Eric Buehler

Early Results

# Accessible Approach to Blood Pressure Prediction and Anomaly Detection

## Introduction:

High blood pressure, or hypertension is one of the most important risk factors for morbidity and mortality. Unfortunately, the measurement of blood pressure is not as straightforward as one might expect due to its inherent variability and inconsistencies in measurement techniques. New approaches have emerged to meet this problem. However, many of them are largely tailored towards privileged groups, ignoring the frequently cited observations that those in rural communities, low-income communities and of minority status have some of the highest risk of hypertension. Thus, the primary goal of this study is to investigate how machine learning might be used in a straightforward and accessible manner to assist healthcare workers in measuring blood pressure. The study accomplishes this by comparing newer modeling techniques against traditional OLS regression and using these techniques to design a simple tool that can determine if an observed blood pressure is anomalous or not.

## Methodology:

Blood pressures varies unequally across variables, so the stratification of individuals into different blood pressure groups has been shown to improve model quality (Stamler, Jeremiah, et al., 1975). Typically, this is done with a few demographic variables such as age and gender. However, the ability of K-modes and K-prototype to cluster on categorical and continuous data provides an attractive alternative to automate and potentially improve upon the standard method of stratification (Huang, Z., 1998).

In terms of predictions methods, the study calls for approaches whose predictions are easily interpreted. Thus, methods such as linear regression, LASSO regression, decision trees and, to a lesser extent, K-nearest neighbors are prime candidates. While many studies opt to

predict blood pressure labels such as “pre-hypertension” and “hypertension”, this study aims to predict continuous systolic and diastolic blood pressure. The prediction of continuous values is necessary for the goal of anomaly detection. However, it does introduce additional complexities to the study.

Once the best predictive technique is chosen, it will be used for the purpose of anomaly detection. This can be accomplished by comparing the blood pressure predicted by the model to an observed blood pressure. If the observed is over some predetermined cut-off distance from the predicted, it will be considered erroneous. If it is under, it will be considered valid. The distance between predicted and observed will be considered in terms of both systolic and diastolic using mahalanobis distance.

**Early Results:**

Recent progress for the study has largely been in two areas: establishing a predictive baseline and clustering. The predictive baseline for blood pressure was established by fitting an OLS regression model to both systolic and diastolic blood pressure. Predictors were mainly selected drawing on support from the literature on blood pressure prediction (Whelton, P. K., et al., 2018). Interaction terms were added to the model based on support from the literature and looking at two-way interactions using ANOVA. Assumptions of homoscedasticity, normality and multicollinearity were checked and corrected for as best as possible using transformations (log, squares, etc.) and interaction terms. Ultimately the following equations for systolic and diastolic blood pressure were produced:

- $systolic\ blood\ pressure \sim bmxarmc + bmxsad1 + bmxht + bmxwaist + bmxwt + riagendr + ridageyr + I(ridageyr\text{^}2) + hispanic + white + black + diq010 + lbdldl + log(lbxtc) + log(lbxglu) + bpq020 + bpq040a + bmxarmc{:}ridageyr + riagendr{:}lbdldl + riagendr{:}lbxglu + bmxsad1{:}diq010 + riagendr{:}ridageyr + bmxsad1{:}bpq040a + ridageyr{:}white + ridageyr{:}diq010 + ridageyr{:}lbdldl$

- $diastolic\ blood\ pressure \sim bmxarmc + bmxsad1 + log(bmxwt) + riagendr + ridageyr + I(ridageyr\text{^}2) + white + black + asian + diq010 + log(lbdldl) + lbxtc + bpq020 + bpq040a + ridageyr{:}white + bmxsad1{:}ridageyr + ridageyr{:}bpq040a + riagendr{:}lbdldl$

These models have R-squares of 34 and 22, respectively. While the R-squares leave something to be desired, the mean squared errors of the models of 180 and 93, respectively, were reasonable. This means that on average systolic blood pressure was off by 13 mmHg and diastolic, by 9 mmHg. While this isn't particularly impressive, it may be the case that when considering the predictions jointly, they may be more accurate. However, this requires further investigation.

