Heart Disease Prediction Modeling Based on Behavioral Risk Factors

Allie Schneider [1], Francis Villamater [1], Stephan Dupoux [1]

1: College of Computing & Informatics, Drexel University – Philadelphia, PA 19104, USA

ABSTRACT – The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey conducted by the CDC aimed at gathering data on the health status of U.S. residents. We use a subset of the variables generated by the survey that can directly or indirectly influence heart disease. Utilizing PySpark, we've built a random forest model, to classify whether or not a patient carries a diagnosis of heart disease. This random forest employs an ensemble of decision trees trained via the bagging method, favored by the randomness added when growing trees. These techniques yielded a model with a split behavior depending on the class that's being predicted and an overall subpar performance, Area under PR: 0.07 and Area under ROC: 0.5. Identification of appropriate attributes and further tuning is necessary.

INTRODUCTION

According to the CDC, CHD is one of the leading causes of death for people of most races in the United States. 47% of Americans have at least 1 of 3 key risk factors for coronary heart disease (CHD): high blood pressure, high cholesterol, and smoking. Other key indicators include diabetic status, obesity (high BMI), not getting enough physical activity or drinking too much alcohol. We suspect that these variables, outside of the 3 key risk factors, would be impactful towards predicting CHD.

DATASET

The dataset originally comes from the CDC and is a major part of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to gather data on the health status of U.S. residents. A Kaggle user noticed many different factors in the dataset that directly or indirectly influence heart disease and selected the most relevant variables from it and did some cleaning so that it would be usable for machine learning projects. We are using this cleaned dataset from Kaggle which includes key indicators of CHD and reflects data from the 2020 annual CDC survey of 400,000 adults related to their health status. The dataset includes the following 18 attributes and 319,795 instances.

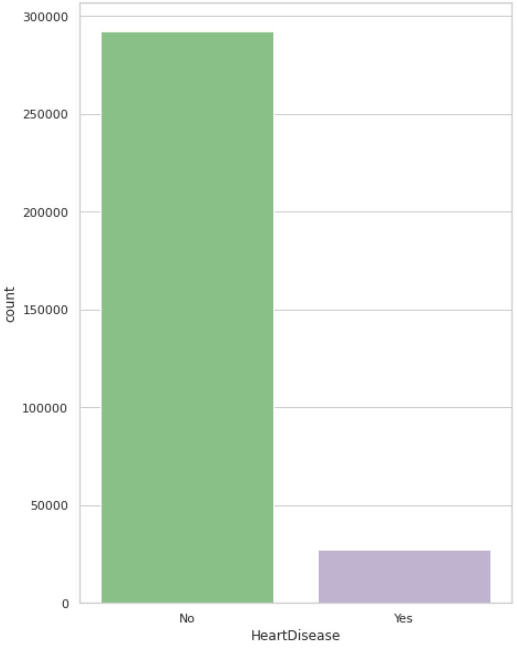

Figure 1: Prevalence of CHD in survey population

Variable descriptions:

- `HeartDisease` (dichotomous): [Target Variable] Respondents that have ever reported having coronary heart disease or myocardial infarction
- `BMI` (continuous): Body Mass Index (BMI)
- `Smoking` (dichotomous): Have they smoked at least 100 cigarettes in their lives?
- `AlcoholDrinking` (dichotomous): Yes if adult men have >14 drinks/week or adult women >7 drinks/week, no otherwise.
- `Stroke` (dichotomous): Have they ever been told they had a stroke?
- `PhysicalHealth` (continuous): including physical illness and injury, for how many days during the past 30 days was their physical health not good?
- `MentalHealth` (continuous): For how many days during the past 30 days was their mental health not good? (0-30)
- `DiffWalking` (dichotomous): Do they have serious difficulty walking or climbing stairs?
- `Sex` (dichotomous): Male or female
- `AgeCategory` (categorical, 13 levels): Age category
- `Race` (categorical, 6 levels): Race/ethnicity
- `Diabetic` (categorical, 4 levels): Have they ever been told they have diabetes?
- `PhysicalActivity` (dichotomous): Adults who reported engaging in physical activity/exercise over the past 30 days outside of their regular job
- `GenHealth` (categorical, 5 levels): How would they rate their overall health?
- `SleepTime` (continuous): On average, how much sleep do they get in a 24-hour period?
- Asthma (dichotomous): Have they ever been told they have asthma?
- `KidneyDisease` (dichotomous): Excluding kidney stones, bladder infection, or incontinence, have they ever been told they have kidney disease?
- `SkinCancer` (dichotomous): Have they ever been told they have skin cancer?

