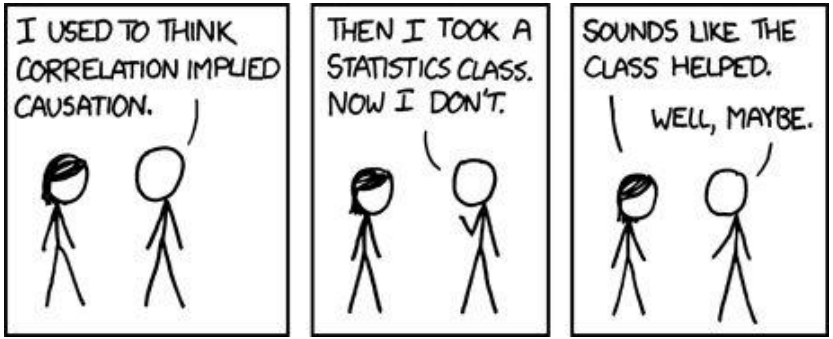
STAT628 Module 2

Group 11

Body Fat Estimation Project

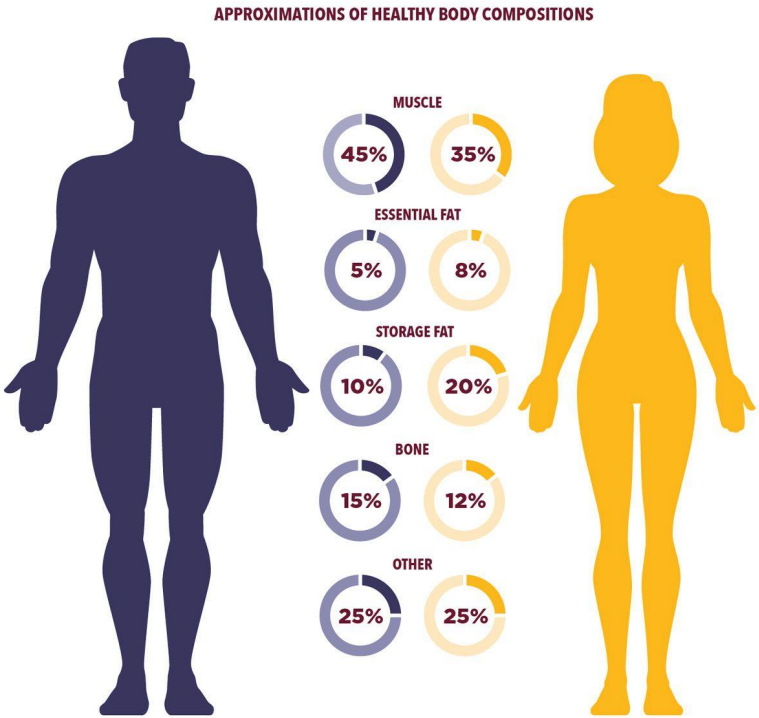
Shoura Maheshwari
Chixu Ni
Sreeja Kodati

