Date of submission March 18, 2020

Building Predictive Models for Heart Disease using the 2015 Behavioral Risk Factor Surveillance System

Alex Teboul
DePaul University, IL 60604 USA

This paper was written for DSC 540 – Advanced Machine Learning at DePaul University with professor Casey Bennett.
Python Code Link: https://colab.research.google.com/drive/1Qonvg6ZK6r8Nfps6GYHB_jW60_VJwY_h

ABSTRACT Heart Disease is among the most prevalent chronic diseases in the United States, impacting millions of Americans each year and exerting a significant financial burden on the economy. This paper explores the use of machine learning techniques towards heart disease prediction using the 2015 Behavioral Risk Factor Surveillance System (BRFSS) survey. Specifically, Random Forests (RF), Gradient Boosting (GB), AdaBoost (Ada), and Neural Networks (NN) are applied to the binary classification task of predicting heart disease based on survey responses to 21 of the questions asked in the BRFSS. Out of the 253,680 survey responses collected from the 2015 BRFSS survey for use in this study, 23,893 had heart disease or had previously had a heart attack. The 4 models tested were assessed with feature selection, without feature selection, and using different parameter settings. To better understand the impact of the class imbalance on model performance through training and testing, two additional datasets were created with different class ratios. These splits were 50-50 and 60-40, of no heart disease to heart disease respondents, respectively. Model evaluation criteria were Accuracy, AUC, and Run-time. The 4 different models had nearly identical performance on the three datasets, with the best overall performances coming in the full dataset with feature selection. Of all the models tests, the simple Neural Network with feature selection, an Adam solver, and logistic activation had the best performance with 91% (+/- 0%) Accuracy, 0.84 (+/- 0.01) AUC, and a 36 second run-time on the full dataset. The features for High Blood Pressure, High Cholesterol, Stroke, General Health, Difficulty Walking, Sex, and Age that were selected in this model are in line with research that identifies these as significant risk factors for heart disease. These 7 selected features, as survey questions, could serve as a useful awareness tool for those at high risk of developing heart disease.

Keywords Random Forests, Gradient Boosting, AdaBoost, Neural Networks, Heart Disease, Heart Attack, Machine Learning

1.INTRODUCTION

In the United States alone, heart disease claims roughly 647,000 lives each year – making it the leading cause of death.[1] The buildup of plaques inside larger coronary arteries, molecular changes associated with aging, chronic inflammation, high blood pressure, and diabetes are all causes of and risk factors for heart disease.[2] While there are different types of coronary heart disease, the majority of individuals only learn they have the disease following symptoms such as chest pain, a heart attack, or sudden cardiac arrest. This fact highlights the importance of preventative measures and tests that can accurately predict heart disease in the population prior to negative outcomes like myocardial infarctions (heart attacks) taking place.

The Centers for Disease Control and Prevention has identified high blood pressure, high blood cholesterol, and smoking as three key risk factors for heart disease. Roughly half of Americans have at least one of these three risk factors.[2] The National Heart, Lung, and Blood Institute highlights a wider array of factors such as Age, Environment and Occupation, Family History and Genetics, Lifestyle Habits, Other Medical Conditions, Race or Ethnicity, and Sex for clinicians to use in diagnosing coronary heart disease. Diagnosis tends to be driven by an initial survey of these common risk factors followed by bloodwork and other tests. [1]

In this paper, I explore the efficacy of different machine learning techniques for heart disease prediction in the general population. To this end, I make use of cross-sectional survey data collected in 2015 through the Behavioral Risk Factor Surveillance System. The aim of this project is to build a model with relatively high accuracy and AUC that could serve as an awareness-tool for those at high risk of developing heart disease. Granted, some features presuppose prior medical consultation, as respondents need to have had blood pressure or cholesterol measured before. With early warning and better diagnostic practices, preventative measures like changes to diet, exercise habits, and medication can reverse or delay onset of heart disease.[3,4]

