Alexia Marcous
Springboard Data Science Career Track
August 5th2019 Cohort

Capstone 1 Consolidated Report: Predicting Hypertension

The problem and proposal

One in three American adults have high blood pressure, which greatly increases the risk for heart disease and stroke, the first and third leading causes of death in the United States[1]. Untreated, close to 50% of people with untreated hypertension die of ischemic heart disease and nearly 30% die of stroke in the United States.[2] Globally, 1 in 4 men and 1 in 5 women world wide have hypertension, with a disproportionate amount, two-thirds, living in low- and middle-income countries.[3]

Given the intense demands on families, communities and healthcare systems in diagnosing and treating this epidemic, the ability to identify predictive factors, particularly preventable factors, could make a significant difference in the quality of for billions of people around the world.

This project will use data from the National Health Interview Survey to create a model for predicting hypertension. This model could be used by governments and healthcare institutions to prioritize public health programs and other resources aimed at screening and prevention. Hospitals could use this data to focus on those features during intake interviews, identifying the opportunity for earlier diagnosis and treatment, as well as education.

Data collection and wrangling

The National Health Interview Survey (NHIS) ranks as one of the largest surveys conducted annually by the U.S. government. On average, the survey covers about 100,000 people in about 42,000 households each year, sampling the civilian, non-institutionalized population (this excludes anyone who is incarcerated, residents of nursing homes and members of the armed forces living in barracks).

The Integrated Public Use Microdata Series (IPUMS) is part of the Institute for Social Research and Data Innovation at the University of Minnesota and houses the world's largest collection of individual-level census and survey data. IPUMS collects, preserves and

harmonizes data from the NHIS and provides easy access to this data with enhanced documentation. They make their data freely available at ipums.org.

Feature and Sample Selection

80 features from the NHIS were selected on the basis of 1) availability since 1997, when these features were consistently available, and 2) relevance to the project topic, predicting hypertension. 22 samples representing yearly surveys taken from 1997 to 2018 were selected. These extract parameters resulted in a 449.2mb csv file containing 2,061,980 rows.

Initial Review of Missing/Unusable Values

A search for missing data in the resulting dataset revealed that one feature was missing from 1997 but was available since 1998. Another feature was actually only collected since 2013. These 2 features were dropped from data extract parameters, leaving 78 features.

As important as considering missing values was the presence of positive responses, which would contribute to the ability to predict hypertension. The reasoning here was that if there were very few positive responses to a particular question, little could be gleaned from its impact on whether a person had hypertension or not. For example, the "Activity limitation from: Alcohol/drug problem" column had no missing values, but only 346 of the responses in the 2,061,980 rows indicated that there was any activity limitation. Many features were found to have very limited numbers of positive responses and 20 features with less than 10,000 usable responses were eliminated, leaving 58 features.

Pandas Profiling, 1st Iteration

This dataset with 58 features and 2,061,980 rows was then evaluated using Pandas Profiling – a library created for exploratory data analysis that can produce a report of individual feature evaluation, summary statistics, correlations and warnings regarding anomalies and missing data. The report finished in 30 hours and 21 minutes, but produced a large amount of very helpful information, which would have taken longer to program and produce from scratch!

Of particular interest was the clarification of the distribution of survey responses to yes/no questions over a greater possible number of responses then yes/no. For example, the target feature, hypertension, had 6 possible values, with this distribution:

Data from Pandas Profiling for Target Variable on dataset with 2 million rows:

	Count	Frequency
0 – Not in Universe	1390284	67.4%
1 - No	470288	22.8%
2 – Yes	200354	9.7%
7 - Unknown-refused	406	< 0.1%
8 - Unknown-not ascertained	30	< 0.1%
9 - Unknown-don't know	618	< 0.1%

To better identify what rows actually were relevant, rows where the value of the target column was 0 (not in universe) or 8 (not ascertained) in this column were dropped, resulting in a dataset with 671,666 rows. This also reframed the approach from predicting a binary outcome to one of 4 possible outcomes, but with the vast majority falling into the yes/no categories.

Pandas Profiling, 2nd Iteration

Given the reduction of rows to 33% of the original dataset and the impact on the prediction results, the Pandas Profiling report was produced again to better understand this new dataset.

