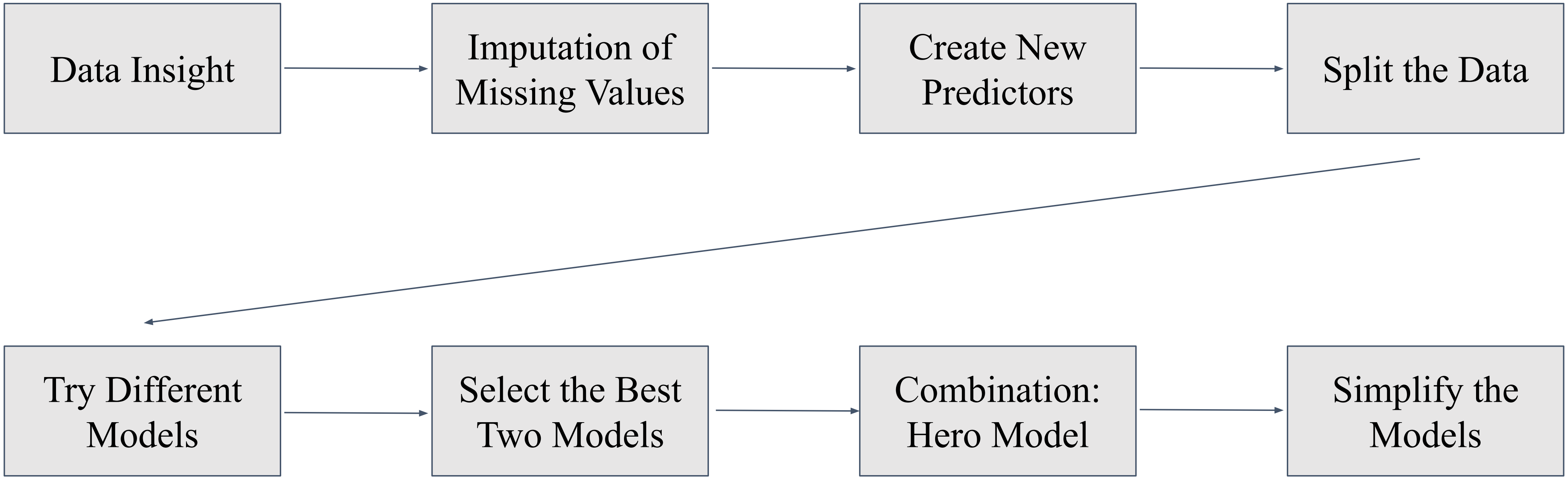
PREDICTING ALCOHOL STATUS USING INDIVIDUALS’ VITALS

Lecture 2, Group 9:
Chenyu Li (echolee0806@g.ucla.edu)
Xiangyuan Meng (alanmeng29@g.ucla.edu)
Jiarui Song (jiarui2002@g.ucla.edu)
Joanna Sun (jsun1101@g.ucla.edu)
Adriana Villavicencio (adrianavllc2@g.ucla.edu)

Agenda

- Introduction
  - Context
  - Data Description
    - Density Plots
    - Stacked Bar Charts
- Data Processing/Manipulation
  - Missing Values
  - Outliers
  - Create New Predictors
- Model Selection
  - Compare Testing Error
- Reduce Predictors
- Summary

Mind Map


Data Insight
Imputation of Missing Values
Create New Predictors
Split the Data
Try Different Models
Select the Best Two Models
Combination: Hero Model
Simplify the Models

Introduction: Context

Context

This project utilizes analytical techniques to predict an individual's **alcohol status** given their health record. These conclusions can be extended to guide healthcare professionals and policymakers into understanding which predictors give us an accurate alcohol status level.

Models

GLM	LDA	QDA
Random Forest	XGBoost	SVM

Predictors

ID, Sex, Age, Height, Weight, Waistline, Sight Left , Sight Right, Hear Left, Hear Right, SBP, DBP, BLDS, Total Cholesterol, HDL Cholesterol, LDL Cholesterol, Triglyceride, Hemoglobin, Urine Protein, Serum Creatinine, SGOT AST, SGOT ALT, Gamma GTP, BMI, BMI Category, Age Category, Smoking Status

Train-Test Split & Response Proportion

- Removed ID column
- **Used 20,000 observations from the training data as our 'unofficial' testing data**
- Calculated proportion of Alcoholic Status in training data

	Rows	**Columns**
X.train	50,000	26
X.test	20,000	26

Train	**N**	**Y**
Alcoholic.Status	50.16%	49.84%

Introduction: Data Description

Density Plots

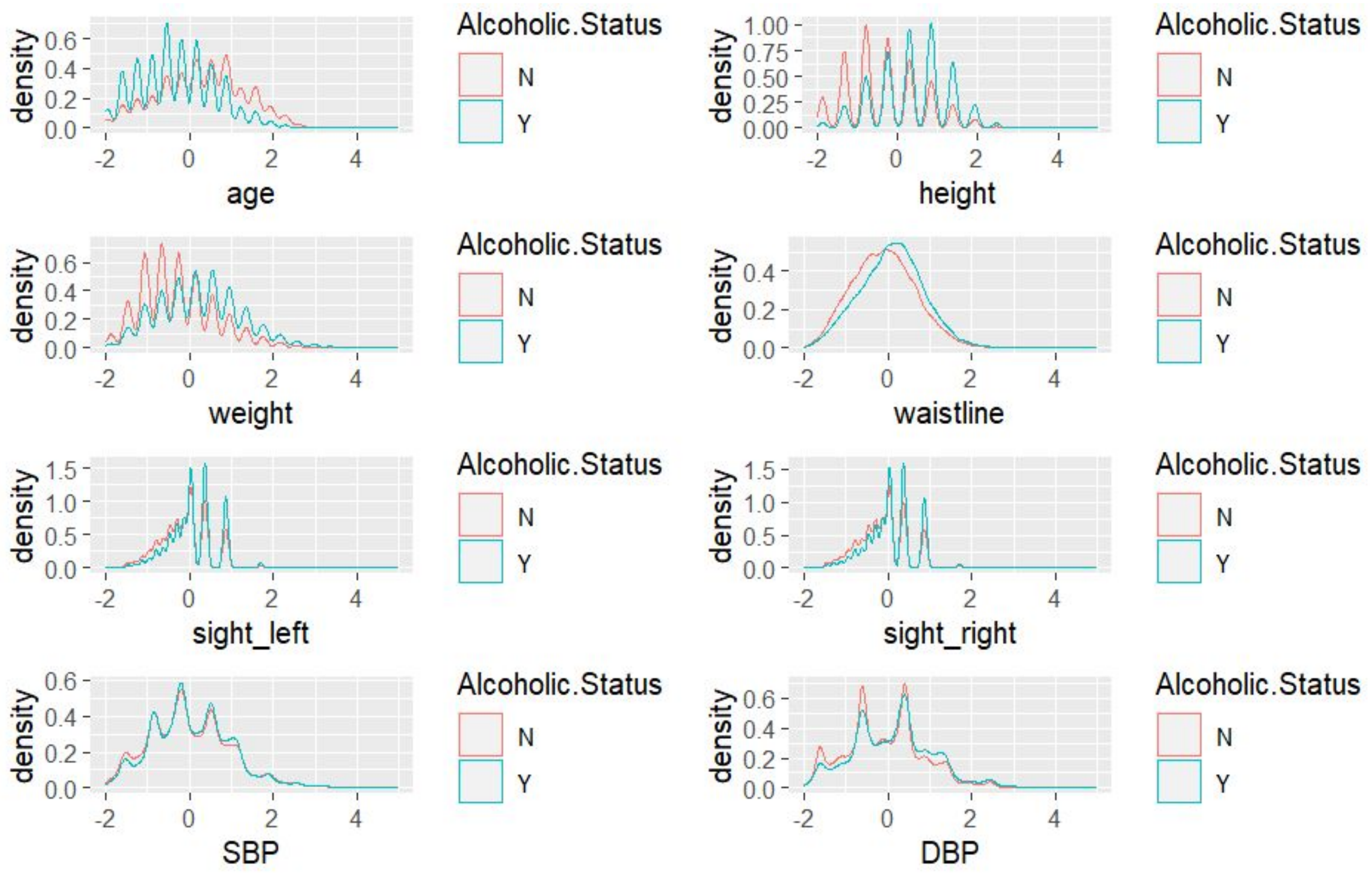
Plots show us which numerical predictors are best/worst for Alcoholic Status variable