2.Literature Review

Predictive and diagnostic models for heart disease have been explored in the fields of cardiology and data mining.[5,6,7] Additionally, researchers have studied the use of the Behavioral Risk Factor Surveillance System Survey (BRFSS) for developing predictive models of chronic diseases.[8]

Researchers have demonstrated that models of high accuracy can be developed to predict heart disease. A 2016 paper published in IEEE, explored models like Naïve Bayes, KNN, Decision Trees, and Neural Networks.[5] The best of their models achieved 80.6% accuracy in predicting heart disease taking into account features like age, gender, blood pressure, cholesterol, and pulse rate. A 2013 paper exploring Neural Networks for heart disease prediction took into account features additional features such as tobacco smoking, alcohol intake, obesity, physical activity, and family history.[6] These models achieved higher accuracy than the comparable neural networks in the 2016 paper. Specifically, their best model achieved 89% accuracy. That being said, neither paper reported on their AUC or sensitivity/specificity. Both studies were also exploring small datasets of less than 500 individuals. My aim in this project is to explore model building on a much larger dataset, to validate the predictive power of these techniques on a cross-sectional, representative sample of the U.S. population.

With this in mind, the CDC's 2015 BRFSS survey was selected for use in this project, as it contains survey responses from over 441,455 Americans. A study by CDC researchers Zidian et. al in 2019, demonstrated machine learning techniques can be applied to the BRFSS datasets to predict chronic disease risk with high Accuracy and AUC.[8] Specifically, the team explored SVMs, decision trees, logistic regression, random forests, neural networks, and Gaussian Naïve Bayes classifiers for Type II Diabetes. They found that for a cross-sectional group of 138,146 survey participants from 2014, neural networks could predict Type II Diabetes incidence with 82.4% accuracy, 90.2% specificity, and 0.795 AUC. My findings in this project suggest that heart disease is another chronic disease that can be well modelled using the BRFSS. For comparison, the best model for heart disease I report in this paper was a neural network that achieved 91% (+/- 0%) accuracy, 0.84 (+/- 0.01) AUC, and a 36 second run-time on the full dataset with 5 fold cross validation on 7 features from the BRFSS.

3.Methodology

This project draws upon existing research into the risk factors that impact heart disease to build binary classifiers using Random Forests, Gradient Boosting, AdaBoost, and Neural Networks. The steps used to accomplish this are as follows:

1) Dataset and Feature Selection
2) Data Preprocessing
3) Create 50-50 and 60-40 Datasets
4) Model Building

3.1. Dataset and Feature Selection

The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that is collected annually by the CDC. Each year, the survey collects responses from over 400,000 Americans on health-related risk behaviors, chronic health conditions, and the use of preventative services. It has been conducted every year since 1984. For this project, I downloaded a csv of the dataset available on Kaggle for the year 2015. This original dataset contains responses from 441,455 individuals and has 330 features. These features are either questions directly asked of participants, or calculated variables based on individual participant responses.

From this original dataset, I selected features that matched the risk factors I identified in my research of heart disease and through the literature review. To help understand what

the variables meant, the response options, and original questions, I consulted the BRFSS 2015 Codebook. This Codebook is available in the Appendix along with the Kaggle dataset link. To start this process, I referenced some of the same features chosen for a research paper by Zidian Xie et al for *Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques* using the 2014 BRFSS.[8] Diabetes and Heart Disease outcomes are strongly correlated, with the primary cause of death for diabetics being heart disease complications. Given this information, it was a useful starting point.

I first chose my dependent or response variable to be a calculated variable _MICHD. This variable combined responses from individual participants to form a new calculated question: Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI). In this way, I determine that any participant that has ever been diagnosed with heart disease or has had a heart attack will be considered as having heart disease.