To make the profiling report more understandable, a mapping for more descriptive names was created and the columns in the dataset were renamed. Also, 30 was maximum readable number of features in the correlation heat maps, after which the axis labels overlapped and could not be deciphered. To make the heat maps readable, profiling was updated to run on batches of columns of a variable number, adding in the target column to each batch so it would be included in the correlations.

Impact on features

The new profiling report on the hypertension-focused dataset revealed interesting changes in prevalence for some features. For example, “ever told had angina pectoris” went from 67.4% NIU values in the 2m dataset to 0% in hypertension dataset. “Needed but couldn't afford dental care, past 12 months” went from 56.3% NIU to also having no zeros in the hypertension-filtered data. Others features changed in the opposite direction, for example “Ever told had a learning disability” went from 86% to 100% NIU.

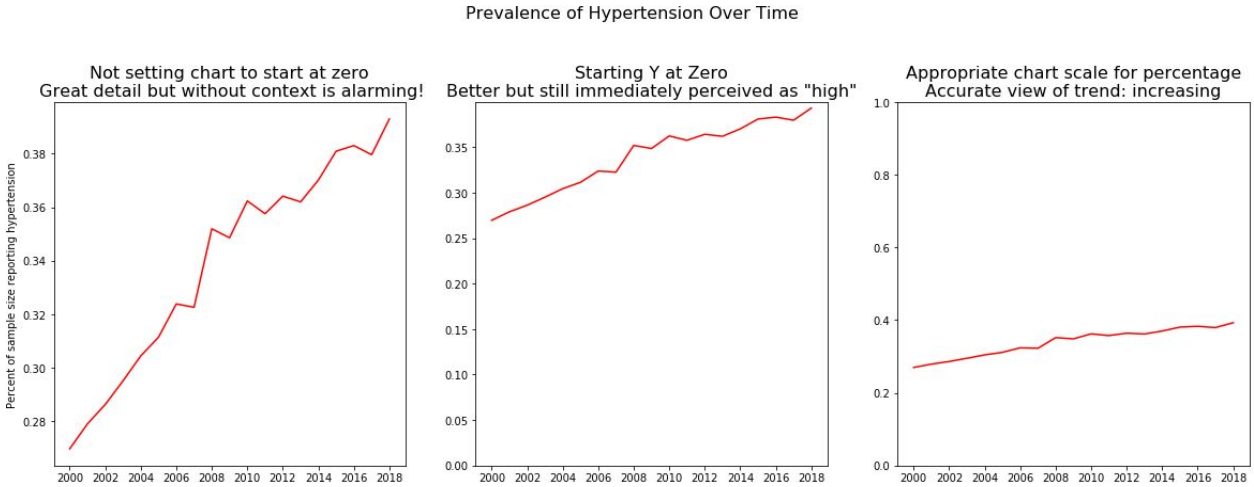
Removing more features

Careful examination of code book was conducted based on the distribution of values in the profiling report. This proved to be extremely important, because while most features were coded as 0 for NIU, some features used different values to indicate that the question was not asked. For example, "Felt everything an effort, past 30 days" showed 23.7% zeros, but in looking at the common values on the report, 67.4% were coded as 6. The codebook revealed that 0 meant "none of the time" and 6 represented NIU, for a total of 91.1% unusable responses. Another example was "Days had 5+ drinks, past year" – which showed 12.8% zeros but most responses were coded as 996, which also represented NIU, and a total of 93% unusable responses. The same went for "Number cigarettes per day (current smokers)", with no zeros but tallying 93.6% NIU with other coded values.

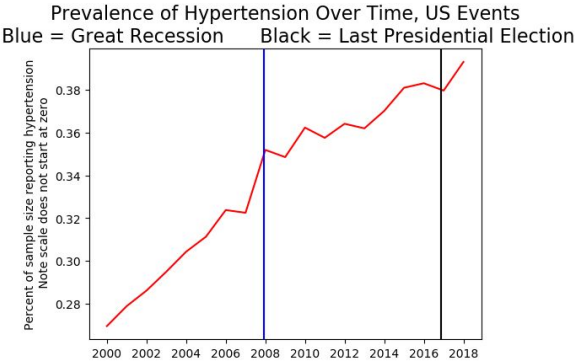
This thorough review resulted in dropping 11 additional columns due to a low percentage of usable data, resulting in 39 columns in the dataset.

