MGT 6203 Project Progress Report

Team #68

Team Members:

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

Problem Statement

- Necessary background information/framing of the problem
    a. 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. This isn't an innocuous diagnosis either, as Hypertension is one of the leading causes of death and disease globally, and can significantly increase the risk of heart, brain, and kidney complications. Seeing as there is a growing number of people that have such a medical condition, 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.
    b. Blood pressure categories used to engineer our response variable were taken from the "American Heart Association's" [2] categorizations.
- An overview of the problem in general
    a. For this project our group sought to shed some light on what factors may correlate and predict elevated level of blood pressure. We used data provided by the National Health and Nutrition Examination Survey elaborated on below, to seek to answer this inquiry. The data source contained everything from income, diet, lab results, to responses to questions about pregnancy, activity levels, and much 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.
- Any initial hypotheses?
    a. Before even touching the data, 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

Data Prepping

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), which 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.

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.)

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.

1. 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.
2. 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
3. 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.
4. 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:

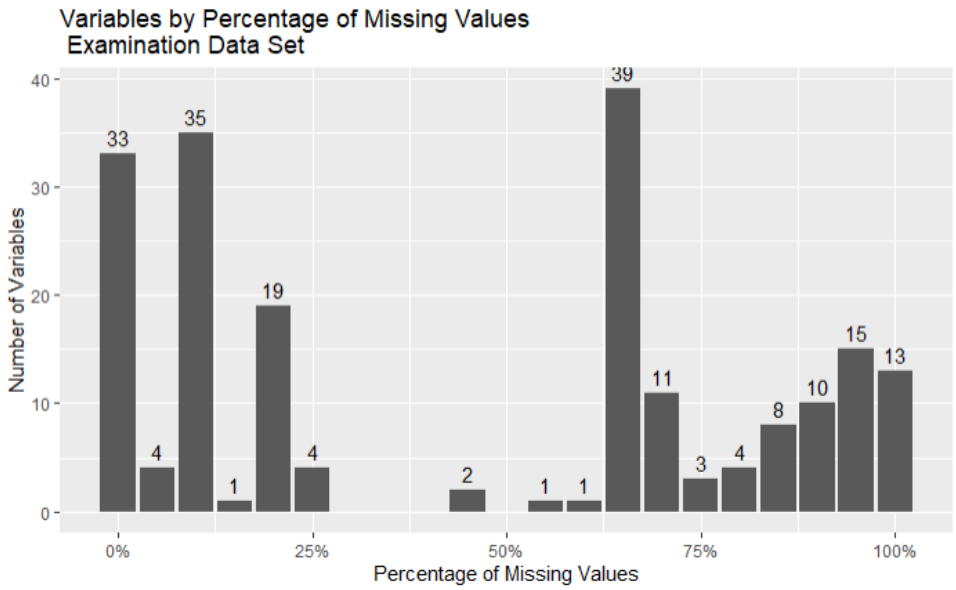
1. Merge with Response Variable - Merging the entries of the dietary dataset with entries of our response variable, blood pressure, automatically removed any entries of our dietary dataset that didn't have a response that it could be modeled with. This alone reduced the number of entries in our dataset from 9813 to 7172.
2. 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.
3. 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.
4. 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:

1. 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.
    a. 203 variables remaining (note: one column is the unique identifier)
2. Sparse Feature Removal - Removed Columns with > 10% missing values. Following a similar reason to Niko's cleaning in the Diet data set.
    a. 72 variables remaining (note: one column is the unique identifier)



3. Feature Creation -
    a. Tooth Quality Count:
        i. 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".
        ii. Instead of looking at each tooth, I've counted how many E's, F's, etc. each row has.
        iii. This reduces 30 old columns to 15 new columns
        iv. 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, the variables "RXDRSC2" (ICD-10-CM code 2), "RXDRSC3" (ICD-10-CM code 3) and "RXDRSD3" (ICD-10-CM code 3 description) 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. For variables "RXDDRUG" (Generic drug name), "RXDDAYS" (For how long have you been using or taking {PRODUCT NAME}?) and "RXDRSC1" (ICD-10-CM code 1), the rows with outlier values "55555" and "99999" were removed. In additions, 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:

1. Sparse Feature Removal - Removed Columns with > 25% missing values as columns with null values cannot be used for drawing meaningful insights and predictions.

2. 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.

Exploratory Data Analysis (EDA)

EDA results from each sub-dataset, if there are any useful insights, are shown below.