Best:

- Age
- Height
- Weight

Worst:

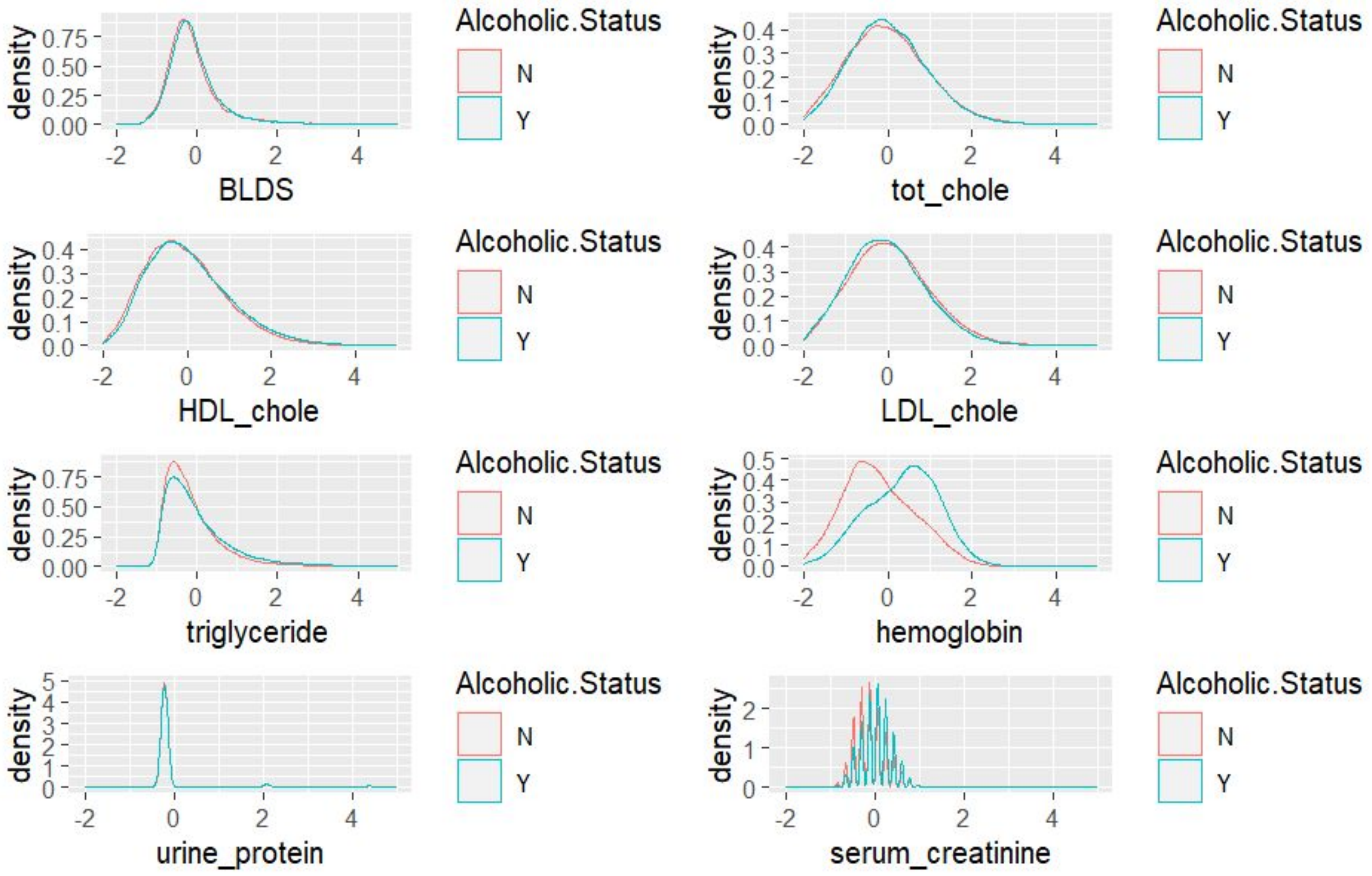
- Waistline
- Sight Left
- Sight Right
- SBP
- DBP



Introduction: Data Description

Best:
- Age
- Height
- Weight
- Hemoglobin

Worst:
- Waistline
- Sight Left
- Sight Right
- SBP
- DBP
- BLDS
- Total Cholesterol
- HDL Cholesterol
- LDL Cholesterol
- Triglyceride
- Urine Protein
- Serum Creatine

