Identifying Key Factors Contributing to Stage II Hypertension: Insights and Implications

Team Name: AGL
Lingqi Huang(hlingqi@umich.edu)
Liangqi Tang(tanglq@umich.edu)

April 2024

1 Introduction

High blood pressure, also known as hypertension, is a prevalent condition affecting the body's arteries. In 2021, hypertension was cited as a primary or contributing cause in 691,095 deaths in the United States.[4] When an individual has high blood pressure, the force of blood pushing against the artery walls is consistently elevated. This places additional strain on the heart, requiring it to work harder to pump blood throughout the body. If left untreated, high blood pressure significantly increases the risk of serious health complications, including heart attacks, strokes, and other cardiovascular issues.[1] According to guidelines established by the American College of Cardiology and the American Heart Association, individuals with a systolic blood pressure (the top number) **greater than 140 mm Hg** or a diastolic blood pressure (the bottom number) **greater than 90 mm Hg** are classified as having stage II hypertension.[1]

In this short paper, we aim to investigate the potential contributions of common variables such as age, gender, and others to the development of **stage II hypertension**. Additionally, we endeavor to utilize existing variables to assess whether an individual may be at risk of stage II hypertension. To achieve this, we analyze data from the National Health and Nutrition Examination Survey (NHANES) spanning from 2017 to March 2020[5], comprising 4,398 observations. The variables under scrutiny include RIAGENDR (gender), RIDAGEYR (age in years), DSD010 (any dietary supplement taken), DSDCOUNT (number of dietary supplements taken), BMXBMI (body mass index), ALQ130 (average number of alcohol use in the past 12 months), SMQ020 (if the individual smoked at least 100 cigarettes in their lifetime), and INDFMMPD (family monthly poverty level index). By leveraging these data and variables, we aim to address the following question:

• What associations exist between our variables and the numerical values of systolic and diastolic blood pressure, and how can we leverage statistical learning methods to make inferences from these associations?

• Which variables among those considered are the most influential in classifying stage II hypertension?

• How can we apply statistical learning methods such as Support Vector Machines (SVM) and XGBoost to classify individuals with known covariates into either stage II hypertension or non-stage II hypertension? What is the predictive performance of these classification models? Additionally, do the existing variables provide sufficient strength and information to effectively classify individuals into these two categories?

2 Exploratory Data Analysis

We initially divided our dataset into two groups based on whether individuals met the criteria for stage 2 hypertension according to the American College of Cardiology and the American Heart Association[1]. These criteria define stage 2 hypertension as having a systolic blood pressure exceeding 140 mmHg or a diastolic blood pressure exceeding 90 mmHg. Out of 4398 individuals, 846 were classified as having stage 2 hypertension. To better understand the impact of various factors and the distribution differences between the two groups, we generated the following pairwise plot.
To gain a deeper understanding of the effects of various variables between the two groups, we performed **Welch's two-sample t-test**[2] for continuous variables (RIDAGEYR, DSDCOUNT, BMXBMI, ALQ130, INDFMMPI) and conducted **two sample z-test of proportion**[3] for categorical variables (RIAGENDR, DSD010, SMQ020).

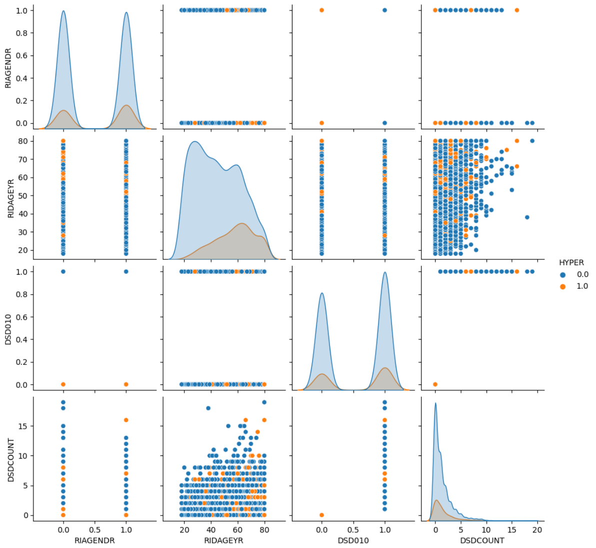

Figure 1: In the pairwise plot, the blue region represents individuals without stage II hypertension, while the yellow region represents those with stage II hypertension. The plot visualizes the relationship between gender (RIAGENDR), age (RIDAGEYR), whether individuals take any dietary supplements (DSD010), and the number of dietary supplements taken (DSDCOUNT).

