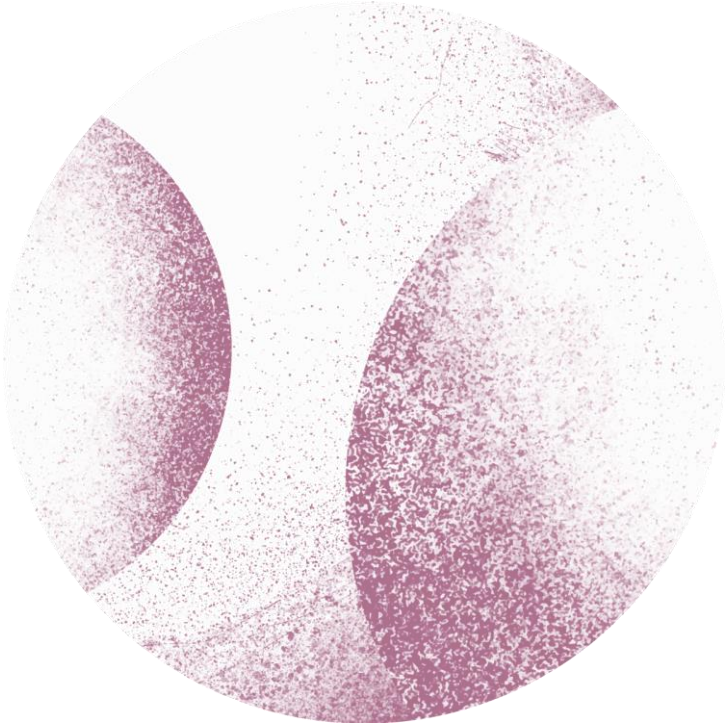
Predicting Hypertension Using Environmental and Heritable Risk Factors



Outline

01

Introduction

02

Data Description & Visualization

03

Model & Empirical Analysis

Forward Stepwise Regression

Logistic Regression

K-Nearest Neighbors

Random Forest

04

Conclusion

01

Introduction

Overview of Risk Factors Associated with Hypertension

What is Hypertension?

An individual is diagnosed with hypertension (high blood pressure) if their

Systolic Blood Pressure ≥ 140 mmHg

OR

Diastolic Blood Pressure ≥ 90 mmHg

Known Risk Factors

- Age
- Alcohol Consumption
- Race
- Gender
- Sodium Intake
- Family History
- Response to Stress

Hypothesis to Test: if heritable factors are more important to predicting hypertension than environmental factors

Interacting Risk Factors

“Men are three times more frequent drinkers and intake about 80% more ethanol than women. These sex-differences tend to decrease among the younger.”

— Teixeira et al., 2018, on ELSA

02

Data Description & Visualization

Exploring Patterns in Hypertension

NHANES DATASET

01

ABOUT THE DATA

Sample: ~5000 people each year
Interviews & Physical Exams Medical, Laboratory, Questionnaire

02

Pre-Processing

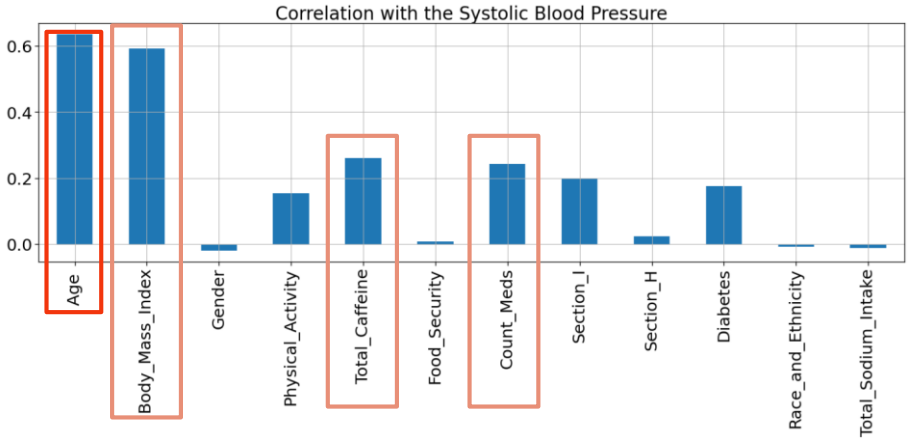
Dropped nulls, duplicates, datatype
Set target variable and defined thresholds

Correlation of Variables

Data Visualization

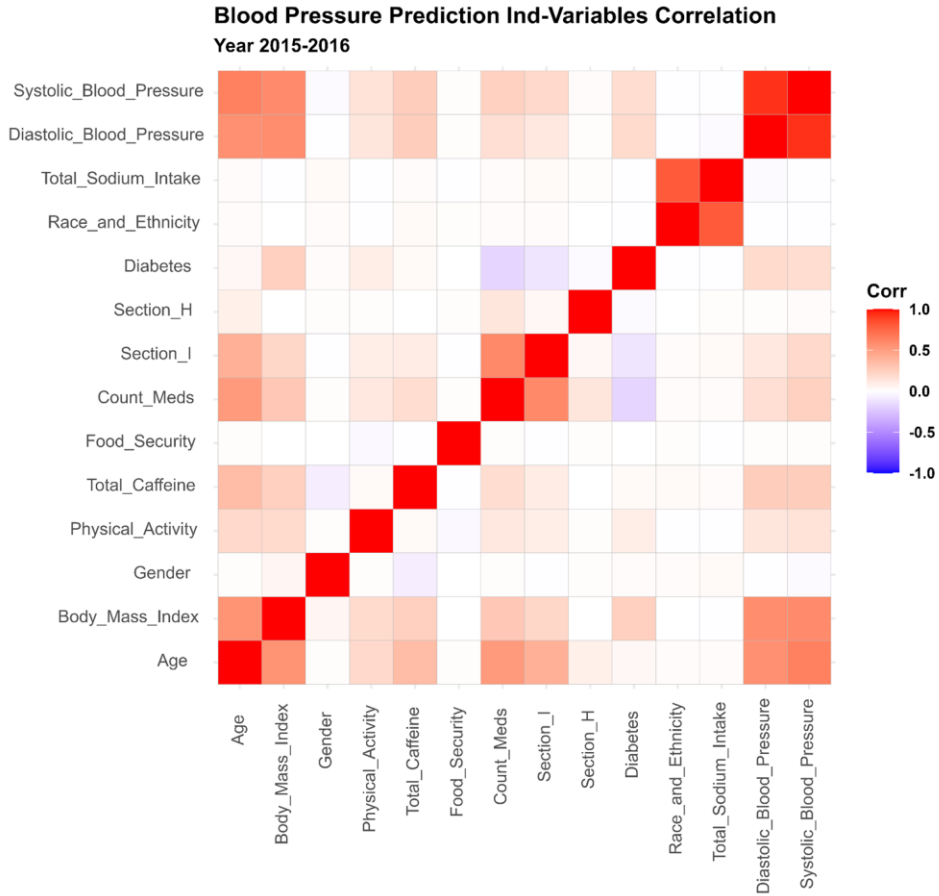
Diastolic | Systolic Blood Pressure <- Predictor

