National Health and Nutrition Examination Survey Analysis

PREPARED BY TEAM #68 FOR MGT6203

Gabriel Mink

Bella (Yifei) Ding

Vincent Pan

Nikolos Lahanis

Rahul Sati

1. Overview of Project

1.1 Team Members

- Gabriel Mink
  - GTID: 903738167: gmink3@gatech.edu
- Bella (Yifei) Ding
  - GTID: 903131776: yding302@gatech.edu
- Name: Vincent Pan
  - GTID: 903847411: vipan@gatech.edu
- Name: Nikolos Lahanis
  - GTID: 903674177: nlahanis3@gatech.edu
- Name: Rahul Sati
  - GT Id: 903549883: rsati3@gatech.edu

1.2. Background and Problem

According to the World Health Organization [1] the number of adults living with elevated levels of blood pressure known as "Hypertension" has doubled since 1990 to 1.28 Billion as of August 25, 2021. Hypertension is one of the leading causes of death and disease globally, and can significantly increase the risk of heart, brain, and kidney complications. Analyzing collected data surrounding the subject, and seeing which signals might be correlated and predictive of it, could lead to not only a better understanding of what may be linked to it, but could lead to ideas of how one one can prevent it.Blood pressure categories used to engineer our response variable were taken from the "American Heart Association's" [2] categorizations.

1.3. Overview of Problem

We endeavoured to find which factors may correlate and predict elevated levels of blood pressure. Data provided by the National Health and Nutrition Examination Survey elaborated on below was used. The data source contained everything from income, diet, lab results, responses to questions about pregnancy, activity levels, and much more. The data had over 1.8K columns, so we were confident in the robustness of the source, and hoped we could mine out some interesting insights. Our goal was to predict whether an individual had elevated blood pressure levels based on a combination of all these other signals.

1.4. General Approach

We first conducted academic research, to see what subject matter experts said about precursors to blood pressure, and how previous Data Scientists worked to solve this predictive task. Then we gleaned feature ideas, and a general framework to model predicting blood pressure as a machine learning task. Our team then conducted heavy exploration, data cleaning, dimensionality reduction techniques, and feature engineering to arrive at one master dataset containing blood pressure labels and our features. This master dataset was partitioned into train, and test, then put through 3 different classifier models. The models were evaluated according to accuracy..

1.5. Initial Hypothesis

Our group spent some time reading through medical websites and journals to get a prior understanding of what is linked to elevated blood pressure. Sites like , cdc.gov[4], mayoclinic.org[3], and a research paper from the journal of the American College of Cardiology[6] were primarily used. From those sources, we hypothesized weight, age, high sodium diet, and alcohol consumption would all be predictive factors

2. Overview of Data

2.1. Data Set

2.1.1 Source

For this project, we chose a dataset about the National Health and Nutrition Examination Survey from Kaggle.com from the Centers of Disease Control and Prevention (CDC) (https://www.kaggle.com/datasets/cdc/national-health-and-nutrition-examination-survey).

“The National Health and Nutrition Examination Survey” (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations. This entire dataset contains several subsets, and all of the subsets contain a common column “SEQN” (Respondent Sequence Number). SEQN is a unique identifier for each patient. This variable was used to join the cleaned subsets together. The description of the subsets are presented in the table below.

2.1.1 Overview

Each of the six subset datasets’ dimensions and a brief description of them is shared below.