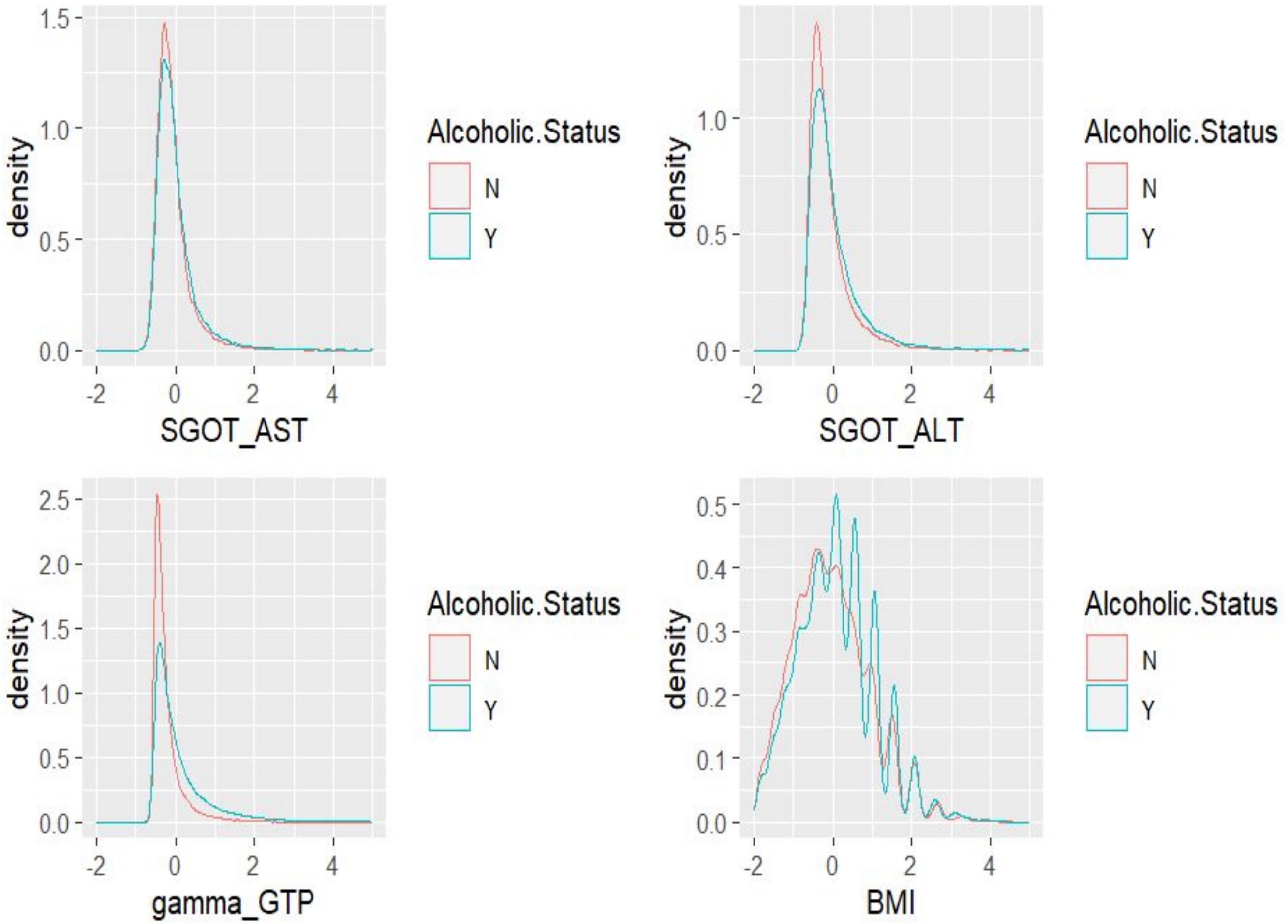


Introduction: Data Description

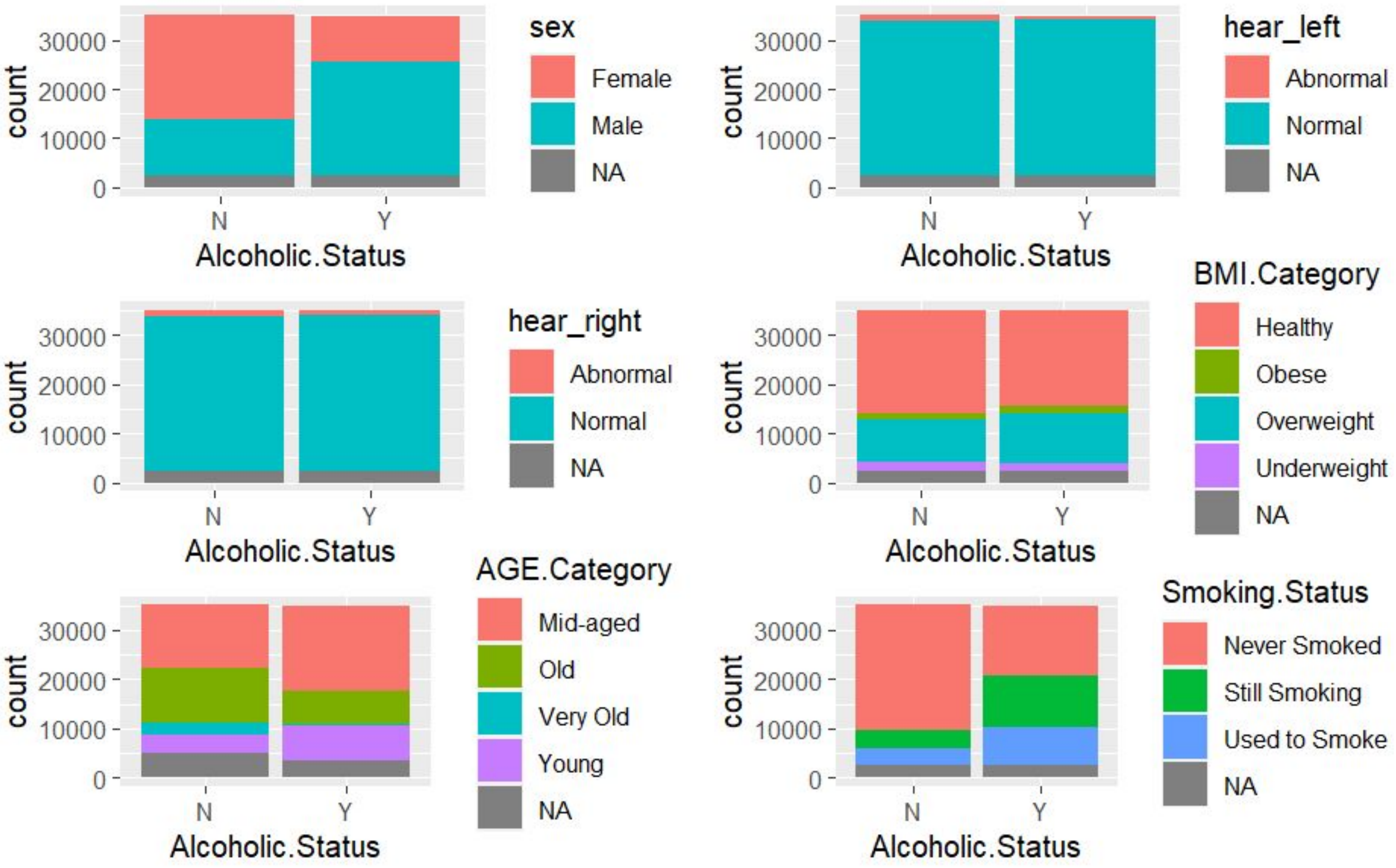
Best:
- Age
- Height
- Weight
- Hemoglobin

Worst:
- Waistline
- Sight Left
- Sight Right
- SBP
- DBP
- BLDS
- Total Cholesterol
- HDL Cholesterol
- LDL Cholesterol
- Triglyceride
- Urine Protein
- Serum Creatine

- SGOT AST
- SGOT ALT
- Gamma GTP
- BMI



Introduction: Data Description



Stacked Bar Charts

Plots show us which categorical predictors are best/worst for Alcoholic Status variable

Best:

- Sex
- Age Category
- Smoking Status

Worst:

- Hear Left
- Hear Right
- BMI Category

Agenda

- Introduction
  - Context
  - Data Description
    - Density Plots
    - Stacked Bar Charts
- Data Processing/Manipulation
  - Missing Values
  - Outliers
  - Create New Predictors
- Model Selection
  - Compare Testing Error
- Reduce Predictors
- Summary

Data Processing/Manipulation: Missing Values

Predictor	NA Percent
Sex	7.15%
Age	7.01%
Height	7.06%
Weight	7.07%
Waistline	7.10%
Sight Left	6.90%
Sight Right	7.04%
Hear Left	6.89%
Hear Right	6.93%
SBP	7.02%
DBP	7.01%
BLDS	6.94%
Total Cholesterol	7.03%

Predictor	NA Percent
HDL Cholesterol	6.92%
LDL Cholesterol	7.02%
Triglyceride	6.88%
Hemoglobin	7.03%
Urine Protein	6.95%
Serum Creatinine	6.93%
SGOT AST	6.97%
SGOT ALT	7.02%
Gamma GTP	7.02%
BMI	7.13%
BMI Category	7.00%
Age Category	11.83% ✱
Smoking Status	6.99%

Analysis

Most NA values reside in **Age Category** predictor (11.83%)

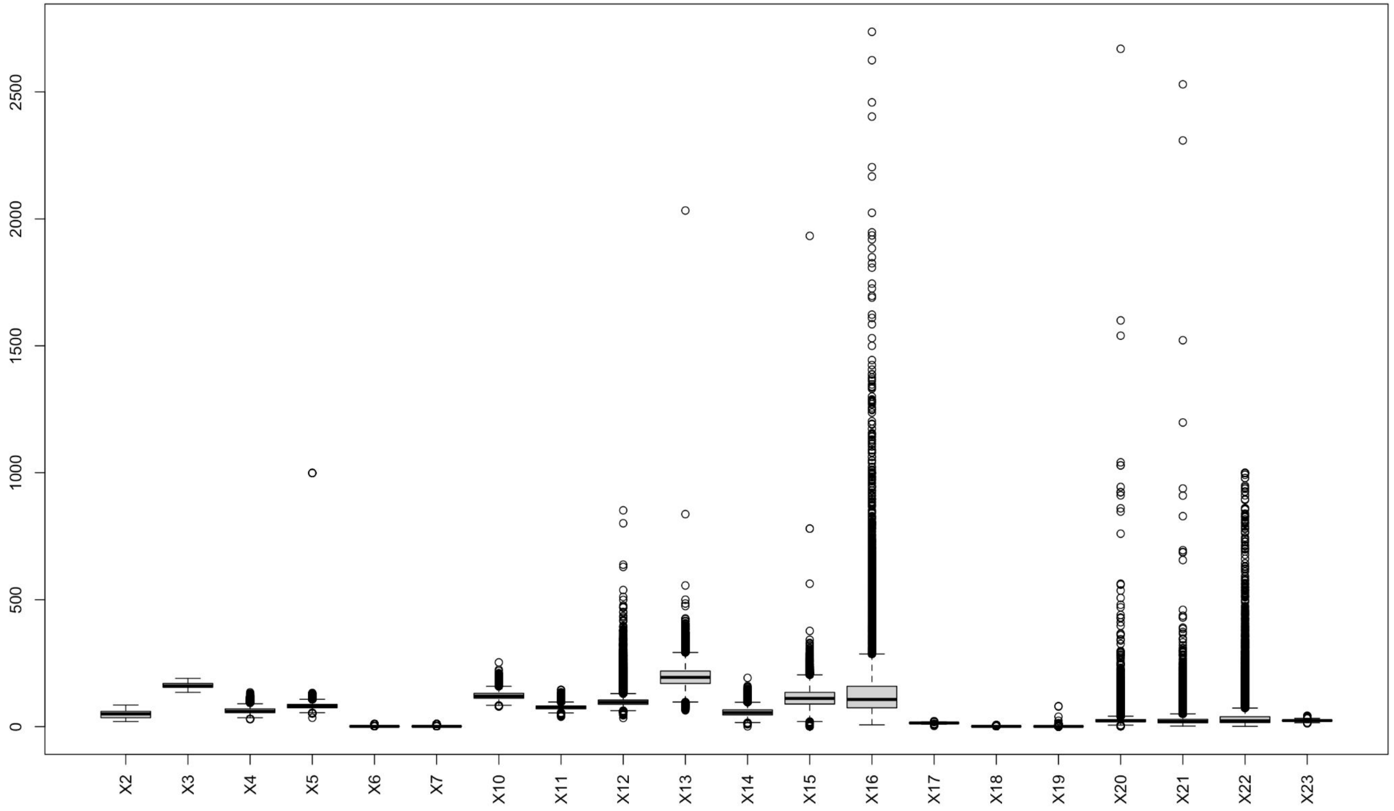
- All other predictors have similar NA percent (6.90% - 7.15%)