[Dark Red] Age is highly correlated with other predictors and predicted variables



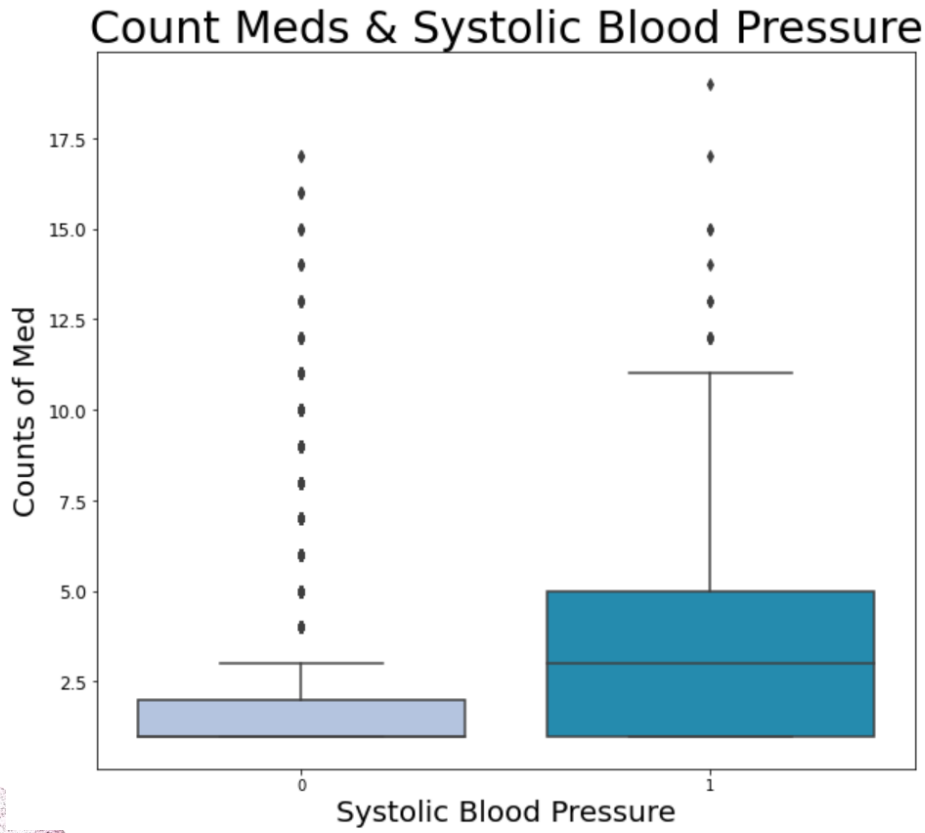
[Light Red] Relatively significant for the prediction

- sns.corr() — bar plot with selected independent-variables and the correlation with target



