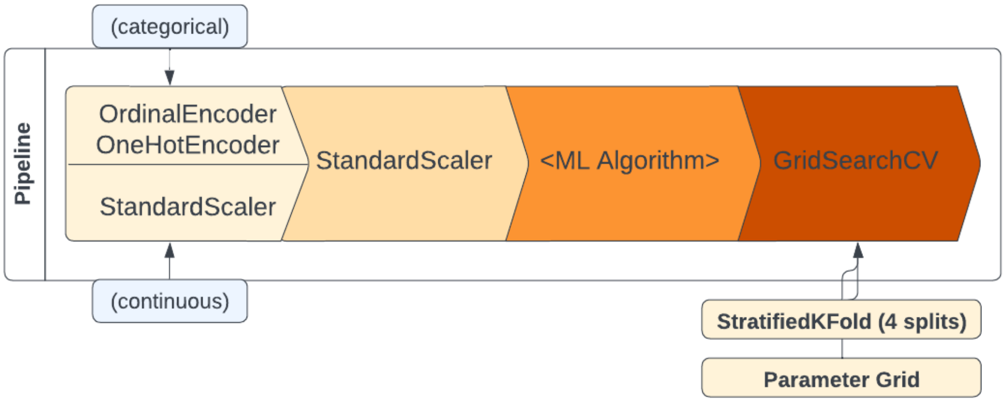
The dataset was initially split 98:2 while stratified across the target variable to provide a training/validation set and testing set appropriate for the large size of the dataset and the unbalanced nature of the target variable. A 4 fold StratifiedKFold split, to account for the unbalanced target variable, was performed upon the training/validation set within the model hyperparameter tuning processes for cross validation purposes.

PREPROCESSING

StandardScaler was used on all continuous variables and OrdinalEncoder and OneHotEncoder were used for ordinal and non-ordinal categorical variables respectively. After this preprocessing was applied, all preprocessed variables were then scaled with StandardScaler for the sake of the coefficients of the linear models trained.

PIPELINE

GridSearchCV was used for hyperparameter tuning. As a result, the pipeline given to GridSearchCV consisted of the preprocessing described above, followed by the machine learning algorithm of choice. The StratifiedKFold as part of splitting the data was performed within GridSearchCv upon the provided parameter grid.



<Fig 6: A visual depiction of the pipeline>

METRIC

The desired metric for these models was the F2 score. This is due to the unbalanced target variable and our aim to reduce the number of false negatives (we don't want to miss those at risk for heart disease or heart attacks, and interventions are not particularly costly nor harmful). The baseline F2 score was calculated by assuming the model only predicted 1.0/"yes".