Exploratory data analysis summary

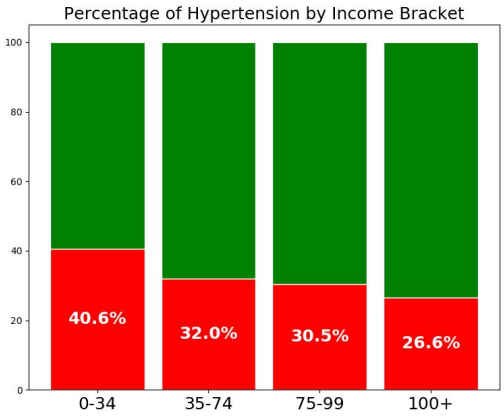
Hypertension was declared an epidemic over twenty years ago.2 Is it still an issue, or is our blood pressure now under control? To answer this question, we can look at the trend of hypertension in the dataset prepared for this project. To adjust for variable numbers of surveys conducted each year, positive hypertension responses were plotted as a percentage of each year size, yielding a more accurate visual for the **hypertension trend**. Graphing this data proved to be an excellent example for the importance of proper scaling:



This trend answers the question about relevance and shows that hypertension is clearly an ongoing, increasing problem, and predicting hypertension would be a valuable contribution to healthcare.



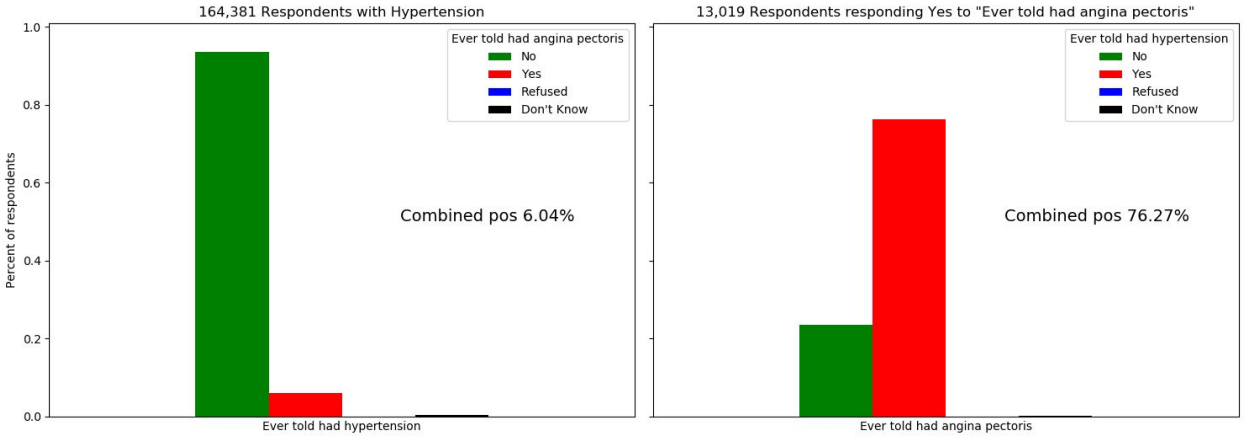
Viewing the trend line also raises questions about what could be causing the spikes. Two major events occurred around points of increase: the great recession and the most recent presidential election. It would be very interesting to look deeper into the correlation of financial/political events and blood pressure. With the current dataset, we can ask **is there a relationship between hypertension and income?**



Looking at the percentage of hypertension by income bracket, we can see that there is a decreasing trend in prevalence rates across income brackets.

Stress has not been definitively confirmed to cause hypertension,[5] but has been linked to risk factors for high blood pressure.[6] Stress from financial instability at the low end of the spectrum and the possible compounding of inadequate access to healthcare, nutrition and other socioeconomic factors could increase the vulnerability of this population to hypertension.

In considering factors that compound hypertension, we can look at the relationship between **hypertension and other chronic conditions**. Here a very interesting phenomenon appeared. When looking at the data from the perspective of respondents with hypertension, there was a low percentage of these respondents with other chronic conditions. However, when looking at the data from the perspective of respondents with a chronic condition, there was a high percentage of these respondents with hypertension. For example, as we see on the left, 6% of people with hypertension said they had experienced chest pain (angina pectoris), but in those with chest pain, 76% had hypertension.