Dataset Ref	**Dataset Name**	**Description**
1	demographic.csv [Data Dictionary] *Number of Columns: 47* *Number of Rows: 10.2k*	*Description:* Demographic data of the individual. This includes Age (DMDHRAGE), Race (RIDRETH1), and Gender (RIAGENDR)
2	diet.csv [Data Dictionary] *Number of Columns: 168* *Number of Rows: 9.8k*	*Description:* Data on the individual's first day diet. This is a questionnaire on how much e.g. salt (DBD100), amount of food was eaten (DR1_300), and tap water was drunk (DR1_320Z).
3	labs.csv [Data Dictionary] *Number of Columns: 224* *Number of Rows: 9.8k*	*Description:* Data on the Labs. Not all tests were conducted for all individuals. However, some examples of test/ detection include: Monocyte number (LBDMONO), Platelet count (LBXPLTSI), and Potassium (LBXSKSI)
4	questionnaire.csv [Data Dictionary] *Number of Columns: 953* *Number of Rows: 10.2k*	*Description:* A questionnaire asking about the health and behaviors of the individual. This includes time spent outdoors (DED125), whether they have been diagnosed with osteoporosis (OSQ060), and whether they are taking prescribed medicine (BPQ050A).
5	examinations.csv [Data Dictionary] *Number of Columns: 424* *Number of Rows: 9.8k*	*Description:* Data on the health examination for the individual. This includes questions on whether they are currently pregnant or breastfeeding (CSQQ241), grip test (MGDEXSTS), and Total abdominal fat mass (DXXTATM)
6	medications.csv [Data Dictionary] *Number of Columns: 13* *Number of Rows: 20.2k*	*Description:* Data on any medication (if any) the individual is taking. This includes the reason they are prescribed for (e.g. Muscle spasm, insomnia etc.)

2.2. Data Cleaning

2.2.1 Methodology and Techniques

After obtaining the raw dataset, each of the subsets are pre-processed in order to get rid of irrelevant columns based on project scope and outlier data points. Detailed description of data preparation for each subset, if there are any outstanding findings, is shown below.

Demographic

This dataset contained demographic data for each observation such as highest level of education, country they were born in, number of people in household, gender, income levels, etc. 47 unique columns and 10.2K rows were present and the following data cleaning and transformations were applied:

Sparse Feature Removal - Removed columns too sparse to obtain proper coverage. Columns missing 10% or more entries were removed. Resulting in 33 total columns kept. Drop columns that don't add usable/interpretable information, such as Data Release cycle, and whether an interpreter was used to conduct the interview. These columns don't add any generalized predictable power to our model. This brought the overall columns down to 28 from 33. String encoding - All of the provided columns were originally encoded as numeric regardless of whether or not the data they captured were categorical, such as country born in and marital status. Each of the leftover 28 columns were referenced with the data dictionary and encoded as factors in they were deemed categorical in nature. Drop nulls - Rather than imputing, we elected to drop the nulls for this dataset, as imputing someone's race or gender isn't a trivial matter. The final table reduced the number of columns from 47 to 28 columns and from 10.2K rows to 8756.

Diet

The initial data set for dietary data, diet.csv, consisted of datatable that constrained 168 unique features and 9813 entries. The following steps were sequentially taken in order to clean the dataset before processing it for feature selection:

Merge with Response Variable - Merging the entries of the dietary dataset with entries of our response variableautomatically removed any entries of our dietary dataset Reducing the number of entries in our dataset from 9813 to 7172. Sparse Feature Removal - columns of this dataframe that have too many blank entries are not complete enough for us to draw any meaningful conclusions from their results. As such, I decided to remove any columns that had greater than 1000 missing rows, or columns missing more than 10% of its entries. This trimmed the number of features in the dietary dataset from 169 to 90. Imputation - Even though the features with a large number of blank entries were removed, features with a smaller amount of blank entries that were kept still needed to be accounted for. The decision was made to replace the null values in these features with the median of its column. Outlier Removal (Cook's Distance) - In order to identify entries that are considered outliers, I used Cook's Distance and removed any entry with a value greater than n/4. This condensed the total number of entries in the dataset from 7172 to 6914. The result of this cleaning resulted in a dataset of an original size of 168 features and 9813 entries, all the way down to a cleaned 90 features and 7172 entries.