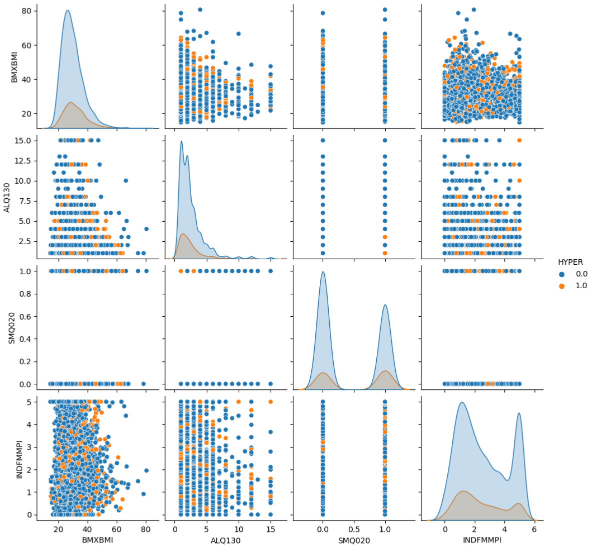

Figure 2: In the pairwise plot, the blue region indicates individuals without stage II hypertension, while the yellow region denotes those with stage II hypertension. This visualization explores the relationships between BMI (BMXBMI), number of alcoholic drinks consumed (ALQ130), smoking status (SMQ020), and the poverty index (INDFMMPI).

Variable	test statistic	P-value
RIDAGEYR	-22.67	< 0.001
DSDCOUNT	-1.684	0.092
BMXBMI	-6.39	< 0.001
ALQ130	0.499	0.617
INDFMMPI	2.151	0.032

Table 1: Welch's Two Sample t-test for Continuous Variable

Based on the test results for both continuous and categorical variables between the two groups, we have identified potential determinants of current stage II hypertension, including RIDAGEYR, BMXBMI, INDFMMPI, RIAGENDR, DSDCOUNT, and SMQ020.(See Table 2 and Table 3)We will now proceed to the inference stage to further explore the relationships between these variables and the two groups.

3 Regression and Inference

After conducting the exploratory data analysis, our next step is to employ regression methods to investigate the impact of various predictors on hypertension. Our primary focus lies in determining the direction and significance of each predictor's influence on hypertension. We aim to discern whether the influence exerted by each predictor is statistically significant. Additionally, we seek to perform inference on the regression results, enabling us to draw meaningful conclusions from our analyses. Moreover, we aspire to develop accurate predictions regarding both the hypertension value and the likelihood of an observation being hypertensive.

We conducted separate regression analyses on BPXOSY (Systolic Pressure) and BPXODI (Diastolic Pressure). Our analysis encompassed various regression methodologies, including tree methods, linear/polynomial regression, and GAM (Generalized Addictive Model). To ensure robustness, we partitioned the cleaned dataset into training (80%) and testing (20%) subsets, stratified by HYPER. This stratification allowed tree methods to optimize parameters based on minimizing MSE (Mean Squared Error) on the test dataset, facilitating a comparative assessment of fitting

Variable	test statistic	P-value
RIAGENDR	-3.383	< 0.001
DSDCOUNT	-1.625	0.104
SMQ020	-5.72	< 0.001

Table 2: Two Sample Z-test of Proportion for Categorical Variables

Variable	Mean(Hyper=0)	Mean(Hyper=1)	Std(Hyper= 0)	Std(Hyper=1)
RIDAGEYR	44.61	57.89	17.21	14.81
DSDCOUNT	1.35	1.49	2.01	2.13
BMXBMI	29.63	31.51	7.49	7.69
ALQ130	2.50	2.46	2.02	2.08
INDFMMPI	2.54	2.41	1.61	1.53

Table 3: Basic Statistics for Two Groups.

results. Additionally, in polynomial regression, we utilized cross-validation techniques to determine the optimal degree of polynomial for model fitting.

