DSCI 5240 – Data Mining and Machine Learning for Business

University of North Texas

Fall 2020

Project Final Report

Identifying Hypertension Cases Using Classification Models

By

Prathima Nuthalapati

November 22, 2020

Executive Summary

Hypertension plays a significant role in a group of risk factors associated with cardiovascular disease and leading causes of death. My goal in this application of data mining research is to predict hypertension in individuals based on health, biomedical, and socio-economic status data. Additionally, i have included study and documentation of findings within the literature related to understanding and predicting hypertension as a means to ground my current project. The business case is straightforward: Societies incur significant costs, expend considerable number of resources, and experience a loss of productivity due to a lack of timely diagnosis of serious health conditions. Except in extreme cases, hypertension is not readily identifiable by individuals due to the lack of recognizable symptoms. Therefore, a predictive tool can help health professionals to identify individuals with high chance of hypertension earlier and before performing the diagnostic clinical procedure.

The data set that i chose for the project comes from the National Health and Nutrition Examination Survey (NHANES). This data is the result of a biannual nationally representative cross-sectional study of adults and children in the United States administered by the Center for Disease Control (CDC). The NHANES studies include several surveys, measurements, and quantification of health-related variables. I've selected 52 variables as input features and synthesized a target binary variable for hypertension class label using average of blood pressure readings in the data set. For the initial round of variable selection, given hundreds of available parameters in the NHANES data sets, we reviewed the scientific literature to determine which features were selected by other scholars and analysts. I then added additional variables by examining each parameter captured in the NHANES data sets. Through my investigation i am able to select variables believed to have some association with blood pressure and hypertension. In this final report, i present methods in data pre-processing, exploratory analysis and conclusions in model classifications and summary.

The core of my analysis consists of the formation and evaluation of five different binary classification models: logistic regression, decision tree, random forest, gradient boosting, and neural networks. To compare the performance of the models i defined two key performance indicators aligned with the business goals: recall score and area under the receiver operating characteristic curve. The best performing model was gradient boosting. I selected this as my final predictive tool.

At the conclusion of this project, I've successfully developed data-driven insights and recommendations to better understand and make decision about hypertension, and an accurate and reliable tool that can be used to predict hypertension based on health, biomedical, and socio-economic status data. The results of my analysis could be utilized in hypertension classification and could be included as inference engines in expert systems for hypertension screening tools.

Contents

1 Introduction

1.1 Background

Hypertension is known as a primary or contributing cause of nearly half a million death throughout the Unites States in 2018 (Center for Disease Control [CDC], 2020). In medical terms, hypertension is defined as a systolic blood pressure equal to or above 130 mm Hg or a diastolic blood pressure equal to or above 80 mm Hg (American Heart Association, 2020). According to statistical results published in 2017 by the World Health Organization (WHO, 2018), more than 1.13 billion people globally are affected with Hypertension. In the literature, Hypertension is considered a primary factor in a group of risk factors associated with cardiovascular disease.

Almost half of American adults (108 million people) have active hypertension or are taking medication for managing it (Whelton et al., 2018). Hypertension is primary in a group of risk factors associated with cardiovascular diseases responsible for 17.9 million deaths worldwide in 2016 (World Health Statistics 2018: Monitoring health for the SDGs, 2020). A prominent application of data mining technologies is the identification of diseases based on the significant amount of health data generated by healthcare systems, personal health monitoring devices, and other relevant data such as socio-economic status data. In this project, we aim to apply data mining techniques to understand and predict patterns of hypertension among American individuals using their health, biomedical, and socio-economic status data.

1.2 Literature Review

I have examined background literature related to understanding and predicting hypertension as a source of health-related information to present grounding in identifying patients with high risks of developing hypertension. I've focused my review of the literature on research that utilized health data and applied different data mining techniques to describe and predict health-related issues.

Throughout the literature, various research utilized neural network models to predict the existence of hypertension in individuals. Other studies compared classification performance and accuracy with logistic regression. Performances of direct human blood pressure have a significant impact on heart disease. Through multiple logistic regression analysis, three factors were identified as having significant influence on the performance of human blood pressure (hypertension). These factors are age, body mass index (BMI), and systolic pressure. The results showed that all three of these factors could affect the performance of blood pressure at risk for hypertension (Ahmad et al., 2014).

According to Polak and Mendyk (2008), an artificial neural network model resulted with greater accuracy in predicting hypertensive patients compared with a logistic regression classification model using age, sex, diet, smoking and drinking habits, physical activity level, and BMI as input features. Lafreniere et al. (2016), conducted research which used an extensive Canadian Primary Care Sentinel Surveillance Network (CPCSSN) data set to predict hypertensive patients using a Backpropagation neural network model. The independent features were age, gender, BMI, systolic and diastolic blood pressure, high and low-density lipoproteins, triglycerides, cholesterol, microalbumin, and urine albumin–creatinine ratio. A confusion matrix and a Receiver Operating Characteristic (ROC) curve were utilized to measure the accurateness of the model. The research resulted with an accuracy of 82%.

According to Golino et al. (2014), classification tree analysis outperformed the traditional logistic regression model in the prediction of increased blood pressure by using body mass index (BMI), waist (WC) and hip circumference (HC), and waist hip ratio (WHR) for women showed best prediction with lowest deviance 87.4 whereas for men BMI, WC, HC, and WHC showed the best prediction with the lowest deviance of 57.25 and misclassification of 0.16.

Chang et al. (2019) proposed a model combining recursive feature elimination with a cross-validation method and classification algorithm. Support vector machine (SVM), C4.5, decision tree, random forest (RF), and extreme gradient

boosting (XGBoost) classification algorithms are used to predict patient outcomes. Results shows that C4.5, RF, and XGBoost can achieve very good prediction results with a small number of features. Among the four classifiers, XGBoost has the best prediction performance, and its accuracy, F1, and area under receiver operating characteristic curve (AUC) values are 94.36%, 0.875, and 0.927, respectively. We reviewed many other studies for predicting hypertension using different classification models, and all of them have advantages and disadvantages. However, the above mentioned are the most relevant.

1.3 The Business Case

Because of the significant role of hypertension in health condition and death, it is considered a public health risk with considerable economic implications. Therefore, it is in the interest of the stakeholders to have actionable knowledge regarding prevention. Although the diagnosis is relatively straightforward, patients may remain undiagnosed for several reasons. Diagnosis requires at least two clinical appointments, and, in some cases, doctors use continuous monitoring to confirm the diagnosis (American Heart Association [AHS], 2020). Data mining models could be used to inform diagnosis based on other health data. They could help find trends and patterns in the greater populations where conventional clinical methods are not feasible (Lopez-Martinez, et al., 2020; Lynn et al., 2009; Polak & Mendyk, 2008). The outcome of this project can be useful for virtually every stakeholder in the healthcare system including doctors, health insurance companies, healthcare providers, and policy makers.

Hypertension, can lead to strokes, heart failures, kidney-related diseases, and other related diseases. The primary goal of treating hypertension is to minimize high blood pressure and shield essential organs such as the brain, heart, and kidneys from further damages. Current reports show that more than 14 million individuals are unaware of hypertension and its underlying effects and therefore do not take medicine for it or participate in other blood pressure management procedures. The medical, economic, and human costs of high blood pressure that is untreated and inadequately monitored are huge.

Efforts on using electronic health records (EHRs) and other health information technologies to classify patients with chronic illnesses who are undiagnosed have recently increased to ensure prompt and effective diagnosis and care. Researchers have developed an algorithm to help detect patients with undiagnosed hypertension that would analyze the EHR data of patients and classify those who had multiple elevated blood pressure levels over 12 months. This algorithm was introduced to ten health centers that provide services for underserved populations. To discover previously undescribed variables correlated with cardiovascular outcomes, more advanced techniques have been used to evaluate data from large randomized controlled trials investigating various blood pressure targets.

The use of neural network models in the present day for the classification of diseases is growing rapidly, not only because of the substantial amount of available data generated by healthcare devices and systems, but also because of the magnitude of available computational resources for statistical analysis and processing. This large volume of data is used to train models and encourages the use of expert systems, natural language processing (NLP), and other techniques in the assessment of many diseases to observe trends and patterns. They may prove helpful for better diagnosis and cost optimization. In conjunction with patient outreach techniques, the use of classification and prediction models to categorize acceptable patient data can be effective ways for health center providers to recognize and assist patients at high risk of undiagnosed hypertension. In addition, patient harm and health care costs may be minimized by adopting this approach.

1.4 Analytical Approaches

In my analysis, I've used data samples collected through the National Health and Nutrition Examination Survey (NHANES) which is one of the most comprehensive publicly available health data sets and a principal source for tracking hypertension in the U.S. population (David et al., 2010). The NHANES data collection campaigns are designed to assess

the health and nutritional status of adults and children across the United States. The data in this survey is unique because it combines social determinants of health data such as smoking, alcohol consumption, dietary habits, and physical examinations with laboratory results. This survey instrument placed emphasis on the data regarding the prevalence of major diseases and risk factors for diseases for a broader population than just data from a medical facility. The latter contains only data for a small subset of the population and does not represent the entire picture of significant disease. In addition, historically, disease trends in the United States have been assessed by surveys.

After data preparation, i probed the data using exploratory analysis techniques such as graphical charts and statistical summaries. The objective of this step is to understand the data better and discover possible meaningful patterns. I'll perform a Chi-squared test to identify the variables that are not independent of the target variable. This step was not performed for the purpose of excluding any variables. But it did serve to better understand the relationship between the features and the target. Because i have calculated average values of the systolic and diastolic pressure measurements, i can build a linear regression model between blood pressure (dependent variable) and input variables. The objective of the project is not to predict the exact blood pressure values but building a linear regression model would be helpful to understand the relationship between the variables better.

At the core of my analysis is the building and evaluation of five different binary classification models: logistic regression, decision tree, random forest, gradient boosting, and neural networks. I've evaluated the performance of each model i calculated by using a confusion matrix, ROC plot, and AUC. Finally, i created an ensemble model to develop the best model out of the five classifiers. My goal was to predict hypertension with minimum false positive rate and maximum recall.

For this project I've used Python and Excel for part of the data pre-processing, Tableau for some of the visualization work, and SAS Enterprise Miner OnDemand for data exploration, model building, and evaluation.

1.5 KPIs

To evaluate and optimize my analytical approach i should define analytical metrics that match with the business objectives and processes. In this case, the business goal is to identify people with hypertension rather than missing a truly positive case (i.e. a person with hypertension) which would be significantly more expensive than misidentifying a negative case. If a person with hypertension is not identified, their condition would not be treated and therefore it might result in severe consequences for the patient and healthcare system. Therefore, in my classification model the false negative rate should be minimized. Figure 1-1 shows basic model performance evaluation metrics in a confusion matrix.

False negative rate is defined by $FNR = \frac{FN}{FN+TP}$ where FN+TP is the total number of true positive cases. Since

$$FNR = \frac{FN}{FN+TP} = 1 - \frac{TP}{TP+FN} = 1 - Sensitivity$$

to minimize the FNR score, the sensitivity or recall score should be maximized. Therefore, i choose the recall score as my first key performance indicator (KPI).

Although false positive rate (FPR) is cheap in this context, but i would want to avoid them as much as it does not hurt the recall score. False positive rate, which is equal to $1 - Specificity$, is the metric that should be minimized to avoid false positive identifications. To achieve both business goals, I use area under the receiver operating characteristic curve (ROC AUC) as my second key performance indicators. Figure 1-2 shows the tradeoff between sensitivity and false positive rate in a typical ROC. The optimum classifier that satisfies the business needs is the one with maximum recall score and ROC AUC.

		Predicted Hypertension Class		
		TRUE	*FALSE*	
Actual Hypertension Class	*TRUE*	True Positive (TP)	False Negative (FN)	**Sensitivity** $\frac{TP}{TP + FN}$
	FALSE	False Positive (FP)	True Negative (TN)	**Specificity** $\frac{TN}{TN + FP}$
		Precision $\frac{TP}{TP + FP}$	**Negative Predictive Rate (NPR)** $\frac{TN}{TN + FN}$	**Accuracy** $\frac{TP + TN}{TP + TN + FP + FN}$

Figure 1-1 Confusion matrix and basic model performance evaluation metrics

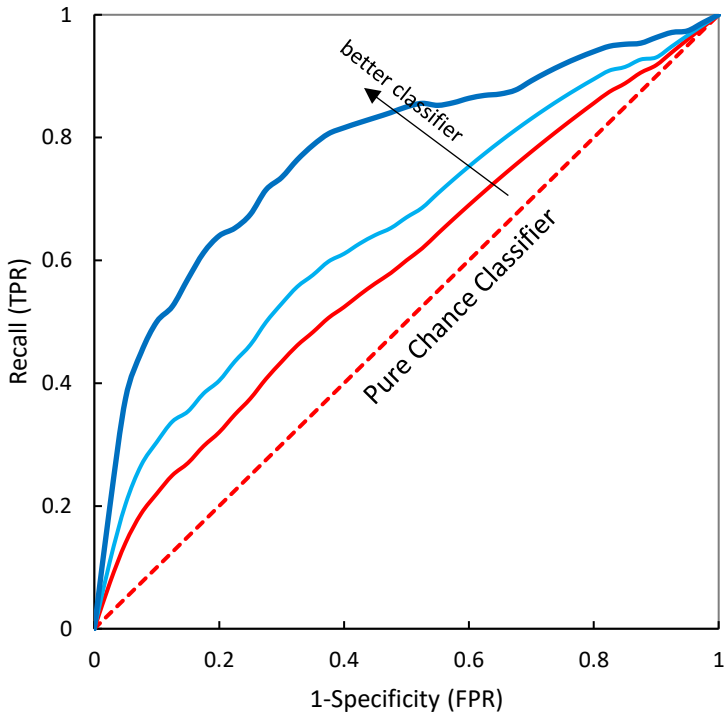


Figure 1-2 Receiver operating characteristic curves

2 Data Preprocessing

In this project I am using date from the National Health and Nutrition Examination Survey (NHANES) which is one of the most comprehensive publicly available health data sets and a principal source for tracking hypertension in the U.S. population (David et al., 2010). NHANES is a biannual nationally representative cross-sectional study of adults and children in the United States administered by the CDC (NHANES, 2020). The NHANES studies include several surveys, measurements, and quantification of health-related variables. Every two years, a sample is selected from the US population to be studied for their health through a rigorous methodology. The sample size has grown from approximately 13,300 in 2009 to almost 16,300 in 2018 (AHA, 2020). Every two years, results of the NHANES study are published through the CDC. The data sets contain many parameters related to demographic, dietary habits, laboratory measurements, health status, and medical examinations of the sample population. For this project I've collected, prepared, and analyzed data for five consecutive periods of NHANES study from 2009 to 2018.

For my study I've selected 52 variables as input features and synthesized six variables into three target variables. For the initial round of variable selection from hundreds of available parameters in the NHANES data sets, I've studied the scientific literature to find out what features have been selected by other scholars and analysts (see Chapter 1). In the next step, I've added more variables by going through all parameters in the NHANES data sets and selecting those that might have some association with blood pressure and hypertension. Out of the 52 selected variables, 25 are interval and the remaining variables are categorical. Descriptions of the input and target parameters are presented in Appendix A

The objective of this analysis is to understand and predict hypertension through data. In the NHANES studies, blood pressure of almost one-third of the sample is measured and recoded. Each blood pressure reading is given in two numbers: systolic pressure and diastolic pressure. The former is the maximum pressure the heart generates while beating and the latter is the amount of pressure in the arteries between heart beats. Each individual whose blood pressure is measured in the NHANES can have up to four reading of systolic and diastolic pressures. I took the average of four reading for each record (individual) and generated two variables named *systolic_ave* and *diastolic_ave*. These are interval variables that are used for data exploration and generating classification labels. According to the American Heart Association (2020) guidelines, if systolic pressure is below 130 mm Hg and diastolic pressure is below 80 mm Hg, the person is not hypertensive; otherwise she has stage 1 hypertension. I've applied this guideline to create a new Boolean variable called *is_hypertension* as the target of classification. When *is_hypertension* is True, the individual has hypertension.

For data preparation I've used Python. The results of each round of NHANES study are published on the CDC website as several SAS Transport File Format (XPT) files. In the NHANES data sets each individual (record) is identified with a unique ID. Therefore, all features of a record can be retrieved by using that ID across multiple files. I've downloaded XPT files and converted them to CSV format. Then, for each NHANES period, I inner joined the CSV files using the unique ID numbers. Finally, I appended all periods of NHANES data to create a single CSV file to use in SAS EM. The prepared file contains 22,151 records and 57 features (52 input features + 1 ID + 1-year marker + 2 blood pressure variables + 1 hypertension class label). Figure 2-1 shows a snapshot of the input data set after being imported to SAS EM. I've performed data exploration in SAS EM through using DMDB, StatExplore, MultiPlot, and Graph Explore as shown in Figure 2-2. Note: I've later added an additional ID filed to the compiled data set that allowed us to randomly select 70% of the records to train our models.

SEQN	year	hypertension	BPXPLS	BPXPULS	BPQ020	BPQ080	RIAGENDR	RIDAGEYR	RIDRETH1	INDHHIN2	INDFMPIR	RIDEXPRG	DIQ010	DIQ160	DIQ1
51624	2009-2010	TRUE	70	1	2	.	1	34	3	6	1.36	4	2	2	
51628	2009-2010	TRUE	72	1	1	1	2	60	4	3	0.69	.	1	.	
51629	2009-2010	FALSE	72	1	2	2	1	26	1	6	1.01	4	2	2	
51630	2009-2010	FALSE	86	1	1	.	2	49	3	7	1.91	.	2	2	
51643	2009-2010	TRUE	86	1	1	1	2	42	4	7	2.35	2	1	.	
51645	2009-2010	TRUE	92	1	1	1	1	66	1	2	0.41	4	2	2	
51647	2009-2010	FALSE	62	1	2	2	2	45	3	14	5	.	2	2	
51648	2009-2010	FALSE	68	1	2	.	1	28	1	3	0.09	4	2	2	
51653	2009-2010	TRUE	88	2	2	1	1	44	3	8	4.43	4	2	2	
51654	2009-2010	FALSE	60	1	1	1	1	66	3	6	2.2	4	2	2	
51655	2009-2010	TRUE	86	1	2	.	2	49	1	6	1.45	.	2	2	
51656	2009-2010	FALSE	62	1	2	1	1	58	3	15	5	4	2	1	
51657	2009-2010	TRUE	76	1	2	.	1	54	3	10	2.2	4	2	2	
51658	2009-2010	TRUE	84	1	2	.	2	26	1	8	1.35	2	2	2	
51660	2009-2010	FALSE	68	1	2	.	1	32	1	6	0.97	4	2	2	
51661	2009-2010	FALSE	62	1	2	1	2	60	1	7	2.75	.	2	2	
51662	2009-2010	TRUE	68	1	2	1	2	44	1	15	3.97	2	2	2	
51664	2009-2010	TRUE	70	1	2	1	2	53	3	15	5	.	2	2	
51666	2009-2010	TRUE	94	1	2	.	2	58	1	14	2.03	.	2	2	
51667	2009-2010	TRUE	74	1	1	1	1	50	3	4	1.24	4	2	2	
51669	2009-2010	TRUE	68	1	1	1	2	65	1	4	0.54	.	2	1	
51670	2009-2010	TRUE	60	1	2	.	1	22	3	8	0.74	4	2	2	
51673	2009-2010	FALSE	96	1	2	.	2	30	3	4	0.68	2	2	2	
51676	2009-2010	FALSE	58	1	1	1	1	27	4	14	2.71	4	2	2	
51677	2009-2010	FALSE	96	1	2	.	1	33	3	6	1.27	4	2	2	
51678	2009-2010	TRUE	84	1	1	1	1	60	3	4	1.03	4	2	1	
51680	2009-2010	FALSE	98	1	2	1	2	60	4	6	2.59	.	2	2	
51683	2009-2010	FALSE	86	1	2	2	2	35	3	14	4.1	2	2	2	
51685	2009-2010	FALSE	64	1	1	2	2	56	3	14	5	.	2	2	
51690	2009-2010	TRUE	80	1	1	1	2	63	2	7	1.59	.	1	.	
51691	2009-2010	TRUE	70	1	2	2	2	57	3	77	.	.	2	2	
51692	2009-2010	FALSE	64	1	1	1	1	54	2	9	3.28	4	2	2	
51694	2009-2010	TRUE	60	1	1	2	1	38	3	5	1.15	4	2	2	
51696	2009-2010	TRUE	70	1	1	.	1	54	4	4	1.39	4	2	2	
51697	2009-2010	FALSE	74	1	2	.	1	27	4	5	1.58	4	2	2	
51699	2009-2010	FALSE	60	1	1	.	2	26	1	6	1.13	2	2	2	
51701	2009-2010	TRUE	68	1	2	2	1	36	3	10	2.2	4	2	2	
51702	2009-2010	TRUE	80	1	1	1	1	44	3	7	1.81	4	1	.	
51704	2009-2010	FALSE	72	1	2	2	1	54	1	9	2.13	4	2	2	
51705	2009-2010	FALSE	80	1	2	.	2	25	1	12	.	2	2	2	
51707	2009-2010	TRUE	94	1	2	2	2	64	5	3	0.69	.	1	.	
51708	2009-2010	FALSE	92	1	2	1	2	39	3	6	1.22	2	2	1	
51709	2009-2010	FALSE	68	1	1	1	1	40	1	7	1.32	4	2	2	