Diet

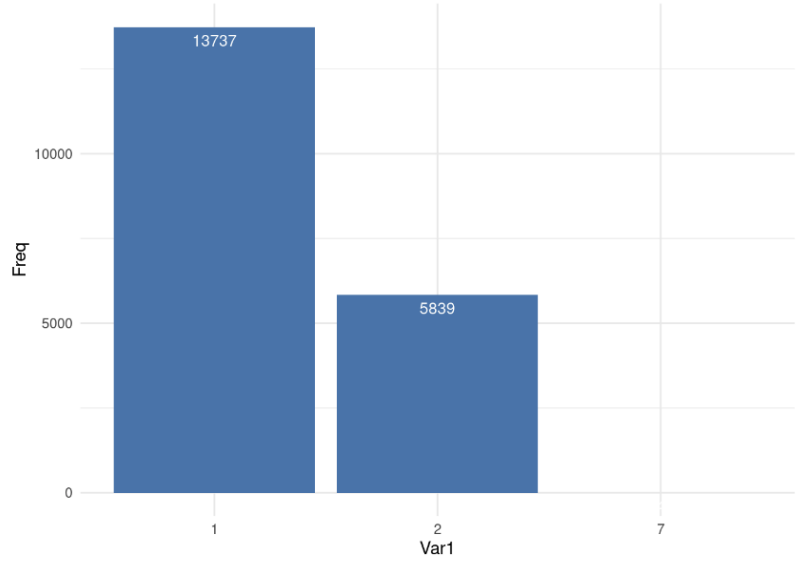
In order to determine which features will be most significant in our dataset, I sequentially ran the following techniques in order to determine which dietary features to keep in our overall blood pressure analysis:

1. Remove Correlated Features - In order to prevent Multicollinearity when running our models, I implemented two main techniques to identify features that are linearly-dependent, and subsequently remove them. First, I created a correlation matrix, and removed any features that had a correlation higher than an absolute value of .6 within that matrix. Next, I calculated the Variance Inflation Factor (VIF) of each column predictor of blood pressure in a linear regression model, and removed any features that had a VIF higher than 5. These two techniques reduced the total number of features in the dataset from 90 to 40.
2. Elastic Net Regression - In order to narrow in on specific features that most directly affect blood pressure I implemented an Elastic Net Regression model. The Elastic-net regression further reduced the number of features from 40 to 33.
3. Remove Features with Low Variance - In order to only narrow in on features that create actionable insights, remove 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 anal

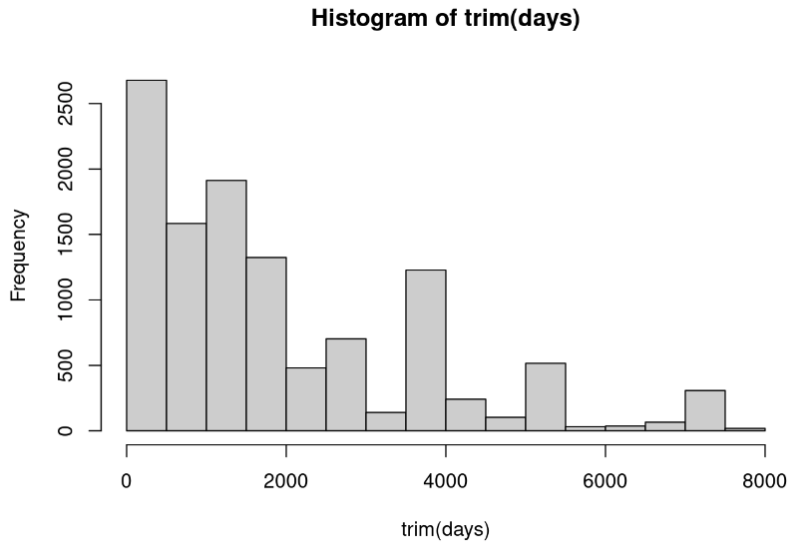
Medications

Using the cleaned dataset using the methods mentioned in the Data Prepping section, some plots were created to have some preliminary conclusions on the dataset. For the variable "RXDUSE" (In the past 30 days, have you used or taken medication for which a prescription is needed?), after plotting the distribution, we can see 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.