3.1 Tree Methods

In our exploration of tree methods, we experimented with Single Tree, Random Forests, and Boosting regression techniques. These approaches provided us with a diverse array of methodologies to analyze and interpret the data.

• Single Tree: Using minimum MSE on test dataset, we choose max_leaf_nodes of BPXOSY and BPXODI to be 7 and 9 respectively.

• Random Forest: Using minimum MSE on test dataset, we choose max_depth of BPXOSY and BPXODI to be 8 and 8 respectively with number of tree to be 1000 and max_feature to be sqrt.

• Boosting: Based on the staged_predict plot, we choose max_iter of BPXOSY and BPXODI to be 1414 and 353 respectively with learning rate = 0.01 and max_depth = 2.

The R^2 scores and RSS of the three regression tree methods are as follows:

Methods	BPXOSY	BPXODI
Single Tree	0.214	0.122
Random Forest	0.244	0.145
Boosting	0.246	0.136

Table 4: R^2 score(test) of Tree Methods on BPXOSY and BPXODI.

Methods	BPXOSY	BPXODI
Single Tree	247.97	116.48
Random Forest	246.78	111.54
Boosting	269.87	104.66

Table 5: MSE(test) of Tree Methods on BPXOSY and BPXODI.

In general, both Random Forest and Boosting methods demonstrate slight improvements over the Single Tree approach. However, the overall R^2 values are small, and the Mean Square Error(MSE) is large, indicating that tree methods are not particularly effective or suitable for this dataset. Further exploration revealed significant overfitting issues associated with tree methods, as evidenced by substantially larger testing losses compared to training losses.

The feature importance plot of Single Tree and Random Forest is as follows:

Upon examination of the feature importance plot, we observed that RIDAGEYR (age), BMXBMI (BMI), and RIAGENDR (gender) exhibit relatively significant influence on blood pressure. Notably, in the Random Forest method, due to the random restriction of the feature set at each split, the feature importance appears to be more evenly distributed across the variables, resulting in a more averaged representation of their impact on blood pressure.