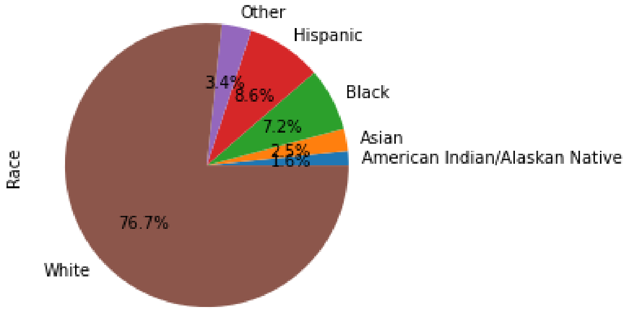

Figure 2: Distribution of Race

EXPLORATORY DATA ANALYSIS

Fortunately, this dataset has been cleaned at the time of compilation and contains no missing values. However, the dataset itself is highly imbalanced. Race shows that over 75% of this survey population was actually "White". One may also notice that `HeartDisease` is also highly imbalanced. But this variable is also our target variable and this disparity is reflective of CHD's prevalence within our survey population. Balancing on `Race` to improve model performance is a much more advantageous strategy. We resampled `Race` to match the number of instances of 'Black' (approx. 22,000 instances) since this would create a favorable dataset size that is neither too large or too small. Arguably, this survey was aimed at representing the American

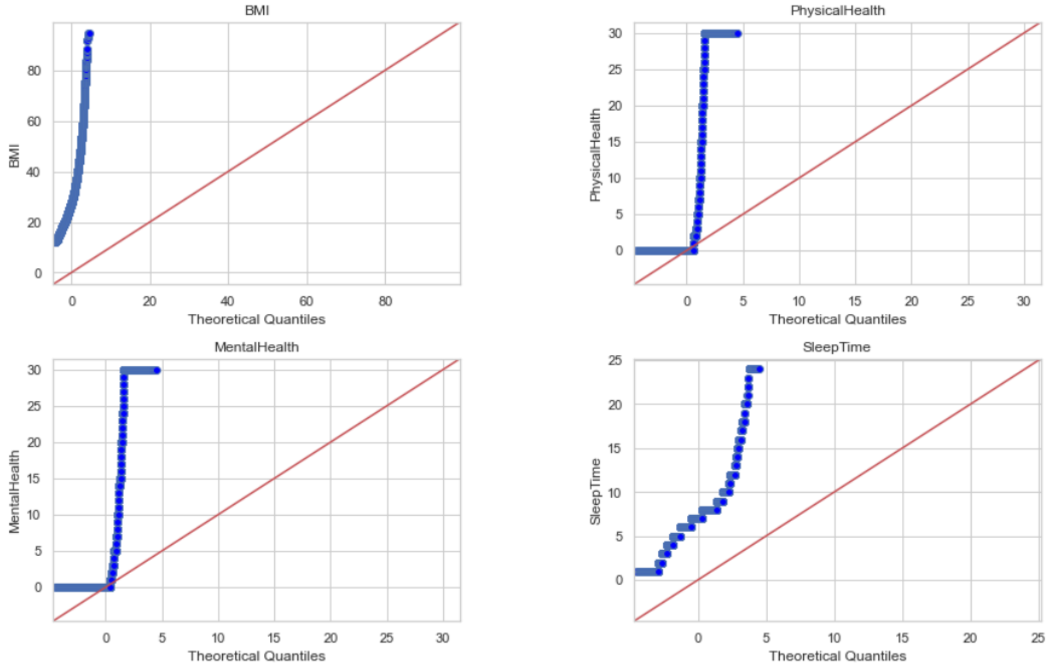

Figure 3: QQ-Plots of Continuous Variables

population, and therefore should have equally surveyed each category of race to avoid underrepresenting any category.

Examining for normality amongst our continuous variables, we can see that none exhibit a normal distribution rather a heavy right skew. However, since we are employing a random forest model, which does not assume normality, these distributions do not need to be transformed.