Examination

The initial data set for examinations data, examinations.csv, consists of 224 variables with 9813 entries. The following steps were taken to clean the data set:

Remove additional Blood Pressure variables - The data set has 21 blood pressure variables (e.g. "BPAARM - Arm Selected", "BPAXSY1 - Systolic: Blood pressure, first reading", "BPAXSY2 - Systolic: Blood pressure second reading" etc.) These measurements, while highly predictive, would not be useful to action on as they are used to find the blood pressure. 203 variables remained Sparse Feature Removal - Removed Columns with > 10% missing values. Following a similar reason to Niko's cleaning in the Diet data set. 72 variables remained. Feature Creation - Tooth Quality Count: 30 of the columns are 'Tooth Count' Related Categorical Responses. The responses include: "E: Missing due to dental disease", "F: Permanent Tooth with a restored surface condition", "S: Sound permanent tooth", and "Unerupted". Instead of

looking at each tooth, I've counted how many E's, F's, etc. each row has. This reduces 30 old columns to 15 new columns, 57 variables remaining

Medication

The initial dataset of medications.csv contains 13 variables with 20194 entries. The following steps are carried out in order to clean the dataset for further analysis:

Get rid of unusable/irrelevant variables: After computing the ratio of missing data for each predictor variable, specific variable are dropped because over 95% of the data in each of them are missing. These columns will not be usable for analysis.Drop rows with outlier data points: For the remaining variables, they were inspected to find potential outlier points. Rows with null values across all variables were also removed. The resulting cleaned table consists of 8 variables (7 predictor variables and 1 unique identifier for joining tables) and 13082 entries, which is a 35.22% row reduction from the original dataset.

Labs

The initial dataset of labs.csv contains 424 variables with 9813 entries. The following steps are carried out in order to clean the dataset for further analysis:

Sparse Feature Removal - Removed Columns with > 25% missing values as columns with null values cannot be used for drawing meaningful insights and predictions.Data imputation – Remaining missing values after the above step were imputed with median value for that column. The result of this cleaning resulted in a dataset of an original size of 424 variables to a cleaned dataset of 47 features.

2.2.2 Challenges

The dataset had a categorical column that was used to label patients with high blood pressure, but if we were to use this our classes would have been highly imbalanced: 317 (high hypertension) vs. 9493 (non-hypertension). We found that this was because they were using the most extreme cases of hypertension for their labels. We reconstrucedt the response variable entirely based on the two factors that comprise blood pressure, systolic, and diastolic readings, and re-engineered the label to delineate normal from elevated blood pressure level. Doing so gave us a far more balanced data set 3027(elevated):4145(normal) . These are the labels we opted to use going forward

2.3. Key Variables

Each of the dataset had a few interesting key variables we investigated and wanted to highlight based on our hypotheses and our general understanding. Blood Pressure is our response predictor. Commonly known variables, such as weight and height, were on our radar as key variables to understand, but, overall, we wanted to explore how any of the other health variables contributed to different blood pressure. Furthermore, the results from the first run logistic regression and from feature selection provides us with clearer direction on which variables ended up being important. See the model section for more information.

2.4. EDA Insights

Diet

In order to determine which features will be most significant in our dataset, I sequentially ran the following techniques in order to determine which dietary features to keep in our overall blood pressure analysis:

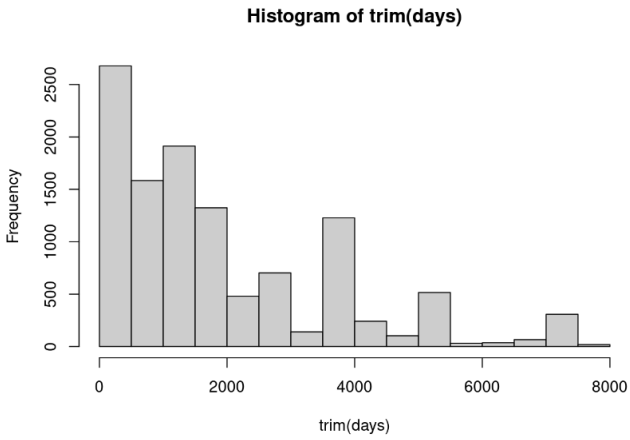
Remove Correlated Features -. First,highly correlated features were removed according to a correlation matrix,. Next any features that had a VIF higher than 5 were removed. Reducing columns from 90 to 40. Elastic Net Regression - The Elastic-net regression further reduced the number of features from 40 to 33. Remove Features with Low Variance - removal of features with low variance allows us to select features that affect blood pressure with noticeable changes in their own values. With the features with low variance removed, the number of features reduces further from 33 to 25.