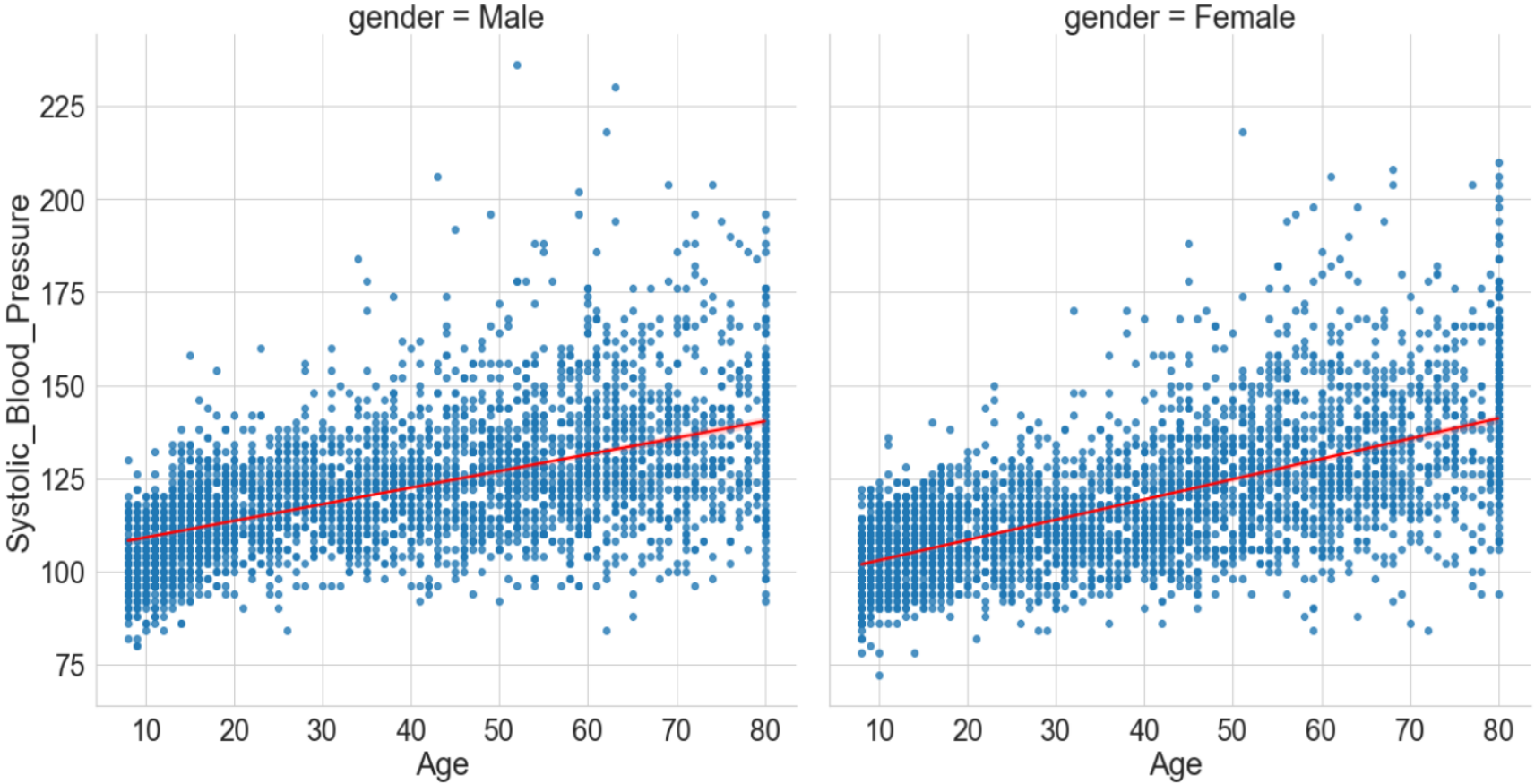
Sample Selected Variables

Data Visualization



Non-hypertension **hypertension**

Systolic Blood Pressure Cross Gender & Age



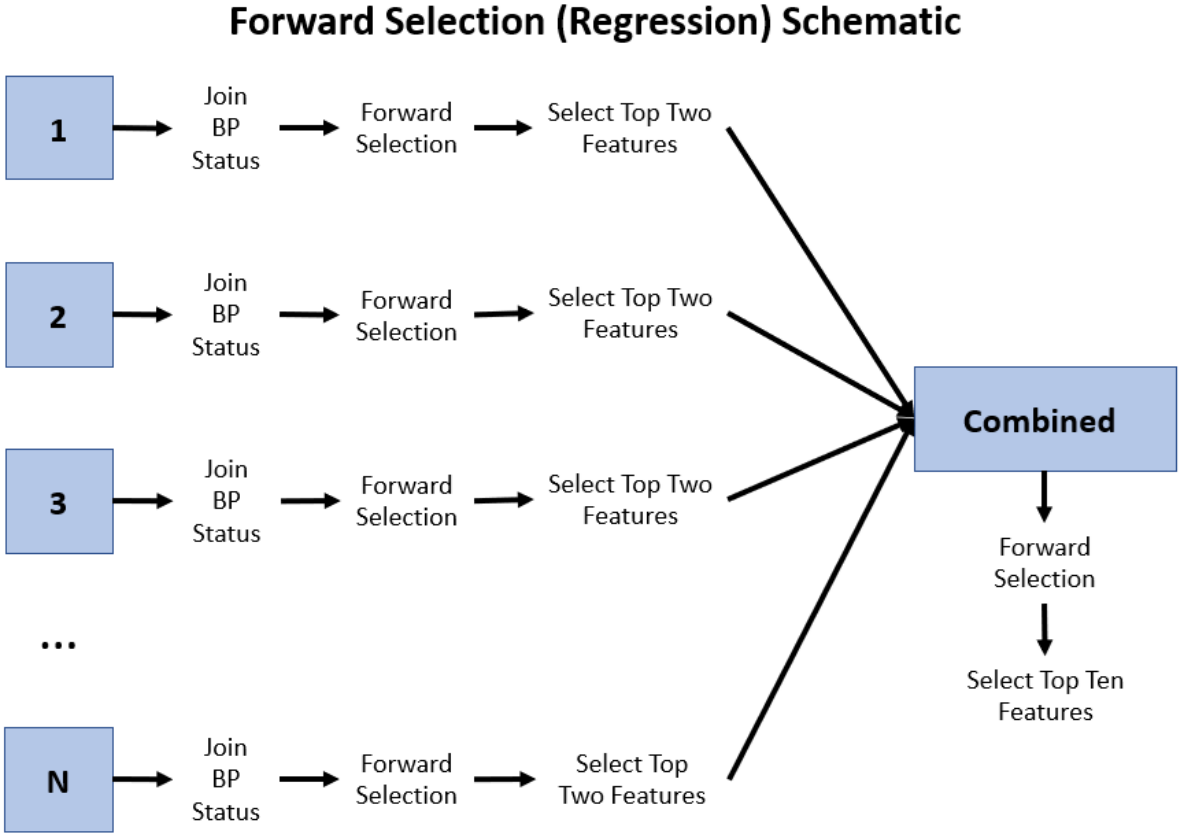
As people get older, they are more likely to develop hypertension

03

Model & Empirical Analysis

Predicting Hypertension

Forward Stepwise Regression



Forward Stepwise Regression Results

- Age
- Body Mass Index
- Gender
- Hours use computer past 30 days
- Caffeine intake
- Food stamp benefits
- Number of medication
- Cardiac medication
- Other medication (Eye/Ear)
- Diabetes