Figure 2-1 A snapshot of the input data set table in SAS EM

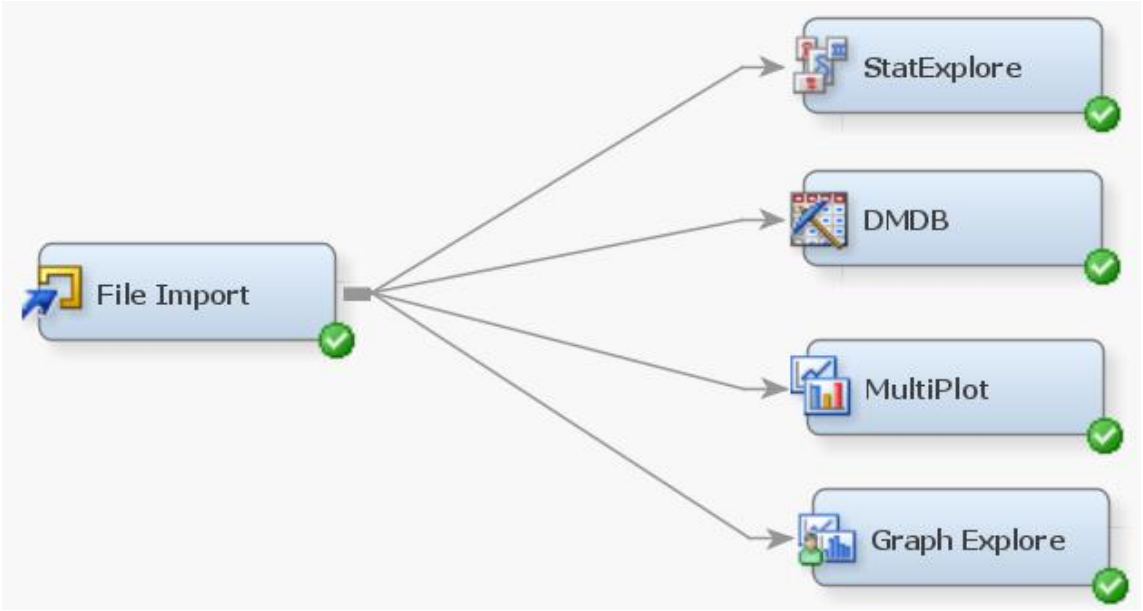


Figure 2-2 Data exploration diagram in SAS EM

2.1 Interval Variables

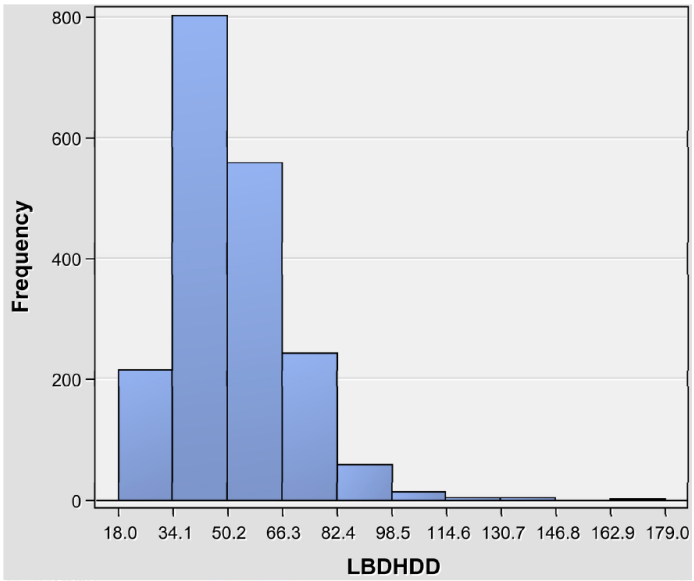
In the input variables, there are 25 interval and 27 categorical parameters. Descriptive statistics of the interval variables are shown in Table 2-1. The percentage of missing data is between 0 and 10%. I feel that amount is within an acceptable range. The majority of the interval variables are the results of lab work conducted on survey participants, and therefore it is expected these results be normally distributed with skewness to the left or right, depending on the measured metric. The remaining variables are expected to exhibit near normal distributions. Figure 2-3 through Figure 2-7-- show the distributions of lab results. The distributions of some of the variables such as Aspartate Aminotransferase (LBXSASSI), Creatinine, refrigerated serum (LBXSCR), and Triglycerides, refrigerated serum (LBXSTR) are extremely left-skewed, while most of the other variables are close to a normal distribution with small left or right skewness.

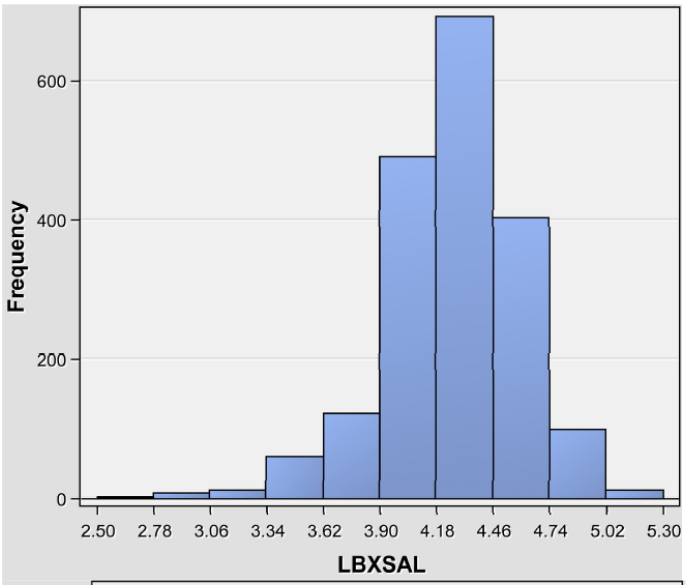
Table 2-1 Descriptive statistics of the interval inputs

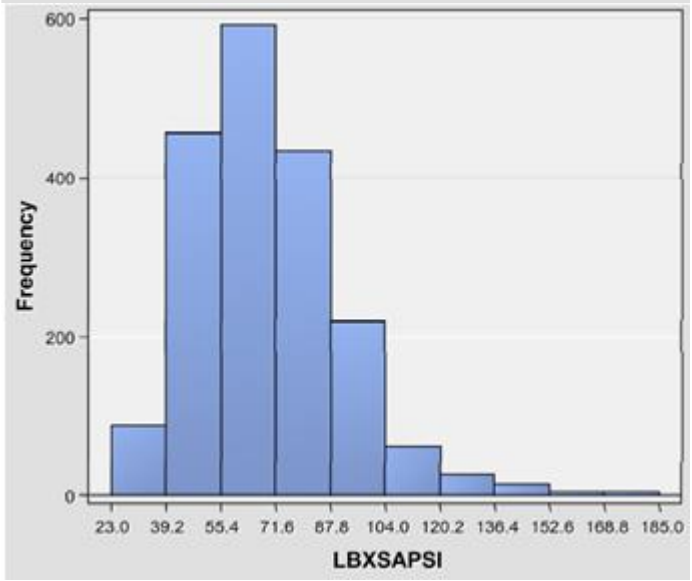
Variable	Mean	Std. Dev.	N	Missing (%)	Min.	Median	Max.	Skewness	Kurtosis
BMXBMI	29.42	7.23	21977	0.79%	13.6	28.2	84.87	1.23	2.84
BPXPLS	73.01	11.86	22151	0.00%	36	72	172	0.62	1.02
DBD895	3.44	3.93	22115	0.16%	0	2	21	2.01	4.71
DR1TSUGR	113.61	80.30	20686	7.08%	0	97.19	1115.5	2.30	12.00
INDFMPIR	2.48	1.65	20113	10.13%	0	2.06	5	0.36	-1.31
LBDHDD	52.61	16.02	20943	5.77%	6	50	226	1.17	3.39
LBXSAL	4.24	0.35	20884	6.07%	2	4.3	5.6	-0.45	1.15
LBXSAPSI	70.13	25.10	20879	6.09%	7	66	907	4.75	96.45
LBXSASSI	25.36	18.00	20862	6.18%	7	22	882	17.92	639.87
LBXSBU	13.01	5.14	20880	6.09%	1	12	96	2.80	22.02
LBXSCA	9.37	0.36	20855	6.21%	6.4	9.4	14.8	0.34	4.29
LBXSCLSI	103.39	3.00	20883	6.07%	70	104	118	-0.49	1.65
LBXSCR	0.87	0.43	20882	6.08%	0.16	0.83	16.64	14.37	333.14
LBXSGB	2.93	0.46	20851	6.23%	1	2.9	7.5	0.79	2.88
LBXSGL	101.91	39.98	20881	6.08%	19	92	777	4.91	35.11
LBXSIR	84.90	36.28	20855	6.21%	2	81	557	1.13	4.55
LBXSKSI	3.97	0.34	20880	6.09%	2.4	3.96	7.3	0.41	1.55
LBXSNASI	139.34	2.33	20883	6.07%	102	139	154	-0.38	5.00
LBXSOSSI	278.36	5.04	20878	6.10%	207	278	315	0.13	3.94
LBXSPH	3.71	0.57	20878	6.10%	1.7	3.7	9.6	0.37	1.71
LBXSTP	7.18	0.46	20852	6.23%	4.7	7.2	11.3	0.21	1.20
LBXSTR	153.56	135.98	20867	6.15%	9	119	6057	8.85	234.56
LBXSUA	5.38	1.43	20876	6.11%	0.7	5.3	18	0.56	0.74
LBXTC	192.68	41.40	20943	5.77%	59	190	813	0.89	4.76
RIDAGEYR	44.43	14.33	22151	0.00%	20	45	69	-0.02	-1.21

Figure 2-8 shows the correlation matrix between all lab work variables. In general, between most pairs there is very weak correlation, which makes them suitable for a linear regression model. Some variables such as LBXSCR (Creatinine, refrigerated serum), LBXSGL (Glucose, refrigerated serum), LBXSTR (Triglycerides, refrigerated serum) do not have any level of correlation with any variable. On the other hand, noticeable correlations are observed between other pairs such as LBXSGB-LBXSTP (Globulin and Total Protein), LBXSNASI-LBXSOSSI (Sodium and Aspartate Aminotransferase), and LBXSTP-LBXSAL (Total Protein and Albumin, refrigerated serum).

The histograms of BMI measurements and heart pulse per minute are shown in Figure 2-9. Both are left skewed, with pulse being close to a normal distribution. Two nutrition related features are total sugar intake (DR1TSUGR) and the number of meals not home prepared (DBD895), for which the distributions are shown in Figure 2-10. Both parameters are left-skewed. Two last interval variables are age (RIDAGEYR) and income Ratio of family income to poverty (INDFMPIR) and their histograms are shown in Figure 2-11. According to the US Census Bureau (2020), INDFMPIR is a metric that indicates poverty level of the household. If it is below 1.0, it means the total family income is less than the poverty threshold. The INDFMPIR of most individuals is above 1.0, with a peak between 4.5-5.0. The age distribution shows that NHANES sample in each period is age stratified.







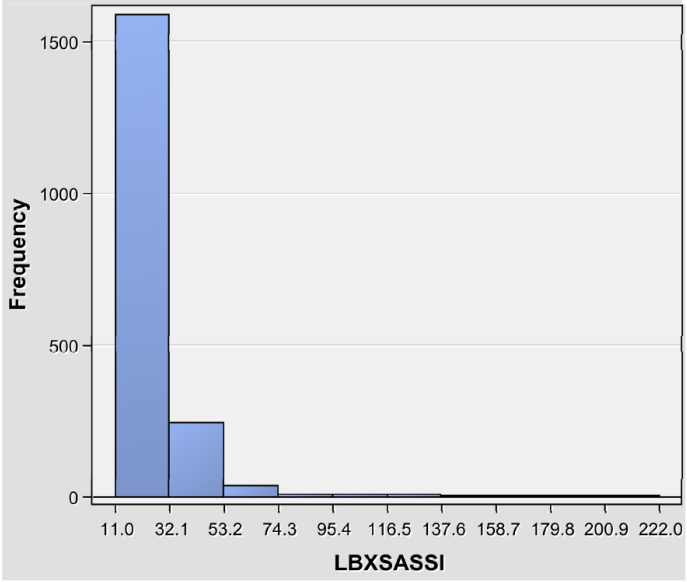


Figure 2-3 Histograms of Direct HDL-Cholesterol (LBDHDD), Albumin, refrigerated serum (LBXSAL), Alkaline Phosphatase (LBXSAPSI), Aspartate Aminotransferase (LBXSASSI)

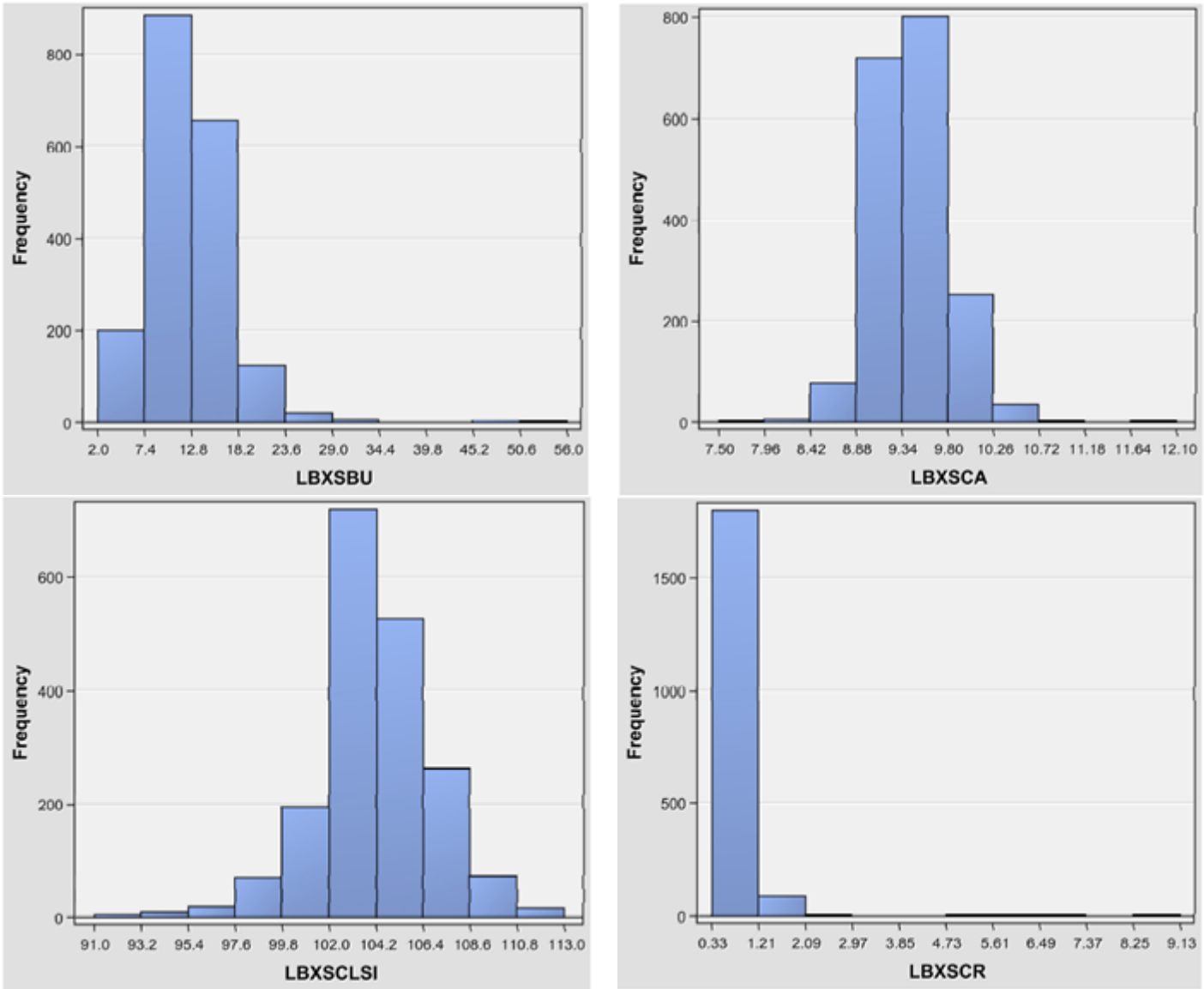


Figure 2-4 Histograms of Blood Urea Nitrogen (LBXSBU), Total Calcium (LBXSCA), Chloride (LBXSCLCI), Creatinine, refrigerated serum (LBXSCR)

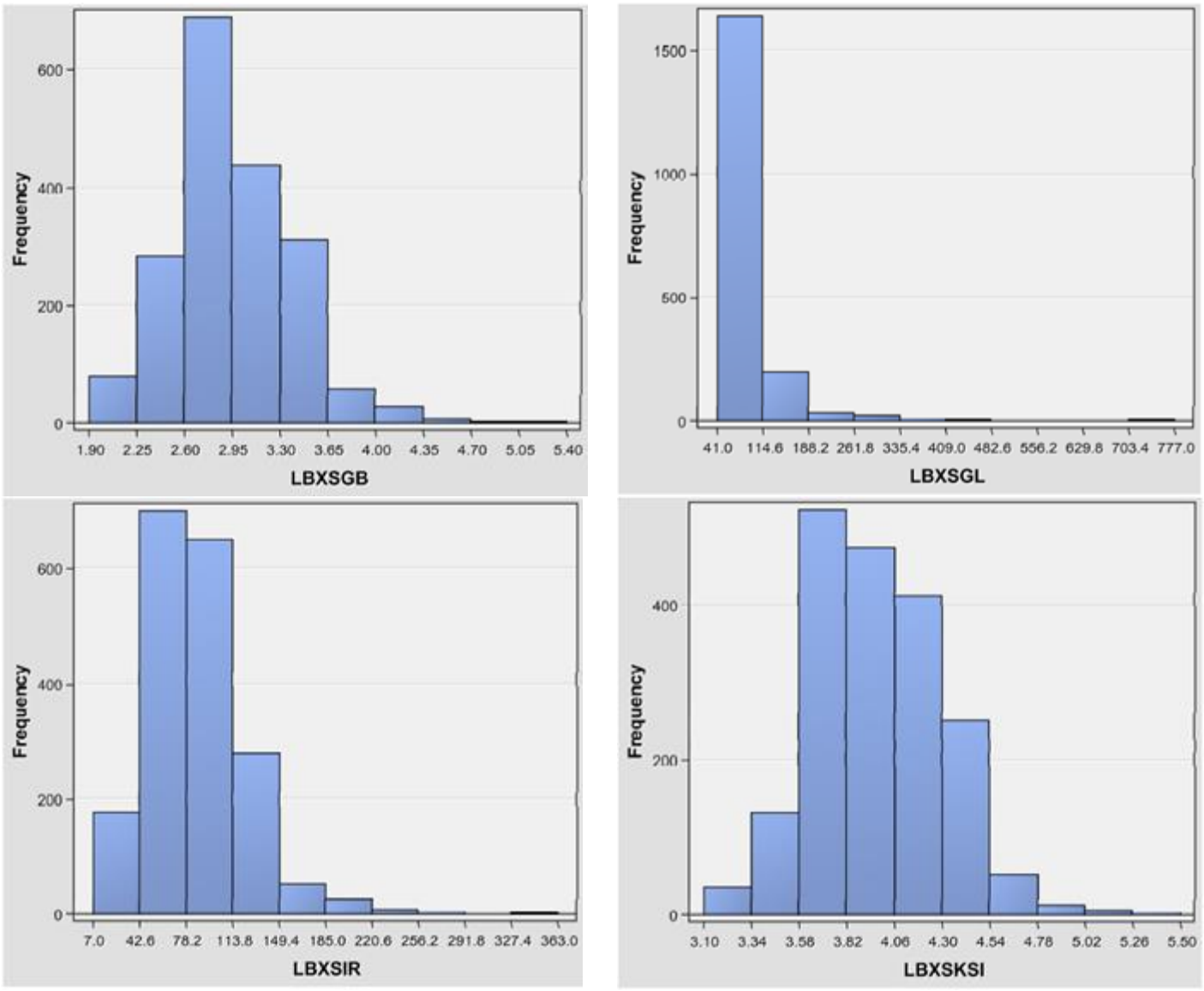


Figure 2-5 Histograms of Globulin (LBXSGB), Glucose, refrigerated serum (LBXSGL), Iron, refrigerated serum (LBXSIR), Potassium (LBXSKSI)