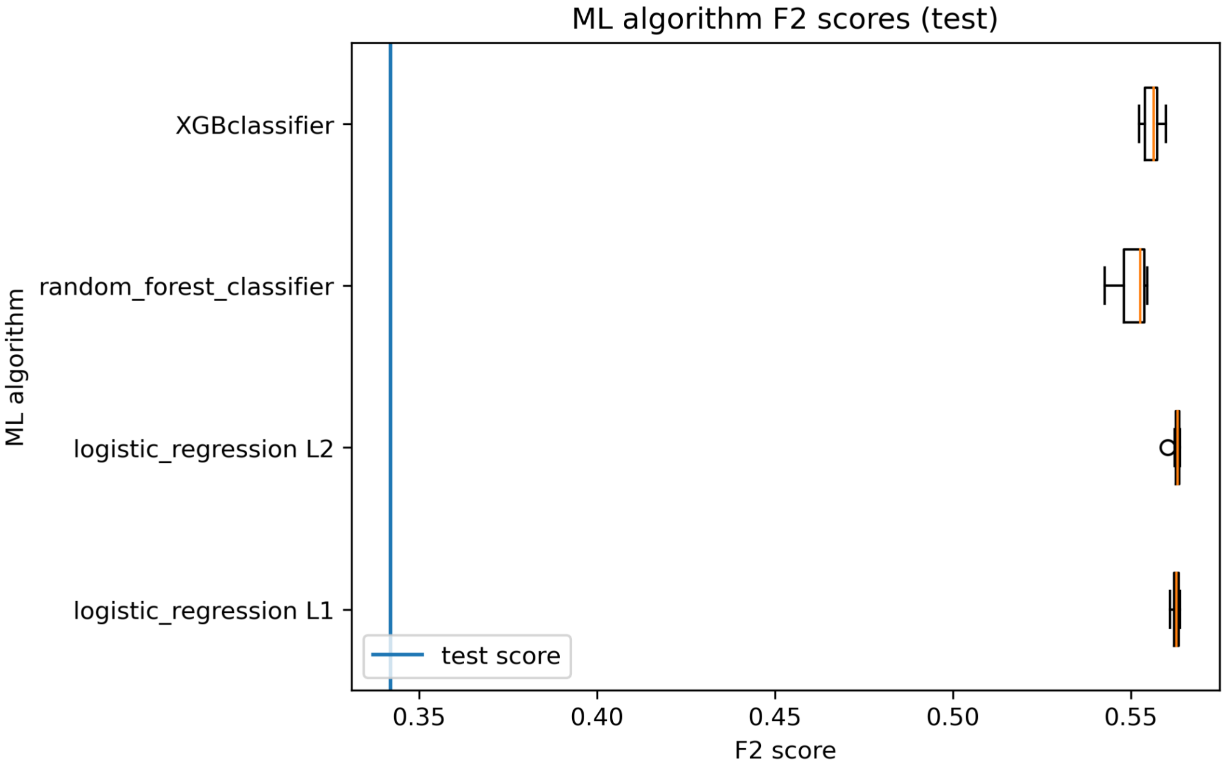
ML Algorithm	**Parameters tuned**
Logistic Regression (L1)	C: [0.01, 0.1, 1, 10, 100] class_weight = [None, 'balanced']
Logistic Regression (L2)	C: [0.01, 0.1, 1, 10, 100] class_weight = [None, 'balanced']
RandomForestClassifier	max_depth: [None, 1, 10, 100] max_features: [None, 0.33, 0.66, 1.0] class_weight: [None, 'balanced', 'balanced_subsample']
XGBClassifier	reg_alpha: [0.1, 1, 10] reg_lambda: [0.1, 1, 10] max_depth: [1, 10, 100] scale_pos_weight: [1, 10]

<Fig 7: A chart of the algorithms trained and the hyperparameters tuned>

RESULTS

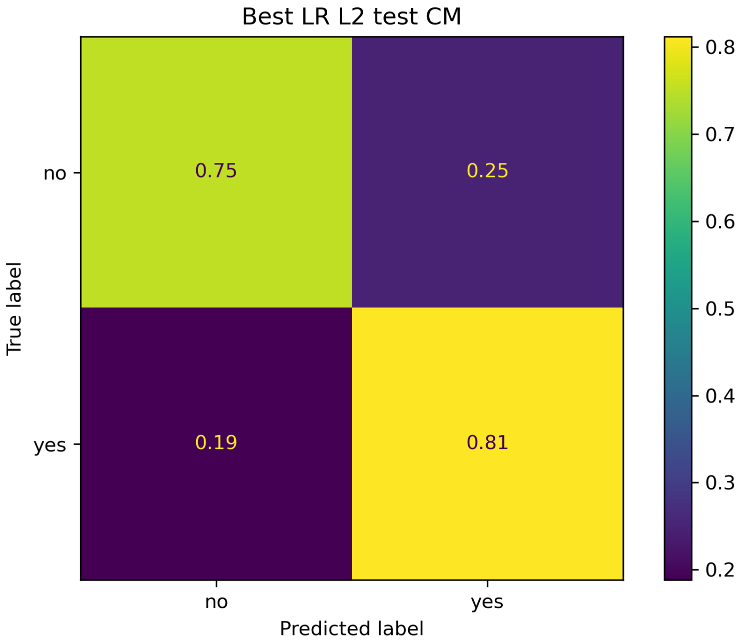
Model	**Mean (test)**	**Standard Deviation (test)**	**Standard deviations from baseline**
LogisticRegression (L1)	0.5626	0.00087	1884
LogisticRegression (L2)	0.5628	0.00100	1653
RandomForestClassifier	0.5509	0.00392	411
XGBClassifier	0.5561	0.00278	586