Logistic Regression

Predictors

- Age
- Body Mass Index
- Gender
- Hours use computer past 30 days
- Caffeine intake
- Food stamp benefits
- Number of medication
- Cardiac medication
- Other medication (Eye/Ear)
- Diabetes

Response

- Binary variable
- Hypertension - class 1
- Non-hypertension - class 0

Logistic Regression

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -5.1206017  0.4075259 -12.565  < 2e-16 ***
ridageyr        0.0623804  0.0029335  21.265  < 2e-16 ***
bmxbmi          0.0331222  0.0068914   4.806 1.54e-06 ***
riagendr       -0.1900635  0.0939234  -2.024 0.043011 *
paq715          0.0293872  0.0103216   2.847 0.004411 **
total_caffeine -0.0001768  0.0002270  -0.779 0.435973
fsd855          0.0110118  0.0775248   0.142 0.887046
count_meds     -0.0818590  0.0231753  -3.532 0.000412 ***
section_I       0.4316776  0.1407836   3.066 0.002168 **
section_H      -0.1865131  0.4563866  -0.409 0.682779
diq010         -0.1397528  0.1087930  -1.285 0.198941
```

	Predicted - 0	Predicted - 1
Actual - 0	1576	53
Actual - 1	218	62

Accuracy: 0.858

Logistic Regression

Imbalanced Dataset

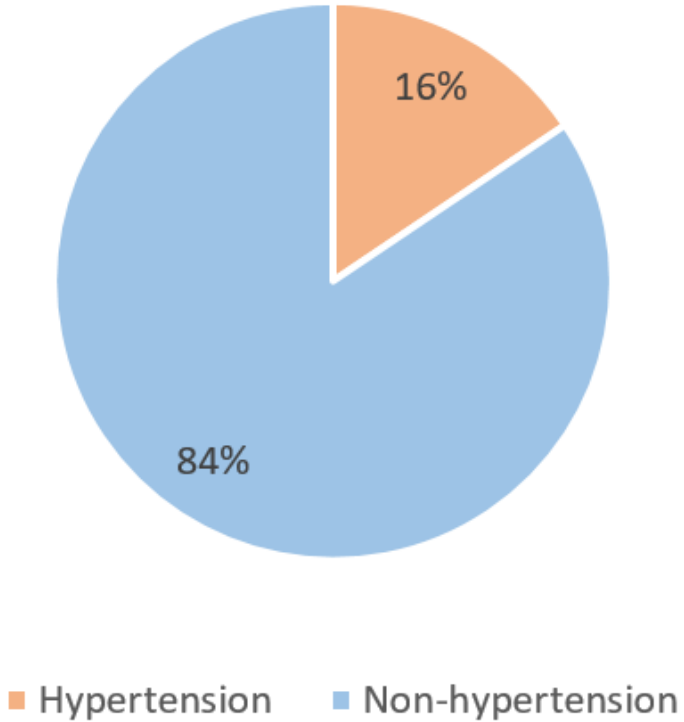
16% - hypertension
84% - non-hypertension

Problem

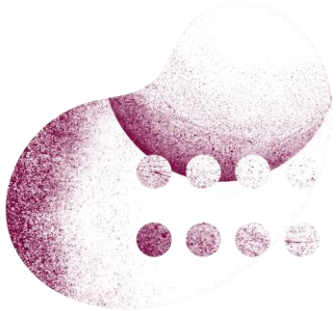
Logistic regression predicts most as non-hypertension

Solution

Resampling
Clustering



Balance the Dataset



Resampling

- Under-sample: only keep a percentage of non-hypertension data.
- Over-sample: duplicate hypertension data.

Clustering

- K-means clustering
- Cluster non-hypertension data
- Use cluster centroids to replace non-hypertension data

Logistic Regression

Balanced dataset - resample

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.798443   0.518446  -7.327 2.36e-13 ***
ridageyr     0.068721   0.004154  16.544  < 2e-16 ***
bmxbmi       0.047414   0.009598   4.940 7.82e-07 ***
riagendr    -0.343088   0.132078  -2.598 0.009387 **
paq715       0.046692   0.019828   2.355 0.018530 *
count_meds  -0.107838   0.032633  -3.305 0.000951 ***
section_I    0.516532   0.225390   2.292 0.021921 *
diq010      -0.287728   0.151548  -1.899 0.057619 .
```

	Predicted - 0	**Predicted - 1**
Actual - 0	255	79
Actual - 1	69	217

Accuracy: 0.7613

Logistic Regression

Balanced dataset - cluster

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.5759577	0.5662673	-4.549	5.39e-06	***
ridageyr	0.0777491	0.0045383	17.132	< 2e-16	***
bmxbmi	0.0323142	0.0093188	3.468	0.000525	***
riagendr	-0.7324972	0.1713260	-4.275	1.91e-05	***
total_caffeine	-0.0012378	0.0003467	-3.570	0.000357	***
count_meds	-0.1449069	0.0358661	-4.040	5.34e-05	***
section_I	0.8503215	0.2676950	3.176	0.001491	**
diq010	-0.3501939	0.1817966	-1.926	0.054067	.