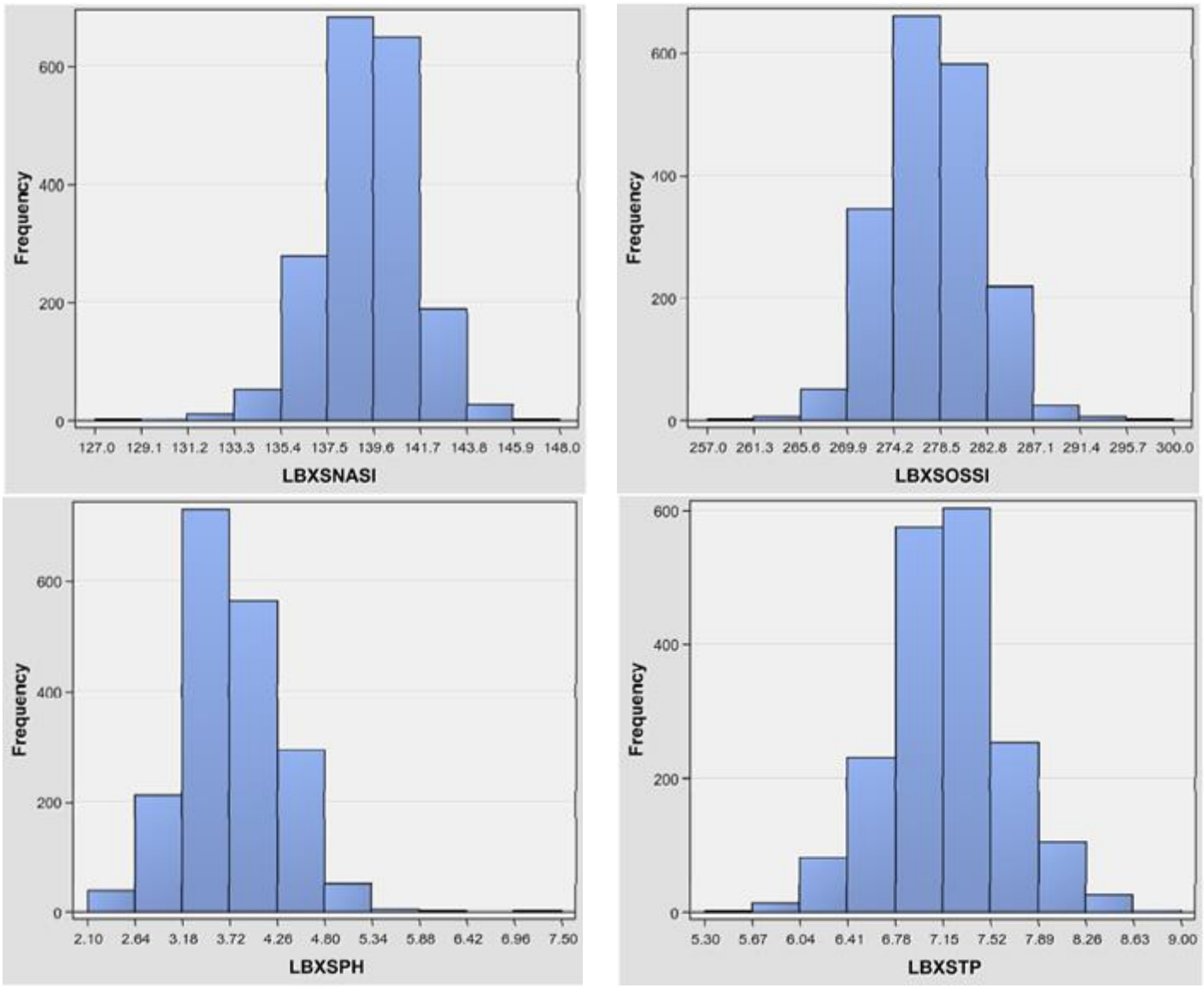


Figure 2-6 Histograms of Sodium (LBXSNASI), Osmolality (LBXSOSSI), Phosphorus (LBXSPH), Total Protein (LBXSTP)

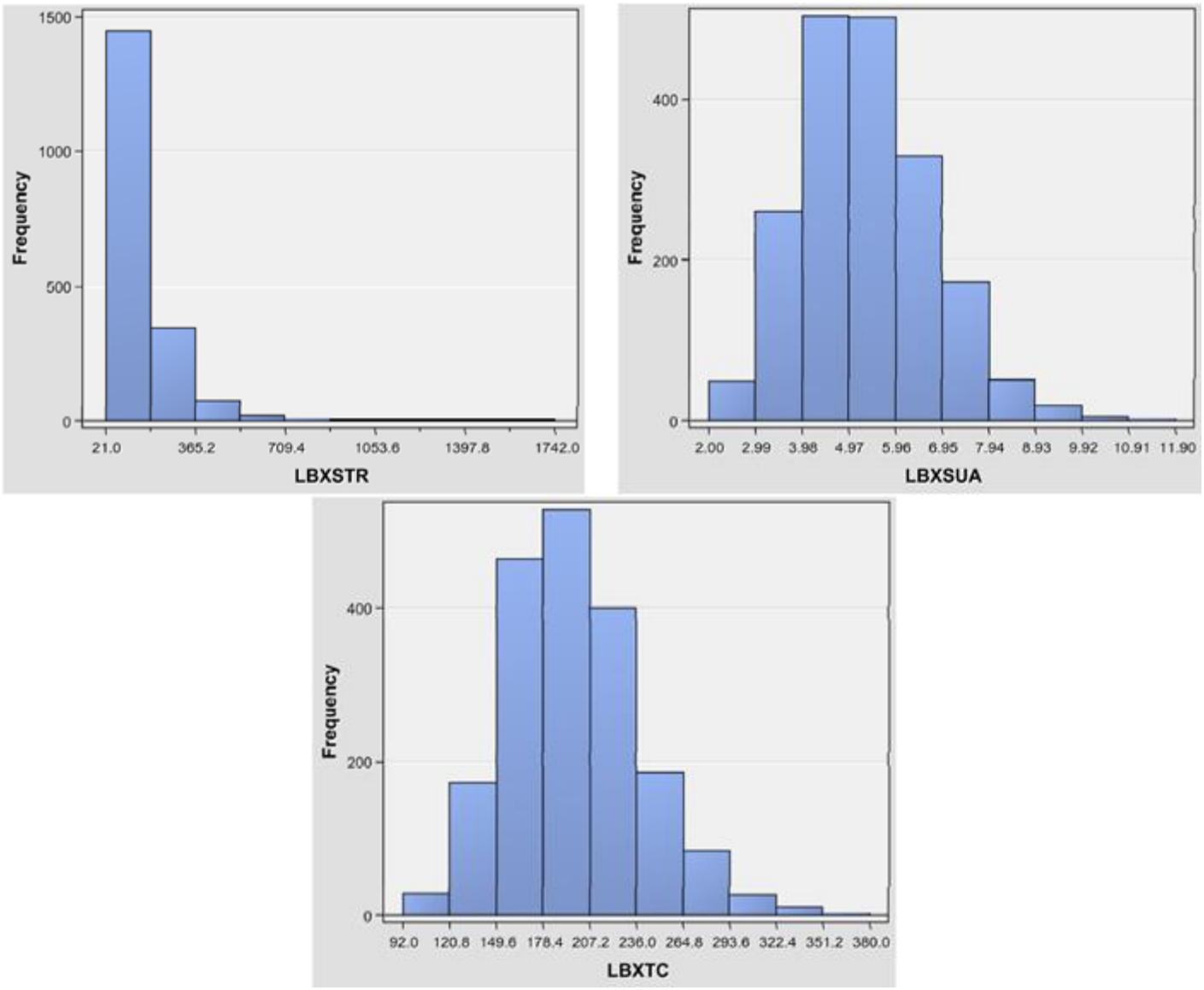


Figure 2-7 Histograms of Triglycerides, refrigerated serum (LBXSTR), Uric acid (mg/dL), Uric acid (mg/dL), Total Cholesterol (LBXTC)

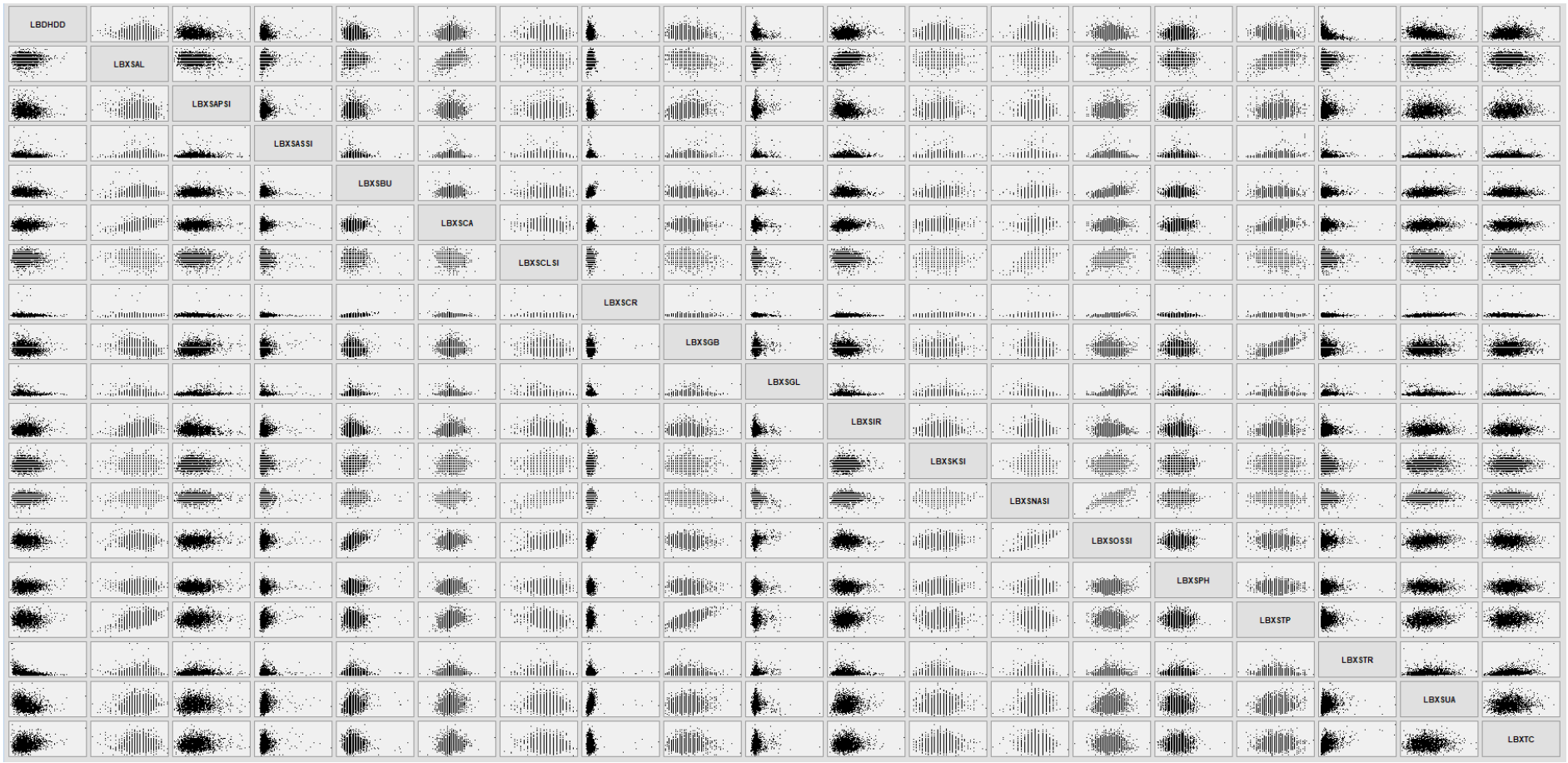


Figure 2-8 Correlation matrix between the lab result variables

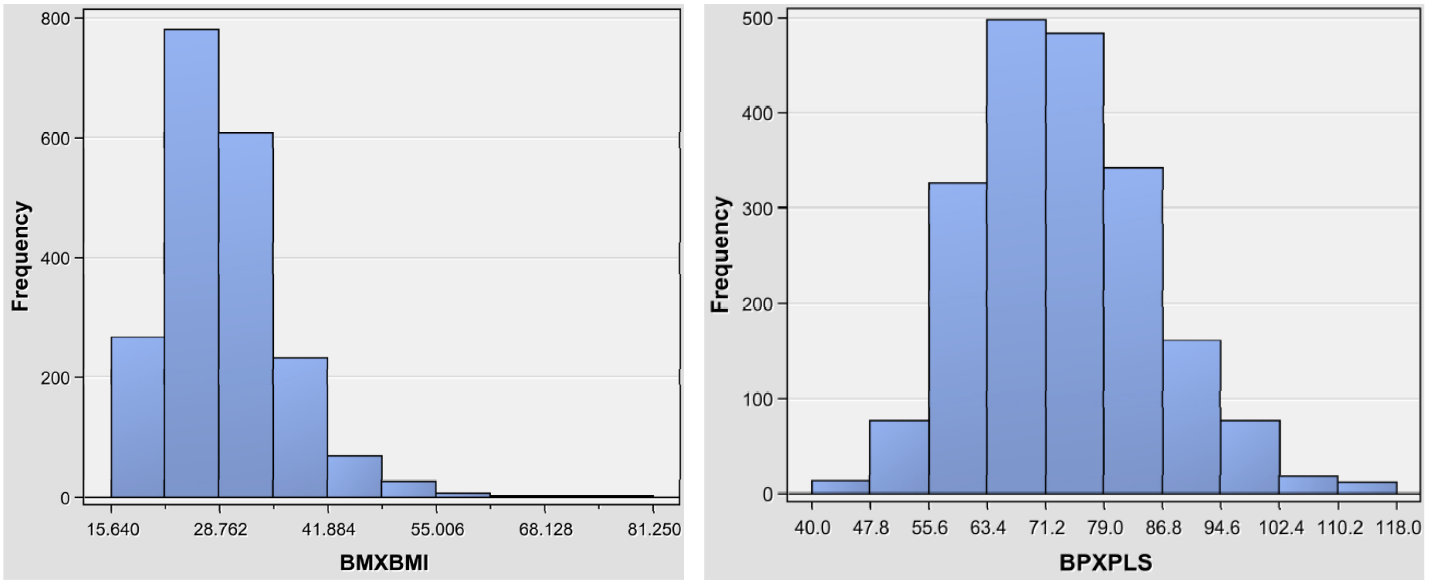


Figure 2-9 Histograms of Body mass index (BMXBMI), Puls per minute (BPXPLS)

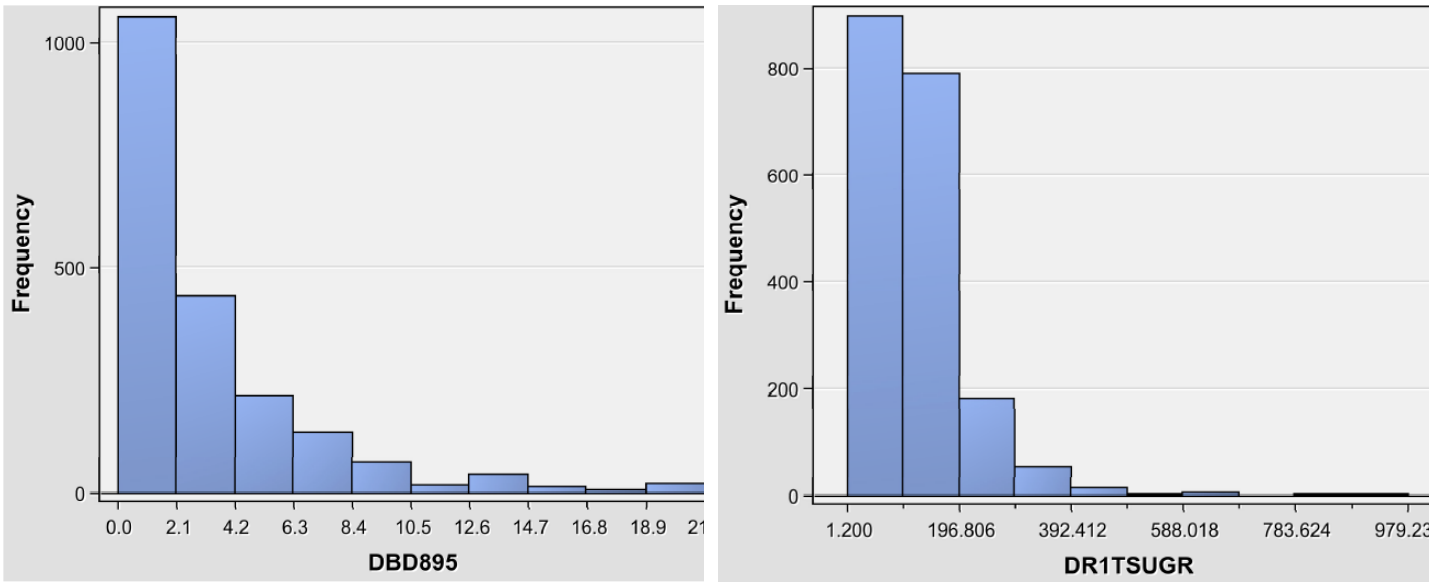


Figure 2-10 Histograms of Number of meals not home prepared (DBD895), Total sugars intake (DR1TSUGR)

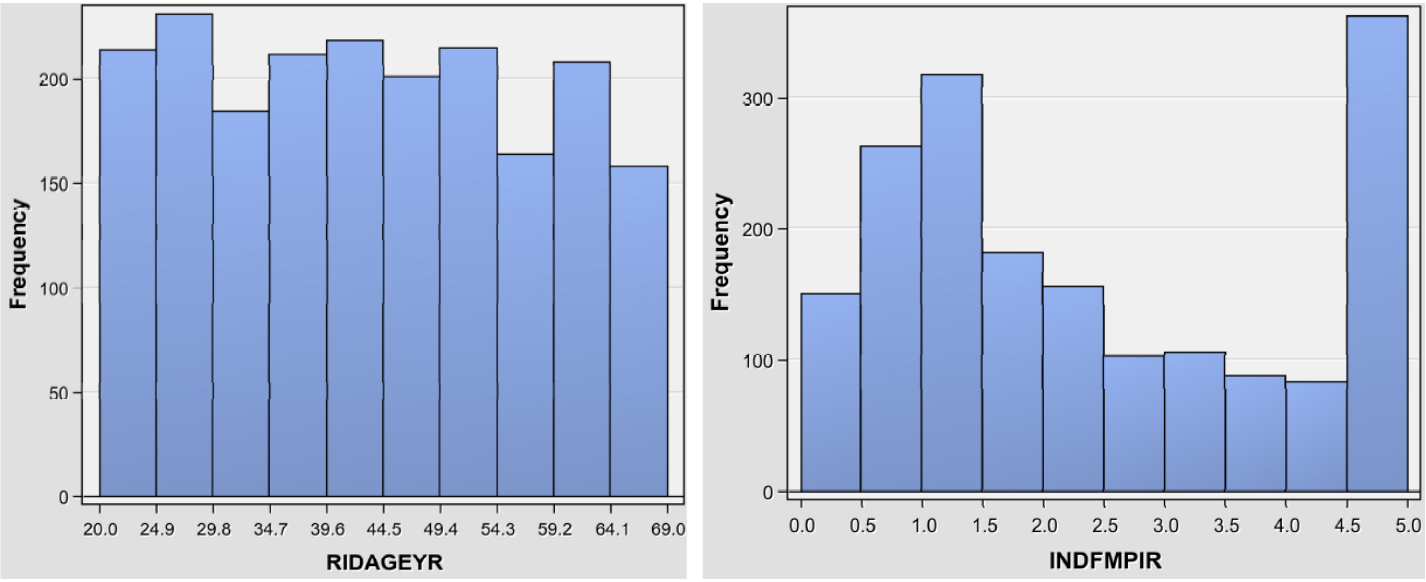


Figure 2-11 Histograms of Age (RIDAGEYR) and Ratio of family income to poverty (INDFMPIR)

2.2 Categorical Variables

There are 27 categorical input variables in my data set. These variables imply various aspects of the participant health and demographic conditions. To explore the patterns in data, I've plotted the average percentage of hypertensive cases for each level in all categorical variables. Figure 2-12 through Figure 2-14 show these patterns. Some of these patterns are interesting and important for my analysis. On average almost 58% of individuals whose "Ever told you had high blood pressure" (BPQ020) feature is True, have hypertension. This means 42% participants who said their doctors have told them they have high blood pressure, do not have hypertension based on their blood pressure readings in the NHANES survey. On the other side, 26% of those to whom no doctor has ever said they have high blood pressure, have hypertension. There are several reasons for inconsistency between BPQ020 data and hypertension status. For instance, participant may not correctly recall what their doctors had told them before the survey, or the blood pressure readings in the survey are not representative of their general blood pressure status.

A person's diet is an important factor for blood pressure. The DBQ700 chart in Figure 2-12 shows that as the quality of healthy diet decreases, the average percentage of hypertension cases increases (1 = Excellent healthy diet and 5 = poor healthy diet). Adding salt to food more often (DBD100) also can contribute to hypertension. Another interesting pattern is the relationship between diabetes and hypertension. Diabetes damages arteries and makes them targets for hardening, which can cause high blood pressure (De Boer et al., 2017). my data set shows that 50% of individuals to whom a doctor has told they have diabetes or are borderline diabetic (DIQ010 = 1 or 3) have hypertension, while only 34% of those whose doctors never told they have diabetes have hypertension. The same pattern is observed with pre-diabetes (DIQ160) and risk of diabetes (DIQ170).

Blood pressure values are important factors in the general health of an individual. High blood pressure can be a participating factor in several complications and diseases. Or it can influence an underlying disease. That is why as the general health condition of individuals get better (HUQ010) there are fewer number of people with hypertension on average.