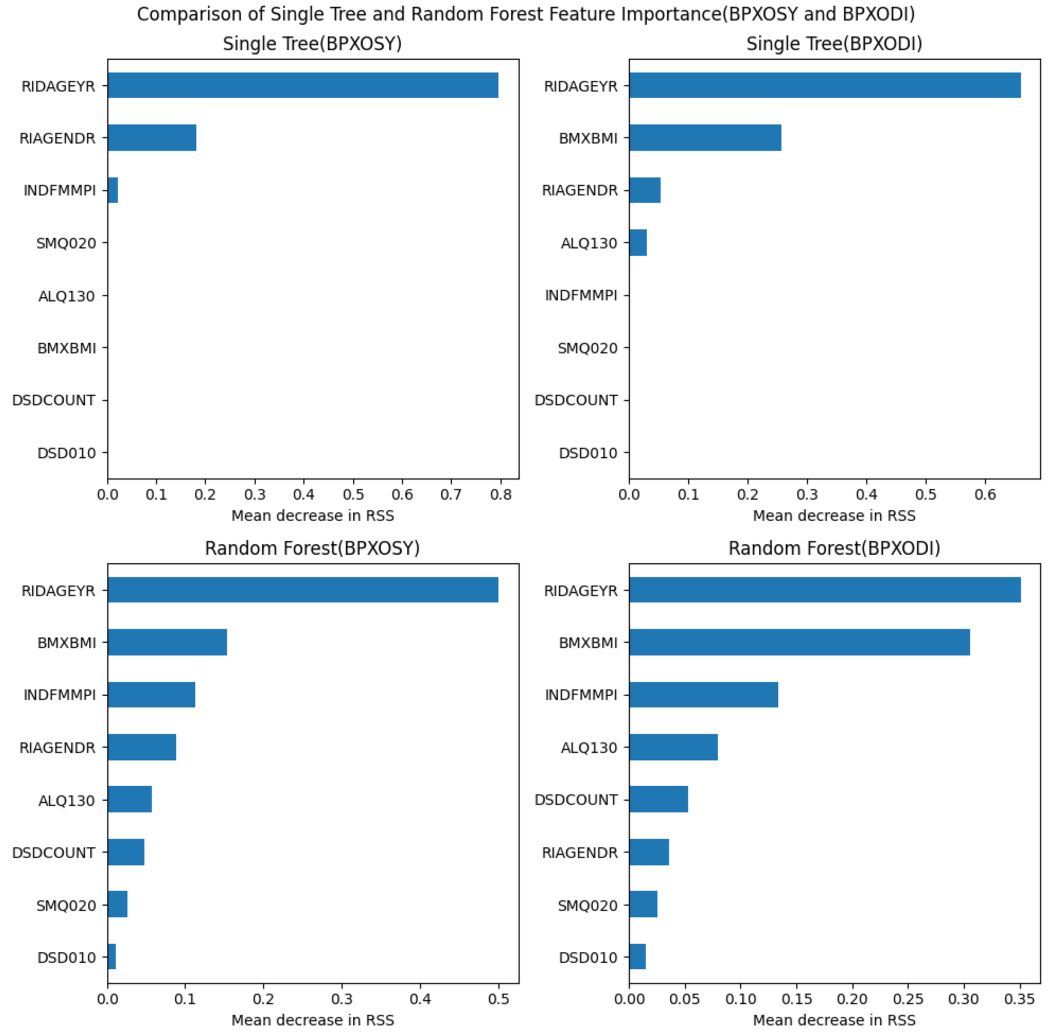

Figure 3: In the feature importance plot, the importance is measured by the mean decrease in RSS. Generally, RIDAGEYR(age), RIAGENDR(gender) and BMXBMI(BMI) seems to play a relatively important role in the regression

3.2 Linear/Polynomial Regression

We then move on to Linear/Polynomial Regression on this dataset as they are the simplest but very effective way to show how blood pressure are affected by these predictors and whether the influence is significant.

Firstly, we show the Linear Regression fitting results of BPXOSY here:

Dep. Variable:	BPXOSY	**R-squared (uncentered):**	0.965
Model:	OLS	**Adj. R-squared (uncentered):**	0.965
Method:	Least Squares	**F-statistic:**	1.199e+04
Date:	Sat, 20 Apr 2024	**Prob (F-statistic):**	0.00
Time:	17:33:05	**Log-Likelihood:**	-16061.
No. Observations:	3517	**AIC:**	3.214e+04
Df Residuals:	3509	**BIC:**	3.219e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P> |t|	[0.025	0.975]
RIAGENDR	9.9116	0.823	12.046	0.000	8.298	11.525
RIDAGEYR	0.9417	0.023	41.300	0.000	0.897	0.986
DSD010	2.9219	1.017	2.872	0.004	0.927	4.917
DSDCOUNT	-0.9001	0.253	-3.561	0.000	-1.396	-0.405
BMXBMI	1.8012	0.036	49.892	0.000	1.730	1.872
ALQ130	3.6797	0.205	17.911	0.000	3.277	4.082
SMQ020	0.2258	0.845	0.267	0.789	-1.431	1.882
INDFMMPI	2.5694	0.247	10.391	0.000	2.085	3.054

Omnibus:	36.513	**Durbin-Watson:**	2.002
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	39.145
Skew:	-0.218	**Prob(JB):**	3.16e-09
Kurtosis:	3.278	**Cond. No.**	153.

Notes:
[1] $\mathbf{R^2}$ is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

It's apparent that all coefficients are statistically significant except for SMQ020 (smoking). Despite achieving a high R^2 score of 0.965 on the training data, the model exhibits a clear indication of overfitting when assessed against the test data. This overfitting hinders the extrapolation and generalization of the linear regression results, undermining their reliability in practical applications.

Utilizing cross-validation, we determined that the minimum MSE for polynomial regression occurred when the degree was set to 2 for both BPXOSY and BPXODI. The final results of the linear/polynomial regression are summarized as follows:

Methods	BPXOSY	BPXODI
Linear	0.226	0.073
Polynomial(degree = 2)	0.244	0.132

Table 6: R^2 score(test) of Linear/Polynomial Methods on BPXOSY and BPXODI.