	Predicted - 0	**Predicted - 1**
Actual - 0	208	99
Actual - 1	60	228

Accuracy: 0.7328

Logistic Regression

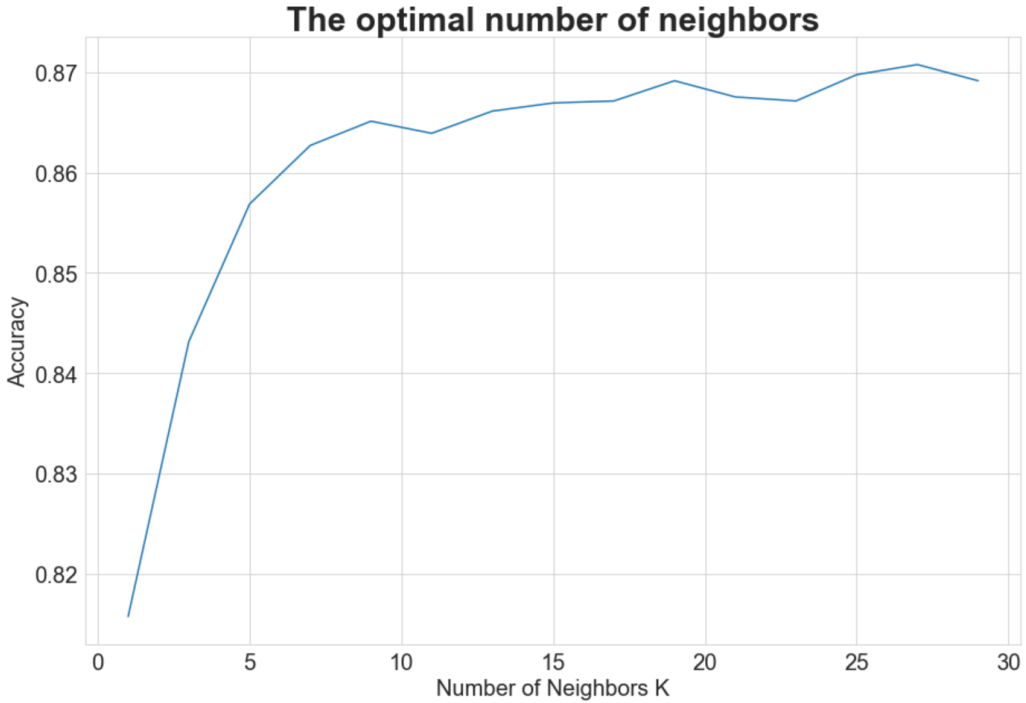
Always Significant Variables

- Age,
- Body Mass Index
- Gender

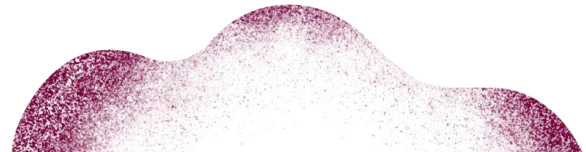
Hypertension Classification Accuracy

~ 75%

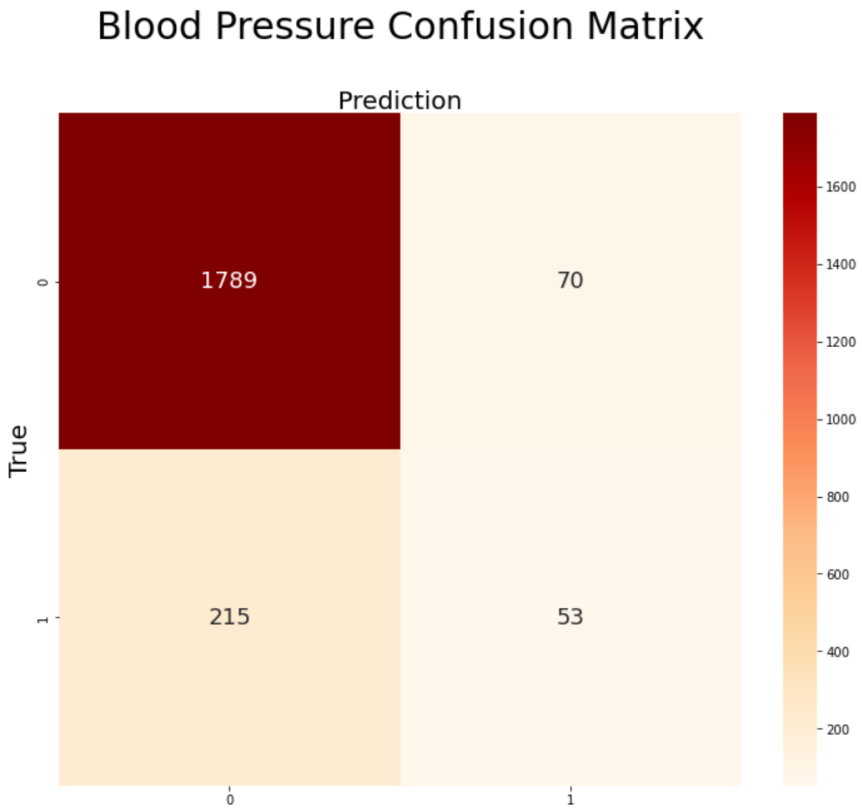
KNN: Optimal K

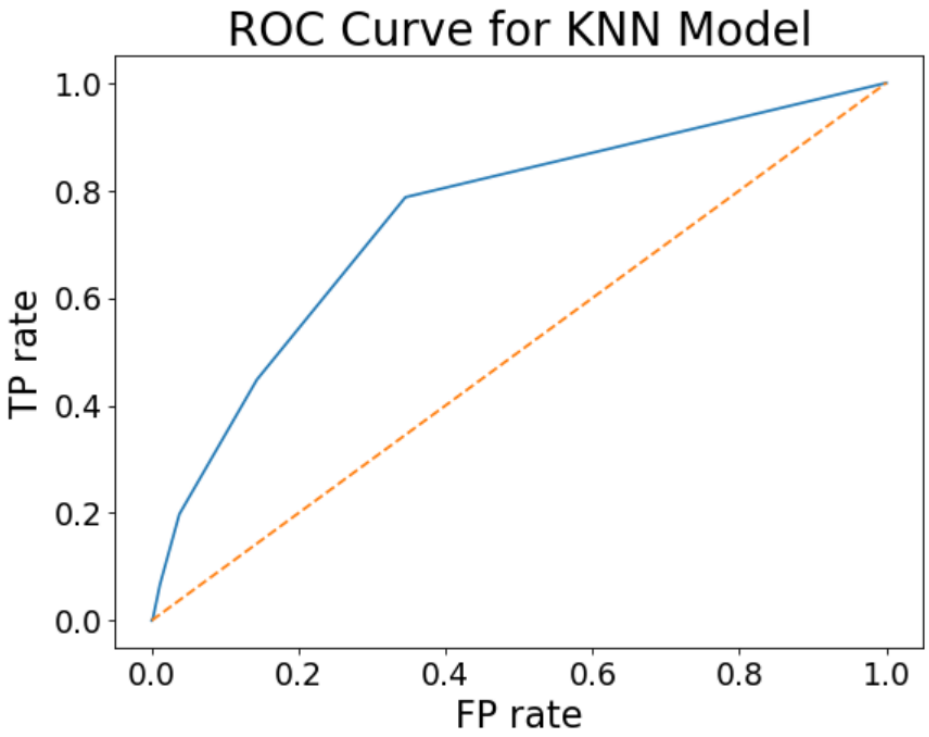


- Potential optimal values of K: {5, 7, 9}
- Risk of a large K:
  - The accuracy curve stays flat after that point