Using various tools at our disposal for feature selection, the total number of features in our dietary dataset reduced from 90 to 25 that are considered significant in the scope of our blood pressure analysis

Medications

For the variable “RXDUSE” (In the past 30 days, have you used or taken medication for which a prescription is needed?) we saw that the majority of patients fall into “1”, which indicates that a patient has used or taken medication for which a prescription is needed, and the patients who do not take prescription medicine are about almost half of that.

A histogram of “RXDDAYS” (For how long have you been using or taking {PRODUCT NAME}?) was also plotted, indicating how long patients have been taking a certain medication. Since there are outliers present in this dataset, in order to get a better distribution, the x-axis values that only fall into the range of [mean - 1.5*IQR, mean + 1.5*IQR] were selected in the plot. From the figure below, the majority of patients have been taking their medicines for less than 2000 days.



Duration of Prescription Medication Usage

Examination

The examination data has two main categories of information: Basic Measurements (e.g. Weight and Height), and Teeth health. We see that a quarter of patients have No Sound Teeth, while the rest appear to be a bit more uniformly distributed. According to the histogram.

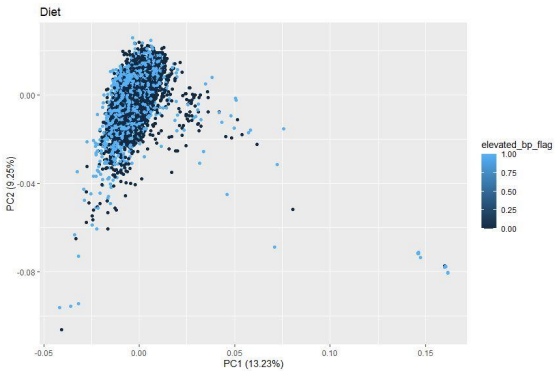
Shown by the scatterplot, as expected, as height increases, weight generally increases. However it is not a perfectly linear relationship.We observe a few data points that share the same height around the 161 cm mark, and the 60 kg mark. This could be an indication of some data accuracy issue.

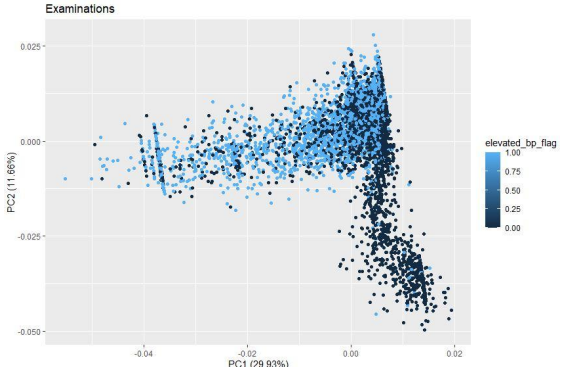
Labs

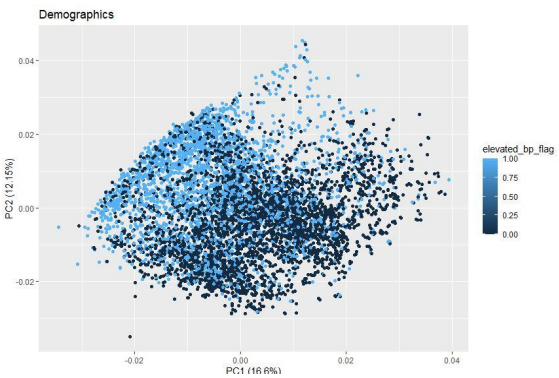
After initial data cleaning I got 47 features out of 424. Now, further, to avoid multicollinearity correlated variables were removed. After creating correlation matrix, features with correlation higher than 80% were removed. This resulted in reducing the factors to 38 from the cleaned 47 data set.