<Fig 8: A chart of the models trained and their mean test scores, standard deviation of test scores, and standard deviations from the baseline when applied to the test set. The means of each model were very similar and the standard deviations were similarly small for each model.>

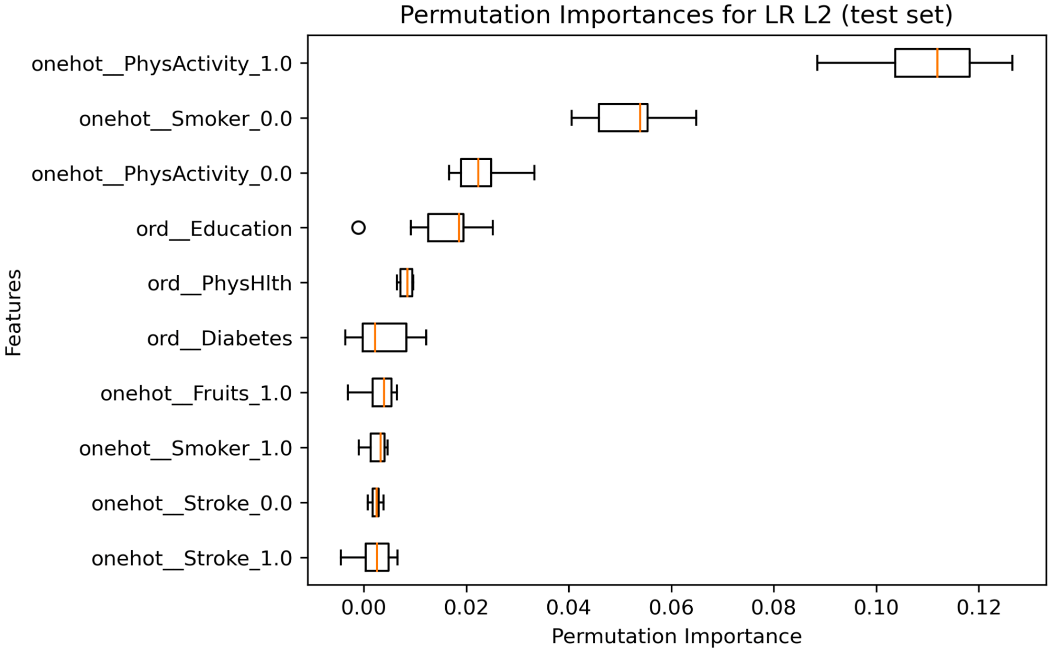


<Fig 9: A chart of the F2 scores for the various algorithms versus the baseline F2 score.>

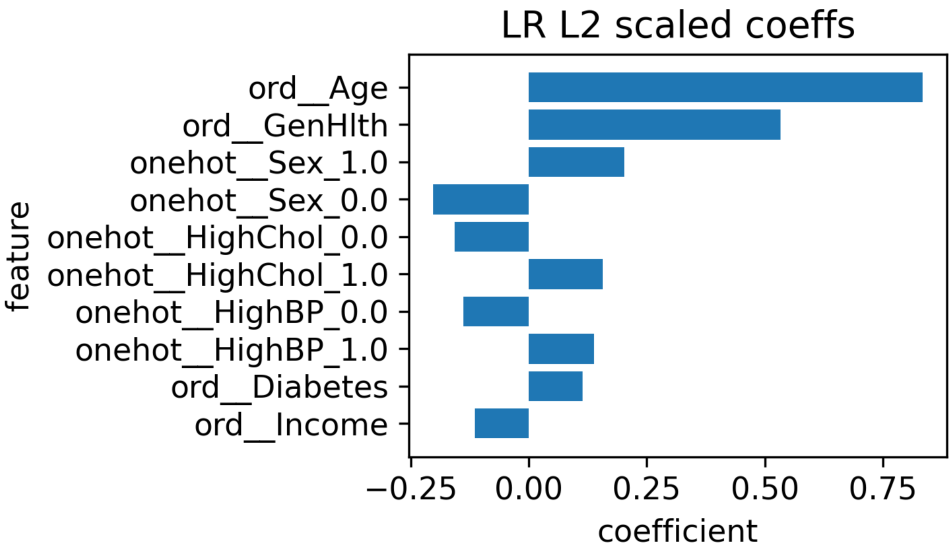
All models showed significantly higher F2 scores than the baseline value of 0.3421. Standard deviations were similarly very low across all models. Based on the test scores, the Logistic Regression model run with L2 regularization was the most predictive model with the largest mean F2 score and an appropriately low standard deviation.