  - Classifying everything as non-hypertension
  - Reducing the model complicity and sacrificing the accuracy

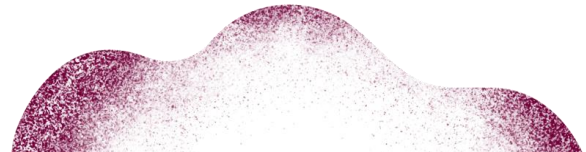


KNN with K = 5



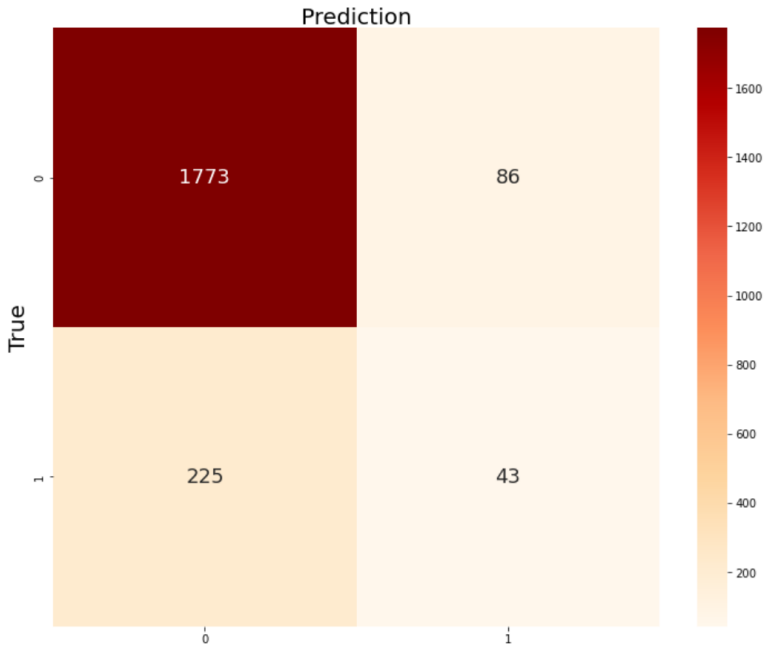


Model Accuracy <- 0.801
Roc_auc_score(y_test, y_pred) <- 0.739



KNN Using the Best Parameters

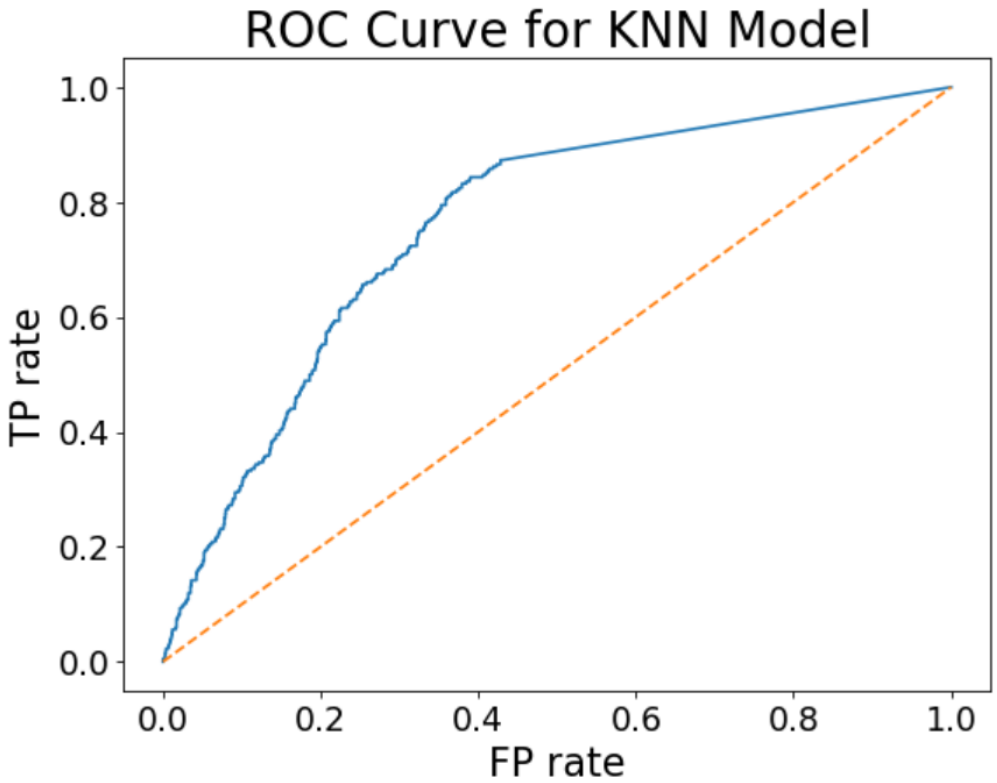
KNN Confusion Matrix with Best Parameter



Model Accuracy <- 0.853
Roc_auc_score(y_test, y_pred) <- 0.759

Parameter :