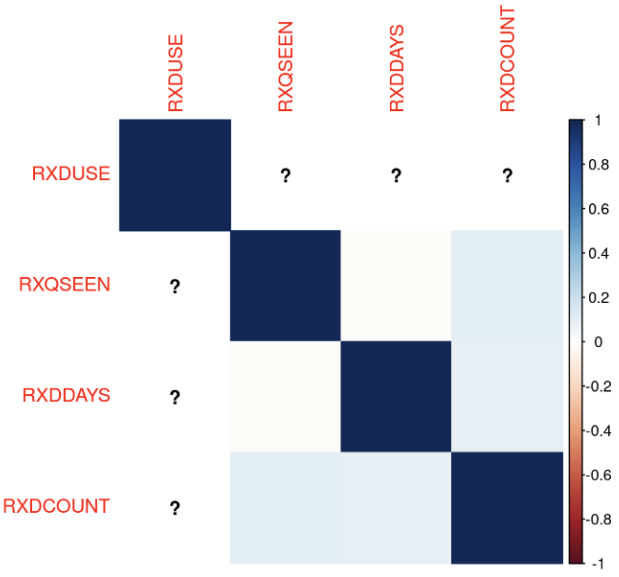
Distribution of Prescription Medication Usage

In addition, 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.



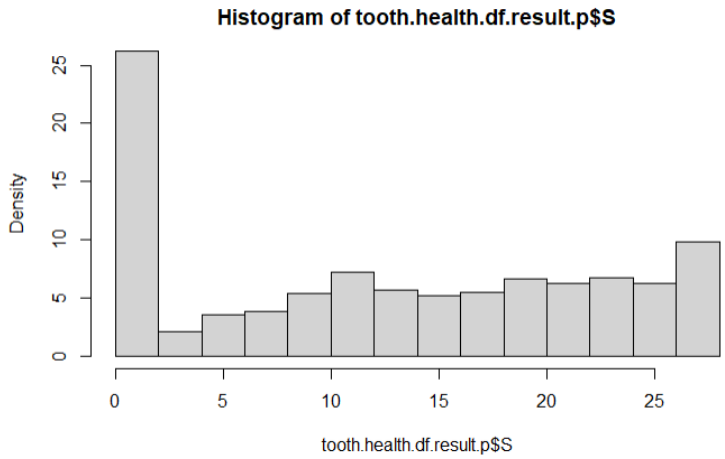
Duration of Prescription Medication Usage

Lastly, the correlation plot for selected variables was also plotted. Overall, there was not much correlation between all variables. There is very low correlation between RXDCOUNT, RXQSEEN and RXDDAYS.



Correlation plot for numeric variables

Examination



Histogram of Distribution of Patients by Number of Tooth Health = S (Sound).

- We see that a quarter of patients have No Sound Teeth, while the rest appear to be a bit more uniformly distributed.

Scatter Plot of Weight (BMXWT) and Height (BMXHT)

- As expected, as height increases, weight generally increases
- 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.

Initial Modeling

After cleaning and combining the datasets, we decided to run a preliminary regression model to obtain some results first.

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.