Best LR L2 test CM
True label
no
yes
0.75
0.25
0.19
0.81
no
yes
Predicted label
0.8
0.7
0.6
0.5
0.4
0.3
0.2

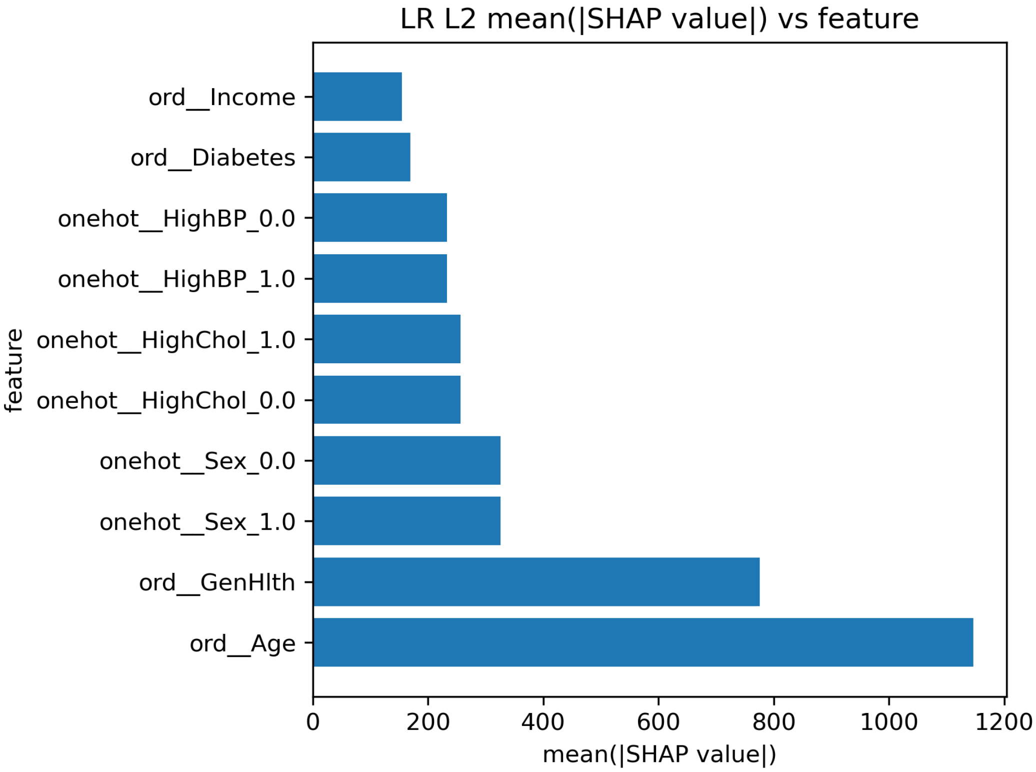
<Fig 10: A confusion matrix for the logistic regression model (L2 regularization) that performed best on the test set. The model performed relatively well on the test set. Only 20% of the true positives were falsely negative.>


Permutation Importances for LR L2 (test set)
Features
onehot__PhysActivity_1.0
onehot__Smoker_0.0
onehot__PhysActivity_0.0
ord__Education
ord__PhysHlth
ord__Diabetes
onehot__Fruits_1.0
onehot__Smoker_1.0
onehot__Stroke_0.0
onehot__Stroke_1.0
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Permutation Importance