Impute NA Values via MICE

- **Used the whole data set**
- Use "**pmm**" for numerical variables
- Use "**polyreg**" for categorical variables
- Factor all variables that need factoring (sex, sight left, etc) and response variable Alcoholic Status

Data Processing/Manipulation: Outliers

X1 = Sex, **X2 = Age**, **X3 = Height**, **X4 = Weight**, **X5 = Waistline**, **X6 = Sight Left** , **X7 = Sight Right**, X8 = Hear Left, X9 = Hear Right, **X10 = SBP**, **X11 = DBP**, **X12 = BLDS**, **X13 = Total Cholesterol**, **X14 = HDL Cholesterol**, **X15 = LDL Cholesterol**, **X16 = Triglyceride**, **X17 = Hemoglobin**, **X18 = Urine Protein**, **X19 = Serum Creatinine**, **X20 = SGOT AST**, **X21 = SGOT ALT**, **X22 = Gamma GTP**, **X23 = BMI**, X24 = BMI Category, X25 = Age Category, X26 = Smoking Status



Analysis

There appear to be many variables with outliers that may drawback our model

Deduct outliers

We used **5% and 95% quantile** to replace outliers

- These quantiles are smaller/larger than 1.5 IQR of the first and the third quartile

Data Processing/Manipulation: Create New Predictors

Analysis

We observed *multicollinearity* within some predictors since their VIF > 5:

- Sex, Waistline, SBP, DBP, Total Cholesterol, HDL Cholesterol, LDL Cholesterol, Triglyceride, SGOT AST, etc

BLOOD PRESSURE CATEGORY	SYSTOLIC mm Hg (upper number)		DIASTOLIC mm Hg (lower number)
NORMAL	LESS THAN 120	and	LESS THAN 80
ELEVATED	120 – 129	and	LESS THAN 80
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 1	130 – 139	or	80 – 89
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 2	140 OR HIGHER	or	90 OR HIGHER

	DESIRABLE	BORDERLINE HIGH	HIGH
Total Cholesterol	Less than 200	200 - 239	240 and higher
LDL Cholesterol	Less than 130	130 - 159	160 and higher
HDL Cholesterol	50 and higher	40 - 49	Less than 40
Triglycerides	Less than 200	200 - 399	400 and higher

WAIST TO HIP RATIO CHART FOR WOMEN

	Lowest risk of MI*					+12-25%		+25-50%		+50-100%		2x risk and above			
HIPS \ WAIST	0.70	0.72	0.74	0.76	0.78	0.80	0.82	0.84	0.86	0.88	0.90	0.92	0.94	0.96	0.98
35	24.5	25.2	25.9	26.6	27.3	28.0	28.7	29.4	30.1	30.8	31.5	32.2	32.9	33.6	34.3
36	25.2	25.9	26.6	27.4	28.1	28.8	29.5	30.2	31.0	31.7	32.4	33.1	33.8	34.6	35.3
37	25.9	26.6	27.4	28.1	28.9	29.6	30.3	31.1	31.8	32.6	33.3	34.0	34.8	35.5	36.3
38	26.6	27.4	28.1	28.9	29.6	30.4	31.2	31.9	32.7	33.4	34.2	35.0	35.7	36.5	37.2
39	27.3	28.1	28.9	29.6	30.4	31.2	32.0	32.8	33.5	34.3	35.1	35.9	36.7	37.4	38.2
40	28.0	28.8	29.6	30.4	31.2	32.0	32.8	33.6	34.4	35.2	36.0	36.8	37.6	38.4	39.2
41	28.7	29.5	30.3	31.2	32.0	32.8	33.6	34.4	35.3	36.1	36.9	37.7	38.5	39.4	40.2
42	29.4	30.2	31.1	31.9	32.8	33.6	34.4	35.3	36.1	37.0	37.8	38.6	39.5	40.3	41.2
43	30.1	31.0	31.8	32.7	33.5	34.4	35.3	36.1	37.0	37.8	38.7☆	39.6	40.4	41.3	42.1
44	30.8	31.7	32.6	33.4	34.3	35.2	36.1	37.0	37.8	38.7	39.6	40.5	41.4	42.2	43.1
45	31.5	32.4	33.3	34.2	35.1	36.0	36.9	37.8	38.7	39.6	40.5	41.4	42.3	43.2	44.1
46	32.2	33.1	34.0	35.0	35.9	36.8	37.7	38.6	39.6	40.5	41.4	42.3	43.2	44.2	45.1
47	32.9	33.8	34.8	35.7	36.7	37.6	38.5	39.5	40.4	41.4	42.3	43.2	44.2	45.1	46.1
48	33.6	34.6	35.5	36.5	37.4	38.4	39.4	40.3	41.3	42.2	43.2	44.2	45.1	46.1	47.0
49	34.3	35.3	36.3	37.2	38.2	39.2	40.2	41.2	42.1	43.1	44.1	45.1	46.1	47.0	48.0
50	35.0	36.0	37.0	38.0	39.0	40.0	41.0	42.0	43.0	44.0	45.0	46.0	47.0	48.0	49.0

Source: Peters, Bots et al 2018
Comparative references: Rost, Freuer et al, 2018; Nalini, Sharafkhah et al 2019, Cameron, Romaniuk et al 2020

☆ average US female

*MI—myocardial infarction (heart attack)

whyiexercise.com

WAIST TO HIP RATIO CHART FOR WOMEN

	Lowest risk of MI*					+12-25%		+25-50%		+50-100%		2x risk and above			
HIPS \ WAIST	0.70	0.72	0.74	0.76	0.78	0.80	0.82	0.84	0.86	0.88	0.90	0.92	0.94	0.96	0.98
35	24.5	25.2	25.9	26.6	27.3	28.0	28.7	29.4	30.1	30.8	31.5	32.2	32.9	33.6	34.3
36	25.2	25.9	26.6	27.4	28.1	28.8	29.5	30.2	31.0	31.7	32.4	33.1	33.8	34.6	35.3
37	25.9	26.6	27.4	28.1	28.9	29.6	30.3	31.1	31.8	32.6	33.3	34.0	34.8	35.5	36.3
38	26.6	27.4	28.1	28.9	29.6	30.4	31.2	31.9	32.7	33.4	34.2	35.0	35.7	36.5	37.2
39	27.3	28.1	28.9	29.6	30.4	31.2	32.0	32.8	33.5	34.3	35.1	35.9	36.7	37.4	38.2
40	28.0	28.8	29.6	30.4	31.2	32.0	32.8	33.6	34.4	35.2	36.0	36.8	37.6	38.4	39.2
41	28.7	29.5	30.3	31.2	32.0	32.8	33.6	34.4	35.3	36.1	36.9	37.7	38.5	39.4	40.2
42	29.4	30.2	31.1	31.9	32.8	33.6	34.4	35.3	36.1	37.0	37.8	38.6	39.5	40.3	41.2
43	30.1	31.0	31.8	32.7	33.5	34.4	35.3	36.1	37.0	37.8	38.7☆	39.6	40.4	41.3	42.1
44	30.8	31.7	32.6	33.4	34.3	35.2	36.1	37.0	37.8	38.7	39.6	40.5	41.4	42.2	43.1
45	31.5	32.4	33.3	34.2	35.1	36.0	36.9	37.8	38.7	39.6	40.5	41.4	42.3	43.2	44.1
46	32.2	33.1	34.0	35.0	35.9	36.8	37.7	38.6	39.6	40.5	41.4	42.3	43.2	44.2	45.1
47	32.9	33.8	34.8	35.7	36.7	37.6	38.5	39.5	40.4	41.4	42.3	43.2	44.2	45.1	46.1
48	33.6	34.6	35.5	36.5	37.4	38.4	39.4	40.3	41.3	42.2	43.2	44.2	45.1	46.1	47.0
49	34.3	35.3	36.3	37.2	38.2	39.2	40.2	41.2	42.1	43.1	44.1	45.1	46.1	47.0	48.0
50	35.0	36.0	37.0	38.0	39.0	40.0	41.0	42.0	43.0	44.0	45.0	46.0	47.0	48.0	49.0