In terms of feature selection, we elected to include all variables in the dataset. Our reasoning being that all of these variables, outside of the 3 key indicators of CHD: high blood pressure, high cholesterol, and smoking, may contribute to CHD in some way.

To build our random forest model, we created the appropriate PySpark pipeline containing a string indexer, one hot encoder, vector assembler, and then the model itself. As mentioned above, we excluded the attribute `Smoking` in an attempt to find other behavioral predictors of CHD.

RESULTS

Our random forest model performed well in its predictions when classifying “No” for CHD, and poorly on “Yes” for CHD. Considering our application, the aim should be to avoid missing any CHD cases and the False-Negative rate should be as low as possible. When classifying “No”, our model performed very well; accuracy was 0.93, precision was 0.92, a recall of 1.0, and an f-score of 0.96. When classifying “Yes”, although accuracy was fairly high at 0.93, precision, recall, and f-score were all zero. Our model managed to attain an overall Area under the ROC curve of 0.5 while the Area under the Precision Recall curve is 0.073.

Metric/Class:	No	Yes
Accuracy	0.93	0.93
Precision	0.92	0.0
Recall	1.0	0.0
F1-Score	0.96	0.0

Table 1: Random Forest Model Evaluation Metrics

CONCLUSION AND FUTURE WORK

The accuracy of our model was greater than 90% under both classifications. When classifying “No”, precision and f-scores both above 90% and perfect recall (1.0), a desirable characteristic to avoid missing any CHD cases. However, when classifying “Yes”, precision, recall, and f-score are all zero. This discrepancy looks to be a result of our model classifying all cases as “No” for CHD without ever classifying any cases as “Yes” for CHD. This is evident in our confusion matrix where the amount of false negatives is 1888, and the amount of both false and true positives is zero. The area under the PR and ROC seem to correlate this abnormal behavior with subpar scores, 0.07 and 0.50, respectively.

Obtaining a low false negative rate is imperative in our application. However, a proper model should not accomplish this by simply classifying all cases as negative. A model that operates in this manner eliminates the advantages of a machine learning model in the first place.

Investigation of a truly appropriate model is needed. An RF model seemed the most appropriate due to its handling of categorical and numerical data and likelihood of remaining unaffected by outliers by binning variables. Additionally, appropriate tuning can benefit this model’s behavior. Feature importance and limiting inclusion to attributes that are statistically influential to our prediction are likely to avoid producing a nonsensically high recall and improve overall predictive performance.

Looking beyond other models and model performance, other considerations include the survey population. We believe that the dataset should include a ratio of individuals that is balanced and close to the true proportion of individuals, in terms of race, age, etc. The dataset that we used did not have a truly representative model of the US population. Only when we have a representative sample of the American population is when we can confidently use this model for healthcare purposes. A model that predicts with the highest degree would ensure confidence with all parties that would be using it. A training dataset of one million rows or CHD dataset that has data between 2010-2022 would ideally create a much better model.

Future studies could address the question of data gathering: the effect of additional data on the model and where would performance plateau. It would be beneficial to identify the point where a model’s training is sufficient. We come from the assumption that there is no such thing as having too much data for training. As long as you have enough computing power, you will create a good enough model. But would there be a statistically significant difference in adding an extra 200,000 data instances for prediction? Is there a minimal amount of data that is required in order to make a model that has learned from the dataset? This is the next question that we would want to answer for future studies.

The answers that we would gain from this study would be somewhat revolutionary to the healthcare industry. Having the ability to predict with confidence if someone will have CHD based on behavioral factors will be beneficial to preventive medicine. This would also improve public health to show that these specific behavioral factors are big predictors in determining if you're going to get heart disease.

Another benefit for the healthcare industry would be that it could help build trust with the public. The public still holds trepidation when it comes to machine learning and healthcare. Many individuals have little trust in either of these industries due to the notion that both fields want as much information as possible on all individuals of the population. If the study shows that there is no difference in the amount of data required for training purposes, the public would accept that level of risk. Theoretically, this could ease concerns that companies want to know everything about an individual and the population. Paradoxically, if more data produces a better model, public opinion must be swayed where more information (through proper regulation in data acquisition such as HIPAA) results in more effective disease detection and creating better preventative healthcare methods to improve population health.