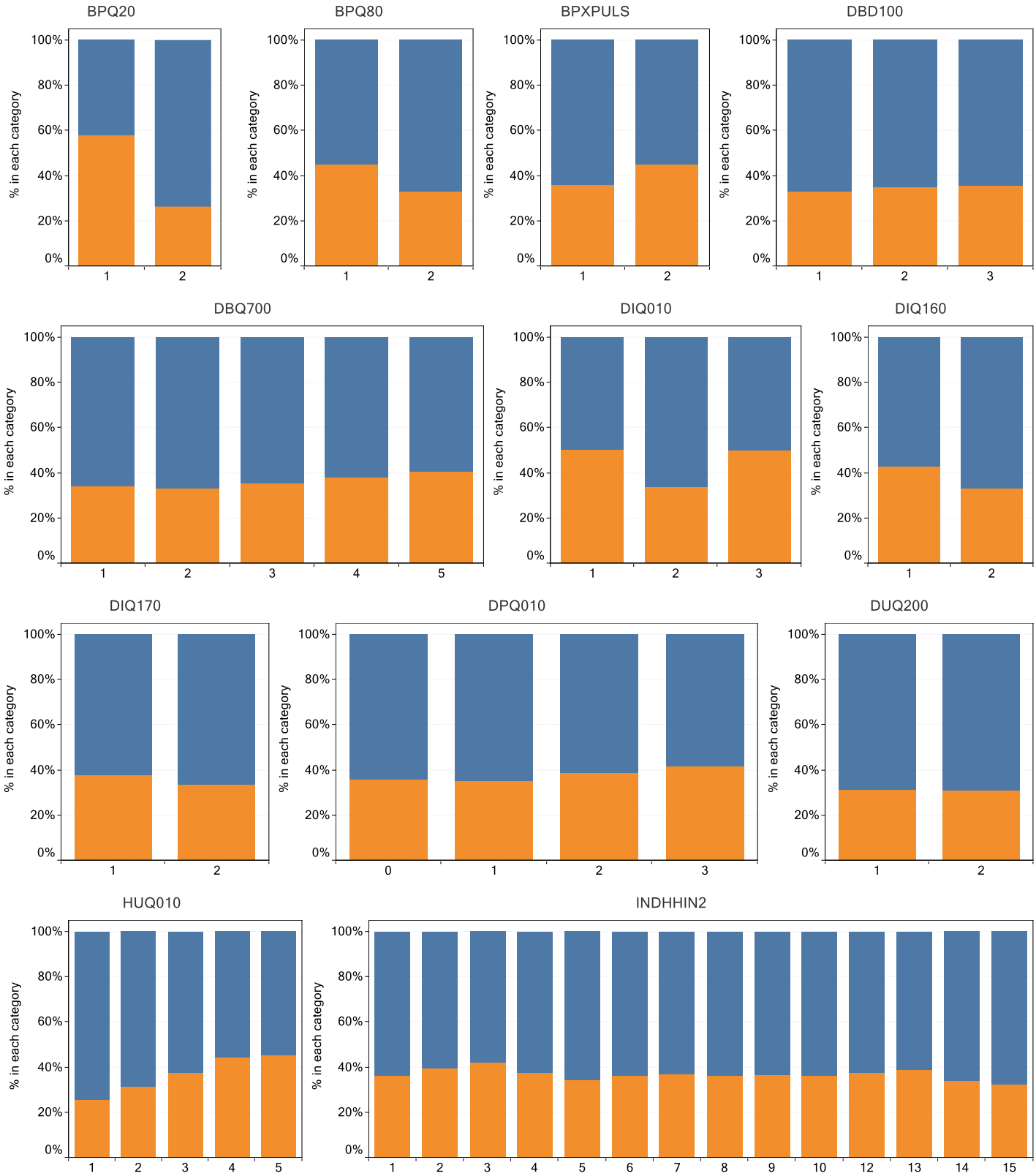


Figure 2-12 Average percentage of hypertension cases by levels of categorical variables (orange: *is_hypertension* True, blue: *is_hypertension* False)

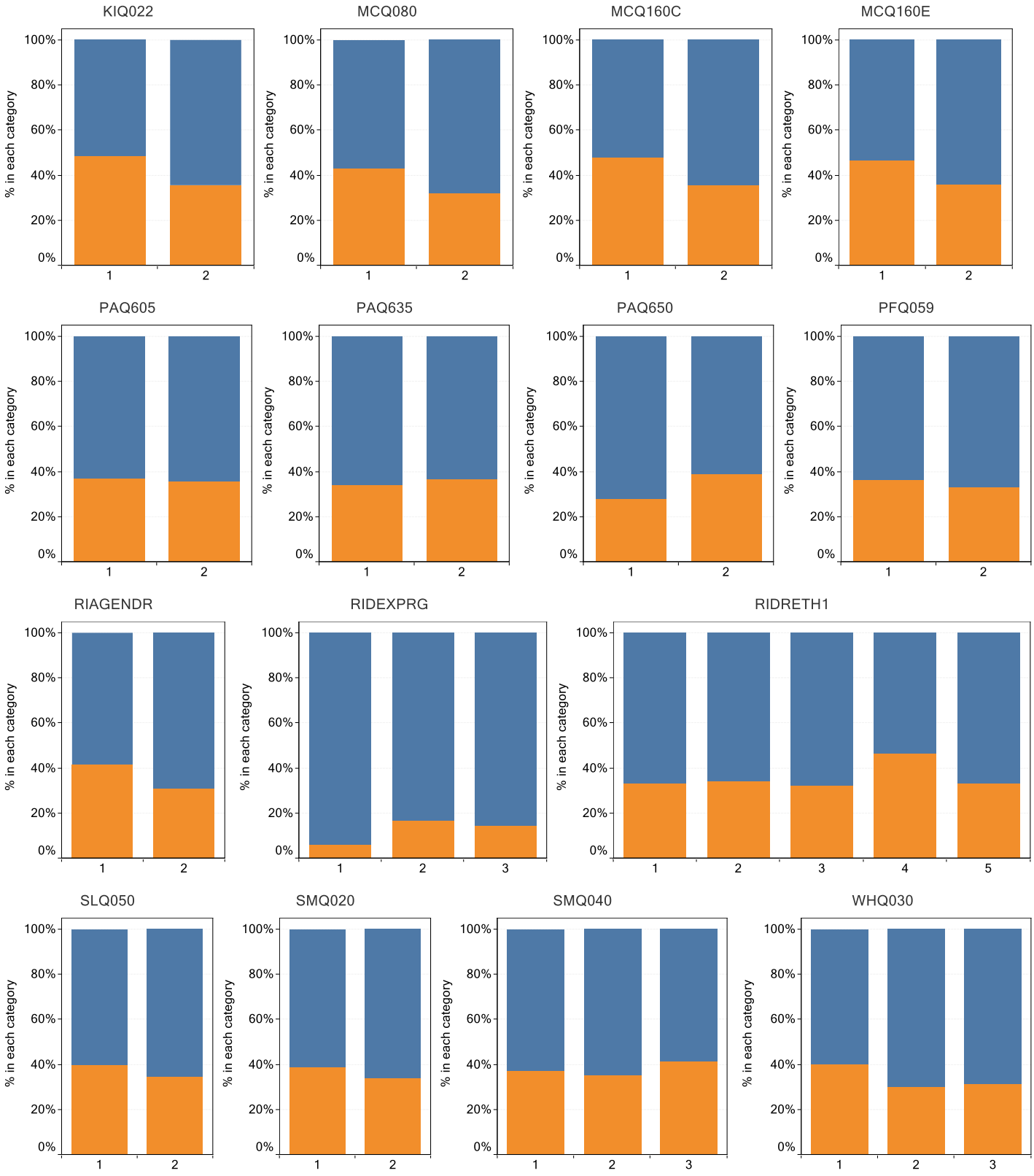


Figure 2-13 Average percentage of hypertension cases by levels of categorical variables (orange: *is_hypertension* True, blue: *is_hypertension* False)

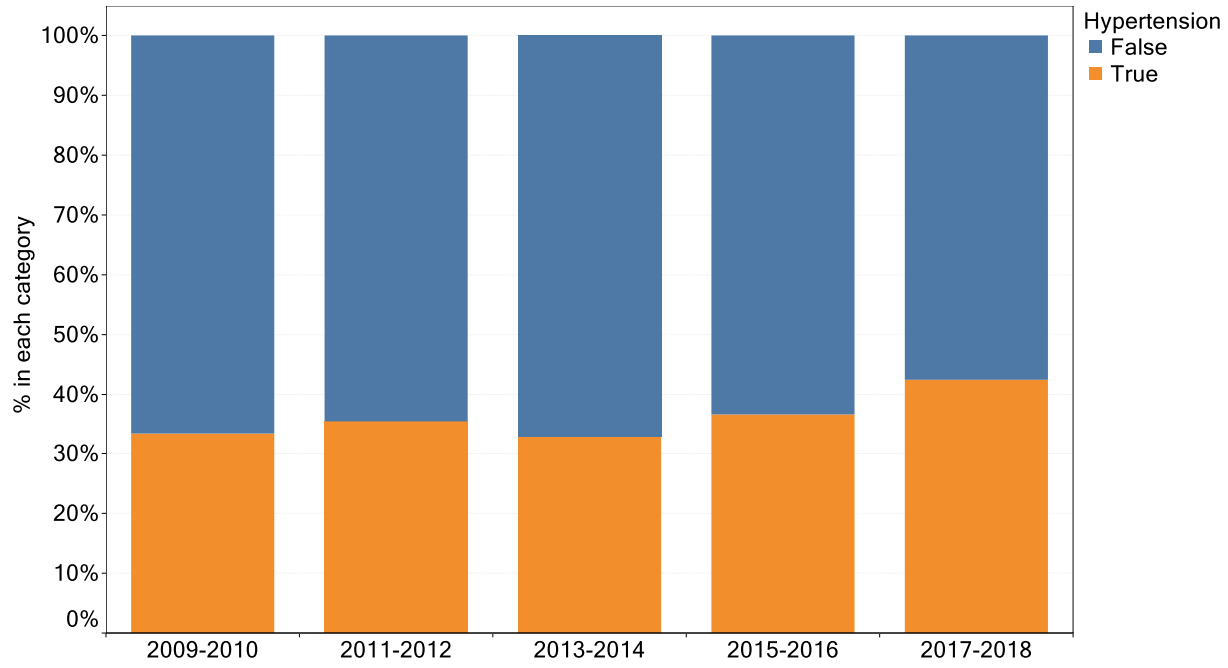


Figure 2-14 Average percentage of hypertension cases by period of NHANES survey (orange: *is_hypertension* True, blue: *is_hypertension* False)

The association between kidney problems and high blood pressure is a well-known fact in the medical sciences. My data shows 48% of participants whose doctor has told they had weak or failing kidneys, are hypertensive. Only 36% of individuals who have never been told they have kidney problem have high blood pressure (see KIQ022 chart in Figure 2-13). Also, higher blood pressure on average is associated with an individual being identified by a doctor as being overweight, is at risk of heart attack, or has had a heart attack (MCQ080, MCQ160C, and E). Weight appears in another feature with the same relationship with hypertension. The WHQ030 chart indicates that individuals who consider themselves overweight are more likely to have hypertension than those who see themselves underweight or fit.

Having vigorous recreational activities (PAQ650) decreases the likelihood of hypertension, while smoking (SMQ020 and SMQ040) increases it. Gender and race also paly important role. Males (level 1 category of RIAGENDR) and those individuals identified as non-Hispanic black (level 4 category of RIDRETH1) have considerably higher average occurrence of hypertension compared to the rest of the population. Another interesting pattern in the data is the increase in the average percentage of individuals with hypertension with time as shown in Figure 2-14.

These observations are derived from comparing the mean values of categorical levels. To find out if the differences are statistically significant, we performed chi-square test of independency between *is_hypertension* (the target variable) and all of the categorical features using SAS EM. Results are shown in Table 2-2. The change in *is_hypertension* is significantly associated with all features except for PAQ605 (Vigorous work activity). For the classification models, we use all features.

Table 2-2 Chi-square test results of the categorical features

Input	Chi-Square	Df	Prob
BPQ020	2056.5	2	<.0001
RIDEXPRG	1416.6	4	<.0001
DUQ200	581.1	4	<.0001
HUQ010	363.8	6	<.0001
BPQ080	359.8	3	<.0001
DIQ160	348.0	3	<.0001
RIDRETH1	309.9	4	<.0001
DIQ010	297.3	3	<.0001
MCQ080	283.8	2	<.0001
RIAGENDR	269.1	1	<.0001
DIQ170	259.9	3	<.0001
PFQ059	221.5	4	<.0001
PAQ650	212.2	2	<.0001
WHQ030	195.0	4	<.0001
year	107.6	4	<.0001
SMQ040	75.9	4	<.0001
DBD100	74.3	4	<.0001
INDHHIN2	62.0	16	<.0001
SMQ020	56.1	4	<.0001
SLQ050	49.5	2	<.0001
DBQ700	49.1	5	<.0001
KIQ022	42.1	3	<.0001
MCQ160C	33.9	3	<.0001
DPQ010	32.8	6	<.0001
MCQ160E	30.6	3	<.0001
PAQ635	12.6	2	0.0019
BPXPULS	11.6	1	0.0007
PAQ605	5.8	3	0.122

2.3 Linear Regression Analysis

The ultimate goal of the analysis is to predict hypertension status of an individual (*is_hypertension*), however because I also have *systolic_ave* and *diastolic_ave* interval variables, a linear regression model can be developed. The purpose of this modeling is not to predict the blood pressure, but to better understand the relationships between the inputs and

two variables that derive the classification target. There is a correlation between systolic and diastolic blood pressure reading (Figure 2-16), and therefore, for this task we only make a linear model for systolic blood pressure. The histogram of systolic blood pressure readings is shown in Figure 2-15 and its relationship with some of the lab result and demographic variables are shown in Figure 2-17.

The analysis of variance of the developed linear model shows it is significant with an F-Value of 9.82, and p-value < 0.0001. The t-Test results for the significant categorical and interval inputs (α=0.05) are shown in Table 2-3 and Table 2-4. The adjusted R^2 of the linear model is 0.28, which means less than 30% of the variability in *systolic_ave* is explained by the linear regression model.

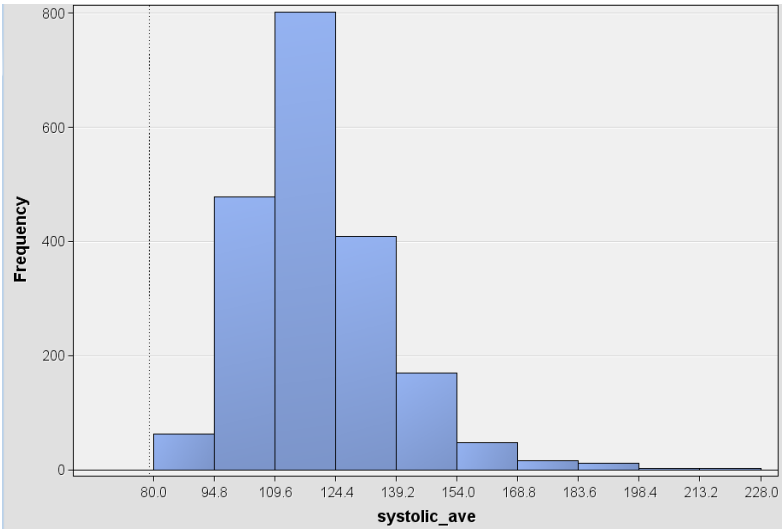


Figure 2-15 Distribution of average systolic blood pressure

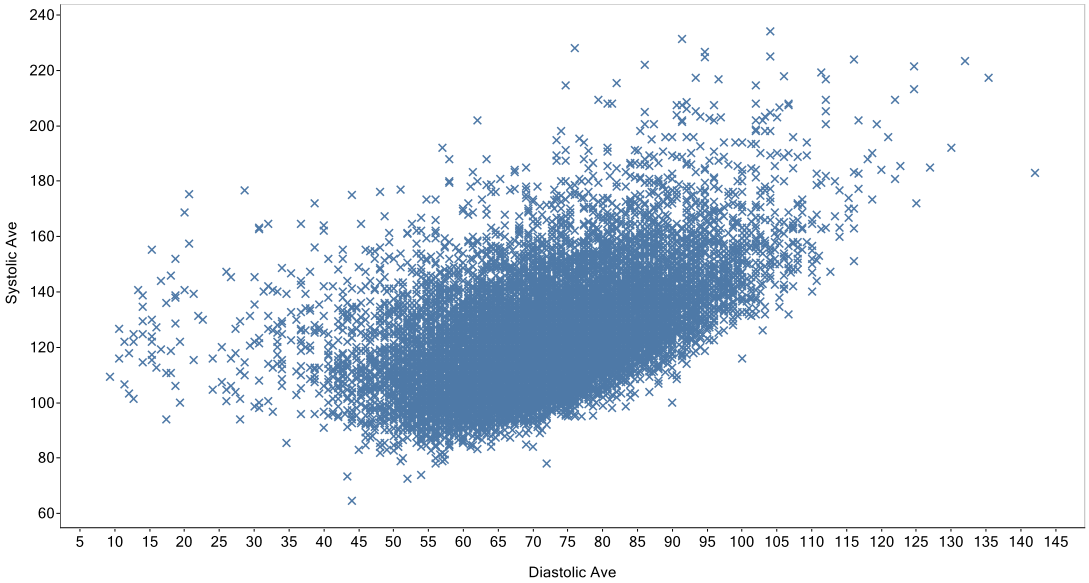


Figure 2-16 Correlation between systolic and diastolic blood pressure readings

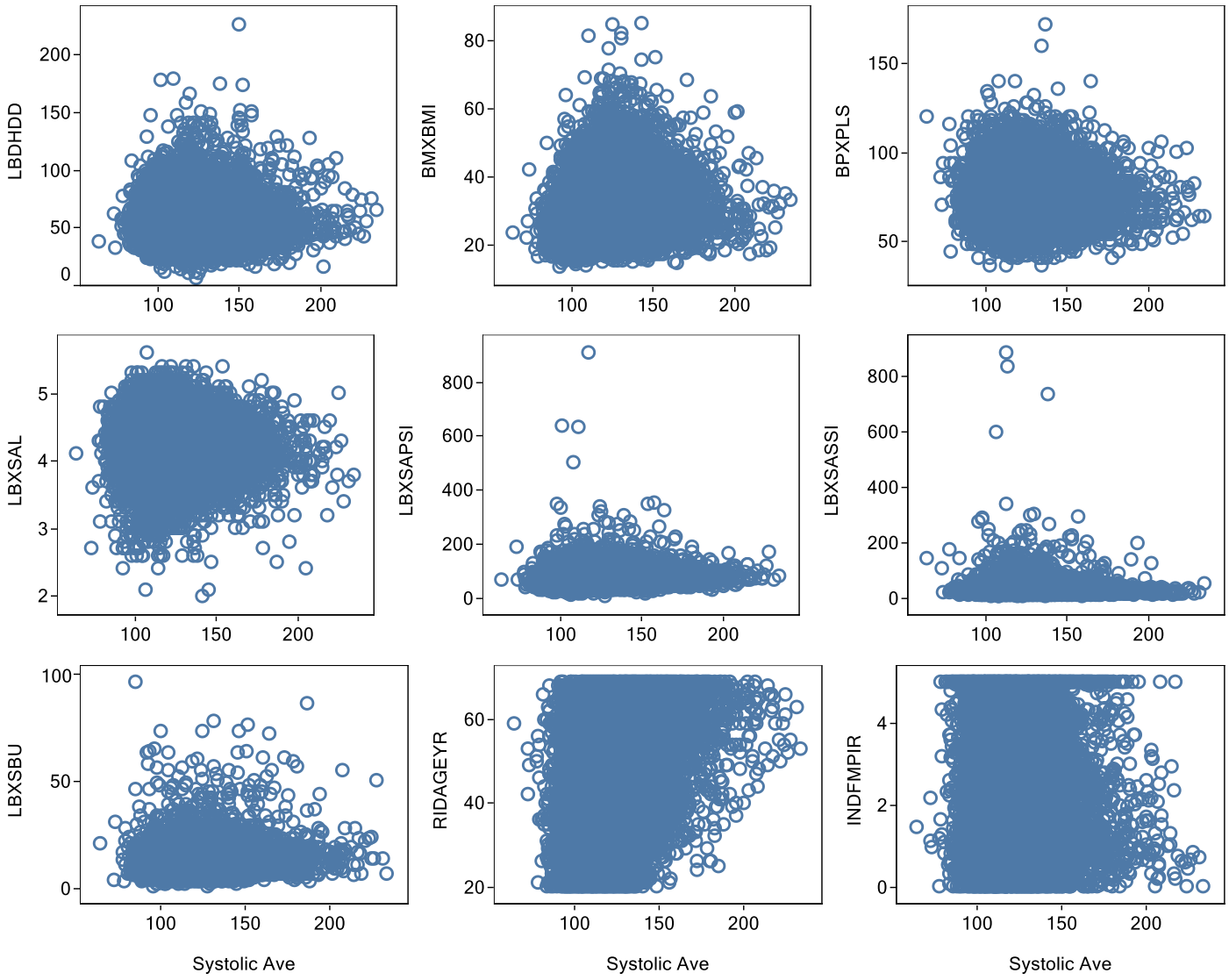


Figure 2-17 Average systolic blood pressure readings versus some of the lab result and demographic features

Table 2-3 t-Test results of the significant categorical variables in the *systolic_ave* linear model

Parameter	Level	Estimate	Error	t Value	Pr > |t|
BPQ020	2	-11.0086	4.0216	-2.74	0.0062
INDHHIN2	12	-5.1079	2.3951	-2.13	0.0331
PAQ605	1	1.2162	0.576	2.11	0.0349
RIDRETH1	2	-1.8636	0.7792	-2.39	0.0169
RIDRETH1	4	3.3407	0.6629	5.04	<.0001

Table 2-4 t-Test results of the significant interval variables in the *systolic_ave* linear model

Parameter	Estimate	Error	t Value	Pr > |t|
BMXBMI	0.3477	0.0657	5.29	<.0001
LBDHDD	0.0705	0.0222	3.18	0.0015
LBXSAL	-31.5549	12.4113	-2.54	0.0111
LBXSCA	2.4305	1.0457	2.32	0.0202
LBXSGB	-30.82	12.4087	-2.48	0.0131
LBXSKSI	-2.522	0.8921	-2.83	0.0047
LBXSTP	32.6456	12.3873	2.64	0.0085
LBXSTR	0.0104	0.00257	4.05	<.0001
RIDAGEYR	0.2004	0.0328	6.11	<.0001

3 Classification Models

The main goal of the analytical part of the current project is to develop a classification model to identify people with hypertension using their health, dietary, demographic, and lab data. I've applied and evaluated several classification algorithms and selected the best one using two KPIs that I defined in section 1.5. The investigated models are logistic regression, decision tree, random forest, gradient boosting, and neural networks. The first step of model building based on each algorithm was hyperparameter tuning, except for the logistic regression because it does not include any hyperparameter. For the logistic regression model, instead of finding the best hyperparameter i found the best input variable set.