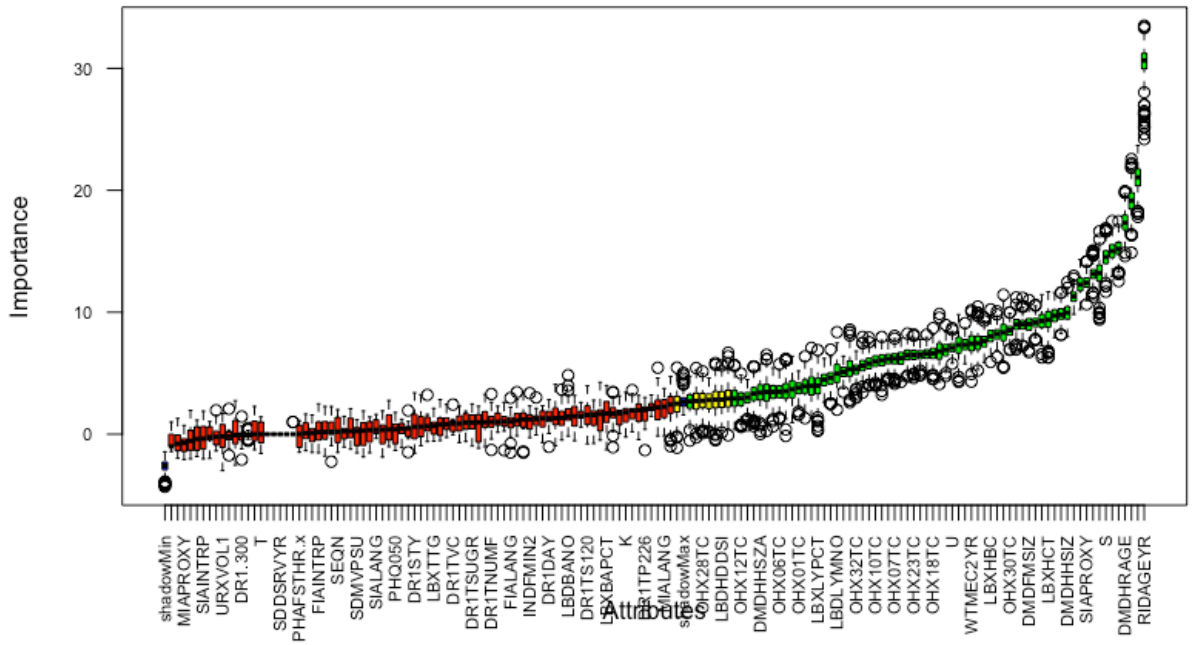
Train Test Split

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

The goal of feature selection is to find the best set of important features for our model. Large number of factors not only increase the computational cost of modeling but also impacts performance of the model. We have 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.




```
> final.boruta
Boruta performed 499 iterations in 13.9262 mins.
Tentatives roughfixed over the last 499 iterations.
 70 attributes confirmed important: BMXARMC, BMXARML, BMXHT, BMXWT, D and 65 more;
 81 attributes confirmed unimportant: AIALANGA, BMDSTATS, DBQ095Z, DMDBORN4, DMDCITZN and 76 more;
```

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

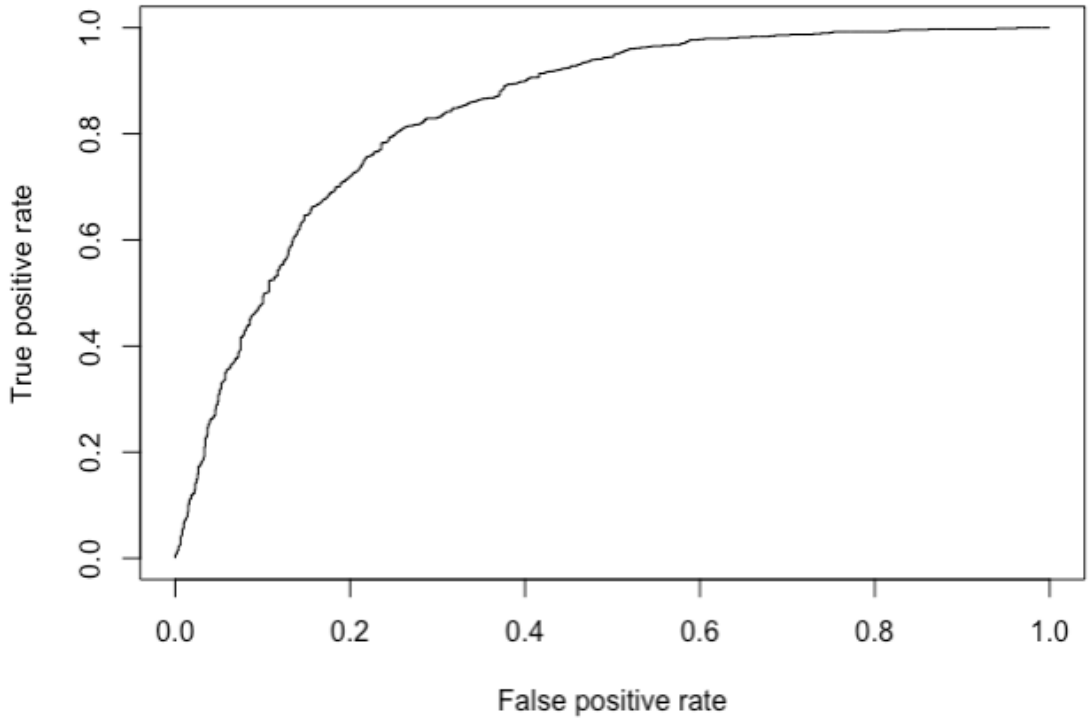
```
> summary(lm)

Call:
glm(formula = target ~ ., family = binomial(link = "logit"),
    data = train_new)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7967  -0.7341  -0.2962   0.7959   2.8775

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.428e+00  3.245e+00  -1.981 0.047619 *
WTDR2D      -9.938e-07  8.377e-07  -1.186 0.235525
DR1TCAFF     1.432e-04  2.430e-04   0.589 0.555818
BMXWT        1.990e-02  5.432e-03   3.663 0.000249 ***
BMXHT       -1.791e-02  8.389e-03  -2.135 0.032785 *
BMXARML      5.916e-02  2.573e-02   2.299 0.021484 *
BMXARMC      6.912e-04  2.011e-02   0.034 0.972585
OHDEXSTS     2.767e-01  1.931e-01   1.440 0.149730


LBXHBC      -1.116e-01  1.689e-01  -0.660 0.508982
LBDTCSI      1.541e-01  3.889e-02   3.961 7.45e-05 ***
RIAGENDR    -4.564e-01  1.231e-01  -3.707 0.000210 ***
RIDAGEYR     3.566e-02  4.076e-03   8.748  < 2e-16 ***
RIDRETH3     6.601e-02  2.677e-02   2.466 0.013653 *
DMQMILIZ     4.495e-01  1.503e-01   2.991 0.002777 **
SIAPROXY     5.116e-01  1.953e-01   2.620 0.008800 **
DMDHHSIZ     2.348e-02  7.009e-02   0.335 0.737659
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6035.1  on 4467  degrees of freedom
Residual deviance: 4264.6  on 4397  degrees of freedom
AIC: 4406.6

Number of Fisher Scoring iterations: 6
```