From there, I selected categories of risk factors and selected questions in the BRFSS that matched those risk factors. I chose features related to the following risk factor categories High Blood Pressure, High Cholesterol, Body Mass Index (BMI), Smoking, Other Chronic Health Conditions, Physical Activity, Diet, Alcohol Consumption, Health Care Access, General Health & Wellbeing, and Demographics. Ultimately, I selected 21 features out of the original 330 in the 2015 BRFSS dataset, which are summarized in Table 1.

TABLE 1
MODEL FEATURES

#	Renamed Features	Categories
*1	Diabetes	Response Variable
2	HighBP	High Blood Pressure
3	HighChol	High Cholesterol
4	CholCheck	High Cholesterol
5	BMI	BMI
6	Smoker	Smoking History
7	Stroke	Chronic Health Conditions
8	HeartDiseaseorAttack	Chronic Health Conditions
9	PhysActivity	Physical Activity
10	Fruits	Diet
11	Veggies	Diet
12	HvyAlcoholConsump	Alcohol Consumption
13	AnyHealthcare	Health Care Access
14	NoDocbcCost	Health Care Access
15	GenHlth	General Health & Wellbeing
16	MentHlth	General Health & Wellbeing
17	PhysHlth	General Health & Wellbeing
18	DiffWalk	General Health & Wellbeing
19	Sex	Demographics
20	Age	Demographics
21	Education	Demographics
22	Income	Demographics

*The response variable used in this project was calculated variable _MICHD in the 2015 BRFSS. A range of features were selected from the broader dataset to encompass relevant categories identified in the literature. The **7 bolded features** were selected via wrapper based feature selection in by the best models.

For more information on these variables, I have included the original variable names and questions that were asked in the BRFSS in the Appendix Table 1. For reference, the original variable names are: _RFHYPE5, TOLDHI2, _CHOLCHK, _BMI5, SMOKE100, CVDSTRK3, DIABETE3, _TOTINDA, _FRTLT1, _VEGLT1, _RFDRHV5, HLTHPLN1, MEDCOST, GENHLTH, MENTHLTH, PHYSHLTH, DIFFWALK, SEX, _AGEG5YR, EDUCA, and INCOME2. These variables covered a wide range of risk factors and contain a higher number of data points than have been explored in past publications involving machine learning techniques applied to heart disease prediction.

3.2. Data Preprocessing

Once I had selected the appropriate features based on risk factor category, I had a dataset of 441,455 responses with my dependent variable and 21 features. As part of the data cleaning process, I removed all missing values, bringing the total number of responses down to 343,605. I then individually went through the variables one by one and modified the values to be binary 0=No, and 1=Yes. Additionally, I removed responses that were marked 'Don't Know/Not Sure' and 'Refused to Answer'. Ordinal variables were changed to a scale '0,1,2,3,4….' for simplicity. To see all of these modifications, my full code is displayed through Google Colaboratory and can be found through the link in the Appendix. Finally, I renamed all of the features so it would be easier to interpret the results of my feature selection during the model building process.

Following these pre-processing steps, I ended up with a dataset with 253,679 responses to 21 features and my 1 response variable for HeartDiseaseorAttack.

3.3. Create 50-50 and 60-40 datasets

As part of my initial exploration of the data, it was discovered that there was a strong class imbalance between cases of respondents with heart disease and those without. Specifically, about 9.4% of participants had heart disease to 90.6% without heart disease. I was curious to see the effects of class imbalance on the accuracy and AUC of my models, so I created a dataset with a 50% heart disease to 50% no heart disease split, as well and one that was 60% no heart

disease to 40% heart disease. These datasets are summarized in Table 2 below. A caveat of these randomly selected splits is they don't quite match the individual variable value distributions among the respondents in the full dataset. That said, with 20,000+ data points, I figured it would be sufficient for an interesting comparison to the full dataset.

TABLE 2
CLEANED DATASETS

#	Class Split (0 – 1)	Counts (0 – 1)
*1	90 - 10	229,787 – 23,893
2	50 - 50	23,893 – 23,893
3	60 - 40	47,786 – 23,893

*For the remainder of the paper, the 90-10 split is referred to as the original or full dataset. Note that 0 is for participant does not have heart disease and 1 is for does have heart disease.