TABLE OF CONTENTS



Introduction

Laying the foundation: A Fat Chance at Simplicity, Robustness, and Accuracy



Data Preprocessing

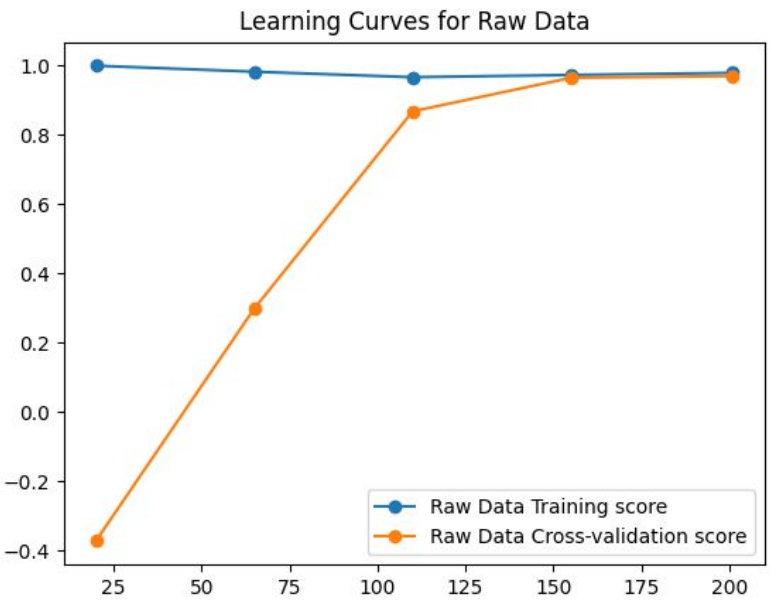
From Raw to Refined: The Data Makeover

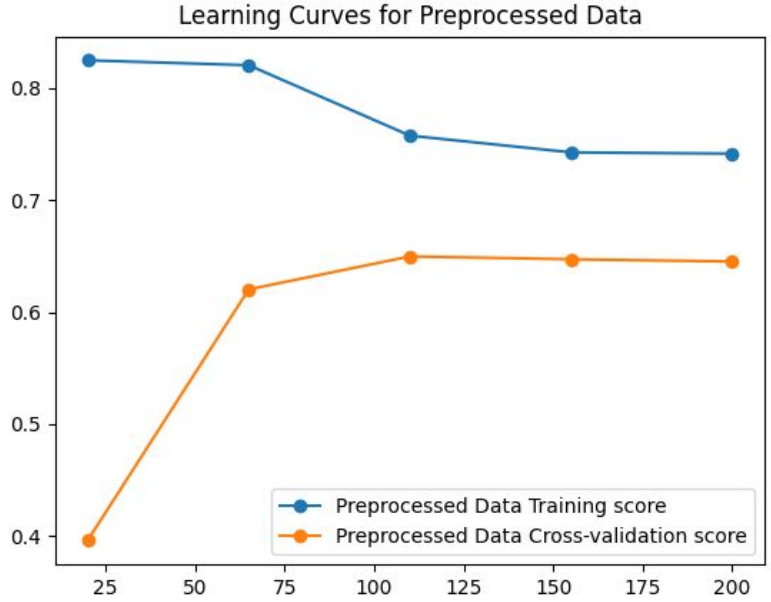
. . . Because if Garbage goes in, Garbage comes out

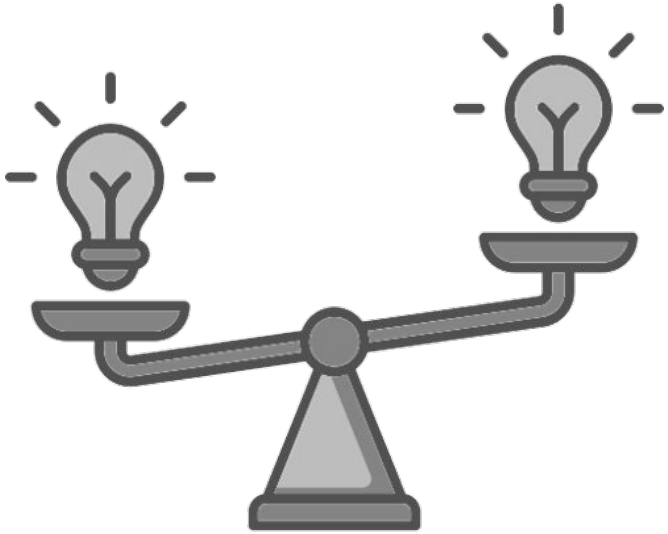
Simplicity	Robustness	Accuracy
Dropping irrelevance	Outlier Management	Skewness Transformation
Handling anomalies, one by one.	Imputation	Robust Scaling
Reducing dimensionality	Robust Scaling	KNN-based Imputation
	Variance Inflation Factor	

Is Preprocessing Important?