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%



Next Steps

We can further try to improve the performance of the model using other ensemble approaches like random forest and XGBoost.

Once we finalize the model, we will tune the parameters and use the best parameters on the predicted dataset.

Findings

Any significant findings on unexpected problems, challenges, or other things from either the overall dataset or specific sub-datasets were recorded below.

Response Variable

One of the first problems we encountered was how we wanted to construct the response variable. 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 opted to reconstruct the response variable entirely based on the two factors that comprise blood pressure, systolic, and diastolic readings, and re-engineer 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

Diet

Following an analysis of the diet.csv dataset, the following features were considered significant as it pertains to our response variable, blood pressure:

Variable.Name	Variable.Description
SEQN	Respondent sequence number.
WTDR2D	Dietary two-day sample weight
DR1DRSTZ	Dietary recall status
DR1EXMER	Interviewer ID code
DRDINT	Indicates whether the sample person has intake data for one or two days.
DR1DAY	Intake day of the week
DR1LANG	The respondent spoke mostly:
DBQ095Z	What type of salt {do you/does SP} usually add to {your/his/her/SP's} food at the table? Would you say . . .
DR1STY	Did {you/SP} add any salt to {your/her/his} food at the table yesterday? Salt includes ordinary or seasoned salt, lite salt, or a salt substitute.
DRQSDIET	Are you currently on any kind of diet, either to lose weight or for some other health-related reason?
DR1TNUMF	Total number of foods/beverages reported in the individual foods file
DR1TSUGR	Total sugars (gm)
DR1TLYCO	Lycopene (mcg)
DR1TVC	Vitamin C (mg)
DR1TVK	Vitamin K (mcg)
DR1TCAFF	Caffeine (mg)
DR1TS120	SFA 12:0 (Dodecanoic) (gm)
DR1TP183	PFA 18:3 (Octadecatrienoic) (gm)
DR1TP184	PFA 18:4 (Octadecatetraenoic) (gm)
DR1TP204	PFA 20:4 (Eicosatetraenoic) (gm)
DR1TP226	PFA 22:6 (Docosahexaenoic) (gm)
DR1BWATZ	Total bottled water drank yesterday (gm)
DR1TWS	When you drink tap water, what is the main source of the tap water? Is the city water supply (community water supply); a well or rain cistern; a spring; or something else?
DRD340	Please look at this list of shellfish. During the past 30 days did you eat any types of shellfish listed on this card? Include any foods that had shellfish in them such as sandwiches, soups, or salads.
DRD360	Please look at this list of fish. During the past 30 days did you eat any types of fish listed on this card? Include any foods that had fish in them such as sandwiches, soups, or salads.