To account for biases due to the random splitting of the training data set and to develop a stable classification model I performed hyperparameter tuning in a 5-fold cross-validation schema. In SAS EM cross-validation can be done though a concept that is called Group Processing. The idea is to segment the input data into different group configurations and process them independently (Schubert, 2010). In SAS EM this can be done by wrapping the model node between a Start Group and an End Group node as shown in Figure 3-1. In this schema, a random integer number between 1 and 5 is generated and assigned to each tuple in the original data set through Transfer Variables node. In the next step the input data set will be segmented into 5 approximately equal and mutually exclusive folds using the random integer numbers generated in the previous step. The Start and End Group nodes perform a 5-fold Cross-validation on each model node and at the end they are all compared. For better computational speed and more efficient optimization, hyperparameters of each model are optimized independently.

The data set that I've used for cross-validation training and model selection includes 70% of the entire original data set. We split the data set randomly into a 30% segment for testing and a 70% segment for model development and selection.

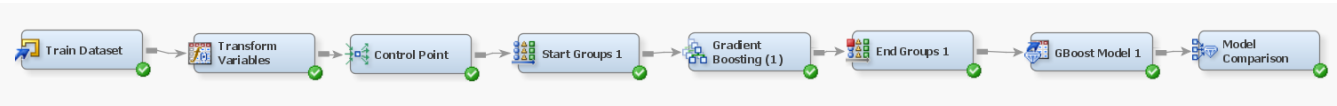


Figure 3-1 SAS EM Diagram with Start and End Group nodes to perform cross-validation through five models

3.1 Logistic Regression

Since my target variable in this data set is binary, I chose to use the logistic regression variant of linear regression. Logistic regression models produce an S-shaped curve instead of a straight line. The Logistic regression curve will force probabilities to fall between 0 and 1. Instead of ordinary least squares, logistic regression uses the maximum likelihood estimation (MLE) method.

Logistic regression appears similar in several ways to linear regression. The main difference being that the y intercept in linear regression mathematical model is replaced by the log odds--or logit--in logistic regression:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x + \varepsilon$$

As opposed to probability, odds are a ratio of the probability of the target event over the probability that the target event does not occur: p/1-p. I chose logit as my link function over the other available options in SAS Enterprise Miner (CLOGLOG and PROBIT) as it is the default function for logistic regression models within SAS EM.

For this model, I've started the logistic regression with all the variables in my data set represented. The model performed quite poorly, but several variables were identified as significant within the results (see Table 3-1). On the second pass, the non-significant variables were removed (set to 'no' in the Use column of the model's variable list) and the model was run again. By eliminating the non-significant variables from consideration, the model performed much better. This second pass also uncovered another variable that was deemed to be non-significant, so the model was run for a third time. That produced the three models shown on the comparison graph displayed in Figure 3-3. Each model was cross-validated 5-fold as shown in the SAS Diagram layout in Figure 3-2.

The list of all the variables, as well as which were removed in the subsequent passes, are documented below in Table 3-1 through Table 3-3.

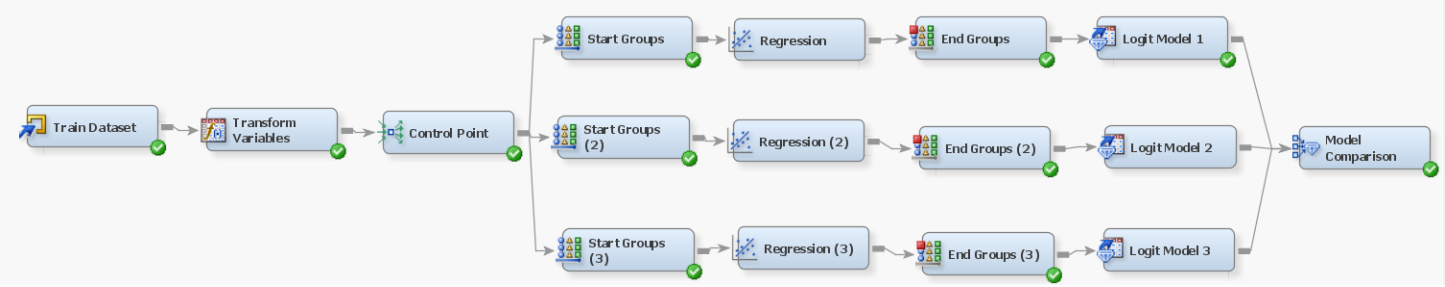


Figure 3-2 SAS EM Diagram for Logistic Regression

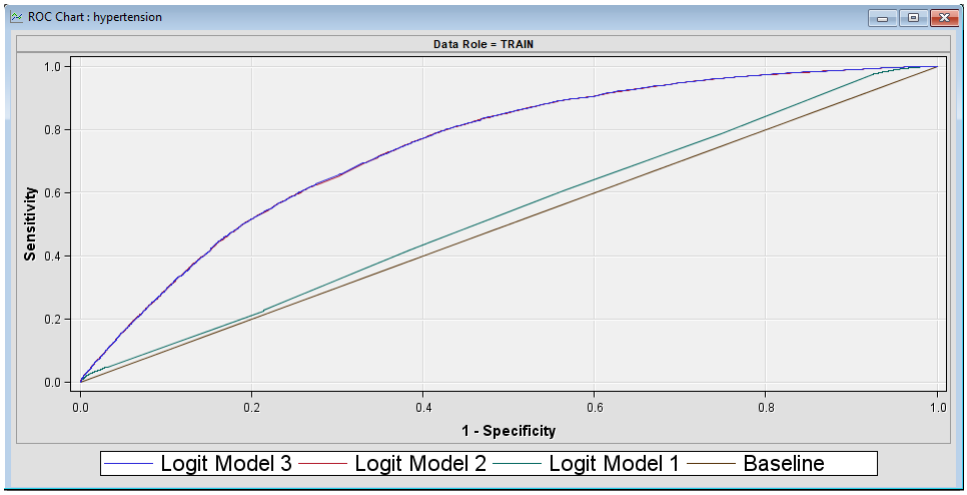


Figure 3-3 Model comparison of the baseline vs the three passes through logistic regression

Table 3-1 Logit variables on first pass before removing insignificant variables from consideration

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
BMXBMI	1	0.0612	0.0126	23.73	<.0001	0.2137	1.063
BPQ020	1	-0.6865	0.1317	27.16	<.0001	-0.1597	0.503
LBXSAL	1	-8.621	0.2387	1304.9	<.0001	-1.6344	0
LBXSGB	1	-8.1662	0.1538	2817.56	<.0001	-1.8336	0

RIDAGEYR	1	0.0399	0.00635	39.53	<.0001	0.2324	1.041
BPXPLS	1	0.0153	0.00488	9.88	0.0017	0.0974	1.015
RIDRETH1	1	0.1473	0.0491	9	0.0027	0.0931	1.159
LBXSCA	1	0.4523	0.2019	5.02	0.025	0.0852	1.572
LBXSTR	1	0.00117	0.00053	4.87	0.0274	0.0936	1.001
LBXSBU	1	-0.0848	0.0397	4.56	0.0328	-0.1868	0.919
SLQ050	1	0.2979	0.1409	4.47	0.0345	0.068	1.347
HUQ010	1	0.1193	0.0677	3.11	0.078	0.0604	1.127
DBD895	1	-0.0219	0.014	2.44	0.1184	-0.049	0.978
DIQ170	1	0.159	0.1021	2.43	0.1192	0.0452	1.172
LBXSCLSI	1	-0.0346	0.0228	2.3	0.1293	-0.0529	0.966
DPQ010	1	-0.1259	0.0884	2.03	0.1545	-0.0454	0.882
PAQ605	1	-0.1464	0.1143	1.64	0.2002	-0.0387	0.864
LBDHDD	1	0.00529	0.00437	1.47	0.2254	0.0456	1.005
LBXSKSI	1	-0.2172	0.1792	1.47	0.2255	-0.038	0.805
PAQ635	1	-0.1429	0.1193	1.43	0.231	-0.0359	0.867
INDHHIN2	1	0.0132	0.0113	1.37	0.2414	0.04	1.013
DIQ160	1	-0.2783	0.238	1.37	0.2422	-0.0334	0.757
WHQ030	1	-0.0798	0.0701	1.3	0.2546	-0.0429	0.923
LBXSPH	1	-0.1038	0.0985	1.11	0.2917	-0.0329	0.901
INDFMPIR	1	-0.0472	0.0448	1.11	0.2921	-0.0414	0.954
RIAGENDR	1	-0.9256	0.886	1.09	0.2961	-0.2372	0.396
PFQ059	1	-0.2384	0.257	0.86	0.3536	-0.0301	0.788
LBXSGL	1	-0.00588	0.00653	0.81	0.3676	-0.0608	0.994
LBXSOSSI	1	0.0945	0.1056	0.8	0.371	0.2248	1.099
LBXSNASI	1	-0.1765	0.199	0.79	0.3753	-0.204	0.838
LBXSASSI	1	0.00347	0.00393	0.78	0.3766	0.0282	1.003
MCQ080	1	0.1267	0.147	0.74	0.3887	0.0316	1.135
BPXPULS	1	-0.3899	0.4799	0.66	0.4165	-0.024	0.677
KIQ022	1	-0.3193	0.4959	0.41	0.5196	-0.0306	0.727
MCQ160E	1	-0.205	0.366	0.31	0.5753	-0.0197	0.815
LBXSCR	1	0.1346	0.2463	0.3	0.5848	0.0195	1.144
BPQ080	1	-0.0413	0.0807	0.26	0.6092	-0.0141	0.96
PAQ650	1	-0.0639	0.1252	0.26	0.6096	-0.0164	0.938
LBXTC	1	0.000803	0.00158	0.26	0.612	0.0182	1.001
DR1TSUGR	1	-0.00027	0.000593	0.2	0.6539	-0.0137	1
DBD100	1	-0.0264	0.0657	0.16	0.6882	-0.012	0.974
Intercept	1	-1.9464	5.0927	0.15	0.7023		0.143
LBXSUA	1	0.018	0.0493	0.13	0.7146	0.0141	1.018
DBQ700	1	-0.0199	0.0638	0.1	0.7553	-0.0107	0.98
LBXSIR	1	0.000343	0.00152	0.05	0.8214	0.0073	1
MCQ160C	1	-0.0239	0.1978	0.01	0.9039	-0.00302	0.976
RIDEXPRG	1	0.0333	0.4375	0.01	0.9394	0.0175	1.034
LBXSAPSI	1	0.000133	0.00236	0	0.9548	0.00176	1
DUQ200	1	0.00662	0.1375	0	0.9616	0.00146	1.007

SMQ040	1	-0.00041	0.0652	0	0.995	-0.00021	1
DIQ010	0	0	.	.	.	.	.
LBXSTP	1	8.4665	.	.	.	.	999
SMQ020	0	0	.	.	.	.	.

Table 3-2 Logit variables on first pass before removing insignificant variables from consideration

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
BMXBMI	1	0.0453	0.00245	340.89	<.0001	0.1801	1.046
BPQ020	1	-0.7461	0.0362	423.88	<.0001	-0.2132	0.474
BPXPLS	1	0.0126	0.00138	82.88	<.0001	0.0819	1.013
LBXSAL	1	0.4552	0.0565	64.86	<.0001	0.0888	1.576
LBXSBU	1	-0.0165	0.0032	26.49	<.0001	-0.0463	0.984
LBXSTR	1	0.0012	0.000128	86.94	<.0001	0.0896	1.001
RIDAGEYR	1	0.0447	0.00132	1141.59	<.0001	0.3522	1.046
RIDRETH1	1	0.1298	0.013	99.09	<.0001	0.0894	1.139
SLQ050	1	0.188	0.0357	27.76	<.0001	0.0468	1.207
LBXSCA	1	-0.0317	0.051	0.39	0.5346	-0.00627	0.969
Intercept	1	-5.9969	0.4677	164.4	<.0001		0.002

Table 3-3 Logit variables on first pass before removing insignificant variables from consideration

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
BMXBMI	1	0.0452	0.00245	340.68	<.0001	0.1799	1.046
BPQ020	1	-0.7463	0.0361	427.66	<.0001	-0.2133	0.474
BPXPLS	1	0.0125	0.00138	82.3	<.0001	0.0813	1.013
LBXSAL	1	0.4382	0.0493	78.98	<.0001	0.0855	1.55
LBXSBU	1	-0.0162	0.00319	25.6	<.0001	-0.0455	0.984
LBXSTR	1	0.00119	0.000128	86.24	<.0001	0.089	1.001
RIDAGEYR	1	0.0445	0.00132	1145.19	<.0001	0.3511	1.046
RIDRETH1	1	0.1294	0.013	99.03	<.0001	0.0891	1.138
SLQ050	1	0.1886	0.0357	27.97	<.0001	0.0469	1.208
Intercept	1	-6.2095	0.311	398.7	<.0001		0.002

As displayed in Table 3-4, the confusion matrix shows an unfortunate sensitivity score (for forecasting true positives).. The ROC AUC score places my logistic regression model in an acceptable range to test for a disease or condition, in my case hypertension (Mandrekar 2010).

Additional tuning of the model was not possible due to the lack of hyperparameters within logistic regression models (Brownlee 2020). Hyperparameters differ from conventional parameters in that they are not simply coefficients given value or weight by the algorithm, but instead are specified by the constructor of the model in advance. Hyperparameters will be explored in more details in subsequent models.

Table 3-4 Output from logistic regression model comparison

		Predicted	
		TRUE	**FALSE**
Actual	**TRUE**	2257	3323
	FALSE	1458	8470

Sensitivity:	0.40
Specificity:	0.85
Precision:	0.61
NPR:	0.72
Accuracy:	0.69
ROC AUC:	0.74

3.2 Decision Trees

Decision trees (DT) are recursive classification algorithm that are easily performed and mirror the notion of a categorical decision due to the visual structure of the results. The visual representation of a DT is that of a tree, hence the name. Determinations of branches is made utilizing a locally optimal approach and is in distinct contrast to models such as logistic regression (Torres, 2020). The value a DT can bring to the concept of hypertension is the straightforward way they can be used to make a determination regarding the independent variable. While less optimal, the ability for DTs to readily model multiple variables make it a good candidate for modeling my data set based on the dependent variable, hypertension. The diagram constructed and eventually used to build a DT model in SAS EM is shown in Figure 3-4.

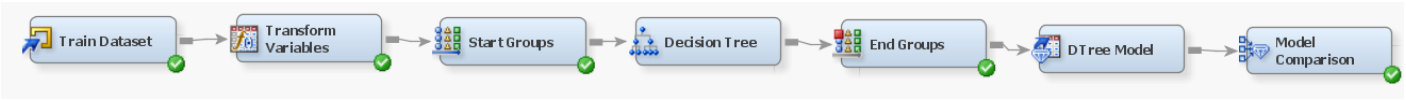


Figure 3-4 SAS EM diagram constructed for Decision Tree analysis

The hyperparameters adjusted to obtain the optimal DT model include maximum number of branches, maximum depth, and assessment measure. Table 3-5 highlights these parameters and provides available ranges. Where maximum branches determine how many branches are created from each splitting rule. Maximum depth is the number of levels the tree will reach when splitting on variables. Each level is determined by the modeler but can be significant based on the number of variables to build the tree. The hyperparameters selected to create the DT and optimize AUC was to set the assessment method to average squared mean, maximum branch to three, and maximum depth of seven. Table 3-6 indicates the optimal hyperparameters selected for the final DT model.

Table 3-5 Hyperparameters for Decision Tree

Hyperparameter Name in SAS	Range
Splitting Rules Hyperparameters	
Assessment Measure	NA
Maximum Branch	2 -10
Maximum Depth	2 - 50

Table 3-6 Optimal parameters for Decision Tree

Hyperparameter Name in SAS	Optimum Value
Assessment Method	Avg. Squared Error
Maximum Branch	3
Maximum Depth	7

The DT model determines the split on each node based on calculated entropy. The model determines a split by comparing the entropy value to the ratio determined by the node observations and the total observations. Once a comparison is made the node is split based on this value. This process is repeated throughout the DT (Torres, 2020). Figure 3-5 displays the results of this process in the form of a tree. The optimal DT model resulted in a total of 76 leaves or pure nodes (see Figure 3-6). That is, those nodes which include those from the dependent variable. The confidence in a pure node increases as the entropy is closer to zero. The AUC in Figure 3-7 provides a means to analysis the ability of the DT model to predict the dependent variable. Summarize the results from the analysis and provide a means of comparison for the most appropriate model for my independent variable.

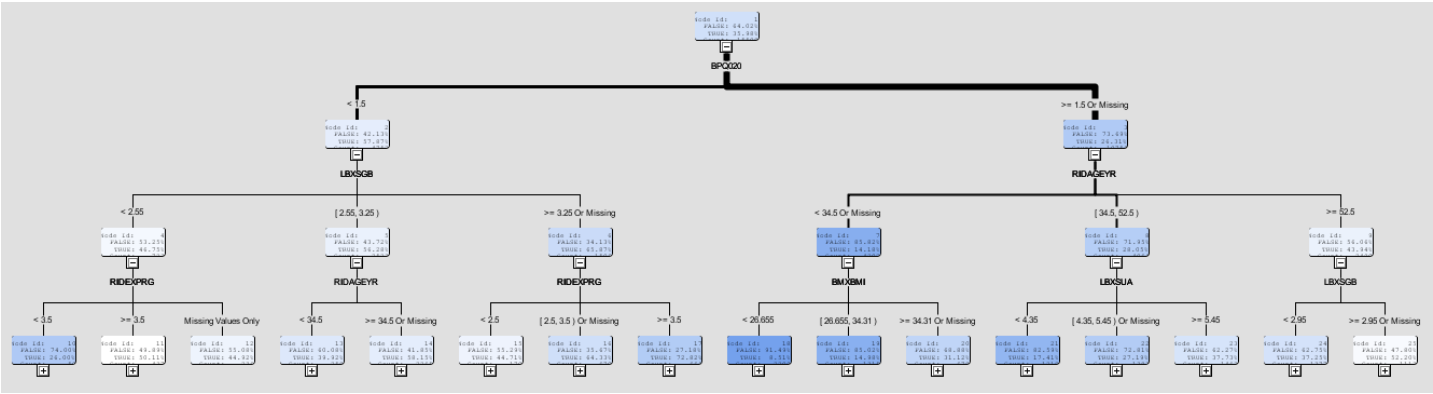


Figure 3-5 SAS EM Decision Tree – Partial

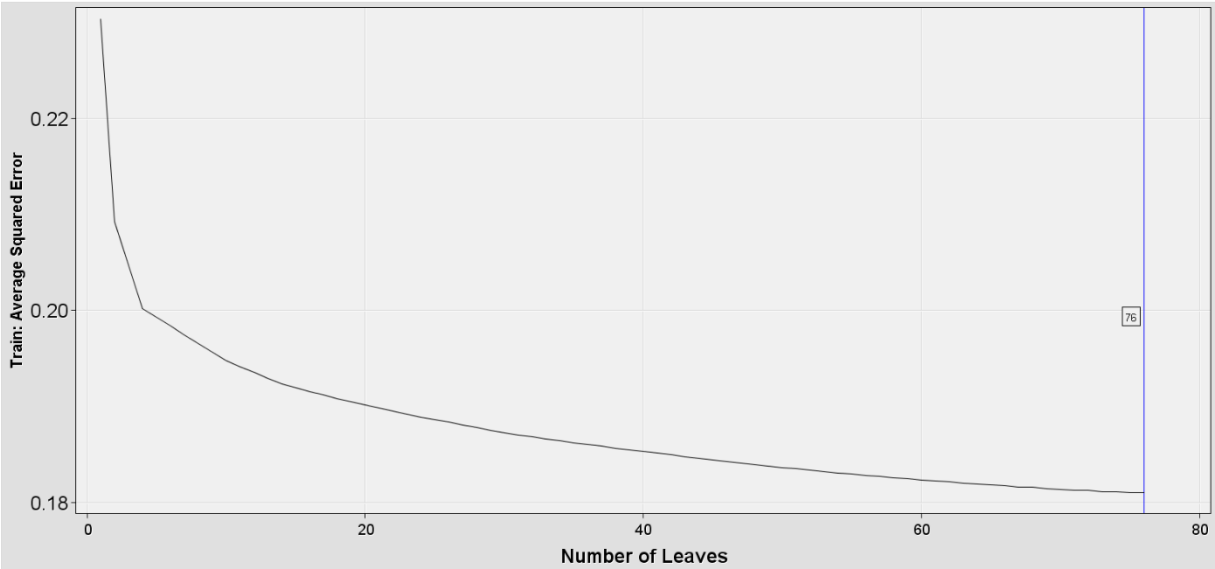


Figure 3-6 SAS EM Decision Tree Assessment Plot

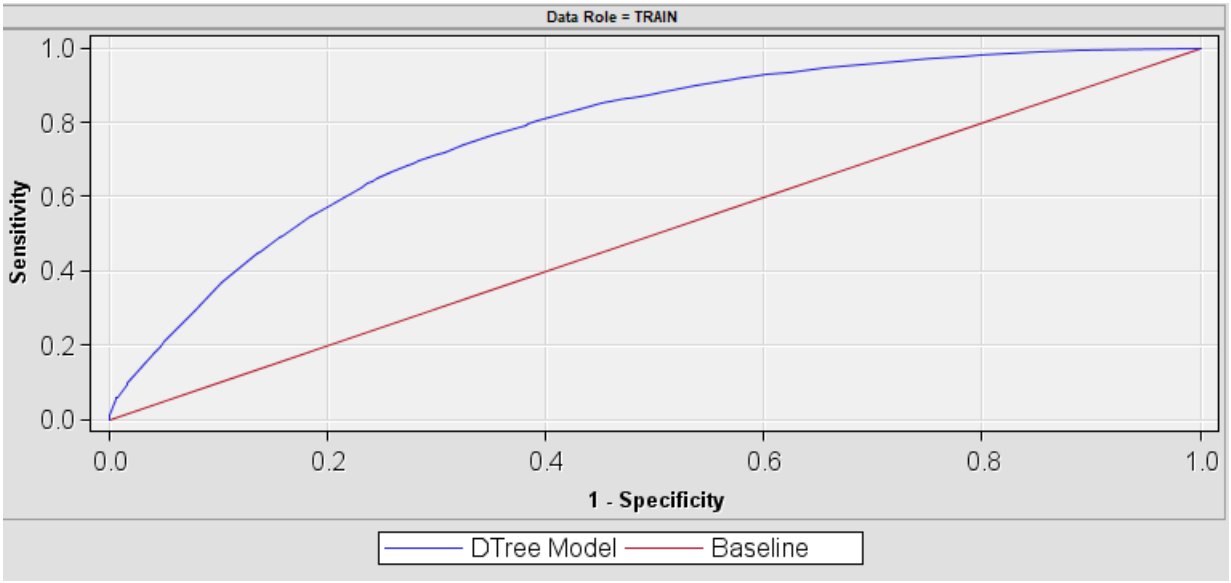


Figure 3-7 SAS EM ROC Diagram for Decision Tree

The confusion matrix, Table 3-7, displays similar results to the logistic regression analysis, with less than optimal sensitivity and a fairly respectable specificity. The AUC is around 0.77 which is good, but getting this value closer to one would provide more confidence in predicting the dependent variable.