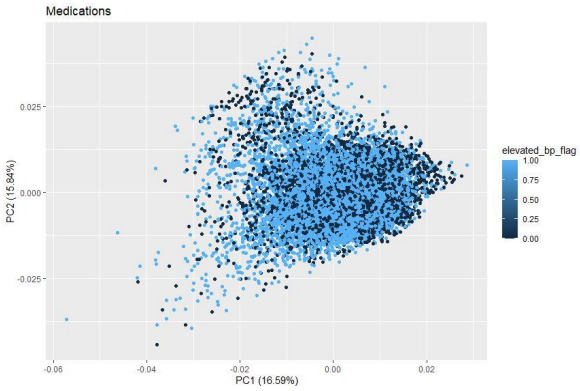
PCA:

Additionally, PCAs were performed on each of the datasets. Below are the results, where we have colored the data points based on our response Blood Pressure.









Just based on these charts, it looks like a soft margin SVM could be used on Demographics or Diet PCA to get a quick result. For Examination, we can see on the bottom right that there is a cluster of similarly colored datapoints. By understanding the PCA of the Examination data, we could understand what kind of patients that cluster is describing.

2.5. Feature Engineering

2.5.1 Methodology and Techniques

ELASTIC NET REGRESSION

By incorporating an Elastic Net Regression model onto each of our cleaned datasets, we implemented feature selection and enhanced the reliability of our predictive models for blood pressure to focus on only the most significant predictable features. Elastic net regression is similar to most regression models, with the one exception that it assigns weights to coefficients, and in this way inherently performs feature selection, because we can decide that we will only incorporate features with high weights (highest predictive ability) into our future models.

Our final deliverable is also in the regression model class, making Elastic Net Regression as a feature selection tool extremely favourable. By using Elastic Net Regression, we successfully reduced the total number of features in our model by a further 35%, while only keeping the ones with the most significant predictive ability.

3. Modelling

Data Preparation

The entire data has different subsets and initial data cleaning is performed for each data set. After this, demographic, diet, labs and examinations data sets are merged together. Next, the response variable called "elevated_bp_flag" is created and merged with the data set. Response variable is 0 when the individual has normal BPXSY1 BPXDI1 readings, and 1 when it isn't (i.e. elevated and high blood pressure according to the blood pressure levels defined by the American Heart Association).Data cleaning and imputation is also performed after we get the merged data set to remove features with > 25% null values. The final cleaned data set has 152 features and 6383 rows.

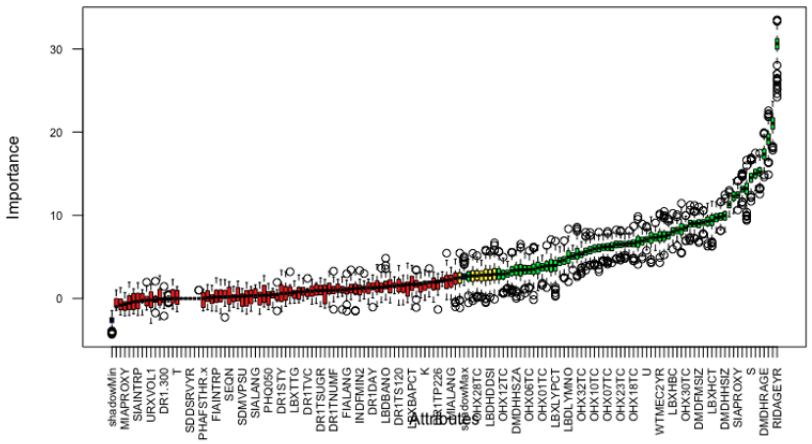
Train Test Split

We split the data into a training and testing set in the 70% train and 30% test set. The training set will be used to fit our model and we will be testing the model over the testing set.