Table : Significant Feature*s from Diet.csv*

Very generally, the above features can be segmented into the following groups: General Interviewer Details, Salt intake, Food Variety/Diet Type, Nutrient Intake, Water Intake, and Seafood Intake. By broadly defining these features into general groups, they can be more easily incorporated into our overall model for blood pressure health outcomes.

Examination

- Unexpected problems:
  - Discarding Predictors: Some predictors (21) had to be discarded due to being intrinsically tied to blood pressure (i.e. the actual measurement of blood pressure)
  - Unbalanced sample: The Blood Pressure variable, PEASCST1, had the distribution of:
  -

	PEASCST1 = 1	PEASCST1 = 2	PEASCST1 = 3
Row Count	9493	3	317
Percentage Total	96.7%	0.00%	3.23%

-

- Blood Pressure Column

Medications

In addition to some insights interpreted from the EDA, we still encountered some problems throughout the process. In this dataset, the correlation between certain variables could not be calculated for some reason. For example, in the correlation plot above, the correlation between “RXDDAYS” and “RXDUSE” could not be viewed. This could be caused by the fact that there are so many boolean variables in this dataset, or there is not much correlation between variables in this particular dataset. It is possible that once this dataset is merged with all other datasets, additional useful findings might be discovered because some variables from this dataset might have correlation with variables from other datasets.

Approach

Feature Selection

After all sub-datasets were cleaned up, we combined all datasets into our final comprehensive dataset to use in our analysis. However, there are still quite a lot of predictor variables left, which is inefficient for modeling and analysis. As a result, we decided to select only features/predictors that would have a significant impact on the response variable. In our preliminary regression model, we chose the Boruta algorithm for this, so we could also continue using this mechanism for further analysis or use other methods such as interpreting p-values in order to achieve better results.

Model Building

After we select the final list of predictors that we would use, we would start on modeling the data. Since we will be comparing various models, we will use a train, validation and test split with ratios around 60%, 20% and 20% respectively, with the training data being used to enable the model to learn the feature weights, and the validation data being used to tune the hyperparameters.

In order to achieve the best results, we would need to explore several types of regression models aside from linear regression such as linear-log, log-linear, log-log, and polynomial. More advanced models like XGBRegression might even be considered if we really cannot achieve usable results in the end. In addition, to improve model accuracy and prevent overfitting, we will also conduct cross-validation on the model. We would explore a few different numbers of folds to find the best value.

Refinements and Iterations

As mentioned above, we would try out several types of models and different numbers of CV folds to optimize the model. In addition to that, we might also consider data transformations such as log transformations, min-max scaling, and square root transformations in order to keep the data points on a relative scale to one another. Other transformations such as one-hot encoding categorical variables, creating ratios of multiple numeric fields, and leveraging

dimensionality reduction techniques such as PCA or K-mean clustering to reduce high-dimensional data sources may also be used.

Also, for the model itself, we could also tune hyperparameters to increase accuracy. Tuning will mainly be done manually with careful attention being paid to prevent over or under fitting of the model. Depending on the time constraint of the project, automatic methods such as gridsearch and hyperopt could also be considered. Finally we would report out training and testing errors of the final model.

Literature Survey

- Our group gleaned their subject matter knowledge as to potential precursors of high blood pressure from a paper published in the Journal of the American College of Cardiology which looked at blood pressure holistically from prevention to management [6]. The paper cited that obesity may be responsible for 40% of all hypertension, which would be a staggeringly predictive feature in our model. The paper also often hinted at varying levels of correlation between these factors and hypertension by sex, indicating gender may play a role in a subject's predisposition. Other factors such as fitness levels, diet, alcohol usage, and harder to gather data about gut microbiome were all also cited to be correlated to elevated levels of blood pressure.
- A research paper published by the Institute of Electrical and Electronics Engineers[5], used artificial neural nets to try to predict systolic blood pressure. Although it predicted a continuous variable, systolic blood pressure is a factor in determining our goal of predicting a categorically elevated blood pressure, as such we sought to glean insights as to which features they found most predictive along with what research they did prior. The paper used features such as age, sex, height, weight, exercise, and alcohol to predict their target variable. variables we also found to be present within our dataset. They had the best results when using BMI, Age, Exercise, and Stress to predict on the 20% of data they withheld. The results of their model had better performance on men than women, again leading us to believe these two genders may have different behaviors in relation to predictive factors for hypertension. Overall the paper was successful in emboldening our decision to use our dataset for the purposes of predicting blood pressure, as this research team was able to do something similar using overlapping signals.

References

- 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