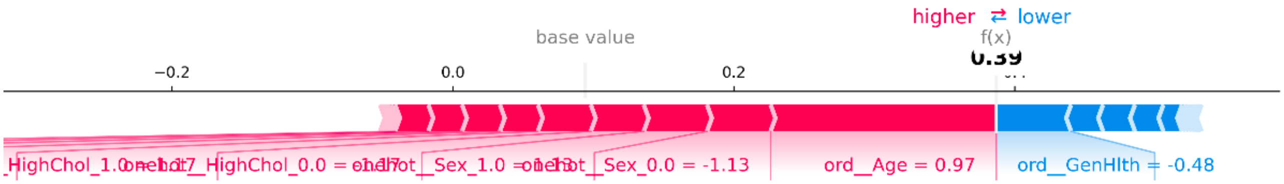
<Fig 11: Permutation importances for the features of the logistic regression model (L2 regularization) that performed best on the test set. Physical Activity, if one is a Smoker, Education, Physical Health, and if one has Diabetes are the most important features.>


LR L2 scaled coeffs
feature
ord__Age
ord__GenHlth
onehot__Sex_1.0
onehot__Sex_0.0
onehot__HighChol_0.0
onehot__HighChol_1.0
onehot__HighBP_0.0
onehot__HighBP_1.0
ord__Diabetes
ord__Income
−0.25
0.00
0.25
0.50
0.75
coefficient

<Fig 12: Scaled coefficients for the logistic regression model (L2 regularization) that performed best on the test set. Age, General Health, Sex, if one has High Cholesterol, and if one has High Blood Pressure are the most important features.>


LR L2 mean(|SHAP value|) vs feature
feature
ord__Income
ord__Diabetes
onehot__HighBP_0.0
onehot__HighBP_1.0
onehot__HighChol_1.0
onehot__HighChol_0.0
onehot__Sex_0.0
onehot__Sex_1.0
ord__GenHlth
ord__Age
0
200
400
600
800
1000
1200
mean(|SHAP value|)

<Fig 13: mean|SHAP values| for features of the logistic regression model (L2 regularization) that performed best on the test set. Age, General Health, Sex, High Cholesterol, and High BP are the most important features>



<Fig 14: A SHAP force plot for index 0 in the test set. This shows that Age, General Health, and Sex play significant roles in the final prediction, especially in the context for this local datapoint. Lower/better General Health appears to push the probability of a positive prediction downwards while higher age and a respondent being male push that probability upwards.>

Across the assessments for global feature importance within the logistic regression model (L2 regularization) that performed best on the test set, common features that appeared at the top included Age, General Health, Sex, High Cholesterol, High BP. This was expected as these features are commonly associated as risk factors for heart disease and heart attacks. Only some of the features, such as general health, high cholesterol, and high blood pressure, are those that clinicians can target to improve outcomes. Age is expected to be significant as chronic conditions become exacerbated with age. The impact of sex upon heart disease and heart attack is surprising. All together, it is suggested that these features had the greatest effect on the final prediction. The permutation importance had the most unique features present at the top including Physical Activity, if one is a Smoker, and Education, of which education is a variable not often directly linked to health outcomes. It's possible that education is linked to an individual's income and ability to maintain a healthy lifestyle and receive care, emphasizing the importance of social determinants of health alongside an individual's choices.

OUTLOOK

To improve the models' predictive power and the efficiency of model training I may remove correlated figures following preprocessing due to numerous one hot encoded features having only two possible values. I would also hope to implement early_stopping_rounds in the XGBClassifier in the future as I struggled to do so with the GridSearchCV. KNeighborsClassifier could be an additional model trained on the data. Additionally, these models can also be trained on data from successive years of the BRFSS and train on a wider variety of the features in the BRFSS as the much more successful models from Dolezel et al. do to improve the predictive power. The current dataset used does not capture the entirety of the data present within the full BRFSS.

REFERENCES

1. Heart Disease | MedlinePlus
2. Heart disease - Symptoms and causes - Mayo Clinic
3. Heart disease - Diagnosis and treatment - Mayo Clinic
4. Health and Economic Benefits of High Blood Pressure Interventions | Power of Prevention
5. Heart Disease and Stroke Statistics-2018 Update: A Report From the American Heart Association

6. ASCVD Risk Estimator +
7. Heart Disease Health Indicators Dataset
8. Examining Predictors of Myocardial Infarction - PMC