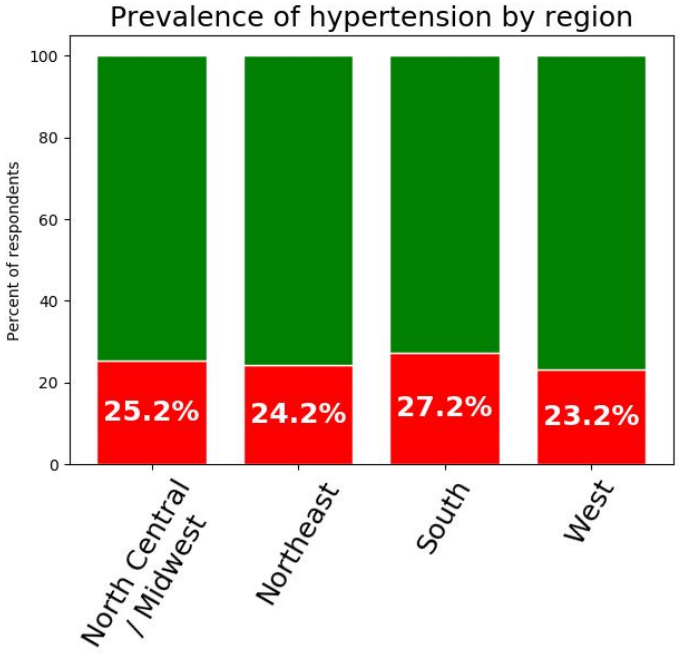


This was also true for cancer, diabetes, emphysema, heart attacks, heart conditions and vision problems. Hopefully this is the result of people being diagnosed and treated for hypertension before it leads to other problems. This would be a great area to explore further.

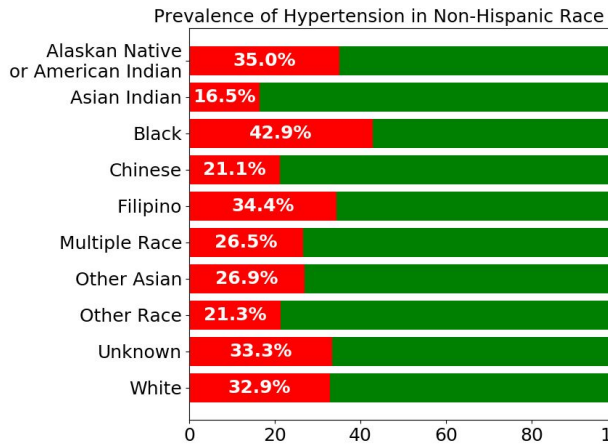
Given the goal of predicting hypertension we can also ask **do hypertension prevalence rates differ by demographics?** In this dataset, differences were found among gender, region and race features.

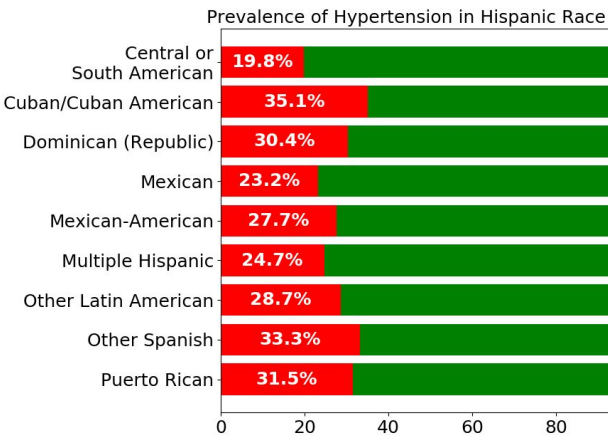
Prevalence rates are higher among men than women, and higher in the southeastern region - not surprising since it was coined the "stroke belt" in the 1940's and maintains disproportionately high stroke mortality rates compared to the rest of the country.[7]

	Male	Female
No hypertension	64.93%	66.68%
Have hypertension	34.91%	33.20%
Refused to answer	0.07%	0.06%
Status Unknown	0.10%	0.06%