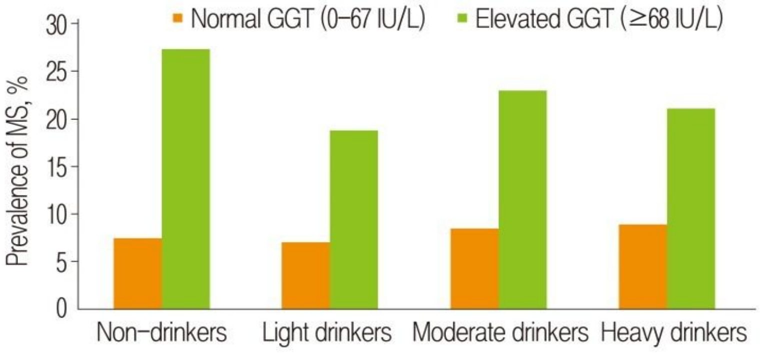
Source: Peters, Bots et al 2018
Comparative references: Rost, Freuer et al, 2018; Nalini, Sharafkhah et al 2019, Cameron, Romaniuk et al 2020

☆ average US female

*MI—myocardial infarction (heart attack)

whyiexercise.com

AST/ALT Ratio	Interpretation
Less than 1	Most likely acute hepatitis
1 to 2	Non-alcoholic fatty liver disease (NAFLD), alcoholic liver disease, cirrhosis, or chronic hepatitis B or C
Greater than 2	Alcoholic hepatitis, cirrhosis, or metastatic liver dis



**Note: References on last page

Data Processing/Manipulation: Create New Predictors

Create New Predictors

We created new predictors based on predictors that may have correlations with each other

- **X27**: Blood Pressure → SBP and DBP
- **X28**: ALT:AST → SGOT AST and SGOT ALT
- **X29**: Cholesterol → Total Cholesterol, HDL Cholesterol, LDL Cholesterol, and Triglyceride
- **X30**: Waist Risk → Sex and Waistline
- **X31**: GTP Category → Sex and GTP

Observations

Next step was to remove predictors used to create new predictors but testing error was large

- *We decided to keep old predictors!*

X27	Normal	Elevated	Stage 1	Stage 2
Blood Pressure	47.65%	11.82%	33.15%	7.38%

X28	Amino-transferases	Cirrhosis	Without Cirrhosis
ALT:AST Ratio	18.71%	33.71%	47.57%

X29	Desirable	Borderline High	High
Cholesterol	34.29%	41.37%	23.34%

X30	Low risk	High risk	Very high
Waist Risk	78.07%	16.00%	5.93%

X31	Normal	Abnormal
GTP Category	83.04%	16.96%

Agenda

- Introduction
  - Context
  - Data Description
    - Density Plots
    - Stacked Bar Charts
- Data Processing/Manipulation
  - Missing Values
  - Outliers
  - Create New Predictors
- Model Selection
  - Compare Testing Error
- Reduce Predictors
- Summary

Model Selection: Confusion Matrices

GLM	Predict	
Original	N	Y
N	7324	2708
Y	2758	7160
Accuracy	**72.60%**	
Error	**27.40%**	

SVM	Predict	
Original	N	Y
N	7013	3069
Y	2432	7486
Accuracy	**72.50%**	
Error	**27.50%**	

LDA	Predict	
Original	N	Y
N	7380	2702
Y	2783	7135
Accuracy	**72.58%**	
Error	**27.42%**	

QDA	Predict	
Original	N	Y
N	**6370**	3712
Y	2305	**7613**
Accuracy	**69.92%**	
Error	**30.08%**	

Observations

1. LDA model gives a better accuracy than the QDA model, so the boundaries between classes are **more likely to be linear**

2. Even though QDA gives the lowest accuracy among all models, QDA does the the **best in predicting "Yes/Yes"** and **worst in predicting "No/No"**

Model Selection: Confusion Matrices

Model Description

We tried different number of predictors for each tree

- 5 predictors (mtry = 5) does the best for **full model**
- 4 predictors (mtry = 4) does the best for **reduced model**

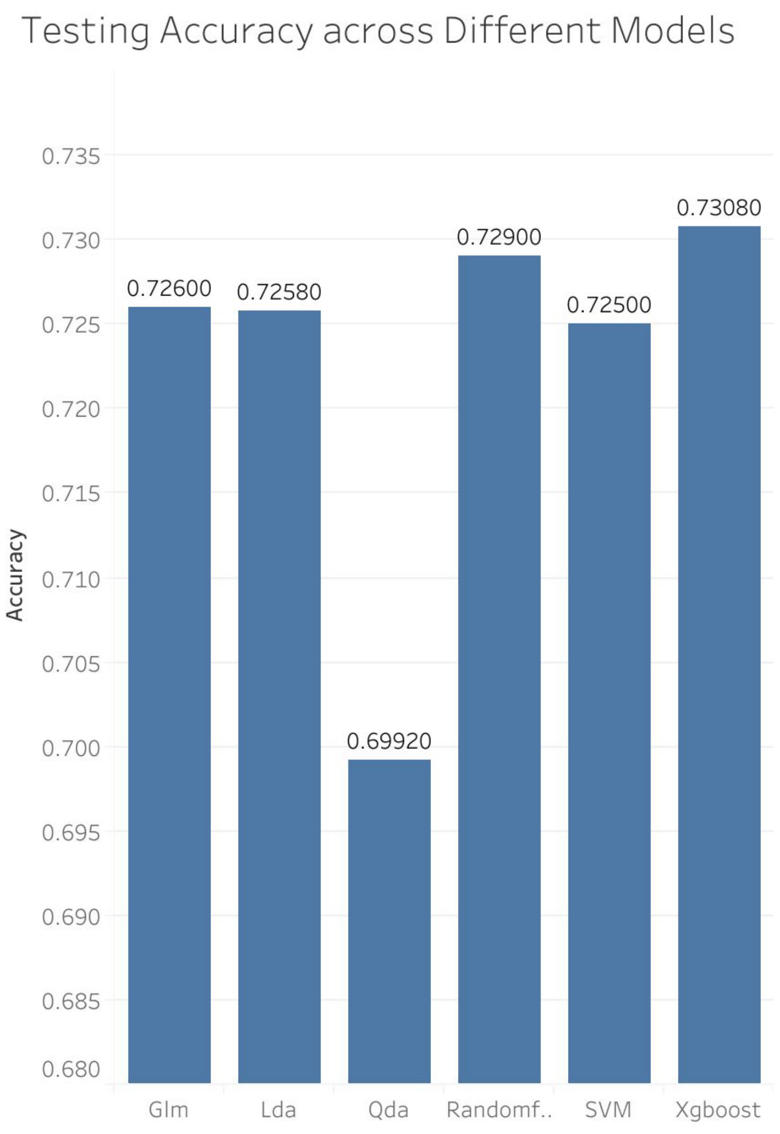
According to the accuracy vs. the number of trees graph, 945 trees gives the best accuracy, but 500 trees gives similar accuracy

- We chose 500 trees (ntree = 500)

Random Forest	Predict	
Original	N	Y
N	7122	2960
Y	2460	7458
Accuracy	**72.90%**	
Error	**27.10%**	

XGBoost	Predict	
Original	N	Y
N	7177	2905
Y	2479	7439
Accuracy	**73.08%**	
Error	**26.92%**	

Model Selection: Compare Testing Errors



Observations

1. QDA has the lowest one, which implies it might have a linear boundary between the classes

2. Random Forest and XGBoost have the better test accuracy. Thus, we proceeded with using these two as our **candidate models** using the entire training data