Methods	BPXOSY	BPXODI
Linear	276.89	112.23
Polynomial(degree = 2)	270.53	105.02

Table 7: MSE(test) of Linear/Polynomial Methods on BPXOSY and BPXODI.

3.3 GAM(Spline and Lasso)

In our Generalized Additive Model (GAM), we designated all quantitative predictors as spline terms and all categorical predictors as factor terms. Additionally, we set the lasso tuning parameter to its default value of 0.6. Below is a summary of the fitting results:

Predictor	Methods	BPXOSY	BPXODI
RIAGENDR	f	***	***
RIDAGEYR	s	***	***
DSD010	f		
DSDCOUNT	s		
BMXBMI	s	.	***
ALQ130	s	**	**
SMQ020	f		
INDFMMPI	s	**	

Table 8: Predictor significance of GAM

Methods	BPXOSY	BPXODI
GAM	278.27	106.87

Table 9: MSE(test) of GAM on BPXOSY and BPXODI.

Methods	BPXOSY	BPXODI
GAM	0.222	0.117

Table 10: R^2 score(test) GAM on BPXOSY and BPXODI.

Notes:
[1] Significance codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
[2] Methods: ’f’: factor, ’s’: spline
[3] $\lambda = 0.6$

These findings align with the results obtained from the previous models: RIAGENDR (Gender) and RIDAGEYR (Age) exhibit significant effects on blood pressure. Notably, BMXBMI plays a more prominent role in BPXODI, while variables like DSD010 and DSDCOUNT demonstrate negligible importance, consistent with the feature importance plot observed in our tree methods analysis. Furthermore, SMQ020 emerges as insignificant, corroborating the observations made in the linear regression analysis.

We got below partial dependence plot:

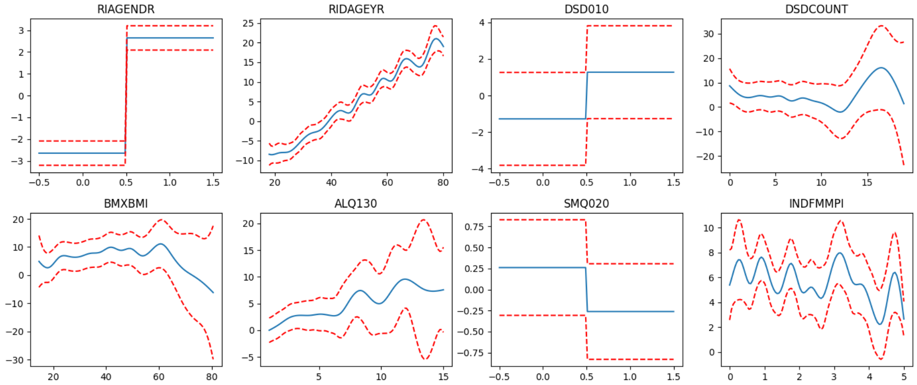

Figure 4: partial dependence plot for BPXOSY

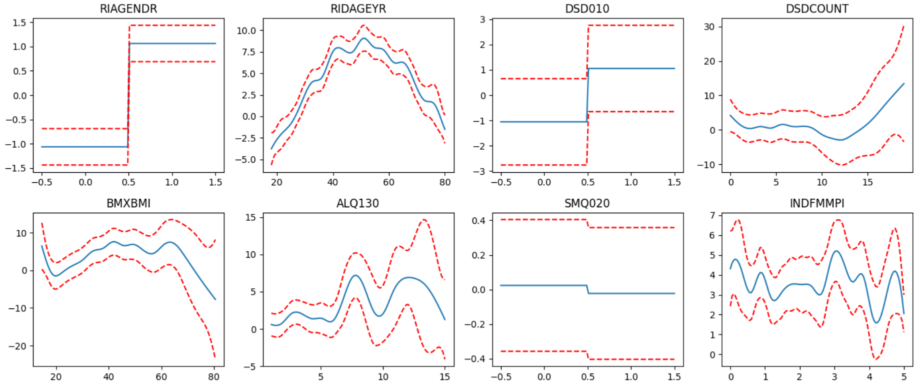

Figure 5: partial dependence plot for BPXODI

We can find there is something interesting in the partial dependence plot: Focused on RIDAGEYR(Age), it seems that BPXOSY increases as Age increases but BPXODI first increases and then decreases.