Feature Selection

We used the Boruta algorithm for variable selection. The Boruta algorithm is a wrapper built around the random forest classification algorithm. What basically happens is that randomly shuffled shadow attributes are defined to establish a

baseline performance for prediction against the target variable. Then a hypothesis test is used to determine, with a certain level of risk (0.05 by default) if each variable is correlated only randomly. Variables that fail to be rejected are removed from consideration.After implementing it, we get 70 features out of 152 features as important features for the model.



3.1 *Logistic Regression*

Model-1

Logistic Regression Model

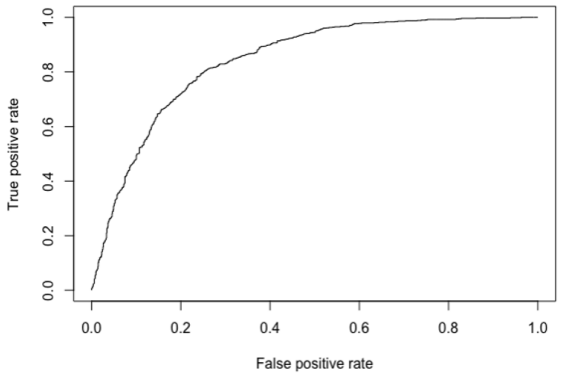
Next step is to build a logistic regression model using all 70 predictor variables that we got after variable selection, and “elevated_bp_flag” is our response variable. The variables found to be most significant by the logistic regression model are BMXWT - Weight (kg) LBDTCSI - Total Cholesterol(mmol/L),RIAGENDR - Gender of the participant, RIDAGEYR - Age in years of the participant

Predict

We then predict the target variable on the test dataset using the probabilistic logistic regression model. Probability of >= 0.5 was indicated as 1 and 0 otherwise.

Performance Measure

We measured the performance on the test dataset using Accuracy and Area under the curve. Accuracy of the model came out to be 76.76%. ROC area under the curve is 84.15%



Cross-validation

A 10-fold cross validation is also conducted on the feature selected dataset. The resulting mode achieved an accuracy of 75.88%, which is slightly lower than the model without cross validation. Due to randomness when splitting into training and testing sets, the cross-validation accuracy could be lower than the basic regression model.

3.2 *Random Forest*

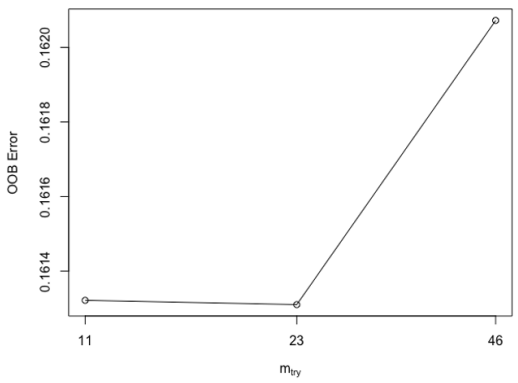
The next model that we have tried is Random Forest using the predictor variables that we got after variable selection, and “elevated_bp_flag” is our response variable. First, we will create a base Random Forest model with default parameters. By default, the number of trees is 500 and the number of variables tried at each split is 23 in this case.

Predict

We then predict the target variable on the test dataset using the base random forest model. Probability of >= 0.5 was indicated as 1 and 0 otherwise.