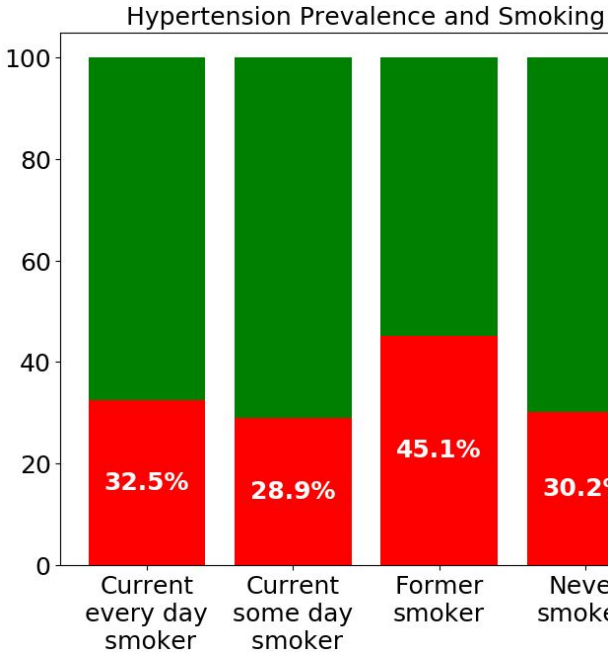
Among race groups, people who identified themselves as Black (non-hispanic) and Cuban/Cuban America had the highest prevalence of hypertension.





Having reviewed hypertension prevalence trends over time, and across income, health and demographic features, a final area to explore are behavioral features, given that lifestyle choices can increase risk for hypertension[8]. This dataset contained information on smoking, allowing for asking the question **are hypertension prevalence rates different for smokers?** The results showed an interesting finding.

Surprisingly, people who currently smoke some but not all days of the week have lower prevalence rates of hypertension than people who have never smoked. Using data from this same datasource, a Jan 2019 study found that occasional smokers have a 72% higher risk of death from cancer, heart disease and respiratory disease.[8] Also, the 45% prevalence rate of hypertension among former smokers should serve as a warning for smokers - a history of smoking will likely catch up with you.



Other behavioral features such as diet, exercise and alcohol consumption have also been identified as risk factors for hypertension,[8] but unfortunately were not present in this dataset. Including those features from another dataset could help improve prediction accuracy.

This analysis revealed that there are many features in the dataset with relationships to hypertension, and the hypothesis for this capstone project is that these features can be

used to predict hypertension. The analysis also revealed that many relationships can be explored further, indicating that other datasets with additional features describing those relationships could be useful for more accurate predictions.

Relevant Statistical Inference Techniques

Data used in this project from the National Health Interview Survey is nominal, where the responses are categorical with no ranking (ex. gender, race, and answers to questions such as whether the person ever smoked or had a stroke). Without continuous data, summary statistics like the mean and standard deviation are not appropriate. Instead, counts, percentages and modal values between feature categories are better ways of examining this data. The Data Story document for this project highlighted features that exhibited potential relationships between different features and between features and the target upon review of these summary statistics. Whether those relationships are statistically significant will be tested with Spearman's R, Cramer's V, Theil's U and a nominal version of the Chi-Squared test for independence.

Examining Correlations

Given that collinearity among features can lead to inaccuracies and variability in modeling, the strengths of any correlations were evaluated between all features. Spearman's test for correlation was chosen because it is nonparametric, meaning it does not rely on any assumption regarding the distribution of the data. The test returns a measure of monotonicty, or how much the features change in synch with each other, where -1 and +1 indicate complete negative or positive synch, and 0 indicates no relationship. Features with an absolute test result greater than 0.3 were removed and the machine learning models were re-run. The F1 value for positive hypertension fell for each model type. This could be due to the inclusion of ranking in the Spearman test, and/or that 0.3 was too low of a cutoff to imply a true correlation.

To remove any consideration of ranking, Cramer's V was run between all features, a nominal version of Pearson's Chi-Square Test for independence between categorical data. The test returns values between 0 for complete independence and 1 for complete dependence in terms of the effect a category for a feature has on the probability of a category for another feature occuring. Compared to the 28 values greater than 0.3 in the feature-feature Spearman results, there were 56 Cramer's V results greater than 0.3, including all of the > 0.3 Spearman features. Given the negative impact on the removal features from the Spearman results, no features were removed.

With regard to feature-to-target correlations, the results of the two tests were much closer. The highest correlation value from both tests was 0.28, between "Ever told had diabetes". The near-zero p-values from both tests for all features indicate this is not due to chance.