- **Weights** : uniform & distance
- **Neighbors**: range (1 ~~ 31)
- **power** : range (1~~ 6)



RANDOM FOREST

01 Parameters

n_estimators=100
Max_depth,min_sample_leaf,
max_features=Default

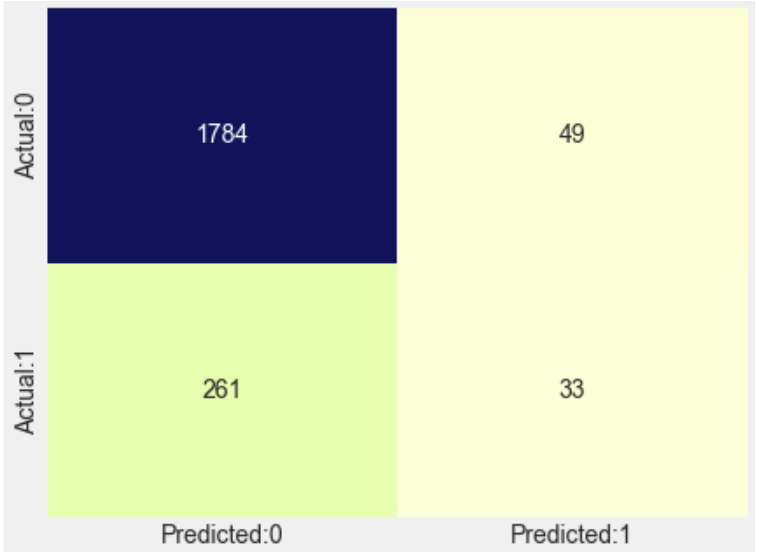
02 Results

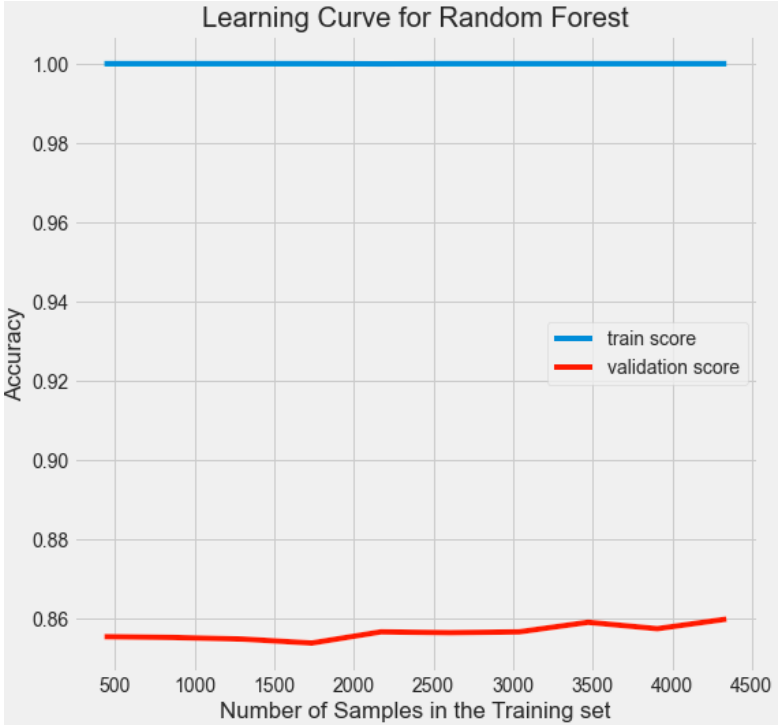
Accuracy:0.854
Precision:0.402
F1 Score:0.175
Recall:0.11

03 Recommendations

The model performed better without enforcing resampling of the training data

RANDOM FOREST





04

Conclusion

What we’ve learned about Hypertension

What We have Learned

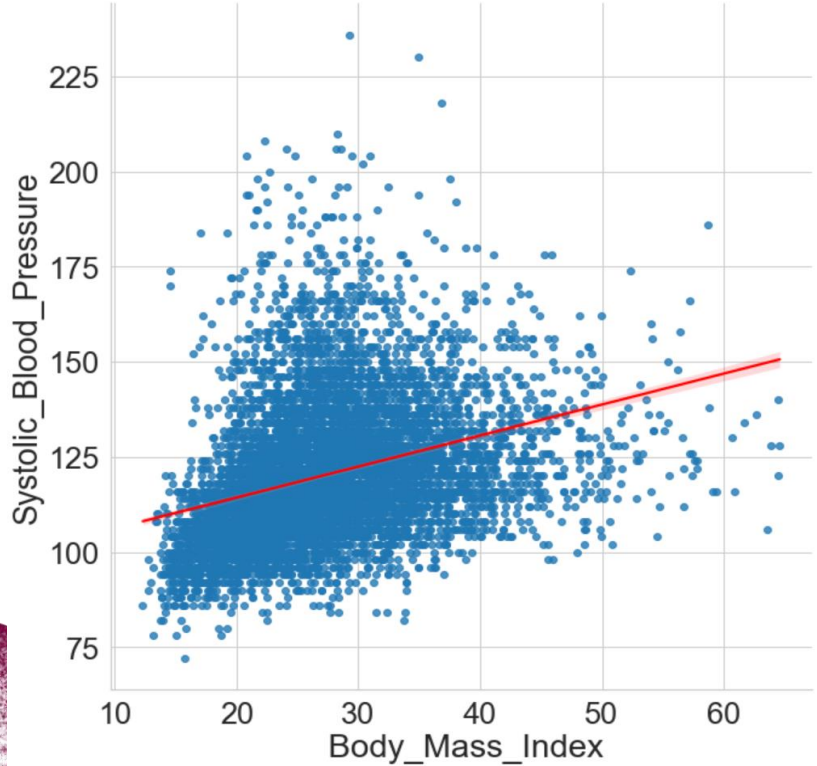
- Hypertension is challenging to model due to a large class imbalance
- The important risk factors for hypertension are challenging to identify
- Picking the right metric
  - The resampled logistic regression classifies the individuals with hypertension the most
  - The random forest has the highest accuracy
- Future Steps
  - Tune the random forest model
  - Try the same resampling methods for all models
- Perhaps more information would be helpful
  - Many of the selected predictors are not significant
- Hypothesis Result: Heritable factors are more significant predictors than environmental factors in our models