Improvement in how the model fits





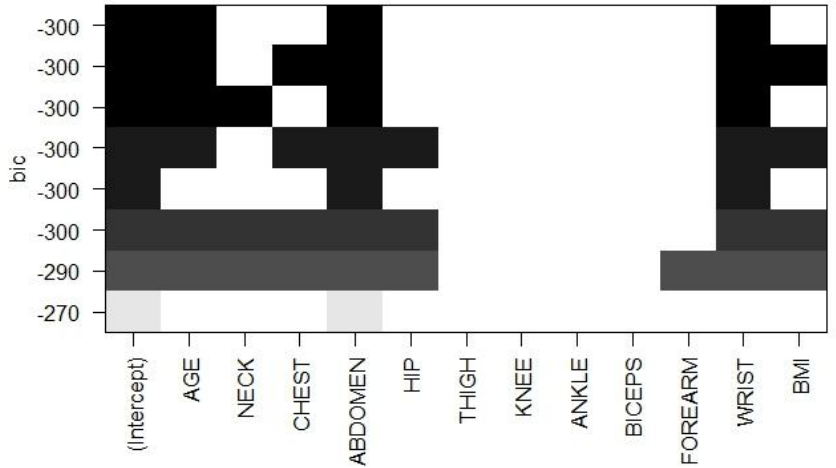


02

Model Selection

The Three that Lasted . . .

- Linear regression
- Robust regression
- Ridge regression



		AGE	NECK	CHEST	ABDOMEN	HIP	THIGH	KNEE	ANKLE	BICEPS	FOREARM	WRIST	BMI
1	(1)	" "	" "	" "	"*"	" "	" "	" "	" "	" "	" "	" "	" "
2	(1)	" "	" "	" "	"*"	" "	" "	" "	" "	" "	" "	"*"	" "
3	(1)	"*"	" "	" "	"*"	" "	" "	" "	" "	" "	" "	"*"	" "
4	(1)	"*"	"*"	" "	"*"	" "	" "	" "	" "	" "	" "	"*"	" "
5	(1)	"*"	" "	"*"	"*"	" "	" "	" "	" "	" "	" "	"*"	"*"
6	(1)	"*"	" "	"*"	"*"	"*"	" "	" "	" "	" "	" "	"*"	"*"
7	(1)	"*"	"*"	"*"	"*"	"*"	" "	" "	" "	" "	" "	"*"	"*"
8	(1)	"*"	"*"	"*"	"*"	"*"	" "	" "	" "	" "	"*"	"*"	"*"

$3 * 8 = 24$ models .

Choosing the Best Model

(Final Model)

	linear	robust	ridge
Number of predictors	3	4	7
R-squared	0.729	0.729	0.714
AIC	187.66	186.92	181.09
BIC	205.27	208.05	205.63
Variance Inflation Factors	1.394	2.194	5.889