Table 3-7 Confusion matrix for Decision Tree

		Predicted	
		TRUE	**FALSE**
Actual	**TRUE**	2659	2921
	FALSE	1920	8008

Sensitivity:	0.48
Specificity:	0.81
Precision:	0.58
NPR:	0.73
Accuracy:	0.69
ROC AUC:	0.77

3.3 Random Forest

A single decision tree model can be very sensitive to noise in the data if it is not pruned (high variance) and it can have a large average error (bias) if it is too simple. A popular technique to improve weak learners such as a single decision tree is to ensemble them in the form of a Random Forest. A Random Forest is composed of many individually grown trees by using only a fraction of data records and attributes. In each iteration data samples are selected by random sampling with replacement (bootstrapping aggregation or bagging) technique. For growing each tree, only a randomly selected sunset of features (input variables) is used at each node. When these two techniques are combined to grow a sufficiently large number of trees, the generalization error will converge to a non-zero value, which means variance of the model approaches to zero. After growing a full forest, typically the averaging or majority vote technique is used to determine the class label of the unseen data samples.

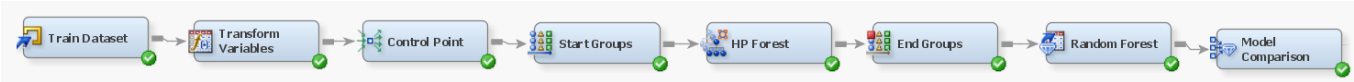


Figure 3-8 SAS EM diagram for random forest model hyperparameter tuning

In the SAS EM platform, the HP Forest node is designated to derive a random forest model. The diagram was built for tuning the random forest model hyperparameters is shown in Figure 3-8 and the list and range of hyperparameters are shown in Table 3-8. In the random forest model, the maximum number of trees and proportion of the observations in each iteration are related to the tree growing part of the algorithm. Generally, higher number of trees improve the performance of the ensemble if the growth of each individual tree is limited by maximum depth. In my grid search for hyperparameters i changed the number of trees between 100 (SAS default value) and 1000. Also, to increase generalization ability of the random forest model we should use only a fraction of records in each iteration which is specified by “Proportion of observations in each sample” parameter in the SAS model.

Table 3-8 Random forest model hyperparameters

Hyperparameter Name in SAS	Range
Splitting Rules Hyperparameters	
Maximum depth	2 - 10
Number of variables to consider in split search	1 - 52
Smallest Percentage of Obs. in Node	1.0E-6 - 1.0E-4
Tree Hyperparameters	
Maximum Number of trees	100 - 1000
Proportion of observations in each sample	≤ 1

The Maximum Depth parameter in the SAS model specifies the maximum depth of each tree grown in the forest. Although Random Forest compensates for overfitted trees through the algorithm, it is better to limit the amount of overfit trees. Therefore, based on my findings in section 3.3 I limited the maximum depth to 10. Another hyperparameter with significant impact on the model performance is the proportion of the features that are randomly selected for growing the trees that is specified by the *Number of variables to consider in split search* parameter in the SAS model. In addition, SAS allows for the setting of the smallest number of training observations a new branch of a tree may have. If the number of observations in a node is less than this value, then the algorithm will not split, preventing the tree from becoming overfit.

Table 3-9 Optimal hyperparameter space for the gradient boosting model

Hyperparameter Name in SAS	Range
Maximum depth	9
Number of variables to consider in split search	40
Smallest Percentage of Obs in Node	1.00E-05
Maximum Number of trees	300
Proportion of observations in each sample	0.8

I performed a grid search for the best hyperparameter set by considering ROC AUC and recall score as my evaluation metrics. Results are shown in Table 3-9. The optimal model is composed of 300 trees and in each iteration, it will use 80% of the observations (through bagging method) and 40 out of 52 original features in the training data set. The confusion matrix and classification chart are shown in Table 3-10 and Figure 3-9. The optimal model has a sensitivity score of 0.65 which is 35% improvement over a single tree model (see Table 3-7). Figure 3-11 shows the ROC plot of the optimal model with the AUC value of 0.88. shows the distribution of the probability scores predicted by the optimum model for each class. The blue curve shows the model is performing well identifying non-events (people without hypertension). This is also reflected in the relatively high negative rate (NPR) show in Table 3-10. However, the red curve does not have a sharp increase in the beginning which shows it is not performing as well for the positive class records.

Table 3-10 Confusion matrix and performance metrics of the tuned random forest model

		Predicted	
		TRUE	**FALSE**
Actual	**TRUE**	3613	1967

FALSE	1133	8795

Sensitivity:	0.65
Specificity:	0.89
Precision:	0.76
NPR:	0.82
Accuracy:	0.80
ROC AUC:	0.88

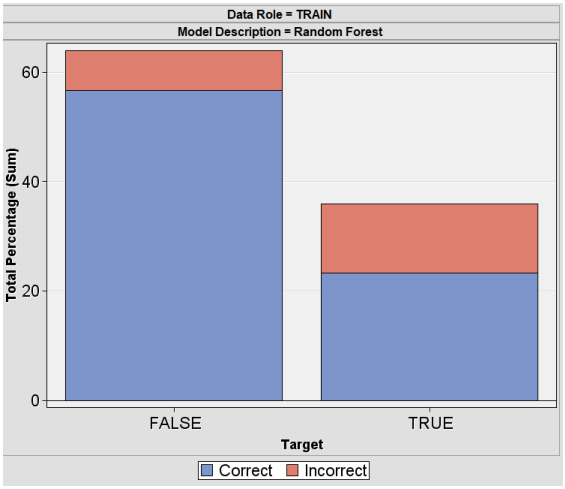


Figure 3-9 Classification chart of the optimum random forest model

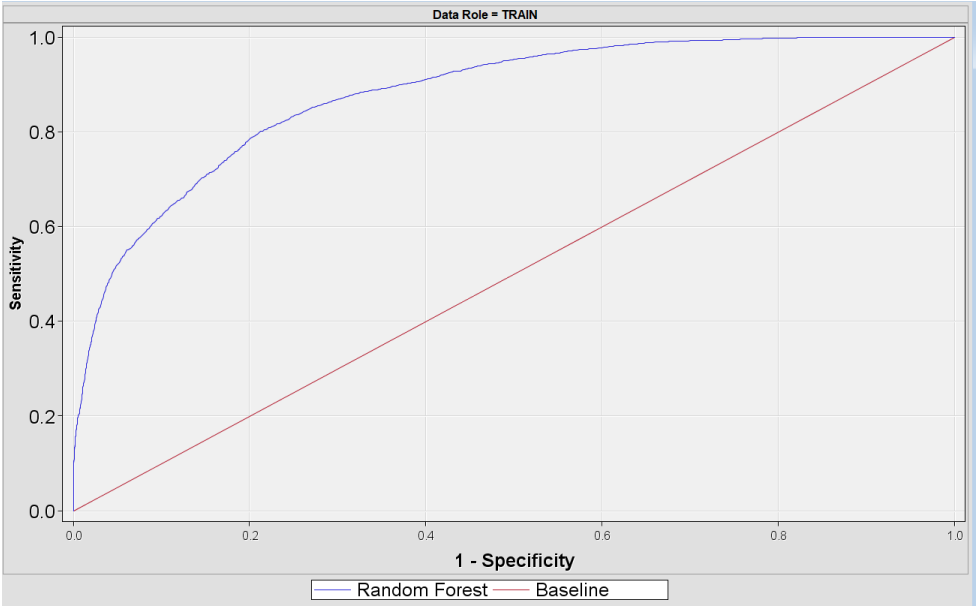


Figure 3-10 ROC plot of the optimum random forest

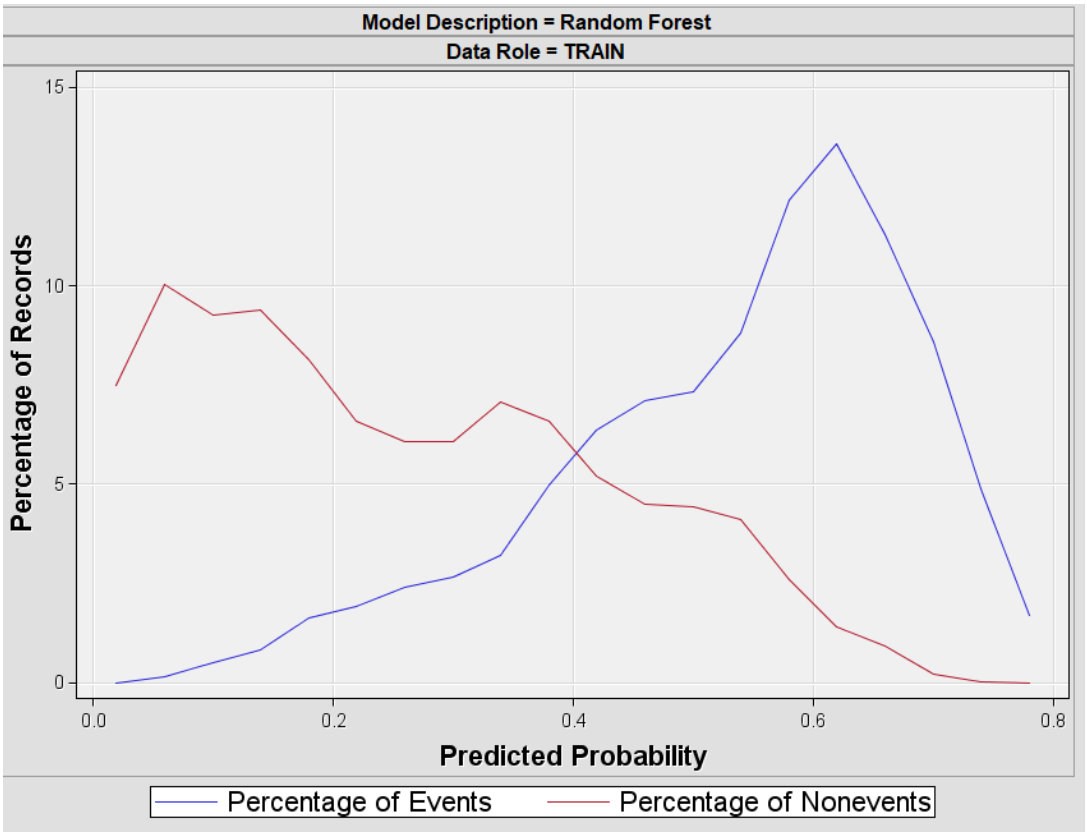


Figure 3-11 Distribution of the predicted probability of each class by the optimum random forest model

3.4 Gradient Boosting

The random forest explained in section 3.3 performed well on negative cases had difficulty identifying the positive class records. One reason for that is that random forest models are not designed to focus on the weakness of the ensemble as it moves forward through the iterations. It does not take into account the result of the previous iteration for improving the next one. Gradient Boosting, on the other hand, takes into account these previous iterations. Gradient Boosting is a sequel technique for linearly combining weak prediction models to produce an ensemble prediction model with higher performance. Gradient Boosting is typically used with Decision Trees of a predefined size as the base weak learner. Its hyperparameters can be divided into two categories as shown in Table 3-12. The first group of parameters control the splitting rules in the base learners (decision trees) and the second group control the boosting process. Maximum branch and depth values determine the maximum size of each base tree. In general, as a tree grows deeper and wider (i.e. more branches), it becomes more complex and it will have more splits that can capture more information about the data. Increasing the depth and width of trees increases the performance of prediction on the training set, however it can also lead to overfitting and therefore both hyperparameter should be limited. Another model hyperparameter is the number of time splitting rules in a path may use the same variable, which is specified in the SAS EM model through the Reuse Variable. Its range is between 1 and the maximum depth of the tree. The leaf fraction parameter specifies the smallest number of training observation a new branch may have, expressed as the fraction of the number of available observations in the training data set. This parameter also controls overfitting, and for imbalanced data sets it is better to keep it lower.

Table 3-11 Gradient boosting model hyperparameters

Hyperparameter Name in SAS	Range
Splitting Rules Hyperparameters	
Maximum Branch	2 - 5
Maximum Depth	2 - 5
Reuse Variable	1 - Max depth
Leaf Fraction	0.01, 0.001, 0.0001
Boosting process Hyperparameters	
N Iterations	50 - 1000
Shrinkage	≤ 0.1
Train Proportion	60 - 90

To control the boosting process there are three hyperparameters to consider. The N Iteration parameter in a binary problem defines the number of sequential trees to be modeled. Although the gradient boosting model is generally robust at higher number of base learners, but to avoid overfitting this parameter should be limited. Shrinkage, or learning rate, specifies how much to reduce the prediction of each tree. An important element of gradient boosting method is regularization of the error term by limiting the amount that new rules in the subsequent iteration are modified through the learning rate variable. Empirically it has been shown that using small shrinkage improves generalization ability of the model, but it needs more computational resources. In SAS EM the default value of shrinkage is 0. 1 and in my project i tried several smaller values in the search for an optimized hyperparameter space.

Similar to other models, I performed a grid search to find the optimum hyperparameter set which is shown in Table 3-12. The confusion matrix and performance metrics of the optimum model are shown in Table 3-12 and the ROC plot is shown in Figure 3-13. The sensitivity score of the model is 0.89 and its area under the curve is 0.98. These are very high scores, however there is a possibility of overfitting. Figure 3-14 shows the distribution of the probability scores predicted by the optimum model for each class. The combined curves make a nice U shape distribution that indicates high performance of the model.

Table 3-12 Optimal hyperparameter space for the gradient boosting model

Hyperparameter Name in SAS	Optimum Value
Maximum Branch	4
Maximum Depth	4
Reuse Variable	3
Leaf Fraction	0.01
N Iterations	400
Shrinkage	0.01
Train Proportion	60

Table 3-13 Confusion matrix and performance metrics of the tuned gradient boosting model

		Predicted	
		TRUE	**FALSE**
Actual	**TRUE**	4973	607
	FALSE	135	9793

Sensitivity:	0.89
Specificity:	0.99
Precision:	0.97
NPR:	0.94
Accuracy:	0.95
ROC AUC:	0.98

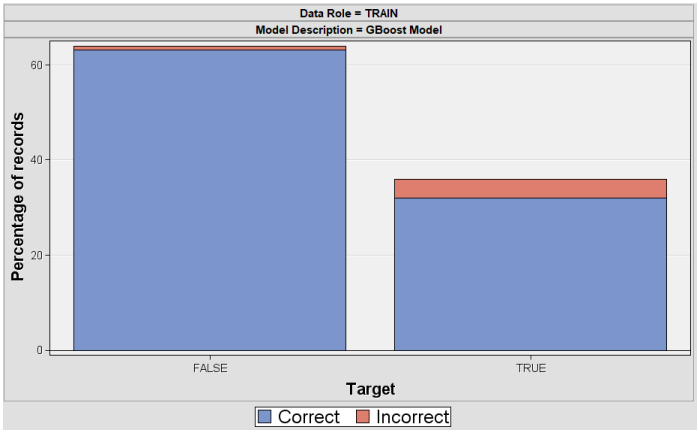


Figure 3-12 Classification chart of the optimum gradient boosting model

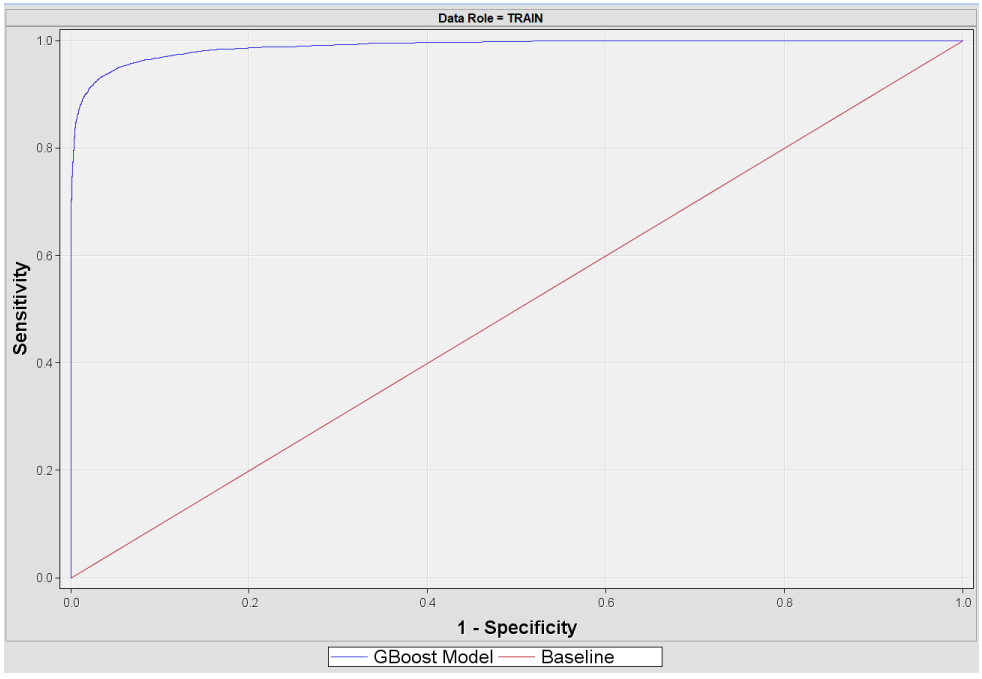


Figure 3-13 ROC plot of the optimum gradient boosting model

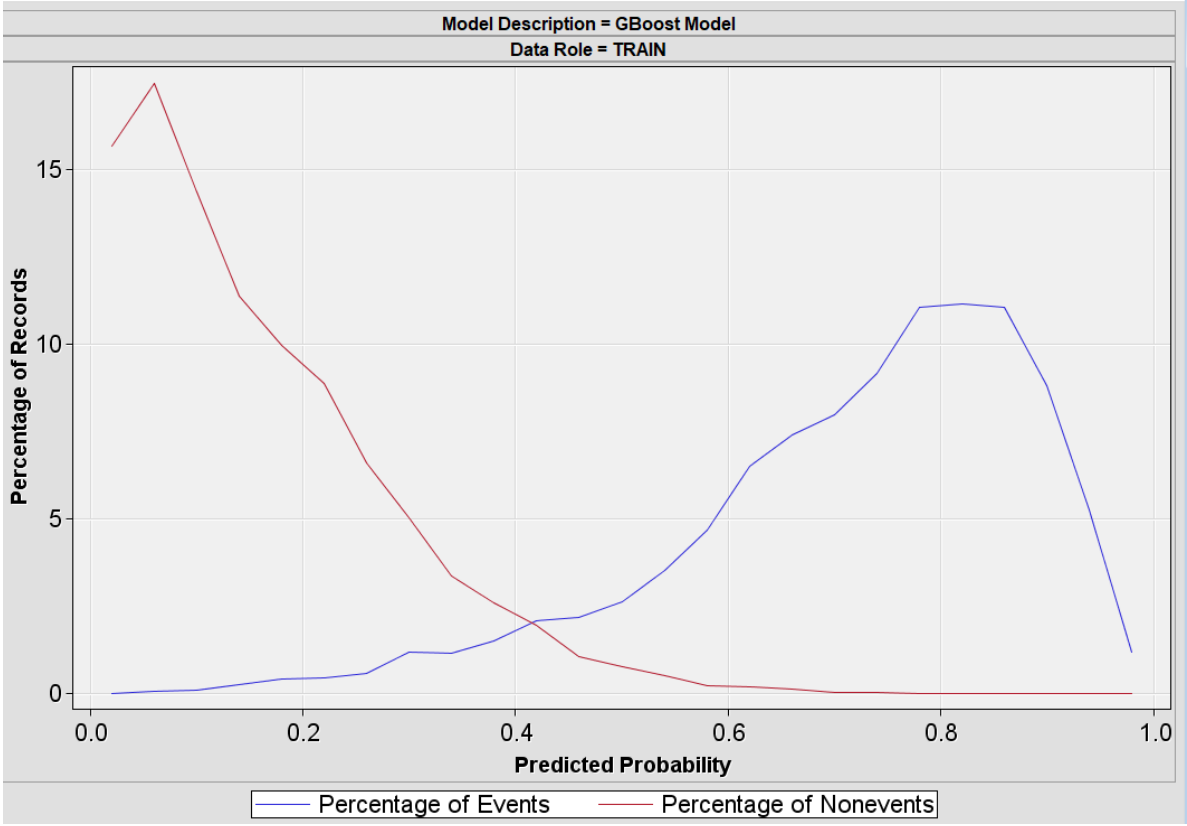


Figure 3-14 Distribution of the predicted probability of each class by the optimum gradient boosting model

3.5 Neural Networks

Another supervised machine learning algorithm that I applied to develop a predictive model was artificial neural networks (ANNs). The idea behind ANNs is to build a network of mathematical neuron-like functions to learn the correct association between a set of input features and output variables. Each neuron in the network has an activation function with a threshold that generates a binary signal. When a large enough number of neurons are combined in the correct structure, they can be trained to solve linear and non-linear problems such as regression and classification. Typically, a neural network consists of three layers, input, hidden, and output. The input layer is connected directly to the input feature from the data set. The number of neurons in the input layer is equal to the number of features selected for the classification or regression task. The hidden layer(s) are added to the model to increase the complexity and flexibility of the model in order to solve non-linear problems. The number of hidden layers, nodes, and hyperparameters within each hidden layer can significantly impact the performance of the entire model. In the output layer, the output signal of the neurons in the final hidden layer are combined to compute the class label (or coefficients in the case of regression problems). An important hyperparameter of any ANN is the activation function assigned to the neurons. Different functions such as rectified-linear, exponential-linear, tanh, or sigmoid have different threshold send out activation signal.