Appendix

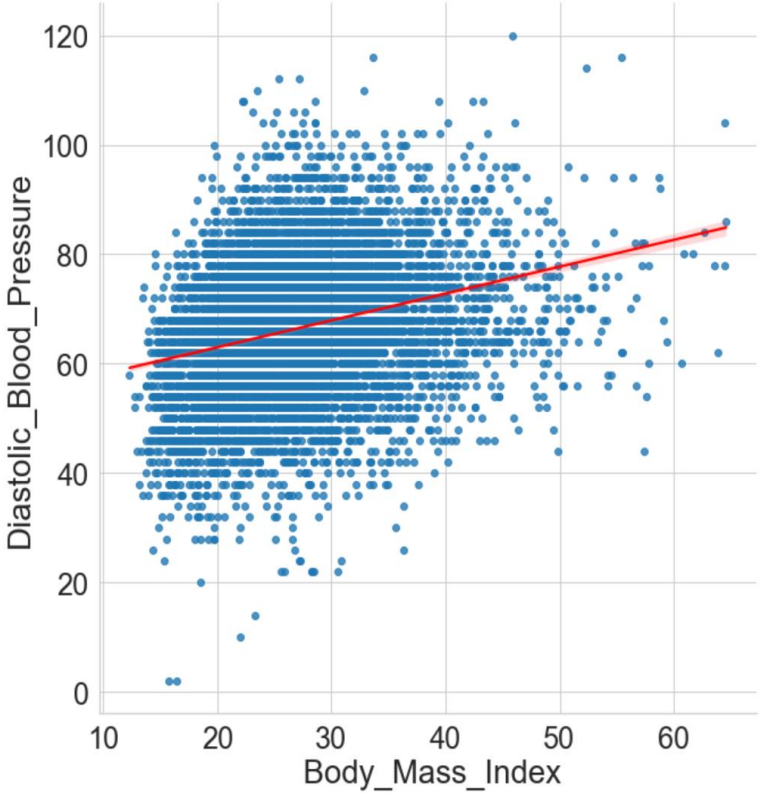
Sample Selected Variables

Data Visualization

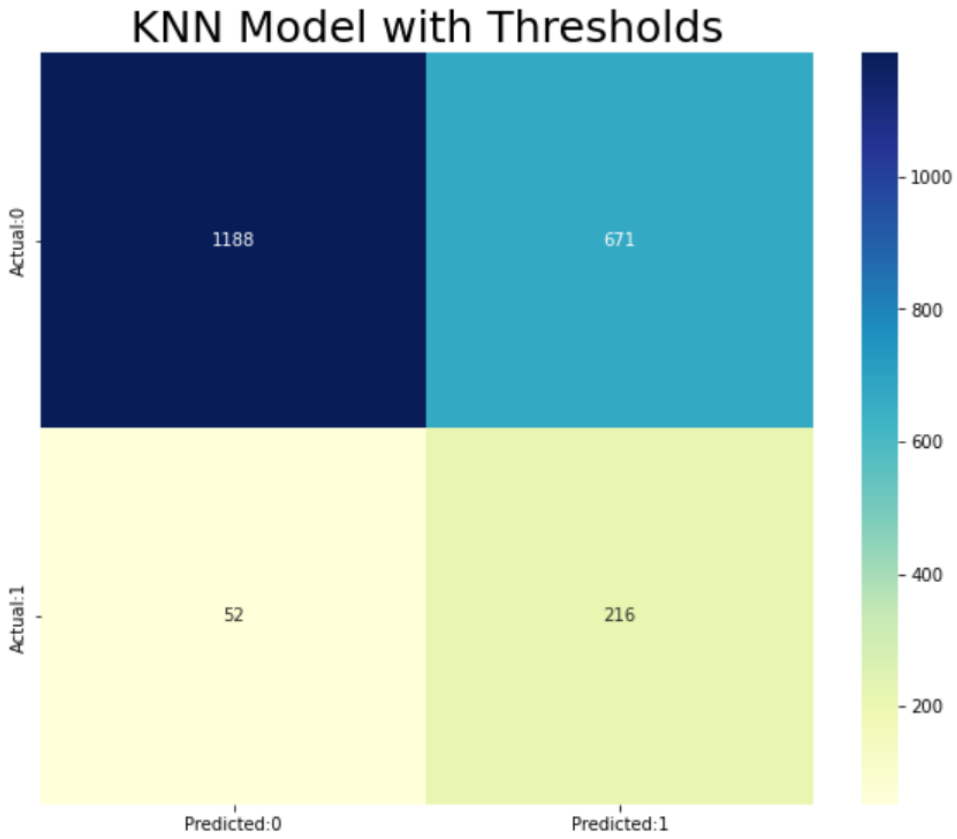
Systolic Blood Pressure with Body Mass Index



Diastolic Blood Pressure with Body Mass Index



KNN with Thresholds



Best threshold: 0.102

Gmeans: 0.719

	Thresholds	Gmeans
0	1.678571	0.000000
1	0.678571	0.000000
2	0.642857	0.061068
3	0.607143	0.086340
4	0.571429	0.172541
5	0.535714	0.250841
6	0.500000	0.264538
7	0.464286	0.341434
8	0.428571	0.388024
9	0.392857	0.467334
10	0.357143	0.546450

Model Accuracy <- 0.660

Roc_auc_score(y_test, y_pred) <- 0.759