03

Final Model

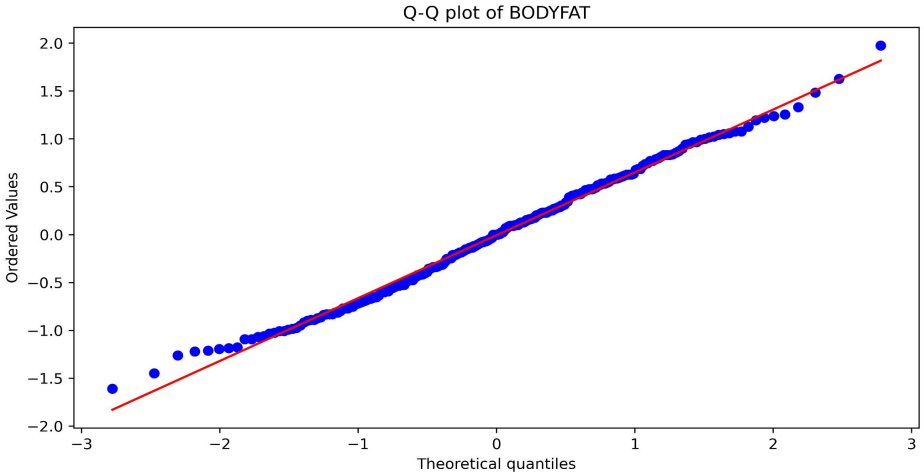
Ultimate Model

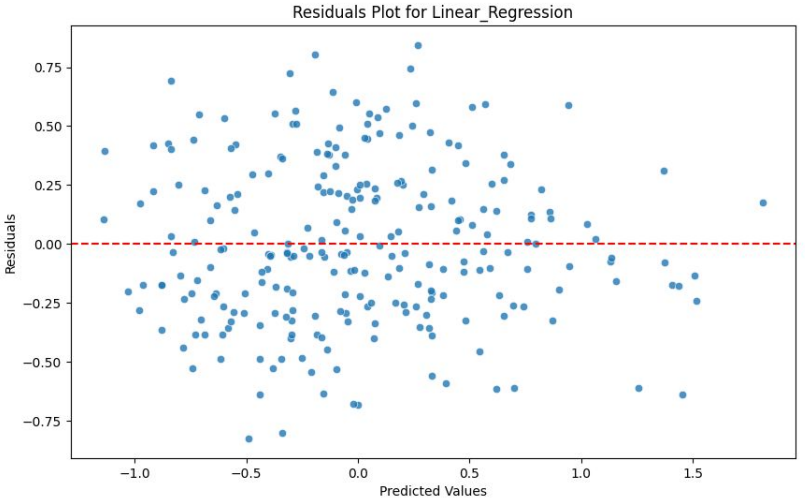
BODYFAT = (0.1156) + (0.1113 * AGE) + (0.8928 * ABDOMEN) + (-0.2328 * WRIST)

	Coefficient	Std. Error	T-value	P-value
AGE	0.11133	0.03247	3.429	0.0007
ABDOMEN	0.89278	0.03899	22.898	<2e-16
WRIST	-0.23283	0.03641	-6.395	8.02e-10

All features are significant through T-test.

Model Diagnostics





The data satisfies all the assumptions for a linear model.

04

Visualizations (Shiny App)

