Alexia Marcous
Springboard Data Science Career Track
August 5th2019 Cohort

Capstone Project 1 – Predicting Hypertension
Statistical Data Analysis

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