3.4. Model Building

The model building processes explored in this paper are the same as the ones explored in the course *DSC 540 – Advanced Machine Learning*. Specifically, the different models are tested on 65% train, 35% test splits with 5-fold cross validation. Performance is measured in terms of accuracy, AUC, and run-time. The python code template used in class was leveraged in this project to run the models for Random Forest, Gradient Boosting, AdaBoost, and Neural Networks. All models were tested on multiple parameter settings with features selection turned on and feature selection turned off for the full dataset with 90-10 split. The best model parameter settings and the with/without feature selection are reported in the Results section of this paper. For the 50-50 and 60-40 datasets, I only report the results with feature selection for the best models. The python code template provided in class by Professor Casey Bennett was used in this process. Finally, I report best overall models and important features identified in these highly predictive models.

4.Results

Relatively high model accuracy and AUC were achieved with all of the models (RF, GB, Ada, and NN) on the full dataset both with and without feature selection turned on. Gradient Boosting, AdaBoost, and Neural Networks performed nearly identically, with performance that was slightly higher than that of the Random Forest across all datasets. Given accuracy and AUC that were comparable across the different models, best performance was selected based on runtime of the model. Though it is important to stress that performance was nearly identical outside of runtime, and all runtimes were quite fast.

The range of accuracies for all full dataset models were between 89% and 91%, AUC between 0.74 and 0.84, and runtime between 66 and 153 seconds. There were drops in accuracy on the models built using the 50-50 and 60-40 datasets, while AUC remained roughly constant between all the models, and runtime was of course cut down as well. In the end, the best model selected was the single layer Neural Network with adam solver, logistic activation, and a learning rate alpha of 0.0001. With feature selection turned on it had 91% (+/- 0%) Accuracy, 0.84 (+/- 0.01) AUC, and a 36 second run-time on the full dataset. That being said, there was nearly identical performance between the Neural Networks and the Gradient Boosting and AdaBoost models, such that the best model was selected on the basis of its slightly faster runtime.

4.1. Random Forests

Random forests are an ensemble learning method capable of performing classification and regression tasks. They are built from ensembles of decision trees, with final classification determined by vote among the trees. Table 3 below presents the results of the random forest model building process. The parameters tested in the model building process were the # of trees and different levels of cross validation. For the models displayed in Table 3, all were built using 50 trees and 5-fold cross validation.

TABLE 3
RANDOM FOREST PERFORMANCE

Dataset	Model	Accuracy	AUC	Runtime
Full Dataset	RF w/ Feature Selection	0.89 (+/- 0.00)	0.74 (+/- 0.01)	48 sec
50-50 Dataset	RF w/ Feature Selection	0.72 (+/- 0.01)	0.78 (+/- 0.01)	10 sec
60-40 Dataset	RF w/ Feature Selection	0.73 (+/- 0.01)	0.78 (+/- 0.01)	15 sec
Full Dataset	RF w/o Feature Selection	0.90 (+/- 0.00)	0.82 (+/- 0.01)	66 sec

The Random Forest model performed well on the full dataset both with and without feature selection. Specifically, with feature selection, 89% accuracy, 0.74 AUC, and 48 second runtime was achieved. When feature selection was not turned on, the dataset had slightly better performance, particularly in terms of AUC. As expected, there was a drop in accuracy for the 50-50 and 60-40 datasets, though AUC remained about the same.

4.2. Gradient Boosting

Gradient Boosting is another ensemble learning method, useful for classification problems. In this case, an ensemble of weak decision tree prediction models is used. As with other boosting methods, the overall model is built in stages, generalizing by optimizing a loss function. The parameters tested in my model for heart disease were the number of estimators, loss function, and the max depth of the weak models used in the ensemble. For the models displayed in Table 4, parameter settings are 100 estimators, a loss set to deviance and max depth of 3 are used, along with 5-fold cross validation.