Figure 3-15 SAS EM diagram for Neural network model

To derive my ANN model in SAS EM I used the AutoNeural node in the same cross-validation configuration as the previous models. The SAS EM diagram used in my ANN is shown in Figure 3-15. Model hyperparameters are shown in Table 3-14. The AutoNeural node does limited searches to find better network configurations. Hidden layers are added one after the other which may contain one or more hidden neuron units.

Table 3-14 AutoNeural network model hyperparameters

Hyperparameter Name in SAS	Range
Maximum Iterations	3 - 50
Hidden Units	1 - 8
Train Action	Train, Increment and Search
Activation Functions	Direct, Normal, Softmax, Tanh, Logistic, Exponential, Identity, Sine, Square and Reciprocal
Final Iterations	5 - 40
Total Number of Hidden Units	5-100

i used an architecture with a single hidden layer in which all hidden nodes are added. A node in this model can include one or more neurons. The number of hidden units can vary between 1 and 8. Another hyperparameter is Train Action, a method used during model training. In my SAS EM model, all selected functions are trained until stopping if the method is set to TRAIN. If the method is INCREMENT, the nodes are added one at a time. No activation function is reused during the training. The layer would be retained if it reduces the average error in the output, otherwise it would be dropped from the model. If the method is SEARCH, the nodes are added according to the architecture and the best network is retained. In this case, the *Total Number of Hidden Units* variable determines total hidden units ceiling. SAS EM gives a wide range of option for activation function: direct, normal, softmax, tanh, logistic, exponential, identity, sine, square, and reciprocal functions. I examined all of them in my grid search for the best hyperparameters.

Results of the search for the best hyperparameters are shown in Table 3-15 and the confusion matrix and performance metrics are presented in Table 3-16. The sensitivity of the model is lower than a random model and the AUC is 0.68. The ROC plot of the neural networks with different hyperparameters are shown in Figure 3-16.

Table 3-15 Optimum hyperparameter values for Neural Network

Hyperparameter Name in SAS	Optimum Value
Maximum Iterations	8
Hidden Units	3
Train Action	Search
Activation Functions	Softmax, Tanh, Normal and Direct
Final Iterations	10
Total number of Hidden Units	30

Table 3-16 Confusion matrix for Neural Network

		Predicted	
		TRUE	**FALSE**
Actual	**TRUE**	2423	4157
	FALSE	1528	7400

Sensitivity:	0.37
Specificity:	0.83
Precision:	0.61
NPR:	0.64
Accuracy:	0.63
ROC AUC:	0.68

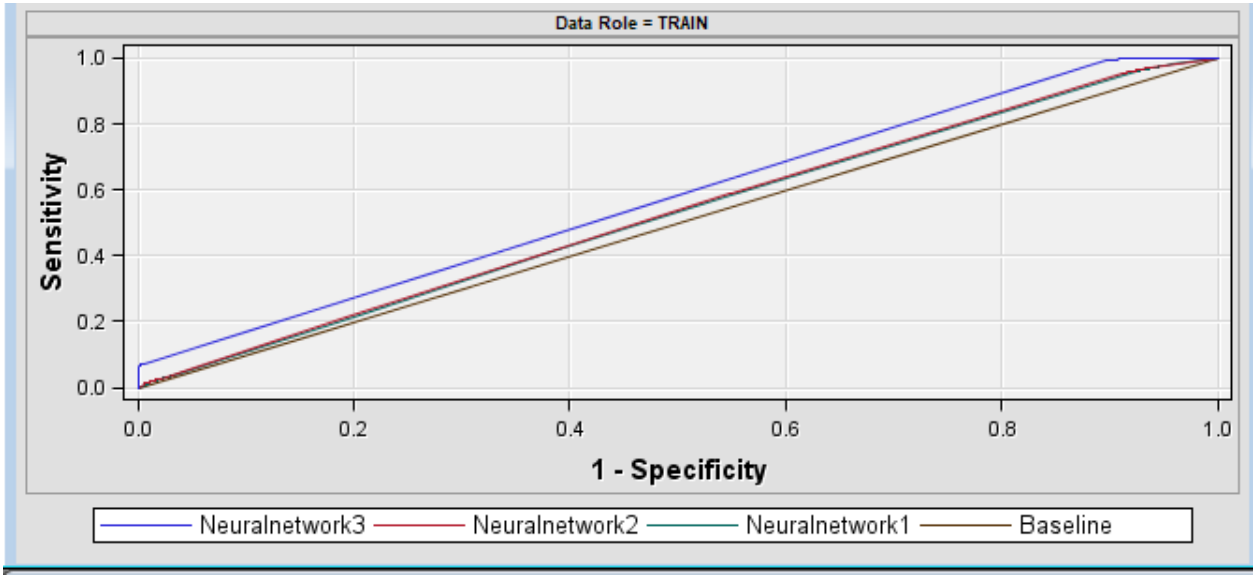


Figure 3-16 ROC Plot of three neural network models

3.6 Model Selection

For the model selection I used sensitivity (recall) and ROC AUC scores of the optimized models in the cross-validation training phase. Table 3-17 shows a comparison between the optimum model scores. My final model is gradient boosting with the hyperparameters shown in Table 3-12. The AUC of the selected model is 0.98 and the recall scores is 0.89 which shows very few missing true positive cases. To evaluate the performance of the selected model on unseen data, we applied it on the test data set that was put aside at the beginning of the model development process. The ROC plot of the application of the final model on the train and test data sets are shown in Figure 3-17. The recall score of the test data set is 0.71. Loss of the classification performance which is expected when the model is applied on new data is in a reasonable range.

Table 3-17 Comparison of five classification model performance metrics

Model	Recall Score	AUC
Logistic Regression	0.41	0.74
Decision Trees	0.48	0.77
Random Forest	0.65	0.88
Gradient Boosting	0.89	0.98
Neural Networks	0.37	0.68

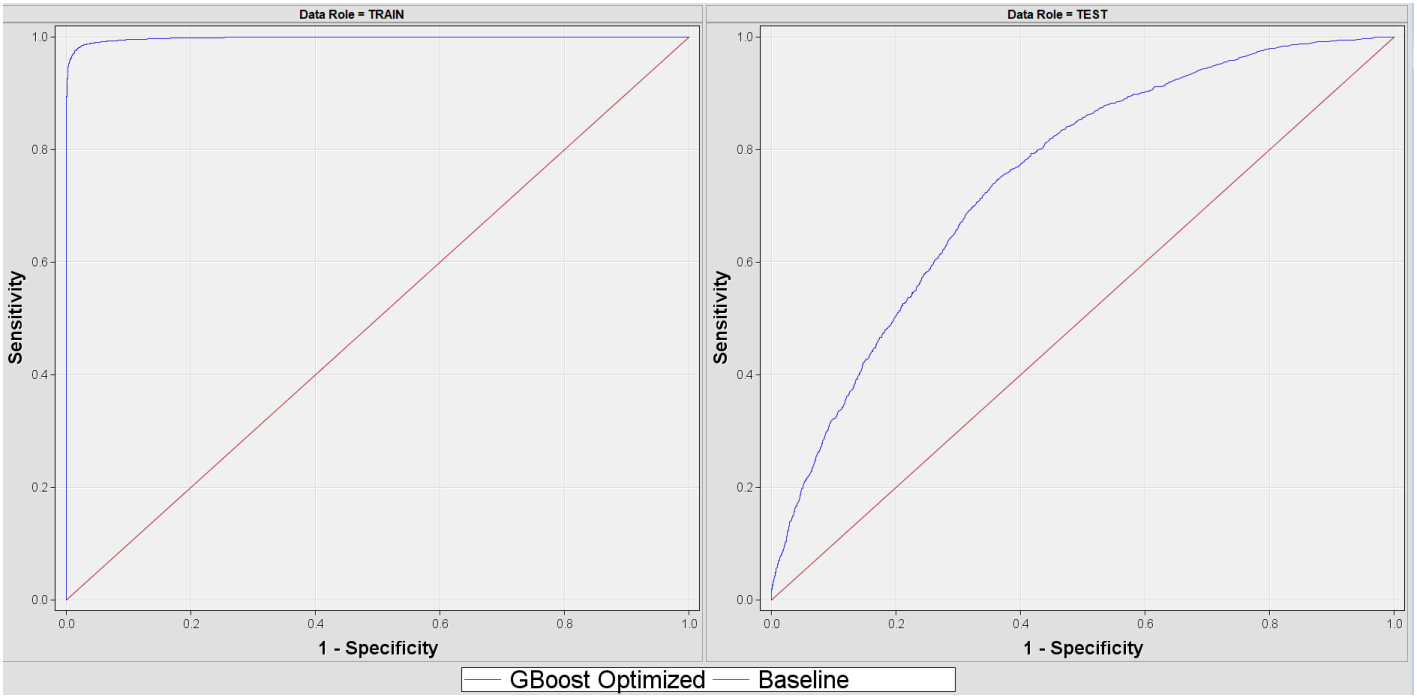


Figure 3-17 Performance of the selected gradient boosting model on the training set and test set

4 Summary

In this project, I've developed a classification model to identify people with hypertension using their health, demographic, dietary habits, and lab data. Hypertension is a primary or contributing cause of approximately half a million death in the US and it can exist in a person undetected without significant symptoms. Early identification of hypertension by healthcare professionals can help physicians and patients start treatment earlier to prevent or reduce serious consequences. For this project, I've used five periods of NHANES data for data exploration and predictive tool building. This data set is collected by CDC and is one of the most comprehensive and representative publicly available health data sets and a principal source for tracking conditions such as hypertension in the U.S. population. The data set in this project consisted of 22,151 records, each represents one person in the surveys.

My exploratory data analysis revealed interesting and useful patterns. On average 42% of the people whose doctor has told them they have high blood pressure, do not have hypertension based on their blood pressure readings in the NHANES survey. Also, 26% of those whom no doctor has ever said they have high blood pressure, have hypertension. While the underlying causes of this inconsistency should be investigated, a reliable predictive tool such as the one I have developed in this project can help healthcare professionals identify positive and negative cases more efficiently. My data also showed the habit of adding salt to food and eating poorly can increase the likelihood of hypertension. The associations of kidney problems, diabetes, smoking, low physical activity, and obesity with hypertension are also reflected in data. Also, being male and black is associated with high blood pressure.

For developing a predictive tool, I've built and tuned five classification models in SAS EM. I've chose sensitivity score and AUC as my key performance indicators because missing a positive case can be significantly more expensive than mis-identifying a negative case. Out of the five models, performance metrics of the Gradient Boosting model performed best, and was selected to be my final predictive model.

The exploratory data analysis and predictive tool that i developed in this project can be directly used by healthcare professionals to both understand the patterns of hypertension and predict it in a faster, more reliable, and cheaper fashion.

References

Ahmad, W. M. A. W., Nawi, M. A. B. A., Aleng, N., Halim, N., Mamat, M., Hamzah, M., & Ali, Z. (2014). Association of hypertension with risk factors using logistic regression. *Applied Mathematical Sciences*, *8*(52), 2563-2572. http://dx.doi.org/10.12988/ams.2014.42130

American Heart Association. (2020). What is High Blood Pressure. https://www.heart.org/-/media/data-import/downloadables/f/9/8/pe-abh-what-is-high-blood-pressure-ucm_300310.pdf?la=en

Bates, D., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big Data in Health Care: Using Analytics to Identify and Manage High-Risk and High-Cost Patients. Health Affairs, 33(7), 1123-1131. https://doi.org/10.1377/hlthaff.2014.0041

Brownlee, Jason. (August 2020). Tune Hyperparameters for Classification Machine Learning Algorithms. https://machinelearningmastery.com/hyperparameters-for-classification-machine-learning-algorithms/

Centers for Disease Control and Prevention. (September 2017). About the National Health and Nutrition Examination Survey. https://www.cdc.gov/nchs/nhanes/about_nhanes.htm

Centers for Disease Control and Prevention. (September 2020). Facts About Hypertension. https://www.cdc.gov/bloodpressure/facts.htm

Center for Disease Control. (n.d.). *Unweighted Response Rates for NHANES 2017-2018 by Age and Gender.* https://wwwn.cdc.gov/Nchs/data/nhanes3/ResponseRates/NHANES-2017-2018-Response-Rates-508.pdf

Chang, W., Liu, Y., Xiao, Y., Yuan, X., Xu, X., Zhang, S., & Zhou, S. (2019). A Machine-Learning-Based Prediction Method for Hypertension Outcomes Based on Medical Data. *Diagnostics*, *9*(4), 178. https://doi.org/10.3390/diagnostics9040178

David, F., Howard, K., Roux Ana, D. & Jiang, H. (2010). A Population-Based Policy and Systems Change Approach to Prevent and Control Hypertension. National Academies Press. Washington, DC. ISBN: 0-309-14810-3, 236 pages, http://www.nap.edu/catalog/12819.html

De Boer, I. H., Bangalore, S., Benetos, A., Davis, A. M., Michos, E. D., Muntner, P., ... & Bakris, G. (2017). Diabetes and hypertension: a position statement by the American Diabetes Association. *Diabetes Care*, *40*(9), 1273-1284.

Golino, H. F., Amaral, L. S. D. B., Duarte, S. F. P., Gomes, C. M. A., Soares, T. D. J., Reis, L. A. D., & Santos, J. (2014). Predicting increased blood pressure using machine learning. *Journal of Obesity*, *2014*., http://dx.doi.org/10.1155/2014/637635

Koren, G., Nordon, G., Radinsky, K., & Shalev, V. (2018). Machine learning of big data in gaining insight into successful treatment of hypertension. Pharmacology Research & Perspectives, 6(3), e00396. https://doi.org/10.1002/prp2.396

Lafreniere, D., Zulkernine, F., Barber, D. & Martin, K. (2016). Using machine learning to predict hypertension from a clinical dataset. In 2016 IEEE Symposium Series on Computational Intelligence (SSCI), 1–7. https://doi.org/10.1109/SSCI.2016.7849886 .

López-Martínez, F., Núñez-Valdez, E. R., Crespo, R. G., & García-Díaz, V. (2020). An artificial neural network approach for predicting hypertension using NHANES data. Scientific Reports, 10(1), 1-14. https://doi.org/10.1038/s41598-020-67640-z

Lynn, K. S., Li, L. L., Lin, Y. J., Wang, C. H., Sheng, S. H., Lin, J. H., ... & Pan, W. H. (2009). A neural network model for constructing endophenotypes of common complex diseases: an application to male young-onset hypertension microarray data. Bioinformatics, 25(8), 981-988.https://doi.org/10.1093/bioinformatics/btp106

Mandrekar, Jayawant N (2010). Receiver Operating Characteristic Curve in Diagnostic Test Assessment. Journal of Thoracic Oncology. Volume 5. Issue 9. Pages 1315-1316. https://doi.org/10.1097/JTO.0b013e3181ec173d.

Mayo Clinic. (May 2018). High Blood Pressure (Hypertension) https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/diagnosis-treatment/drc-20373417

Polak, S. & Mendyk, A. (2008). Artificial neural networks-based Internet hypertension prediction tool development and validation. Appl. Soft Comput. **8**, 734-739. https://doi.org/10.1016/j.asoc.2007.06.001

Schubert, S. (2010). The Power of the Group Processing Facility in SAS® Enterprise Miner. SAS Institute Inc. Cary, NC. https://support.sas.com/resources/papers/proceedings10/123-2010.pdf

Torres, R. (2020, August). *Decision Trees and Random Forests.* UNT Canvas. https://unt.instructure.com/courses/39368/pages/module-9-decision-trees-and-random-forests?module_item_id=2163845

US Census Bureau. (2020). *How the Census Bureau measures poverty*. https://www.census.gov/topics/income-poverty/poverty/guidance/poverty-measures.html#:~:text=The%20total%20family%20income%20divided,Ratio%20of%20Income%20to%20Poverty.&text=The%20difference%20in%20dollars%20between,(for%20families%20above%20poverty).

Whelton PK, Carey RM, Aronow WS, Casey DE, Collins KJ, Dennison C, et al. (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. Hypertension. 2018;71(19): e13–115.https://doi.org/10.1161/HYP.0000000000000065

World Health Organization. (2018). World Health Statistics 2018: Monitoring health for the SDGs. https://www.who.int/gho/publications/world_health_statistics/2018/en/

Appendix A

After reviewing scientific literature and the components of NHANES data sets we have selected 39 variables to include in the classification models.

The target of classification (class label) is synthesized from the Blood Pressure data set of the NHANES. This data set includes values from three separate measurements of blood pressure of each person in the NHANES study. We took the averages of systolic and diastolic values and categorized them according to the American Heart Association guidelines for hypertension (AHA,2020).

Variable	Description	Category Code	Meaning
Hypertension	Is the person hypertensive?	0	No. Average systolic < 130 and average diastolic < 80 mm Hg
		1	Yes. Average systolic ≥ 130 or average diastolic ≥ 80 mm Hg

The following table shows the selected variables, their name in the NHANES database, their group, and categories.