**Figure 1:**

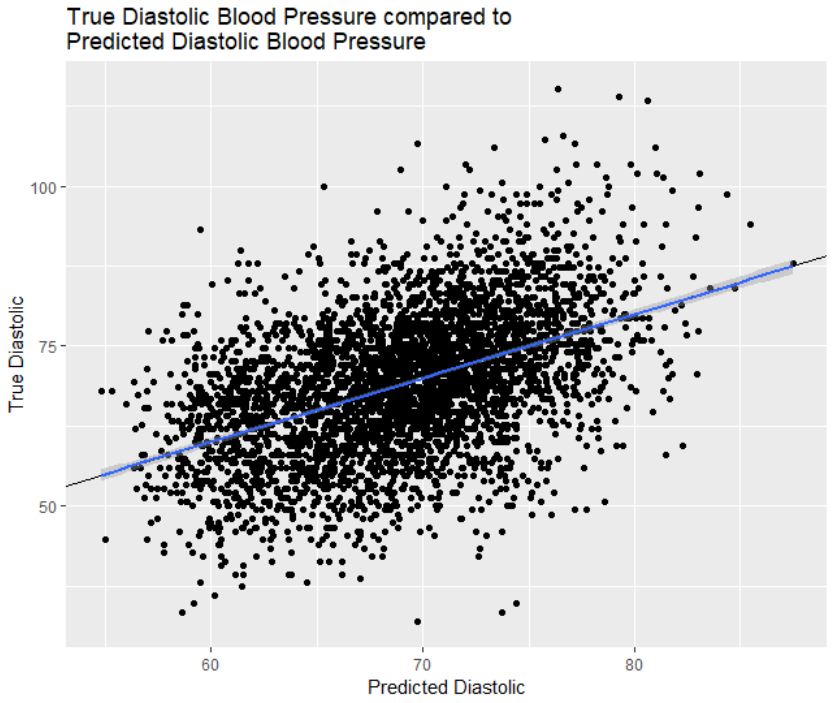
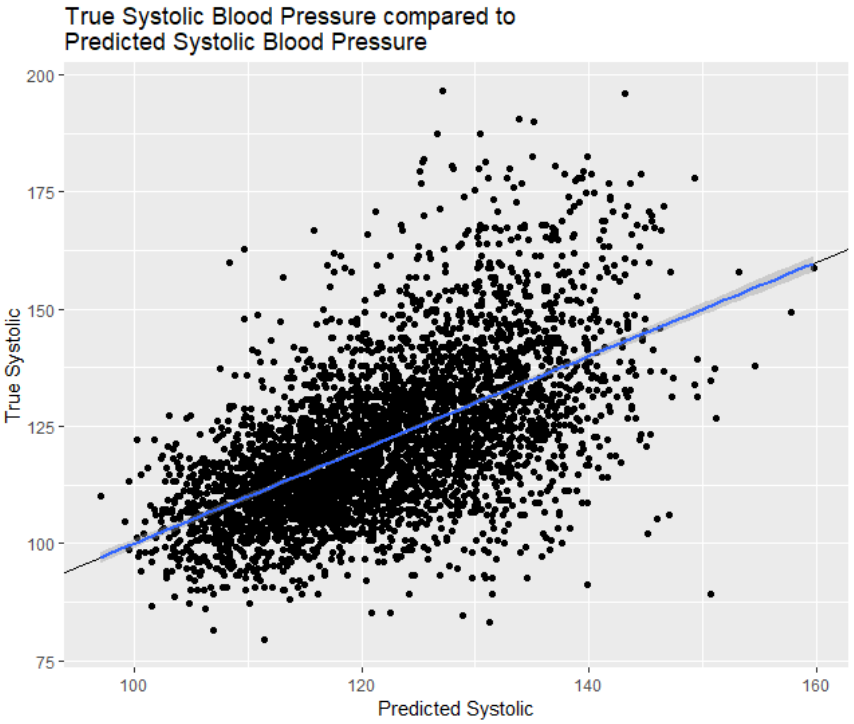


**Figure 2:**




Figures 1 and 2 helps illustrate the error of these models by showing the predicted systolic and diastolic values against the true value. Although the predicted vary from the true by about ten mmHg, the imposed line on both graphs shows that the predicted blood pressure values linearly predict the true values, showing that the general trend is maintained. Nonetheless, the models still leave something to be desired with high levels of variation at higher systolic values, low R-squares, and decent MSEs. Hopefully, more advanced techniques will provide the solution.

Clustering is the other area of progress. Since the NHANES data consists of a mix of both categorical and continuous data, traditional K-means clustering does not quite do the job. However, K-prototype clustering is an extension of K-means that allows both categorical and continuous data to be used for clustering. Cluster prototypes are computed as cluster means for