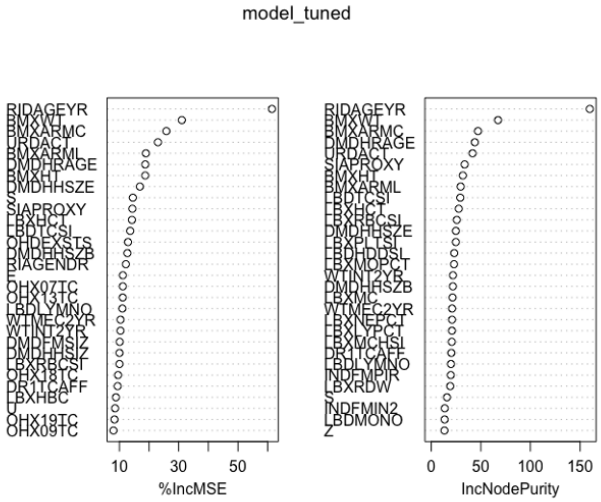
Performance Measure

We measured the performance on the test dataset using Accuracy and Area under the curve. Accuracy of the model came out to be 77.65%. ROC



area under the curve is 84.29%. Next, we will fine tune the parameters of the base model. We can tune the random forest model by changing the number of variables randomly sampled at each stage (mtry). Mtry is the number of variables randomly sampled as candidates at each split.
From the OOB error graph we can see that Mtry = 23 gives us the min error.

Next a tuned random forest model using Mtry = 23 was created, with a % var explained at 33.22.



Predict and Performance Measure

We then predict the target variable on the test dataset using the base random forest model. Probability of >= 0.5 was indicated as 1 and 0 otherwise.We measured the performance on the test dataset using Accuracy and Area under the curve. Accuracy of the model came out to be 77.81%. ROC area under the curve is 84.28%.

Importance of variables

Let's create the plot to visualise our variables of tuned random forest models by importance.RIDAGEYR (Age) , BMXWT (Weight kg), BMXARMC (Arm Circumference cm), URDACT (Albumin creatinine ratio mg/g), BMXARML (Upper Arm Length cm) are top 4 important features.

Cross-validation

The 10-fold cross-validation achieved an accuracy of 77.33%, which is very similar to the accuracy achieved without it. Similarly, the higher accuracy without cross-validation could be because of the randomness in data splitting.

3.3 *XGBoost*

The next model that we have tried is XGBoost using the predictor variables that we got after variable selection, and "elevated_bp_flag" is our response variable.First, we will create a base XGBoost model with default parameters. Specifying the watchlist is an important step here because it is the parameter that tells XGBoost to stop iterating when the validation accuracy (1-error) does not improve anymore.

Predict

We then predict the target variable on the test dataset using the base XGBoost model. Probability of >= 0.5 was indicated as 1 and 0 otherwise.

Performance Measure

We measured the performance on the test dataset using Accuracy and Area under the curve. Accuracy of the model came out to be 76.76%. ROC area under the curve is 83.61%.

Hyperparameter-Tuning for XGBoost Model

Next, we tune the hyperparameters of the XGBoost to get optimal model parameters and then train the tuned model using those parameters.We have created 1,000 random hyperparameter-value sets using a for-loop in a given range. Then we have executed the XGBoost algorithm 1,000 times with the predefined hyperparameter value sets.78.79% is the highest accuracy that we are achieving and now we use the hyperparameters of the best hyperparameter value set into our XGBoost algorithm and again check the model's performance.

Predict and Performance Measure

We then predict the target variable on the test dataset using the base random forest model. Probability of >= 0.5 was indicated as 1 and 0 otherwise.We measured the performance on the test dataset using Accuracy and Area under the curve. Accuracy of the model came out to be 78.8%. ROC area under the curve is 84.31%. The accuracy of the tuned model is better than our base model.

3.4 *Modelling Conclusion*

Comparing the model performance based on accuracy and AUC for all 3 models (Logistic regression, Random Forest and XGBoost) on the test set, XGBoost model with hyperparameter tuning having 78.8% accuracy and 84.31% AUC is selected.