**Note: these accuracy rates are based on our testing data NOT Kaggle

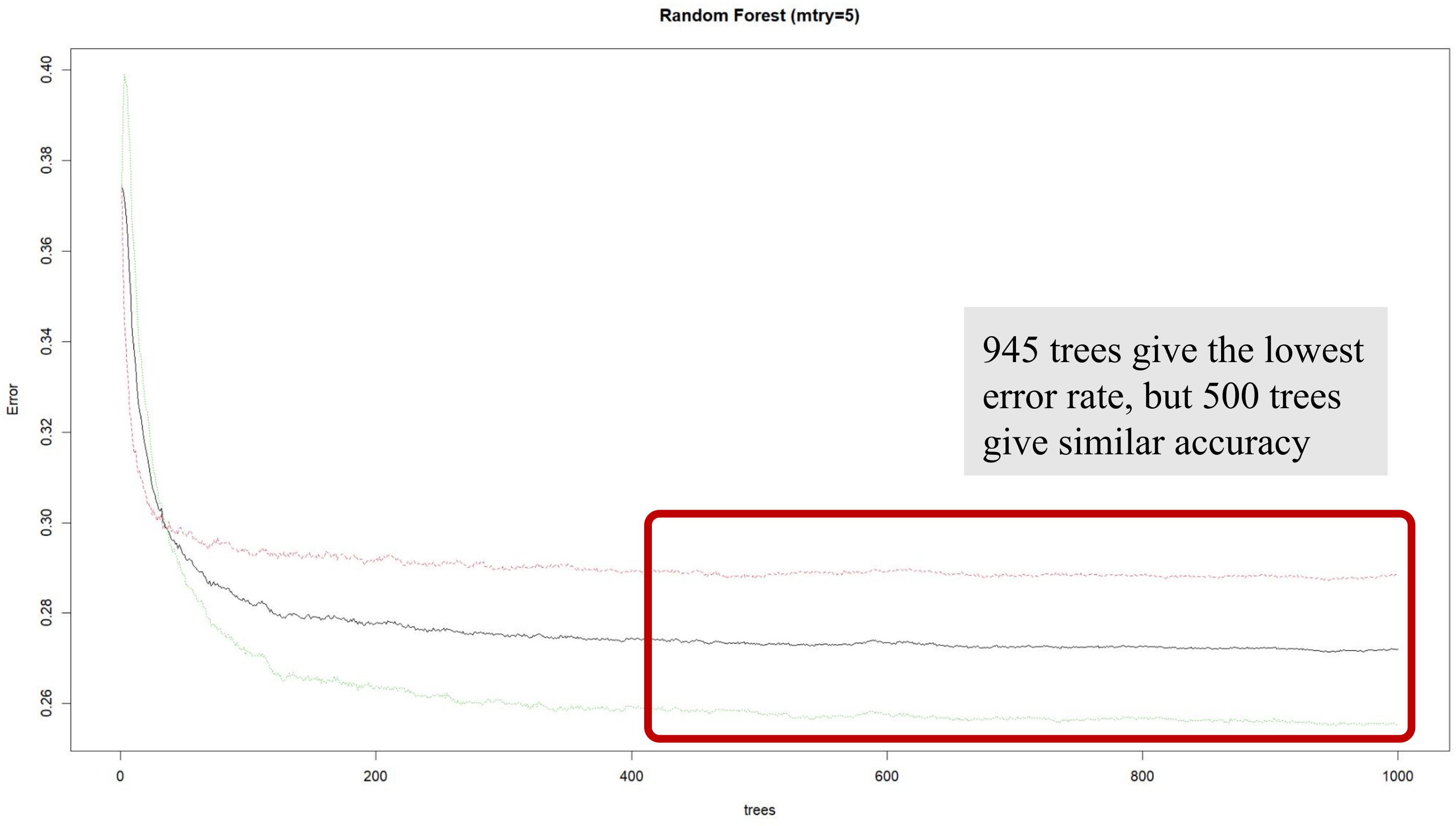
Candidate Model 1 - Random Forest

Note

From the precious model selection, we observed that mtry = 5 and ntree = 500 is best for the full model. Therefore, we use this hyperparameter to train our model

Whole Train Data (RF)	Predict	
Original	N	Y
N	24986	10127
Y	8918	25969
Accuracy	**72.79%**	
Error	**27.21%**	

Error rate vs. Number of Trees



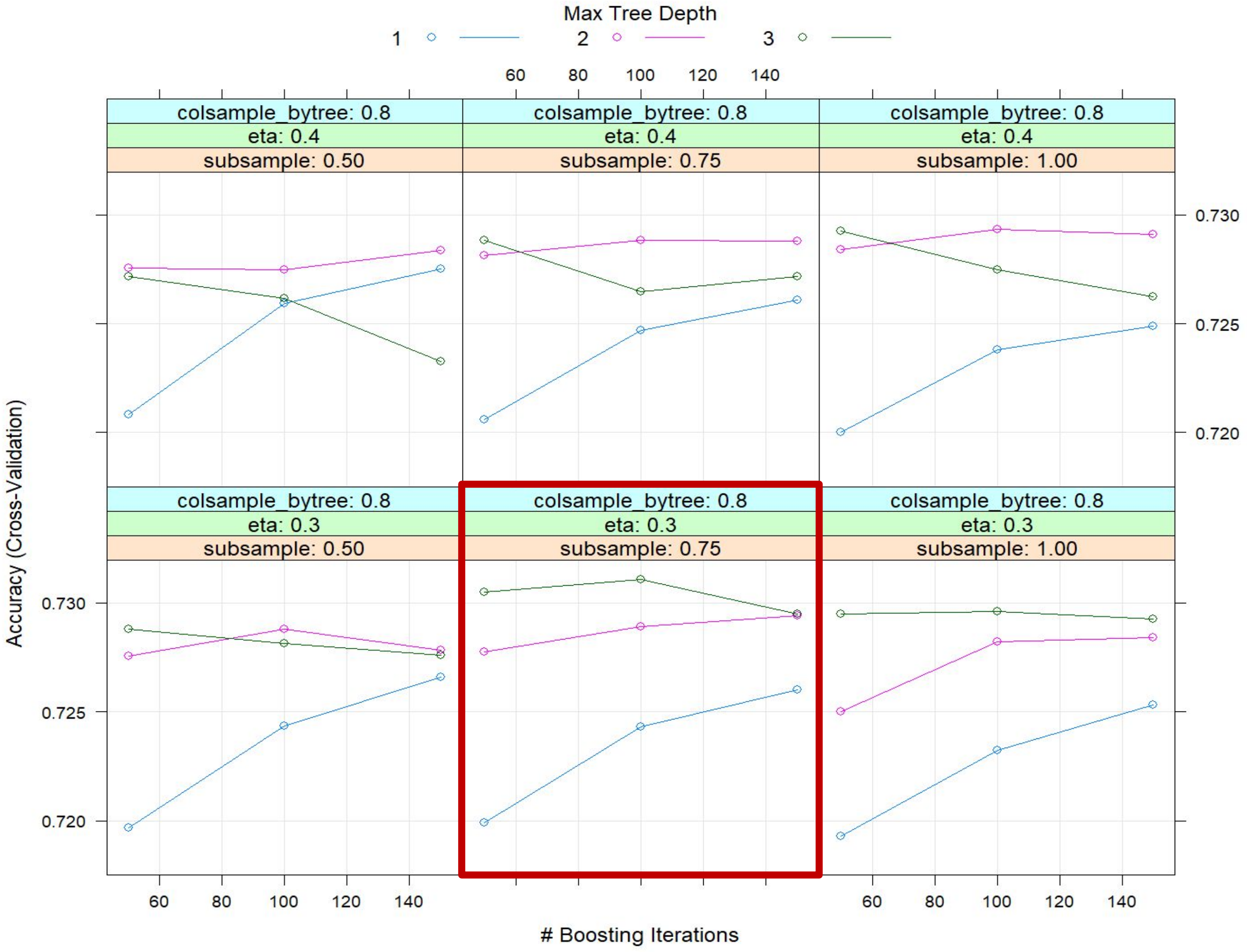
Candidate Model 2 - XGBoost

Best parameters based on 10-fold CV:

Based off # boosting iterations given accuracy rates per Max Tree Depth:

- η = 0.3
- max_depth = 3
- subsample = 0.75
- olsample_bytree = 0.8

Whole Train Data (XGB)	Predict	
Original	N	Y
N	25367	9746
Y	8757	26130
Accuracy	**73.57%**	
Error	**26.43%**	



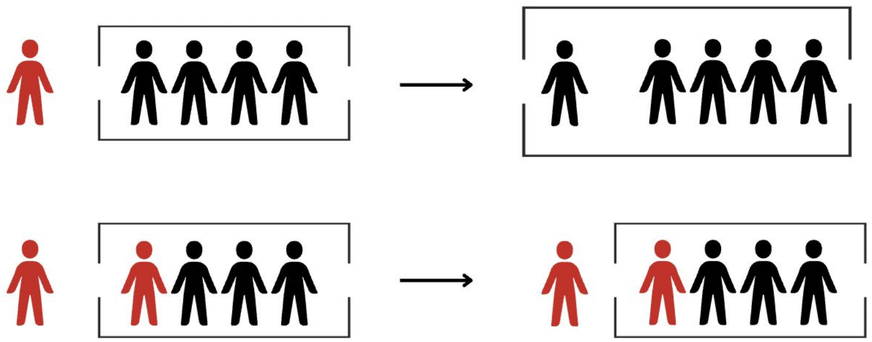
Hero Model

Combination Model Logic

We obtained multiple prediction sets by running XGBoost several times. Due to the model's inherent randomness, our predictions are not consistently identical; they exhibit slight variations and demonstrate different prediction accuracies. Therefore, we decided to **combine the prediction of two Candidate Models**. From these runs, we selected the two sets of predictions with the highest training data test scores. This process was replicated for the Random Forest model.

We used a third XGBoost prediction with an accuracy of 73.26% based on Kaggle as our base. **If the prediction from the base model differs from all the other four predictions from the two models, we adjust the base prediction to align with the consensus of the other four predictions.** For example, if the base prediction is 'Y' and all the other three predictions are 'N,' we change the base prediction to 'N' for consistency.

Kaggle Accuracy : **73.46%**



Agenda

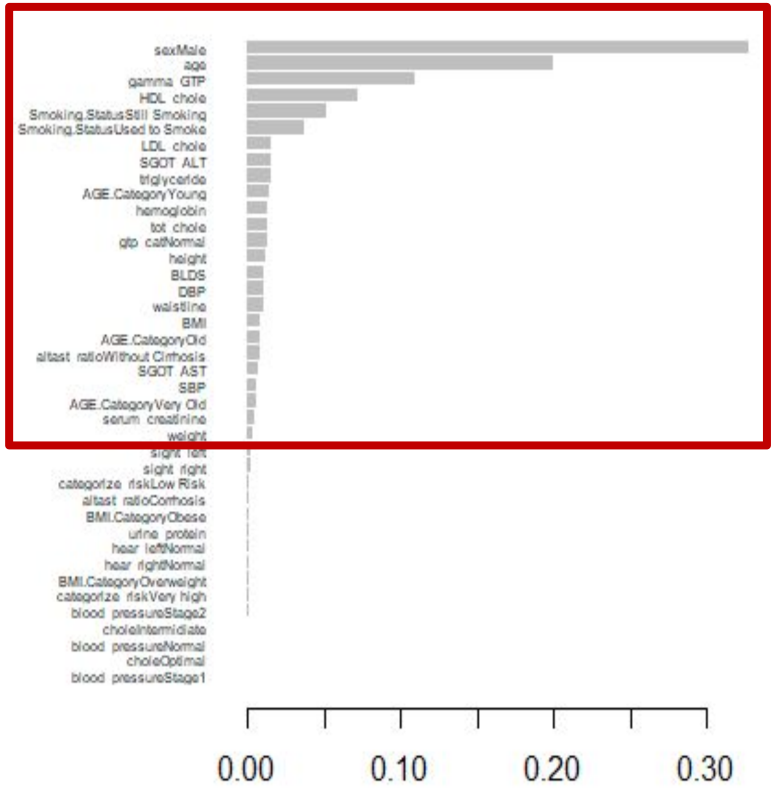
- Introduction
  - Context
  - Data Description
    - Density Plots
    - Stacked Bar Charts
- Data Processing/Manipulation
  - Missing Values
  - Outliers
  - Create New Predictors
- Model Selection
  - Compare Testing Error
- Reduce Predictors
- Summary

Reduce Predictors: Step/XGBoost/PCA

We use backward stepwise regression method and AIC as the criteria to select the **24** predictors

Predictors: Age, Height, Waistline, DBP, BLDS, HDL Cholesterol, LDL Cholesterol, Triglyceride, Hemoglobin, Serum Creatinine, SGOT AST, SGOT ALT, Gamma GTP, BMI, Sex, Hear Left, Hear Right, BMI Category, Age Category, Smoking Status, Blood Pressure, Risk Category, ALT:AST Ratio, GTP Category

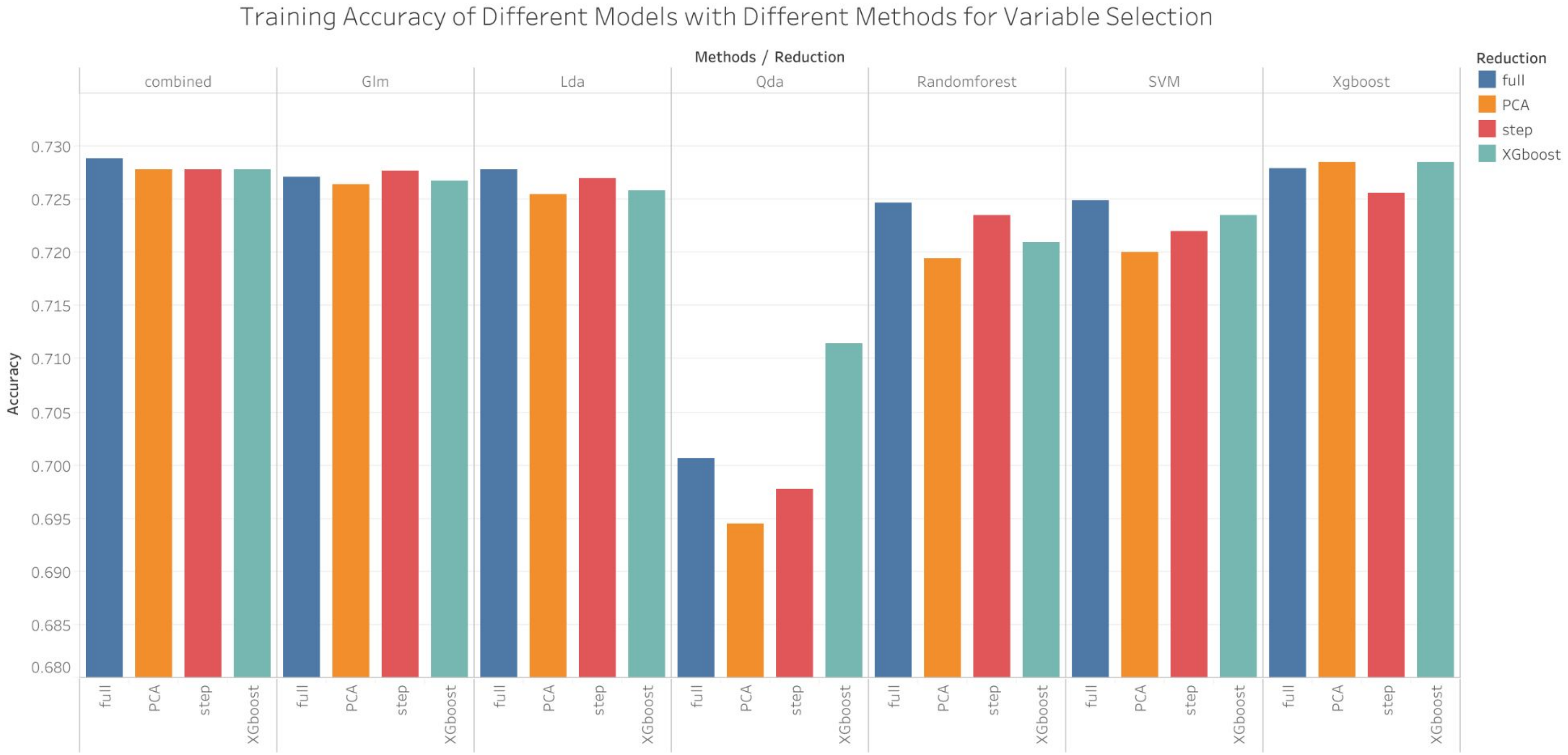
Given the importance graph from the XGBoost, we select the top **16** predictors



