Alexia Marcous
Springboard Data Science Career Track
August 5th2019 Cohort

Capstone Project 1 – Predicting Hypertension
Milestone Report

Importance of Topic

One in three American adults have high blood pressure.[1] Untreated, half of these adults will die of heart disease and another third will die of stroke.[2] Globally, 1 in 4 men and 1 in 5 women have this condition, with two-thirds of those affected living in low- and middle-income countries.[3] The ability to predict hypertension from data could allow for targeted prevention, diagnosis and treatment programs, reducing the burden of disease and improving the lives of billions of people.

Data Source

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. 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.

Data Wrangling

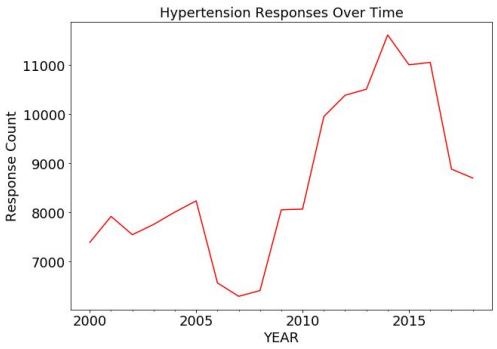
80 features from the NHIS were selected on the basis of availability and relevance. Surveys from 1997 to 2018 amounted to 2,061,980 observations. Analysis of missing data revealed 22 features with many years of missing observations, which were removed.

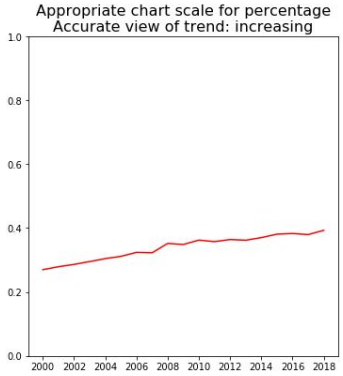
Since all survey data was encoded, there were no other missing values, which meant the usability of the data needed to be evaluated for this particular application. In this case, as important as considering missing values was the presence of positive responses, which would contribute to the ability to predict hypertension. 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.

The most important result of this evaluation was the discovery that the majority of the target variable observations could not be used. In 1,861,626 surveys, the question as to whether the respondent had ever been told they had hypertension was not asked. After unusable observations were removed, the final dataset contained 51 features and 670,642 observations, 33% of the original data extract.

Data Story

Exploratory data analysis was conducted on the dataset to learn more about the features and observations. Accurately assessing the trend of hypertension in the data required normalizing the data due to a deliberate reduction in the number of surveys completed between 2006-2008 in attempt “to achieve cost savings”.[4] The corrected trend shows an increase in the prevalence of hypertension.

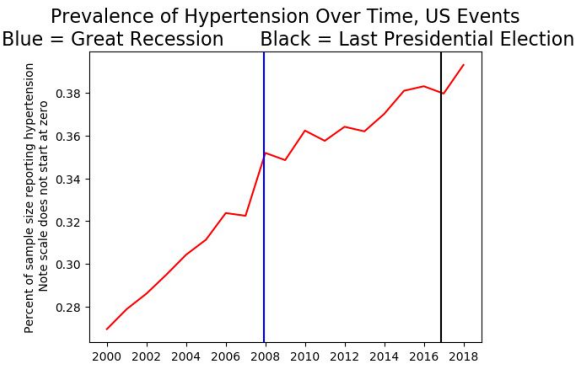


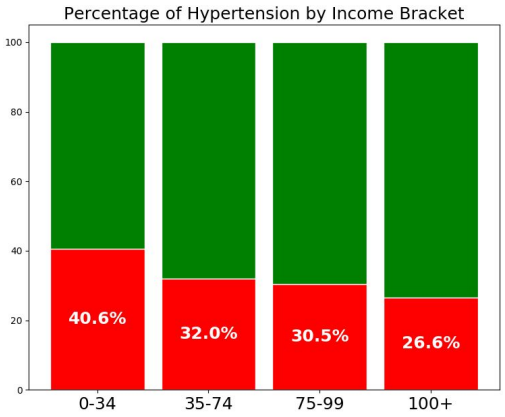


A 2018 study published in the Journal of the American Heart Association corroborated this trend, finding that the absolute burden of hypertension has consistently increased, from 87.0 million in 1999–2000 to 108.2 million in 2015–2016.[5]