Group	Var #	Variable Name	Description	Category Code	Meaning
ID Number	0	SEQN	Respondent Sequence Number	-	Unique identifier
Demographic	1	RIAGENDR	Gender	1	male
				2	female
	2	RIDRETH1	Race/Hispanic origin w/ NH Asian	1	Mexican American
				2	Other Hispanic
				3	Non-Hispanic White
				4	Non-Hispanic Black
				5	Other Race - Including Multi-Racial
	3	RIDAGEYR	Age at screening adjudicated	NA	NA
	4	RIDEXPRG	Pregnancy status at exam	1	Yes, positive lab pregnancy test or self-reported pregnant at exam
				2	The participant was not pregnant at exam
				3	Cannot ascertain if the participant is pregnant at exam
				4	Participant is male, unable to get pregnant
Laboratory Data	5	LBDHDD	Direct HDL-Cholesterol (mg/dL)	NA	NA
	6	LBXTC	Total Cholesterol (mg/dL)	NA	NA

	7	LBXSTR	Triglycerides, refrig serum (mg/dL)	NA	NA
	8	LBXSUA	Uric acid (mg/dL)	NA	NA
	9	LBXSNASI	Sodium (mmol/L)	NA	NA
	10	LBXSCA	Total Calcium(mg/dL)	NA	NA
	11	LBXSTP	Total Protein (g/dL)	NA	NA
	12	LBXSAL	Albumin, refrigerated serum (g/dL)	NA	NA
	13	LBXSAPSI	Alkaline Phosphatase (ALP) (IU/L)	NA	NA
	14	LBXSASSI	Aspartate Aminotransferase (AST) (IU/L)	NA	NA
	15	LBXSBU	Blood Urea Nitrogen (mg/dL)	NA	NA
	16	LBXSCLSI	Chloride (mmol/L)	NA	NA
	17	LBXSCR	Creatinine, refrigerated serum (mg/dL)	NA	NA
	18	LBXSGB	Globulin (g/dL)	NA	NA
	19	LBXSGL	Glucose, refrigerated serum (mg/dL)	NA	NA
	20	LBXSIR	Iron, refrigerated serum (ug/dL)	NA	NA
	21	LBXSOSSI	Osmolality (mmol/Kg)	NA	NA
	22	LBXSPH	Phosphorus (mg/dL)	NA	NA
	23	LBXSKSI	Potassium (mmol/L)	NA	NA
Dietary Data	24	DBD100	How often add salt to food at table	1	Rarely
				2	Occasionally
				3	Very often
	25	DR1TSUGR	Total sugars (gm) - Total Nutrient Intakes	NA	NA
Diet Behavior & Nutrition	26	DBQ700	How healthy is diet	1	Excellent
				2	Very good
				3	Good
				4	Fair
				5	Poor
	27	DBD895	Number of meals not home prepared	NA	NA
Kidney Conditions	28	KIQ022	Ever told you had weak/failing kidneys	0	No
				1	Yes
Blood Pressure & Cholesterol	29	BPQ020	Ever told you had high blood pressure	1	Yes
				2	No
	30	BPQ080	Doctor told you - high cholesterol level	1	Yes
				2	No

	31	BPXPLS	60 sec. pulse (30 sec. pulse * 2)	NA	NA
	32	BPXPULS	Pulse regular or irregular?	1	Regular
				2	Irregular
Current Health Status	33	HSQ010	General health condition	1	Excellent
				2	Very good
				3	Good
				4	Fair
				5	Poor
Diabetes	34	DIQ010	Doctor told you that you have diabetes	1	Yes
				2	No
				3	Borderline
	35	DIQ160	Ever told you have prediabetes	1	Yes
				2	No
	36	DIQ170	Ever told have health risk for diabetes	1	Yes
				2	No
Sleep Disorders	37	SLQ050	Ever told doctor had trouble sleeping?	1	Yes
				2	No
Medical Conditions	38	MCQ080	Doctor ever said you were overweight?	1	Yes
				2	No
	39	MCQ160C	Ever told you had coronary heart disease	1	Yes
				2	No
	40	MCQ160E	Ever told you had heart attack	1	Yes
				2	No
Weight History	41	WHQ030	How do you consider your weight?	1	Overweight,
				2	Underweight, or
				3	About the right weight
Physical Activity	42	PAQ605	Vigorous work activity	1	Yes
				2	No
	43	PAQ635	Walk or bicycle	1	Yes
				2	No
	44	PAQ650	Vigorous recreational activities	1	Yes
				2	No
	45	PFQ059	Limited in any way in any activity because of a physical, mental, or emotional problem?	1	Yes
				2	No
Mental Health	46	DPQ010	Depression - Little interest doing things	0	Not at all
				1	Several days
				2	More than half the days
				3	Nearly every day

Smoking	47	SMQ020	Smoked at least 100 cigarettes in life	1	Yes
				2	No
	48	SMQ040	Do you now smoke cigarettes?	1	Every day
				2	Some days
				3	Not at all
Drug Use	49	DUQ200	Ever used marijuana or hashish	1	Yes
				2	No
Income	50	INDHHIN2	Annual Family Income	1	$ 0 to $ 4,999
				2	$ 5,000 to $ 9,999
				3	$10,000 to $14,999
				4	$15,000 to $19,999
				5	$20,000 to $24,999
				6	$25,000 to $34,999
				7	$35,000 to $44,999
				8	$45,000 to $54,999
				9	$55,000 to $64,999
				10	$65,000 to $74,999
				12	$20,000 and Over
				13	Under $20,000
				14	$75,000 to $99,999
				15	$100,000 and Over
	51	INDFMPIR	Ratio of family income to poverty	NA	NA
Examination	52	BMXBMI	Body mass index (kg/m2)	NA	NA

Appendix B

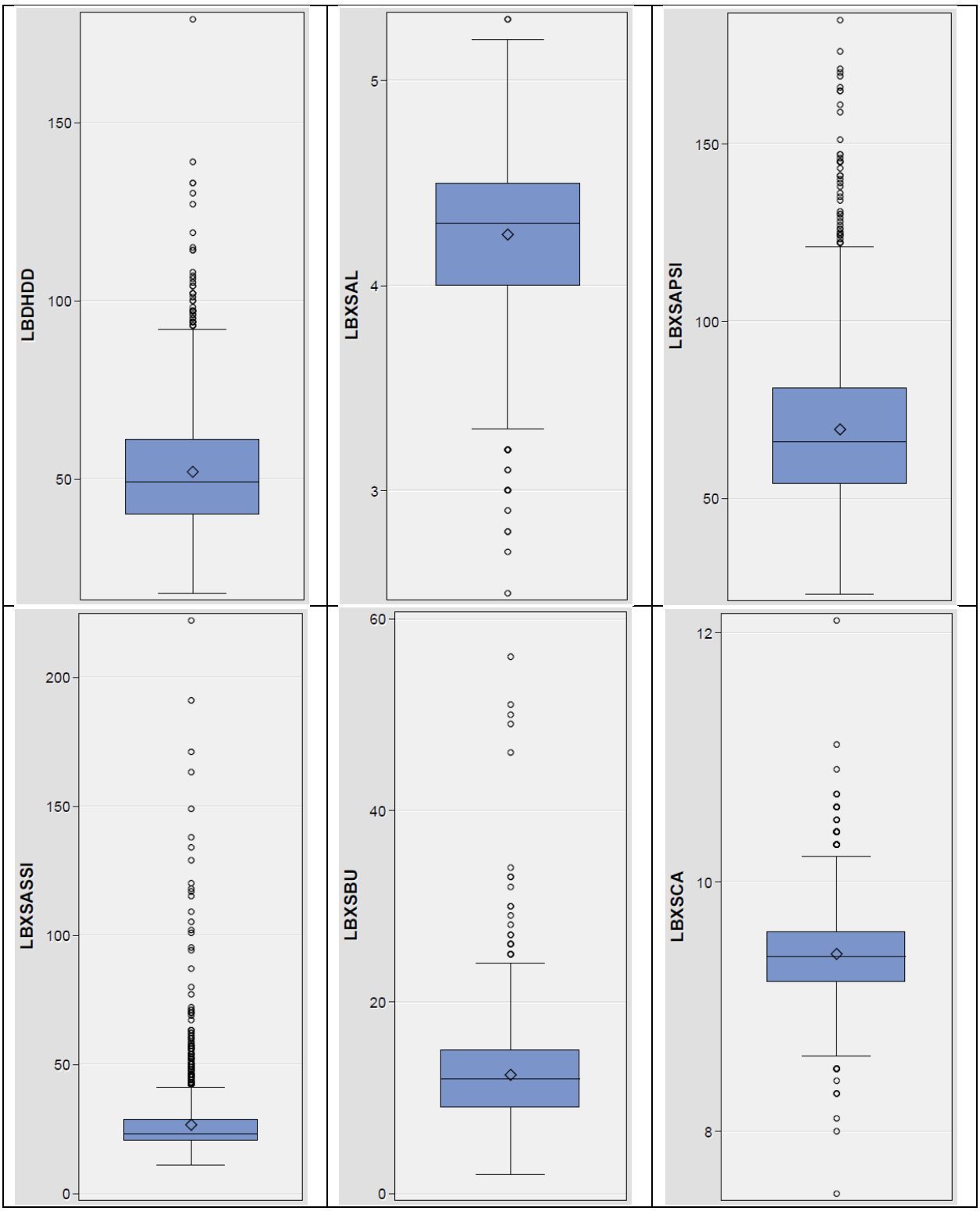


Figure B-1 Box Plots of Direct HDL-Cholesterol (LBDHDD), Albumin, refrigerated serum (LBXSAL), Alkaline Phosphatase (LBXSAPSI), Aspartate Aminotransferase (LBXSASSI), *Blood Urea Nitrogen (LBXSBU), Total Calcium (LBXSCA)*

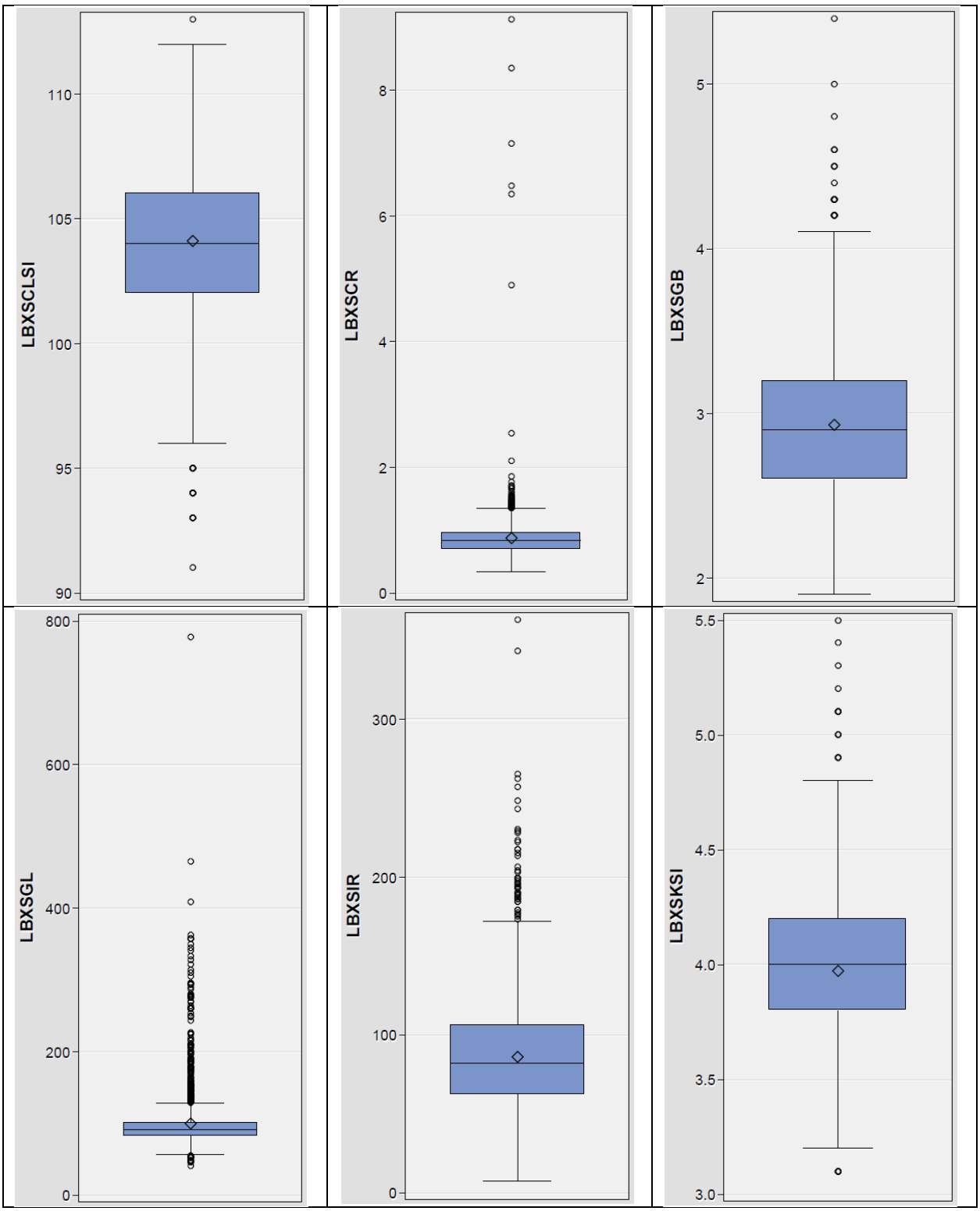


Figure B-2 Box Plots of Chloride (LBXSCLCI), Creatinine, refrigerated serum (LBXSCR), Globulin (LBXSGB), Glucose, refrigerated serum (LBXSGL), Iron, refrigerated serum (LBXSIR), Potassium (LBXSKSI)

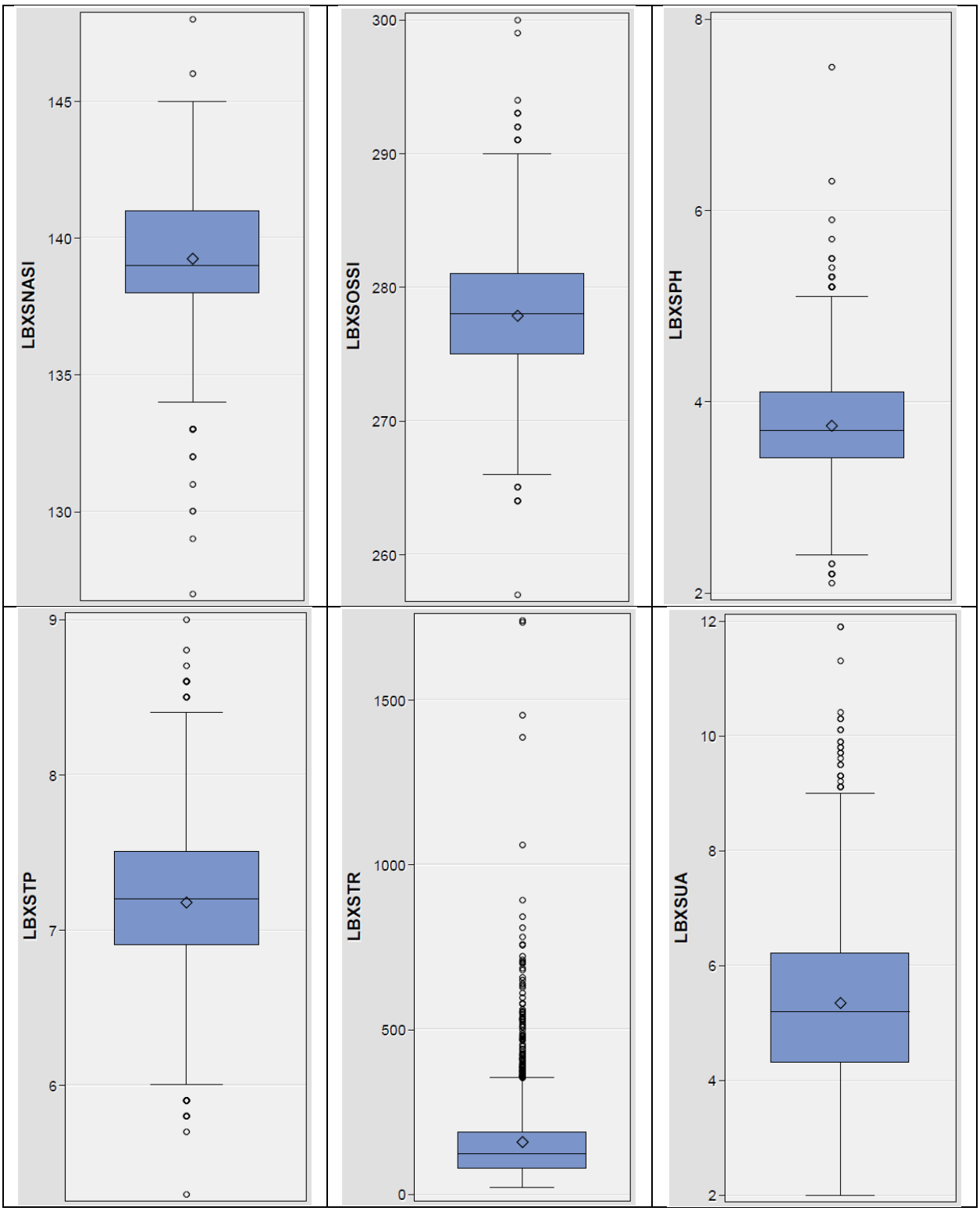


Figure B-3 Box Plots of Sodium (LBXSNASI), Osmolality (LBXSOSSI), Phosphorus (LBXSPH), Total Protein (LBXSTP), Triglycerides, refrigerated serum (LBXSTR), Uric acid (mg/dL), Uric acid (mg/dL)

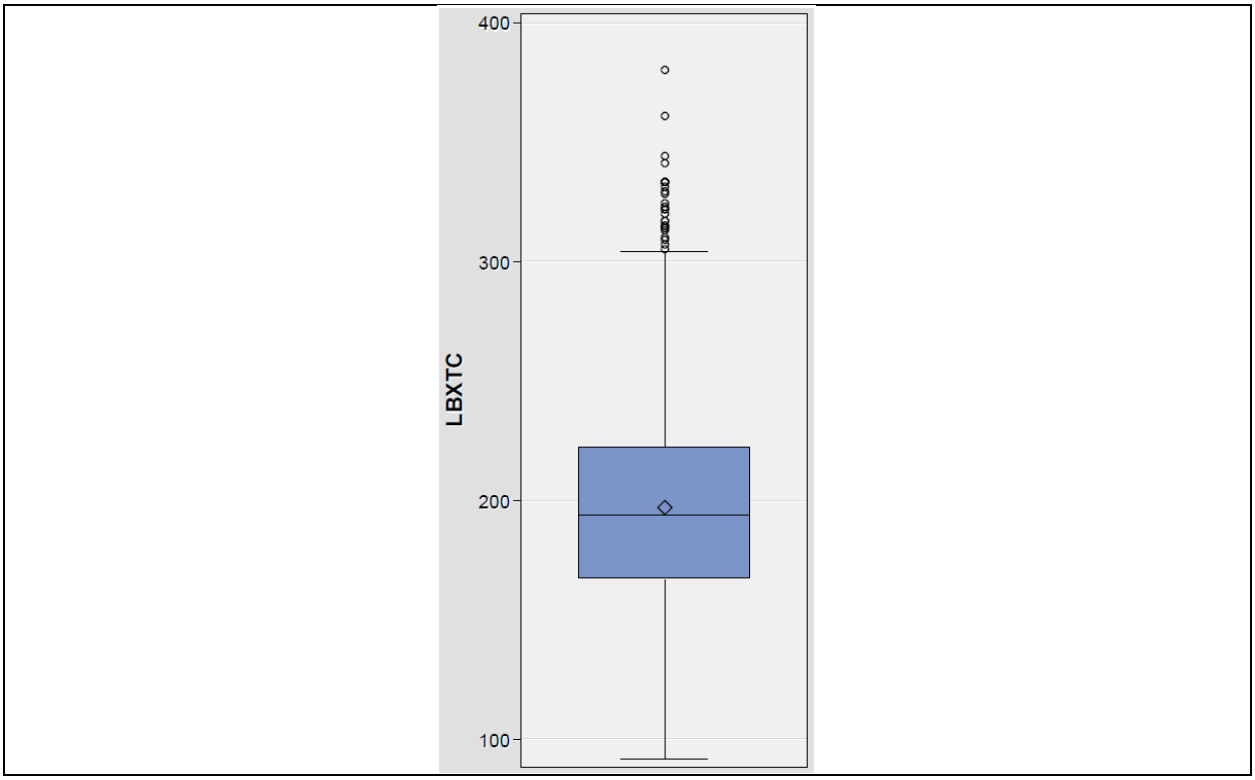


Figure B-4 Box Plots of Total Cholesterol (LBXTC)

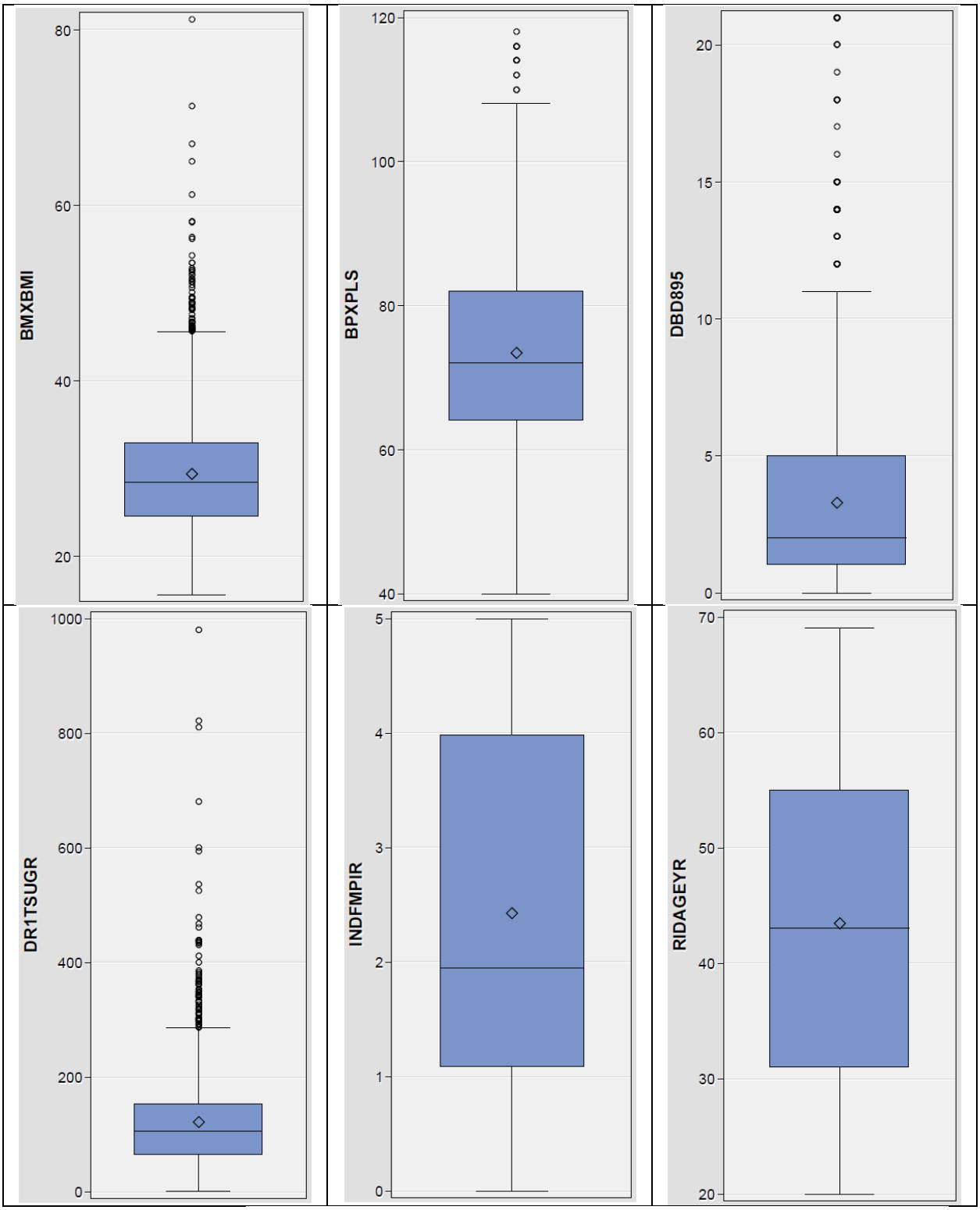


Figure B-5 Box Plots of Body mass index, 60 sec. pulse, Number of meals not home prepared, Total sugars (gm) - Total Nutrient Intakes, Ratio of family income to poverty, Age at screening adjudicated