TABLE 4

GRADIENT BOOSTING PERFORMANCE

Dataset	Model	Accuracy	AUC	Runtime
Full Dataset	GB w/ Feature Selection	0.91 (+/- 0.00)	0.85 (+/- 0.01)	54 sec
50-50 Dataset	GB w/ Feature Selection	0.76 (+/- 0.01)	0.84 (+/- 0.01)	8 sec
60-40 Dataset	GB w/ Feature Selection	0.78 (+/- 0.01)	0.84 (+/- 0.01)	13 sec
Full Dataset	GB w/o Feature Selection	0.91 (+/- 0.00)	0.85 (+/- 0.01)	153 sec

The Gradient Boosting models performed better overall compared to the Random Forests. In particular, AUC remains high across all of the models tested on the different datasets and with/without feature selection. With feature selection, 91% accuracy, 0.85 (+/- 0.01) AUC, and 54 second runtime was achieved.

4.3. AdaBoost

AdaBoost is another ensemble model that leverages weak learners. Different from Gradient Boosting, it combines weak learners into a weighted sum that leads to the final output of the boosted classifier. The combination of weak learners help the final model converge to a relatively strong learner – in theory. In practice, AdaBoost proved just as effective at modeling heart disease as Gradient Boosting did. The parameters tested in my model for heart disease were the number of estimators and the learning rate. For the models displayed in Table 5, parameter settings are 100 estimators at a learning rate of 0.1 with 5-fold cross validation.

TABLE 5
ADABOOST PERFORMANCE

Dataset	Model	Accuracy	AUC	Runtime
Full Dataset	Ada w/ Feature Selection	0.91 (+/- 0.00)	0.84 (+/- 0.01)	49 sec
50-50 Dataset	Ada w/ Feature Selection	0.76 (+/- 0.01)	0.83 (+/- 0.01)	8 sec
60-40 Dataset	Ada w/ Feature Selection	0.77 (+/- 0.01)	0.84 (+/- 0.01)	13 sec
Full Dataset	Ada w/o Feature Selection	0.91 (+/- 0.00)	0.84 (+/- 0.01)	92 sec

The AdaBoost models performed better overall compared to the Random Forests, but just as well as the Gradient Boosting models did. Again, AUC in particular remains high across all of the models tested on the different datasets and with/without feature selection. With feature selection, 91% accuracy, 0.84 (+/- 0.01) AUC, and 49 second runtime was achieved.

4.4. Neural Networks

Neural networks have been shown to work remarkably well for a variety of machine learning tasks. In this project, I explore the performance of a simple neural network with 2 layers. The multi-layer perceptron classifier used here was tested for different solvers, activation functions, and learning rate alphas. For the models displayed in Table 6, the stochastic gradient-based optimized 'adam' is used as the solver for weight optimization. The learning rate was set to 0.0001 and the activation function used was logistic. These parameter settings were determined by the best models produced at the different parameter settings.

TABLE 6
NEURAL NETWORK PERFORMANCE

Dataset	Model	Accuracy	AUC	Runtime
Full Dataset	NN w/ Feature Selection	0.91 (+/- 0.00)	0.84 (+/- 0.01)	36 sec
50-50 Dataset	NN w/ Feature Selection	0.76 (+/- 0.01)	0.84 (+/- 0.01)	18 sec
60-40 Dataset	NN w/ Feature Selection	0.78 (+/- 0.01)	0.84 (+/- 0.01)	21 sec
Full Dataset	NN w/o Feature Selection	0.91 (+/- 0.00)	0.85 (+/- 0.01)	113 sec

Overall, the neural network models performed about the same as the Gradient Boosting and AdaBoost models. There was slightly better performance that with the Random Forests though. Same as with Gradient Boosting and AdaBoost, AUC remains high across all of the models tested on the different datasets and with/without feature selection. With feature selection, 91% accuracy, 0.84 (+/- 0.01) AUC, and 36 second runtime was achieved. It is important to note that there was essentially no loss in performance for the Neural Network with feature selection turned on. This highlights the importance of those 7 key features extracted in the feature selected models. We can cut runtime to 1/3 of the model without feature selection, without sacrificing performance.