Examining Significance of Data Story Relationships

The following table shows the Cramer's V test results for the features showing a potential relationship to hypertension the data story for this project:

Feature	**Cramer's V**
Income	0.095459
Gender	0.011612
Race	0.093175
Region	0.261307
Smoking	0.132648

The test statistics do not show strong correlations, with all p-values near zero.

The other features that showed a potential relationship with hypertension had to do with comorbidities, where the relationship appeared to be asymmetrical: when looking at the data from the perspective of respondents with hypertension, there was a low percentage of these respondents with other chronic conditions. However, when looking at the data from the perspective of respondents with a chronic condition, there was a high percentage of these respondents with hypertension.

The asymmetrical relationship was evaluated with Theil's U test, also referred to as the Uncertainty Coefficient. This test is based on conditional entropy - given the value of one feature, how many possible states does another feature have, and how often do they occur. The test output is in the range of [0,1], where 0 means no association and 1 is full association. This asymmetrical relationship was confirmed with different results whether hypertension was being tested with the feature, or the feature was being tested with hypertension. As illustrated in the data story, the associations were stronger from the perspective of the comorbidity being tested for the presence of hypertension.

Sample results from Theil's U Test with Comorbidities		Theil's U
Ever told had hypertension	Ever told had angina pectoris	0.018354
Ever told had angina pectoris	Ever told had hypertension	0.087731
Ever told had hypertension	Ever told had cancer	0.015805
Ever told had cancer	Ever told had hypertension	0.033008
Ever told had hypertension	Ever told had diabetes	0.061704
Ever told had diabetes	Ever told had hypertension	0.100838
Ever told had hypertension	Ever told had heart attack	0.024073
Ever told had heart attack	Ever told had hypertension	0.091252

Results and In-depth analysis using machine learning

Choosing Algorithms
Algorithms were chosen based on the properties of the data and target variable. The data from the National Health Interview Survey and is labeled and contains the target variable, hypertension, making this a supervised learning problem. The target variable has two categories, yes/no, making this a binary classification problem. Five algorithms were chosen for this problem type, including K-Nearest Neighbors (KNN), Logistic Regression, Random Forest Classifier, Adaptive Boosting (AdaBoost) and Extreme Gradient Boosting (XGBoost).

Choosing Evaluation Metrics
Evaluation metrics were chosen based on the properties of the data and the implication of false predictions. In this case, the data is imbalanced, with 39% of responses positive for hypertension. For this reason, F1 score was chosen over accuracy.

The F1 score was also appropriate due to a difference in the cost of false positives (predicting hypertension when the condition does not exist) and false negatives (predicting no hypertension when the condition does exist), making it better to focus on the true positive rate, also known as sensitivity or recall, and precision, or how "precise" the classifier is when predicting positive instances. Since the F1 score is the weighted average of precision and recall, if the F1 score is high, both precision and recall of the classifier indicate good results.

Initial Results
To establish a baseline for the results, the first four algorithms were run with default parameters on the dataset produced after data wrangling. Parameters were changed only to resolve issues, such as increasing the number of iterations to resolve a convergence warning for logistic regression. The results were:

	Positive F1	Negative F1
KNN	0.38	0.82
Logistic Regression	0.45	0.84
Random Forest	0.49	0.83
AdaBoost	0.47	0.88

Improving Results – Additional Data Wrangling and Feature Selection
The first approach at improving these results was to review the data again. Based on re-examining usable data, one feature was dropped and missing data for two features was removed, resulting in a dataset of 199,569 rows and 38 columns. Feature importance was reviewed but values were inconsistent among models. The second approach at improving these results was to test the impact of scaling the data. The scaling preprocessing method was tested with KNN, improving the positive F1 score to 0.47, so scaling was implemented. The models were re-run on this set, showing improvements on the positive F1 score:

	Positive F1	Negative F1
KNN	0.47	0.78
Logistic Regression	0.54	0.78
Random Forest	0.59	0.77
AdaBoost	0.57	0.78

Improving Results – Tuning Models, Additional Training/Testing
The third approach at improving these results was to adjust the settings on three top-performing algorithms and run them through more training and testing iterations. This was achieved with hyperparameter tuning and cross validation via grid search and random search methods. Multiple rounds of different values for parameters were performed, until the results leveled off here:

	Positive F1	Negative F1
Logistic Regression	0.54 (no improvement)	0.78
Random Forest	0.60 (0.01% improvement)	0.77
AdaBoost	0.58 (0.01% improvement)	0.79