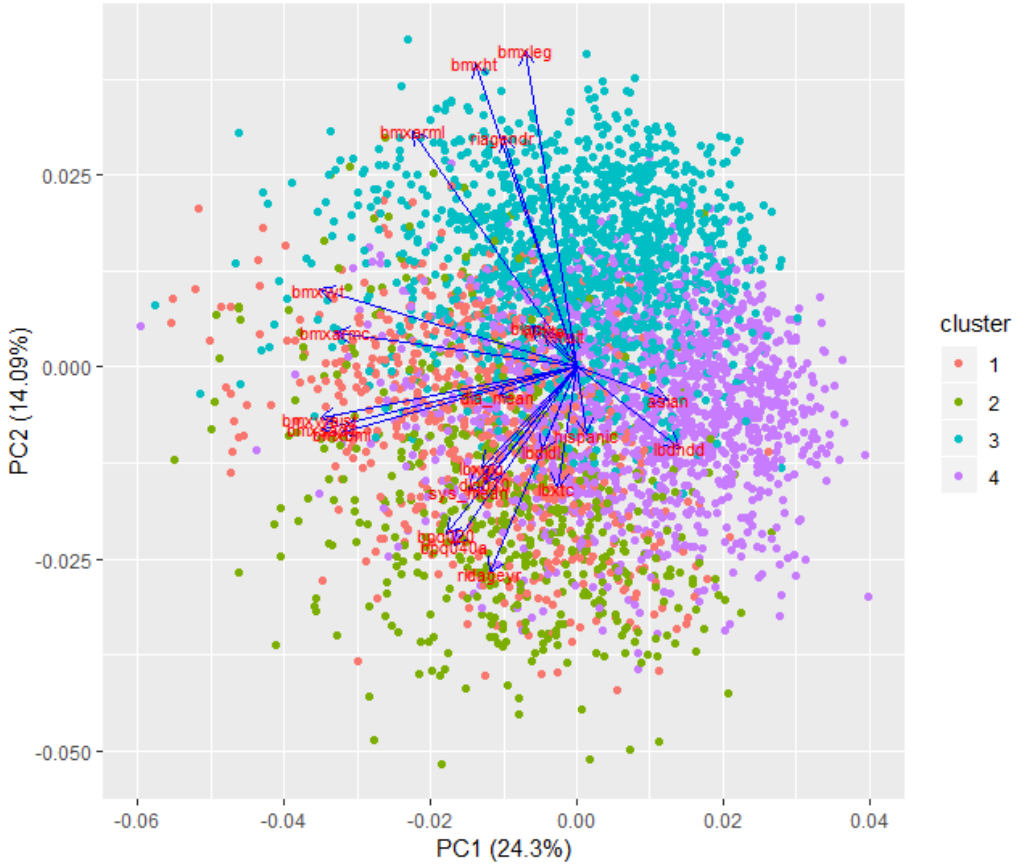
numeric variables and modes for factors (Huang, Z., 1998). Using K-prototypes and examining the total with distances, four clusters were determined to be an appropriate fit for the data. Table 1 shows the centers of these four calculated clusters.

**Table 1:**

| Cluster | Gender | Diabetes | Race | Age | Hypertension Meds | Hypertension |
|---|---|---|---|---|---|---|
| 1 | Male | No | White | 65 | Yes | Yes |
| 2 | Female | No | Black | 55 | Yes | Yes |
| 3 | Male | No | White | 39 | No | No |
| 4 | Female | No | White | 39 | No | No |

It appears that the first cluster generally consists of older white men who have hypertension and are taking hypertension meds. The second cluster consists of older black women who are also hypertensive. Clusters three and four are younger white male and females, respectively, who are not hypertensive and not taking hypertensive meds. Thus, it appears that the clusters identify stratifications that are not inconsistent with the literature in an automated way that can easily be extended to more variables. Figure three further helps visualize these clusters by plotting them on the first two principal components of the blood pressure variables selected from the NHANES data.

**Figure 3:**




The components include additional variables that were excluded from the clustering as they were either dependent variables (systolic and diastolic blood pressure) or lacked clarity in describing the clusters (complex bmi metrics). However, the inclusion of these variables does serve as an additional sanity check for the quality of the clusters. For example, cluster three sits high on the second component, which indicates a high arm (arml) and leg (bmxleg) bmi, and high on the first component, which indicates a lower waist (bmxwaist) and abdominal (bmxsad1) bmi. Thus, cluster three appears to consist of athletic individuals who have strong legs and arms, but slim waists and abdomen. The observation of an athletic individual is consistent with the center of cluster 3, which describes a young male without hypertension. These clusters are subject to

change from the addition of new variables or unforeseen modeling restraints. However, these early findings show a promising proof of concept.

My next steps are to design models for each cluster (Linear, Decision Trees, KNN), compare the models based on mahalanobis distance (comparison of multivariate distributions of diastolic and systolic blood pressure), and determine an appropriate cut off for mahalanobis distance in order to detect outliers from a predicted distribution.

## Work Cited


- Stamler, J., Stamler, R., Riedlinger, W. F., Algera, G., & Roberts, R. H. (1976). Hypertension screening of 1 million Americans: community hypertension evaluation clinic (CHEC) program, 1973 through 1975. *Jama*, *235*(21), 2299-2306.
- Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, *2*(3), 283-304.
- Whelton, P. K., Carey, R. M., Aronow, W. S., Casey, D. E., Collins, K. J., Himmelfarb, C. D., ... & MacLaughlin, E. J. (2018). 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Journal of the American College of Cardiology*, *71*(19), e127-e248.