3.4 Conclusion

Our study reveals that gender and age have the most significant impact on blood pressure, with males exhibiting higher levels compared to females, and older individuals generally showing elevated readings. Interestingly, dietary supplement intake shows no clear influence on blood pressure levels, while Body Mass Index (BMI) and alcohol consumption appear to positively correlate with blood pressure, though the relationship with BMI is notably nuanced, exhibiting fluctuations at extreme values. Smoking does not emerge as a significant contributor to blood pressure variations in our analysis. The role of the poverty index in blood pressure regulation appears complex, possibly intertwined with other factors in the dataset.

Overall, while the regression methods demonstrate good fit on the training data, their performance on the test data is notably poorer. This discrepancy suggests a sign of overfitting, despite our efforts to mitigate it through weight decay methods such as Lasso in GAM. Consequently, it may not be prudent to draw definitive conclusions or inferences solely based on these results. Our next step involves exploring hypertensive classification to determine if we can achieve any improvements and uncover further insights from the data.

4 Classification and Prediction

We are now transitioning to the classification prediction phase. Although our regression predictions may yield relatively low R^2 scores, they still provide valuable insights into how our variables correlate with the numerical values of BPXOSY and BPXODI. In this section, we begin by employing the random forest algorithm with a maximum number of features set to 3 to determine the importance of each variable. Following this, we apply the XGBoost method and SVM to obtain prediction results, enabling us to assess whether individuals are at risk of developing stage 2 hypertension.

4.1 Random Forest and Importance of Variables

We initially generated a plot illustrating the importance of variables, revealing that three variables—RIDAGEYR, BMXBMI, and INDFMMPI—tend to explain variations in our classification problem more effectively. Conversely, categorical variables such as RIAGENDR, SMQ020, and DSD010 showed relatively little importance in our classification predictions. This result appears to contradict the findings of our regression analysis, where RIAGENDR and DSD010 emerged as statistically significant variables positively correlated with the numerical values of both BPXOSY and BPXODI. These results suggest that although these categorical variables may exhibit a positive relationship with BPXOSY and BPXODI, they may not carry sufficient strength to differentiate whether an individual will develop stage II hypertension or not.

RIDAGEYR	**BMXBMI**	**INDFMMPI**	**ALQ130**	**DSDCOUNT**	**RIAGENDR**	**SMQ020**	**DSD010**
0.282	0.262	0.234	0.089	0.073	0.023	0.0218	0.014

Table 11: Importance Table By Random Forest

4.2 XGboost Method and Diagnosis

We proceeded to apply the XGBoost method with specific parameter settings: a learning rate of 0.0072, a maximum depth of 2, and 18,000 estimators. These settings were determined by finding the lowest error rate on the test set through an iterative process, varying the learning rate within a sequence of length 30 from 0.0006 to 0.0015. As a result, we achieved an impressive accuracy score of approximately 0.802, indicating strong performance in our classification task.

Upon scrutinizing our prediction results, we observed a significant occurrence of False-Negative errors, also known as type-II errors, as indicated by the confusion matrix. Table 10 highlights a remarkably low type-I error rate of approximately 0.0057%, but alarmingly, the type-II error rate stands at about 95%. This discrepancy implies that we struggle to classify individuals not in stage II hypertension correctly, while we tend to misclassify those in stage II hypertension as not being in that stage. Given our objective, it becomes imperative to prioritize the reduction of type-II errors, enabling us to achieve more accurate predictions for individuals potentially in stage II hypertension.

		Truth		
		Non-Hyper II	Hyper II	Total
Predicted	Non-Hyper II	697	170	867
	Hyper II	4	9	13
	Total	701	179	880

Table 12: Confusion Matrix for XGboost

4.3 SVM Using Radial Kernel and Diagnosis

In our efforts to reduce type II error, we implemented support vector machine (SVM) using a radial kernel with $C = 200$ and $\gamma = 0.0295$, the latter selected from a search range of 0 to 0.05. While SVM typically yields accuracy scores ranging from 0.7 to 0.75 across various combinations of C and γ, it's important to note that achieving higher accuracy isn't our primary objective. Our focus lies in minimizing type II error. Remarkably, with our chosen settings, we were able to reduce the type II error to 72.6%, while maintaining a relatively low type I error.