Final Results - XGBoost
Having adjusted the data and tuned the models, the final approach was to try a different model. XGBoost was chosen for its consistent outperformance of other models. Using multiple rounds of hyperparameter tuning and cross validation, XGBoost produced the highest positive and negative F1 scores:

	Positive F1	Negative F1
XGBoost	0.60	0.78

XGBoost also produced a model with the highest precision, recall and specificity:
- Precision: if the model predicts hypertension, this result will be correct 68% of the time
- Sensitivity, recall or true positive rate: if the subject has hypertension, the model predicts this accurately 54% of the time
- Specificity or true negative rate: If the subject does not have hypertension, the model predicts this accurately 83% of the time

Summary
The results are less than what was hoped for, and while this model is specific and somewhat precise, it is not sensitive. Reasons for these results could be due to biases in the data. Not all candidates provide responses to the survey, creating selection bias. When the survey is conducted, not all questions are asked, and not questions are answered, creating a bias among the data. In addition, the information is self-reported rather than obtained from verified sources, creating opportunity for inaccuracies in the answers provided.

Fortunately, despite the low sensitivity, the model can still provide benefit. Tests with high specificity provide value when the result is positive, as they can be used for ruling in patients who have a certain disease. With the current results, the usefulness of the model is based less on its ability to predict hypertension, but more on its function as a screening tool for identifying people at risk for hypertension. This model could be used in community outreach efforts to raise awareness about hypertension and identify people who should follow up with their primary care provider to have their blood pressure tested. It could also be used to flag patients in the healthcare setting for a focused assessment of hypertension.

Future Improvements

These results could be potentially be improved with additional efforts on feature selection, such as using methods designed for this purpose available in Scikit-Learn (the SelectKBest method) and pruning features based on their importance in tree-based algorithms.

References, Problem and Proposal

1. https://www.cdc.gov/bloodpressure/infographic.htm
2. https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/in-depth/high-blood-pressure/art-20045868
3. https://www.who.int/news-room/fact-sheets/detail/hypertension

References, Exploratory Data Analysis

1. Zhou D, Xi B, Zhao M, Wang L, Veeranki SP. Uncontrolled hypertension increases risk of all-cause and cardiovascular disease mortality in US adults: the NHANES III Linked Mortality Study. *Sci Rep*. 2018;8(1):9418. Published 2018 Jun 20. doi:10.1038/s41598-018-27377-2
2. Chockalingam A, Campbell NR, Fodor JG. Worldwide epidemic of hypertension. *Can J Cardiol*. 2006;22(7):553–555. doi:10.1016/s0828-282x(06)70275-6
3. 2008 National Health Interview Survey (NHIS) Public Use Data Release. Available here: ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2008/srvydesc.pdf
4. Dorans KS, Mills KT, Liu Y, He J. Trends in Prevalence and Control of Hypertension According to the 2017 American College of Cardiology/American Heart Association (ACC/AHA) Guideline. *J Am Heart Assoc*. 2018;7(11):e008888. Published 2018 Jun 1. doi:10.1161/JAHA.118.008888
5. Mei-Yan Liu, Na Li, William A. Li & Hajra Khan (2017) Association between psychosocial stress and hypertension: a systematic review and meta-analysis, Neurological Research, 39:6, 573-580, DOI: 10.1080/01616412.2017.1317904
6. American Heart Association, "Managing Stress to Control High Blood Pressure", available here: https://www.heart.org/en/health-topics/high-blood-pressure/changes-you-can-make-to-manage-high-blood-pressure/managing-stress-to-control-high-blood-pressure
7. Karp DN, Wolff CS, Wiebe DJ, Branas CC, Carr BG, Mullen MT. Reassessing the Stroke Belt: Using Small Area Spatial Statistics to Identify Clusters of High Stroke Mortality in the United States. *Stroke*. 2016;47(7):1939–1942. doi:10.1161/STROKEAHA.116.012997
8. CDC, "Behaviors That Increase Risk for High Blood Pressure", available here:https://www.cdc.gov/bloodpressure/behavior.htm
9. Non-Daily Cigarette Smokers: Mortality Risks in the U.S.Inoue-Choi, Maki et al. American Journal of Preventive Medicine, Volume 56, Issue 1, 27 - 37