The same drops in accuracy are observable in the 50-50 and 60-40 class split datasets as with the other models. These were an interesting comparison tool, but ultimately because AUC is constant between the datasets, I opt for the full dataset as it has higher accuracy and the same AUC.

4.5. Best Overall Performance

The best overall performance is hard to pin down given nearly identical performance between the Neural Networks, Gradient Boosting, and AdaBoost models. That said, on the basis of runtime alone, the Neural Network beat out the GB and Ada models. The Neural Network with feature selection saw no drop in performance so I selected it as my best overall model. It also had the lowest runtime. Note that Gradient Boosting and AdaBoost were the top models for the other settings.

TABLE 7
BEST OVERALL PERFORMANCE

Dataset	Model	Accuracy	AUC	Runtime
Full Dataset	**NN w/ Feature Selection**	**0.91 (+/- 0.00)**	**0.84 (+/- 0.01)**	**36 sec**
50-50 Dataset	GB w/ Feature Selection	0.76 (+/- 0.01)	0.84 (+/- 0.01)	8 sec
60-40 Dataset	GB w/ Feature Selection	0.78 (+/- 0.01)	0.84 (+/- 0.01)	13 sec
Full Dataset	Ada w/o Feature Selection	0.91 (+/- 0.00)	0.84 (+/- 0.01)	92 sec

TABLE 8
MODEL COMPARISON - WITH FEATURE SELECTION

Dataset	Model	Accuracy	AUC	Runtime
Full Dataset	RF w/ Feature Selection	0.89 (+/- 0.00)	0.74 (+/- 0.01)	48 sec
Full Dataset	GB w/ Feature Selection	0.91 (+/- 0.00)	0.85 (+/- 0.01)	54 sec
Full Dataset	Ada w/ Feature Selection	0.91 (+/- 0.00)	0.84 (+/- 0.01)	49 sec
Full Dataset	**NN w/ Feature Selection**	**0.91 (+/- 0.00)**	**0.84 (+/- 0.01)**	**36 sec**

The argument could be made that Gradient Boosting slightly outperformed the Neural Network, but essentially Gradient Boosting, AdaBoost, and Neural Networks models performed identically on this task of predicting heart disease. The same important features were selected and used in each of these models. Table 8 summarizes the comparison

between the feature selected models for heart disease prediction.

5.Discussion

The goal of this paper was to determine the efficacy of using survey data from the BRFSS to build predictive models for heart disease risk. The secondary motivation was to identify features that are highly predictive of heart disease that could function as key questions for use in an awareness tool for the general public to get informed about their heart disease risk. To this end, it was determined that machine learning models built using the BRFSS for heart disease can be highly predictive of heart disease, as evidenced by models with both high accuracy and AUC.

Note that I emphasize the models with feature selection turned on because they have essentially the same performance as the those without feature selection turned on. In the interest of simplicity and developing an actionable set of survey questions that are highly predictive of heart disease, feature selection is preferred.

The most important features to the models on the full dataset were High Blood Pressure, High Cholesterol, Stroke, General Health, Difficulty Walking, Sex, and Age. For the 50-50 dataset, they were High Blood Pressure, High Cholesterol, General Health, Sex, and Age. For the 60-40 dataset, they were High Blood Pressure, High Cholesterol, Stroke, General Health, Sex, and Age. All make sense given the research into important risk factors for heart disease.