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?**

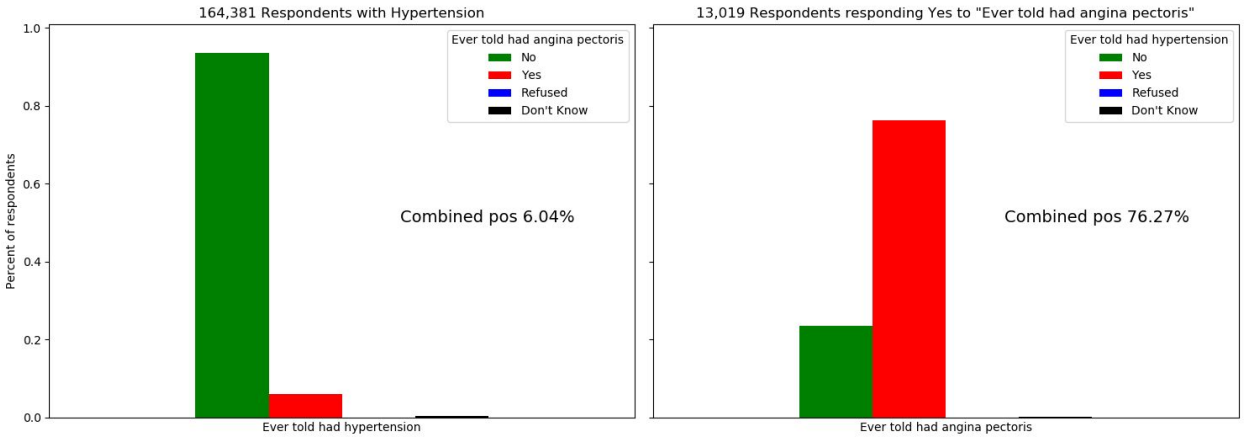




Looking at the percentage of hypertension by income bracket, we can see that there is a decreasing trend in prevalence rates across income brackets.

Stress has not been definitively confirmed to cause hypertension,[6] but has been linked to risk factors for high blood pressure.[7] 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.

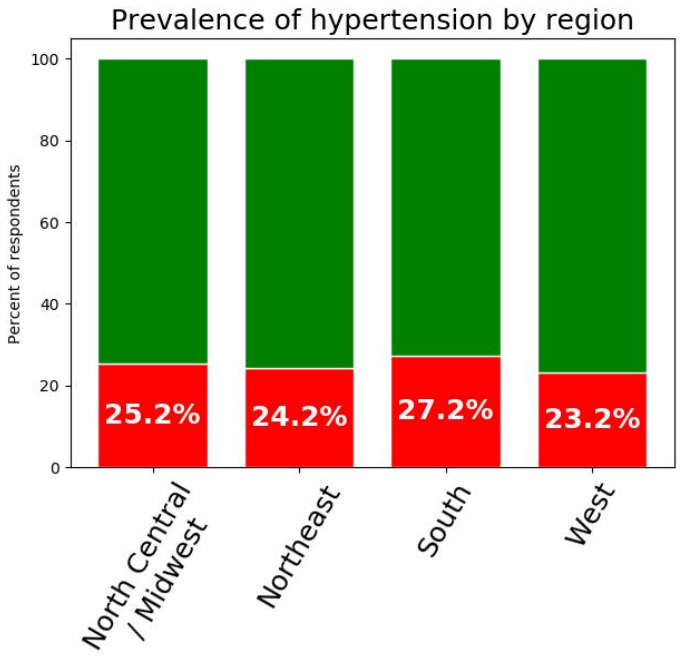


This was also true for cancer, diabetes, emphysema, heart attacks, heart conditions and vision problems. Hopefully this is the result of people being diagnosed and treated for hypertension before it leads to other problems. This would be a great area to explore further.

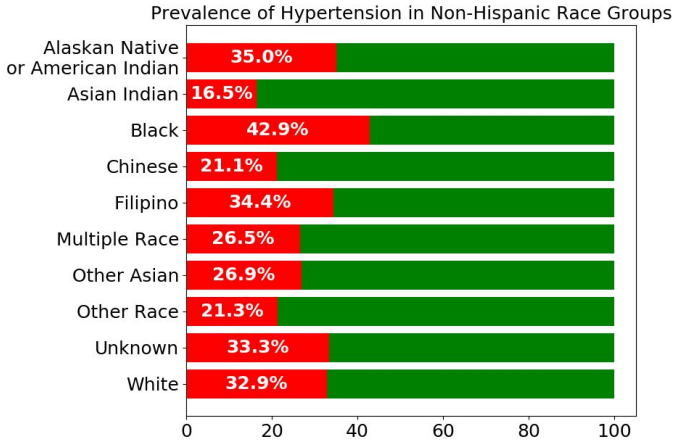
Given the goal of predicting hypertension we can also ask **do hypertension prevalence rates differ by demographics?** In this dataset, differences were found among gender, region and race features.