We decided to use 7 components, which accounts for 96.12% of the variation + 11 categorical predictors = **18** predictors

Comp	Cumu. Proportion
Comp1	55.01%
Comp2	78.45%
Comp3	85.71%
Comp4	89.15%
Comp5	92.13%
Comp6	94.24%
Comp7	**96.12%**
Comp8	97.53%
Comp9	98.62%
Comp10	99.10%
Comp11	99.42%
Comp12	99.64%
……	……

Reduce Predictors: Accuracy Trade-off Based on Training Data



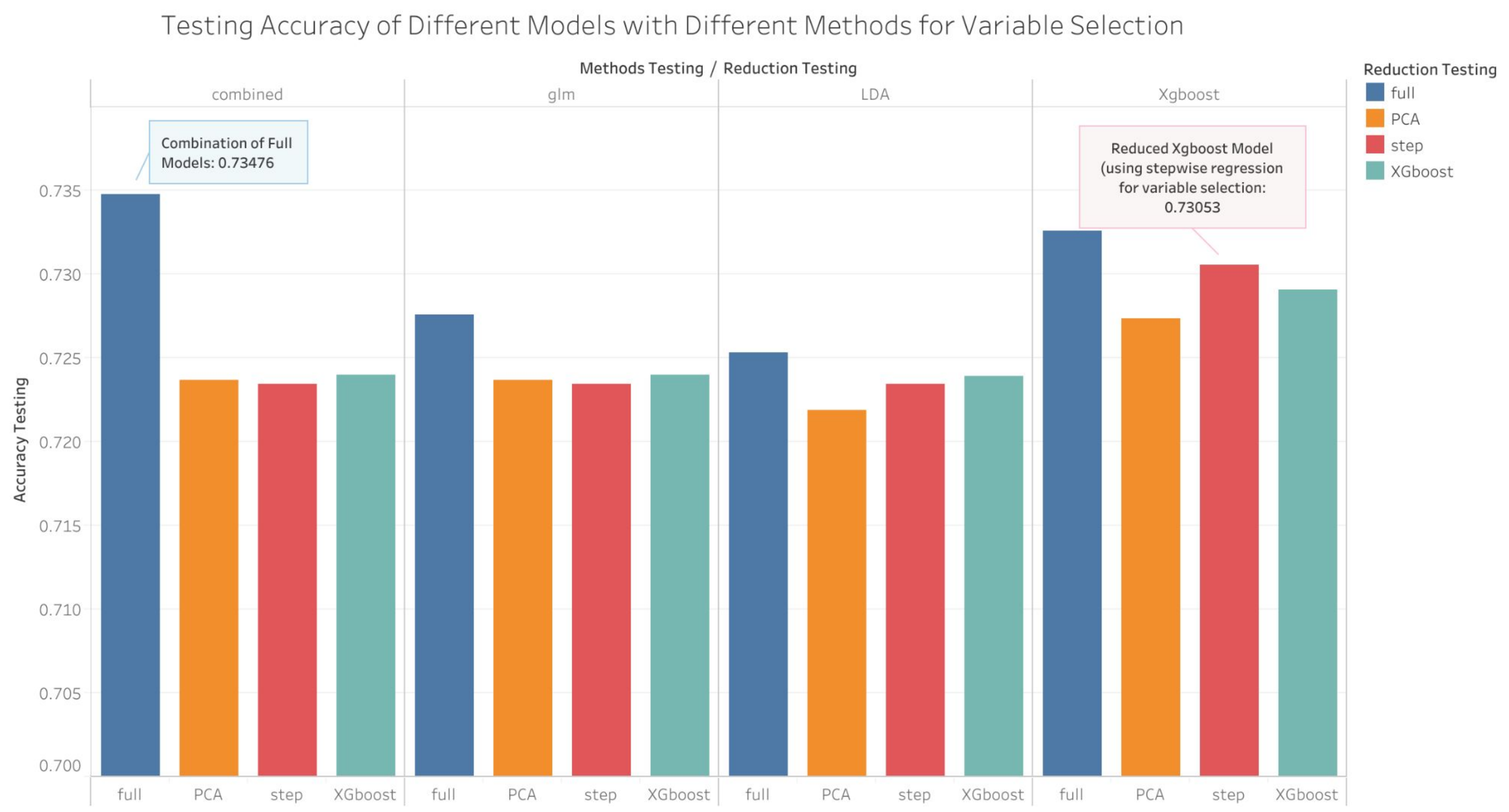
Observations

QDA tends to exhibit poorer performance, suggesting a potential assumption of a **linear boundary** rather than a nonlinear one.

XGBoost, **LDA**, **GLM**, and **Combined models** demonstrate relatively better accuracy in their performance.

**Note: these accuracy rates are based on our training data NOT Kaggle

Reduce Predictors: Accuracy Trade-off Based on Kaggle Testing Data



Observations

The best 'reduced predictors' prediction is using **step predictors and XGBoost methods**

XGBoost - The Best Reduced Model

Predictors: from 31 to 24

Predictors: Age, Height, Waistline, DBP, BLDS, HDL Cholesterol, LDL Cholesterol, Triglyceride, Hemoglobin, Serum Creatinine, SGOT AST, SGOT ALT, Gamma GTP, BMI, Sex, Hear Left, Hear Right, BMI Category, Age Category, Smoking Status, Blood Pressure, Risk Category, ALT:AST Ratio, GTP Category

Confusion Matrix on Testing data split from Training data

Reduced XGBoost	Predict	
Original	N	Y
N	7113	2453
Y	2895	7449
Accuracy	**72.81%**	
Error	**27.19%**	

Tuning Parameters

$\eta = 0.3$; max_depth = 2; subsample = 0.75; colsample_bytree = 0.6

Agenda

- Introduction
  - Context
  - Data Description
    - Density Plots
    - Stacked Bar Charts
- Data Processing/Manipulation
  - Missing Values
  - Outliers
  - Create New Predictors
- Model Selection
  - Compare Testing Error
- Reduce Predictors
- Summary

Summary

Accuracy	→	**Simplicity**
Combination Model		XGBoost Reduced Model

Limitation and Potential Improvement

1. While we created meaningful predictors like 'blood pressure' and 'cholesterol,' they seem to carry **less weight** compared to their original variables in the dataset, especially during the model reduction process (e.g., SDP...).
2. We used Random Forest to impute missing values at first, but it was very computationally extensive. Thus, we resorted to **MICE**; future exploration might involve trying multiple approaches, such as employing techniques like random forest, based on a comprehensive study of the dataset.
3. **Reducing the number of predictors** is often beneficial. Future endeavors could focus on discovering additional methods to decrease predictor size while maintaining accuracy and minimizing rapid drops in performance.
4. Discovering **more relationships** between statistical data and real-world information can enhance the practical application of our model.

Future Insight

Certain predictors, such as “Age”, “Gamma GPT”, “Hemoglobin”, “BMI”, and “Triglyceride,” hold substantial importance and real-world relevance. Individuals exhibiting specific symptoms or falling into risk categories identified by our model may require heightened caution.

References

1. "Understanding Blood Pressure Readings." *Www.Heart.Org*, 17 Oct. 2023, www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings.

2. Hospitals, Medicover. "High Cholesterol Levels: Medicover Hospitals." *Best Hospitals in India | Medicover Hospitals*, www.medicoverhospitals.in/articles/high-cholesterol-levels. Accessed 10 Dec. 2023.

3. "Ast Alt Ratio Calculator: AST to Alt Ratio Chart." *Drlogy*, drlogy.com/calculator/ast-alt-ratio. Accessed 10 Dec. 2023.

4. *Xmlinkhub*, e-cnr.org/ViewImage.php?Type=F&aid=487585&id=F2&afn=9994_CNR_2_1_67&fn=cnr-2-67-g002_9994CNR. Accessed 10 Dec. 2023.

5. "Waist-to-Hip Ratio: Reliable Research Shows If You Need to Lose Weight." *Whyiexercise.Com*, www.whyiexercise.com/waist-to-hip-ratio.html. Accessed 10 Dec. 2023.