Interestingly, Smoking, BMI, Alcohol Consumption, and Diabetes did not factor into these highly predictive models, despite being commonly referred to as very important risk factors for heart disease. It seems that among the 7 most important variables in these heart disease models, 5 are the most useful. These 5 are: High Blood Pressure, High Cholesterol, one's own rating of General Health, Sex (male), and Age (older: 60+ more at risk). If you could only ask 5 questions with respect to assessing someone's risk of heart disease, it could be narrowed down to the 5 questions from the BRFSS, available in the Appendix. The downside to this is that individuals would have to be tested for their cholesterol and blood pressure, and have a doctor indicate to them whether their values were high.

6.Conclusion

The BRFSS is a useful tool for the government to keep track of chronic health condition prevalence in the general population. This project further supports it usefulness in developing machine learning models for the prediction of chronic health conditions. In the work presented in this paper, I have identified 7 features of the BRFSS that can be used to predict heart disease with high performance: high blood pressure, high cholesterol, one's own self rating of general health, past stroke history, having difficult walking, identifying as male, and being older (likely 60+). This paper explored heart disease prediction using a larger dataset than previously published works for heart disease prediction and demonstrated multiple models that also achieve higher accuracy (91%) and AUC (0.84) than in previously published papers on the topic.

7.Future Work

Despite the work presented here, there are additional areas requiring improvement. One specific risk factor not used in the models tested was the demographic: race. In the United States, race can be considered an important risk factor, with different races experiencing slightly different prevalence of heart disease. I did not include it for simplicity and to avoid complications in the model building process. An additional caveat of this work for use in an awareness tool is that not everyone has had their cholesterol or blood pressure measured recently.

An improvement on this project would be to build a Logistic Regression model for heart disease. As perhaps the simplest binary classifier, it would be useful to compare its performance to that the Random Forest, Gradient Boosting, and Neural Network. I did also try to run Support Vector Machines, but following a day of continuous running without any output, I determined that SVMs were not well suited to the large dataset and many features used in this project.

The framework laid out in the Google Colaboratory notebook could be used by others to leverage the BRFSS to build models for other chronic diseases. Within this project, I included BRFSS variables for Stroke and Diabetes incidence as well. In this way, the response variable could very easily be switched to Stroke or Diabetes and have models built for those chronic diseases.

Ultimately, this project has demonstrated a process for using the BRFSS datasets to build machine learning models that have relatively strong performance. Important risk factors identified in the field of cardiology were also important to these highly predictive models. Specifically, having high blood pressure, high cholesterol, past strokes, a low self-rating of your general health, difficulty walking, identifying as male, and being older are highly predictive of heart disease risk in the general American population.

8. APPENDIX

1. The BRFSS 2015 .csv can be downloaded here: https://www.kaggle.com/cdc/behavioral-risk-factor-surveillance-system#2015.csv
2. The BRFSS 2015 Codebook is available here: https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf
3. My code is open sourced using Google Colaboratory. The code can be accessed here: https://colab.research.google.com/drive/1Qonvg6ZK6r8Nfps6GYHB_jW60_VJwY_h

8.1 Original Variable Names and Descriptions

Dependent/Response Variable:

1. Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI) --> _MICHD

Independent Variables:

High Blood Pressure

2. Adults who have been told they have high blood pressure by a doctor, nurse, or other health professional --> _RFHYPE5

High Cholesterol

3. Have you EVER been told by a doctor, nurse or other health professional that your blood cholesterol is high? --> TOLDHI2

4. Cholesterol check within past five years --> _CHOLCHK

BMI

5. Body Mass Index (BMI) --> _BMI5

Smoking

6. Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] --> SMOKE100

Chronic Health Conditions

7. (Ever told) you had a stroke. --> CVDSTRK3

8. (Ever told) you have diabetes (If "Yes" and respondent is female, ask "Was this only when you were pregnant?". If Respondent says pre-diabetes or borderline diabetes, use response code 4.) --> DIABETE3

Physical Activity

9. Adults who reported doing physical activity or exercise during the past 30 days other than their regular job --> _TOTINDA