The most important features are RIDAGEYR (Age in years) & BMXWT (Weight in kg), these are clearly related to high blood pressure. The other variables in order of importance are BMXARMC (Arm Circumference cm), URDACT (Albumin creatinine ratio mg/g) and BMXHT (Standing Height cm).It is interesting to note that one of the important variable URDACT, which was not very obvious factor like age and weight has a proven association with high blood pressure/hypertension, and as per JAHA: Journal of American Heart Association-" Low levels of albuminuria, UACR below 30 mg/g, are associated with increased risk of incident hypertension (JAHA: Journal of American Heart Association Link)"

4. Closing Remarks

4.1. Key Takeaways

- Top factors indicated in current medical research to be correlated with hypertension, also appeared within our model's feature importances. Weight, and age being the most prevalent.
- Choosing a dataset as robust as this with over 1600 columns sounded excellent and provided a plethora of data points, during the EDA portion most columns had a unusable degree of sparsity. The final total number of features utilised was around 151, which was a reduction of over 85% of the total number of columns the dataset initially offered.

4.2. Future Opportunities

- More advanced feature selection algorithms such as GridSearch or HyperOpt could have been explored to marginally improve model performances
- Macro features such as city, state, and region could have been explored. Yielding predictive power for cold-start scenarios.

4.3. Conclusion

- After extensive pre-processing of the Nutritional Health and Examination data set, through PCA feature reduction, sparsity elimination, factorization, and several engineered additional features, a final data set was created to predict blood pressure. XGB, Random Forest, and Logististic regression were all ran and XGB chosen for its accuracy of 78.8%. Weight and age appeared at the top of nearly all of our models, in addition to several other surprising features such as Albumin Creatine, which when further researched[7] was noted as being related to hypertension.
- We hope that this paper promotes further predictive analytic work in nutrition and healthcare in order to prescribe preventative measures and additional insights as to potential correlative factors further upstream of a physical blood pressure test.

00. Appendix

- **Works cited section.**

1. World Health Organization. (n.d.). *More than 700 million people with untreated hypertension*. World Health Organization. Retrieved October 28, 2022, from https://www.who.int/news/item/25-08-2021-more-than-700-million-people-with-untreated-hypertension
2. *Understanding blood pressure readings*. www.heart.org. (2022, September 9). Retrieved October 28, 2022, from https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings
3. Mayo Foundation for Medical Education and Research. (2022, September 15). *High blood pressure (hypertension)*. Mayo Clinic. Retrieved October 28, 2022, from https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/symptoms-causes/syc-20373410
4. Centers for Disease Control and Prevention. (2022, October 14). *Facts about hypertension*. Centers for Disease Control and Prevention. Retrieved October 28, 2022, from https://www.cdc.gov/bloodpressure/facts.htm
5. *Predicting systolic blood pressure using machine learning*. IEEE Xplore. (n.d.). Retrieved October 28, 2022, from https://ieeexplore.ieee.org/document/7069529
6. Whelton PK, Carey RM, Aronow, WS, Casey DE, Collins KJ, Himmelfarb CD, 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: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. J Am Coll Cardiol. 2018;71(19):e127–e248. From https://www.jacc.org/doi/10.1016/j.jacc.2017.11.006?_ga=2.86879320.1182640551.1528306905-1524800955.1528306905
7. Sung, K. C., Ki-Chul Sung Division of Cardiology, Ryu, S., Seungho Ryu Department of Occupational and Environmental Medicine, Lee, J. Y., Jong-Young Lee Division of Cardiology, Lee, S. H., Sung Ho Lee Division of Cardiology, Cheong, E. S., EunSun Cheong Division of Cardiology, Hyun, Y. Y., Young-Youl Hyun Division of Nephrology, Lee, K. B., Kyu-Beck Lee Division of Nephrology, Kim, H., Hyang Kim Division of Nephrology, Byrne, C. D., Christopher D. Byrne Nutrition and Metabolism, & Sung, *C. to: K. C. (2016, September 13). *Urine albumin/creatinine ratio below 30 mg/g is a predictor of incident hypertension and cardiovascular mortality*. Journal of the American Heart Association. Retrieved November 13, 2022, from https://www.ahajournals.org/doi/10.1161/JAHA.116.003245#:~:text=Urine%20Albumin%2FCreatinine%20Ratio%20Below,of%20the%20American%20Heart%20Association