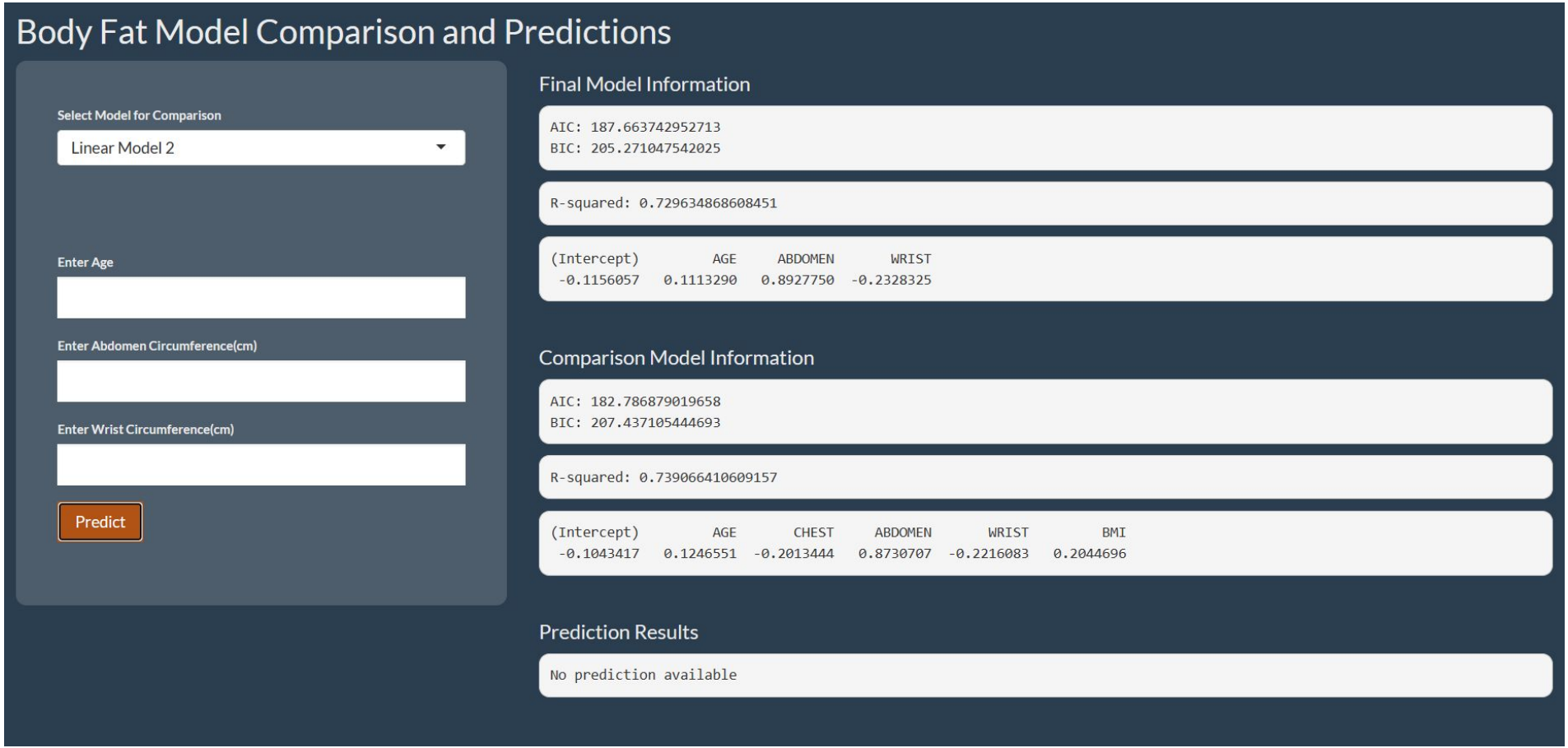
SHINY APP

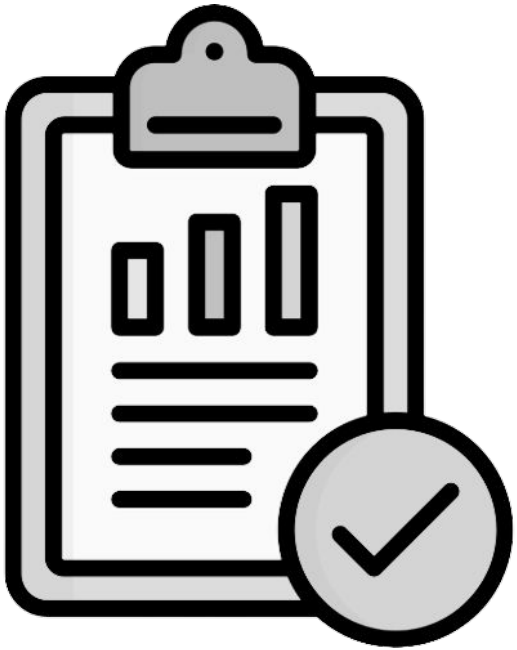
1. Model Comparison
- The Shiny app is designed to compare multiple linear regression models for predicting body fat using different sets of predictors.
- Users can select one of these models for comparison through a drop-down menu in the sidebar.

2. Body Fat Predictions
- Users can input values for age, abdomen circumference, and wrist circumference. These values are used to make predictions using the final model.

Shiny app for this project allows users to explore different linear regression models and compare their performance, and make predictions of body fat based on user-provided input data using the final model.

Visualizing Data with My Shiny App





05

Discussions Of Results

Discussion of Results

1. Most important question - What trade-offs were made?
2. What factors led to the absence of BMI data in our final dataset?
3. Why is the final model, the best model?

Thank you !

Appendix

AIC

Akaike Information Criterion, a kind of information criterion based on entropy

$$AIC = -2\ln(L) + 2k$$

Where L is the maximum likelihood and k is variable's number.

Smaller AIC, better model fits

Appendix

BIC

Bayesian Information Criterions, information criterion similar to AIC, but with higher penalty on sample size to prevent overfit and complexity on large data.

$$BIC = -2\ln(L) + \ln(n) * k$$

Where L is likelihood, n the sample size and k variable's number

Smaller BIC, better model fits.

Appendix

Difference of AIC and BIC

The principles of AIC and BIC are different.
AIC selects a good model for prediction from the perspective of prediction.
BIC selects a model that best fits the existing data from the perspective of fitting, which is the model with the greatest marginal likelihood from the interpretation of Bayes factor.

Appendix

VIF

VIF (Variance Inflation Factor) is an indicator of multicollinearity. Its formula is shown as

$$VIF_k = \frac{1}{1 - R_k^2}$$

Where R here is the coefficient of determination the k-th predictor fitted on other features except the outcome.
When multicollinearity exists, the parameters would be hard to estimate stably and to interpret, since the design matrix becomes ill-conditioned.
Usually multicollinearity can be thought to exist when mean VIF much larger than 1 or max VIF larger than 10.