Diet

10. Consume Fruit 1 or more times per day --> _FRTLT1

11. Consume Vegetables 1 or more times per day --> _VEGLT1

Alcohol Consumption

12. Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) --> _RFDRHV5

Health Care

13. Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare, or Indian Health Service? --> HLTHPLN1

14. Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? --> MEDCOST

General Health & Wellbeing

15. Would you say that in general your health is: --> GENHLTH

16. Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? --> MENTHLTH

17. Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? --> PHYSHLTH

18. Do you have serious difficulty walking or climbing stairs? --> DIFFWALK

Demographics

19. Indicate sex of respondent. --> SEX

20. Fourteen-level age category --> _AGEG5YR

21. What is the highest grade or year of school you completed? --> EDUCA

22. Is your annual household income from all sources: (If respondent refuses at any income level, code "Refused.") --> INCOME2

ACKNOWLEDGMENT

I would like to acknowledge Casey Bennett, who was the professor for this course *DSC 540 – Advanced Machine Learning*. The code framework for building the models was developed by the professor for use in this class and was modified to train/test models in this study. The code and more information can be accessed at www.caseybennett.com.

REFERENCES

1. Heron, M. Deaths: Leading causes for 2017 pdf icon[PDF – 3 M]. National Vital Statistics Reports;68(6). Accessed November 19, 2019.
2. Fryar CD, Chen T-C, Li X. Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999–2010 pdf icon[PDF-494K]. NCHS data brief, no. 103. Hyattsville, MD: National Center for Health Statistics; 2012. Accessed May 9, 2019.
3. Alharbi, M., Gallagher, R., Kirkness, A., Sibbritt, D., & Tofler, G. (2016). Long-term outcomes from Healthy Eating and Exercise Lifestyle Program for overweight people with heart disease and diabetes. European Journal of Cardiovascular Nursing, 15(1), 91–99. https://doi.org/10.1177/1474515114557222
4. Infante, T., Forte, E., Schiano, C., Cavaliere, C., Tedeschi, C., Soricelli, A., Salvatore, M., & Napoli, C. (2017). An integrated approach to coronary heart disease diagnosis and clinical management. *American journal of translational research*, 9(7), 3148–3166.
5. I. Mirza, A. Mahapatra, D. Rego and K. Mascarenhas, "Human Heart Disease Prediction Using Data Mining Techniques," 2019 International Conference on Advances in Computing, Communication and Control (ICAC3), Mumbai, India, 2019, pp. 1-5.
6. S. U. Amin, K. Agarwal and R. Beg, "Genetic neural network based data mining in prediction of heart disease using risk factors," 2013 IEEE Conference on Information & Communication Technologies, Thuckalay, Tamil Nadu, India, 2013, pp. 1227-1231.
7. Manogaran, G., Varatharajan, R. & Priyan, M.K. Hybrid Recommendation System for Heart Disease Diagnosis based on Multiple Kernel Learning with Adaptive Neuro-Fuzzy Inference System. *Multimed Tools Appl* 77, 4379–4399 (2018). https://doi-org.ezproxy.depaul.edu/10.1007/s11042-017-5515-y
8. Xie Z, Nikolayeva O, Luo J, Li D. Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. Prev Chronic Dis 2019;16:190109.
DOI: http://dx.doi.org/10.5888/pcd16.190109

Alex Teboul received his B.S. in Biomedical Engineering in 2017 from the University of Southern California in Los Angeles, CA. Born and raised in Chicago, IL, Alex returned to Chicago following his undergraduate studies to pursue a M.S. Data Science at DePaul University. Working full-time and going to school full time, he expects to graduate in November of 2020.

He is currently employed at Cicero School District 99 as a Data Scientist. District 99 is in Cicero, IL just West of Chicago, IL and services between 10,000 and 12,000 students annually. Past work experience includes working for Northwestern University's Feinberg School of Medicine in Chicago, IL as a Freelance Database Consultant. Internship experience includes working as an Automation Engineering Intern at Volkswagen Autoeuropa in Palmela, Portugal.