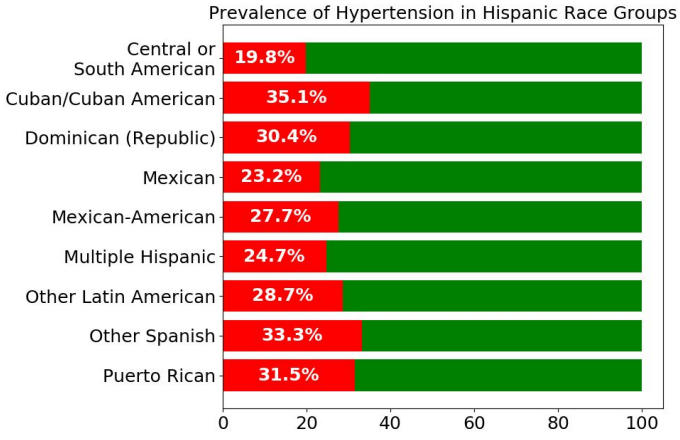
Prevalence rates are higher among men than women, and higher in the southeastern region - not surprising since it was coined the “stroke belt” in the 1940’s and maintains disproportionately high stroke mortality rates compared to the rest of the country.[8]

	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.

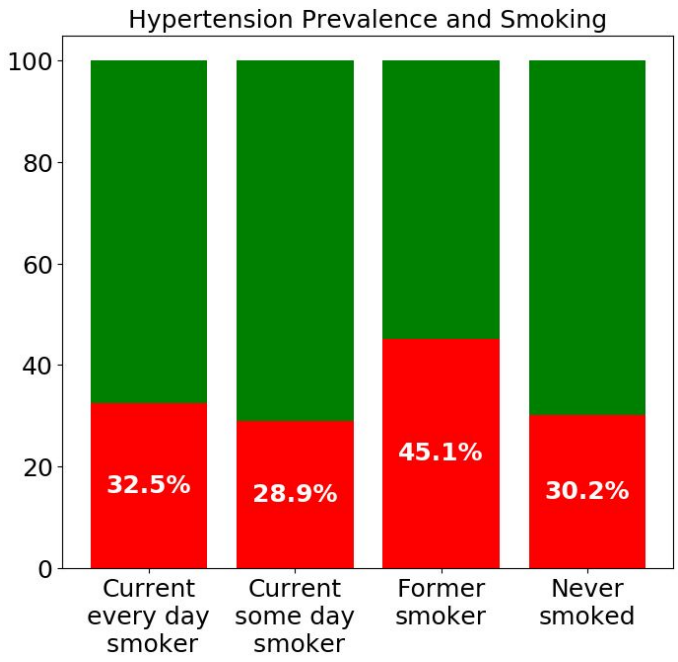




Having reviewed hypertension prevalence trends over time, and across income, health and demographic features, a final area to explore are behavioral features, given that lifestyle choices can increase risk for hypertension[9]. This dataset contained information on smoking, allowing for asking the question **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.[9] 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,[10] but unfortunately were not present in this dataset. Including those features from another dataset could help improve prediction accuracy.



Statistical Analysis

Given that the data from the NHIS is encoded with nominal values, the statistical analysis of the data was focused on testing for collinearity among features and between features and the target. Whether those relationships are statistically significant was tested with Spearman's R, Cramer's V, Theil's U and a nominal version of the Chi-Squared test for independence.

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. Given the reduction in F1 scores, the features were returned.

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 however, 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 visual data analysis revealed potential relationships between hypertension and income, gender, rage, region and smoking. The following table shows the Cramer’s V test results for these features:

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. This underscores the importance of following up visual data analysis with statistical analysis to determine if what looks like a relationship is actually statistically significant.

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

After wrangling, visualizing, and performing statistical analysis on the dataset, it does not appear that there are very strong correlations between any of the features and the target variable. However, machine learning algorithms may be able to uncover as-yet-unseen relationships, and having clean, organized data is the right preparation for this next step.

References

1. Centers for Disease Control and Prevention, “Make Control Your Goal”, www.cdc.gov/bloodpressure/infographic.htm

2. Mayo Clinic, “High Blood Pressure Dangers”, www.mayoclinic.org/diseases-conditions/high-blood-pressure/in-depth/high-blood-pressure/art-20045868

3. World Health Organization, “Hypertension” fact sheet, www.who.int/news-room/fact-sheets/detail/hypertension

4. 2008 National Health Interview Survey (NHIS) Public Use Data Release, ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2008/srvydesc.pdf

5. 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

6. 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

7. American Heart Association, “Managing Stress to Control High Blood Pressure”, 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

8. 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

9. Centers for Disease Control and Prevention, “Behaviors That Increase Risk for High Blood Pressure”, https://www.cdc.gov/bloodpressure/behavior.htm

10. 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