But even though with lower type II error, our final accuracy score is about 0.745 that with higher type I error that is 13.4%, we can see the below confusion matrix.

		Truth		
		Non-Hyper II	Hyper II	Total
Predicted	Non-Hyper II	607	130	737
	Hyper II	94	49	143
	Total	701	179	880

Table 13: Confusion Matrix for SVM

4.4 Conclusion

Despite achieving relatively good prediction with both classification methods, the persistently high rate of type II errors hampers our ability to accurately identify individuals with stage II hypertension. This situation raises concerns about the potential for erroneous conclusions, as some individuals may be misclassified. Such limitations suggest the presence of unaccounted-for variables, prompting the need for more advanced methods to improve classification accuracy.

5 Conclusion and Limitations

In the regression analysis, a significant limitation manifests in the challenge of achieving a high R^2 score on the test dataset, coupled with substantial disparities between the performance of models on the training and testing datasets. This phenomenon, indicative of overfitting, undermines the generalizability of findings and the predictive capacity of the models. Consequently, while insights regarding the influence of individual predictors on blood pressure can be gleaned, the ability to extrapolate these findings to make precise predictions regarding BPXOSY/BPXODI values is notably constrained.

After examining the feature importance plot, we discovered that the classification of stage II hypertension is primarily influenced by age, body mass index (BMI), and the family monthly poverty level index. Surprisingly, variables we initially deemed significant in regression analysis, such as alcohol use and gender, play a minor role in classification. This discrepancy suggests that while these variables may be statistically significant in regression, they lack the strength to effectively contribute to hypertension stage classification. Consequently, it is possible that we overlooked other potentially important variables that could enhance our classification accuracy.

Despite achieving relatively high prediction accuracy with both SVM and XGBoost models, we've observed a troublingly high rate of type II errors. Essentially, our current prediction outcomes suggest that almost all individuals are non-stage II hypertensive, providing a seemed high 80% chance of accurate classification. This highlights the pressing need for further exploration of additional variables that could significantly enhance our classification accuracy and reduce the prevalence of type II errors, thereby leading to overall improved predictions.

Furthermore, our analysis has underscored an important consideration: when striving to identify the best model with the lowest predicted accuracy, algorithms may prioritize minimizing overall error rather than specifically addressing type I and type II errors. However, it is crucial to exercise control over these error types to derive more meaningful conclusions from our analyses.

During the data cleaning process, we encountered numerous missing values (NAs) in our columns. Our approach thus far has been to remove all rows containing NAs, resulting in the loss of approximately 60% of the overall dataset, which originally comprised around 15,000 entries. This significant reduction in sample size may compromise the performance of our prediction methods.

Moving forward, we recognize the importance of implementing more careful missing value handling techniques. For instance, one strategy could involve using existing data to predict and impute missing values, thereby retaining more data and potentially improving the efficacy of our study's predictive models.

6 Group Member Contribution

Lingqi Huang performed some parts of exploratory data analysis, implemented classification methods, and provided interpretations for classification parts.

Liangqi Tang implemented regression methods and provided interpretations for regression parts.

7 Reference

[1] https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/symptoms-causes/syc-20373410

[2] https://en.wikipedia.org/wiki/Welch%27s_t-test

[3] https://sixsigmastudyguide.com/two-sample-test-of-proportions/#:~:text=Two%20sample%20Z%20test%20of%20proportions%20is%20the%20test%20to,that%20have%20some%20single%20characteristic.

[4] https://www.cdc.gov/bloodpressure/facts.htm#:~:text=In%202021%2C%20hypertension%20was%20a,deaths%20in%20the%20United%20States.&text=Nearly%20half%20of%20adults%20have,are%20taking%20medication%20for%20hypertension.

[5] https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2017-2020

[6] https://xgboost.readthedocs.io/en/stable/

[7] *An Introduction to Statistical Learning: with Applications in Python (Springer Texts in Statistics) 1st ed. 2023 Edition*, by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor.