Predicting Body Fat with Decision Tree Model

Group 8

Introduction

Body fat plays a crucial role in our lives. It is important to understand and monitor body fat to determine one's physical well-being. Considering the difficulty of measuring body fat percentage directly from a person, we developed a simple but robust regression Decision Tree model to estimate the body fat percentage based on chest and abdomen circumferences.

Background Information and Data Cleaning

Inspired by previous studies to estimate the percentage of body fat by age and some skin-fold measurements obtained easily with a caliper, we aimed to create a new prediction model for body fat based on a dataset containing the body circumference measurements of 252 men along with their body fat percentage. Firstly, we removed the records with missing values and addressed potential outliers of the body fat, which helped to improve model effectiveness and robustness.

Variable and Model Selection

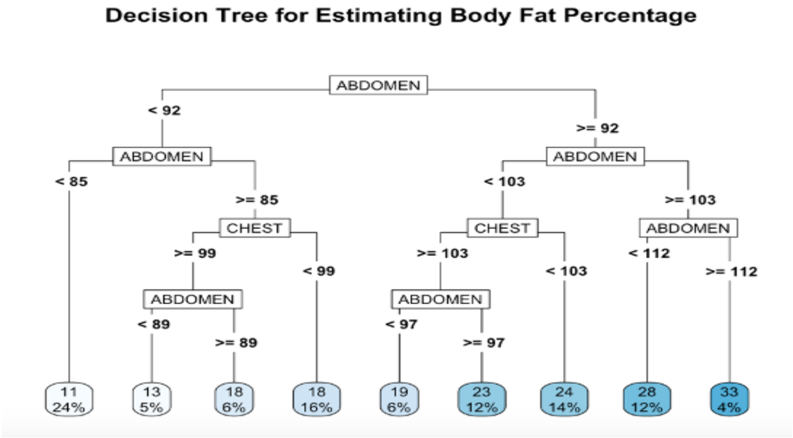
Variable selection is decided by the correlation between body fat and other independent variables. The matrix shows that "Adiposity", "Chest" and "Abdomen" have the top three highest Pearson correlation coefficients. Besides, a relatively strong multicollinearity problem among the 3 variables is observed by calculating the Variance Inflation Factor (>5) among them.

Observing the multicollinearity pattern, we assume tree models will perform better than regression models. Tree models can not only solve multicollinearity problems but also provide feature importance, which can help variable selection further. We opted for a decision tree model over other tree-based counterparts because of its simplicity and high interpretability. Compared to other tree models such as random forests, a decision tree provides a clear visual representation of decision pathways, despite a slight trade-off in accuracy and susceptibility to overfitting. The decision tree's straightforward nature ensures our stakeholders can readily trust and engage with the model's findings.

R-squared, MAE, and MSE are used as performance metrics. Random Forest (rf), Principal Component Regression (PCR), LASSO, and Linear regression (lm) are considered as comparisons to support our final model.

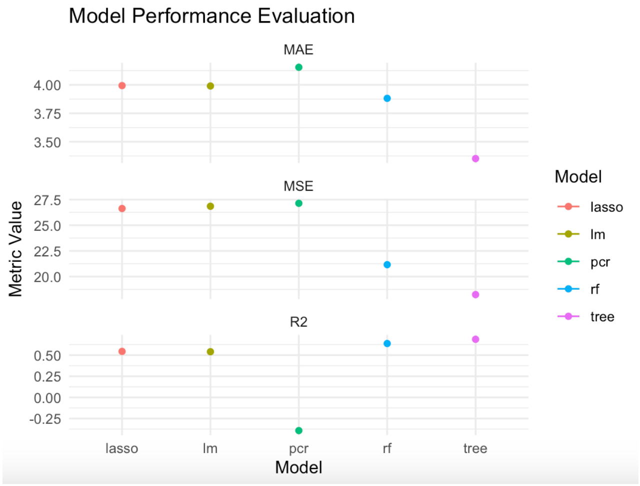
Final Model Statement

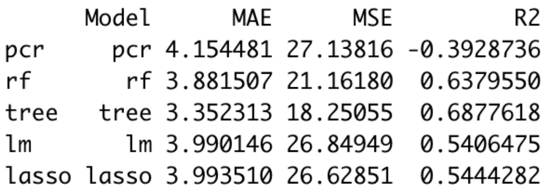
The final model is a regression Decision Tree trained on 250 male measurements. The model shows that a man who has an abdomen value between 92 and 97 and a chest value larger than or equal to 103 is predicted to be 4% of the population and has a body fat percentage of 33. Below is a visualization of our decision tree model.



Model Strengths and Weaknesses

The decision tree is the most compelling model across all evaluation metrics. With the lowest MAE of 3.35 and MSE of 18.25, the decision tree model underscored its best predictive precision. It also achieved an R-squared value of 0.68, demonstrating its capacity to explain approximately 68% of the variability in the target outcome. These robust statistics underscore the decision tree's heightened accuracy and consistency. Here is a table showing some key statistics of our decision tree model and other popular prediction approaches.





```
      Model      MAE      MSE         R2
pcr     pcr 4.154481 27.13816 -0.3928736
rf       rf 3.881507 21.16180  0.6379550
tree   tree 3.352313 18.25055  0.6877618
lm       lm 3.990146 26.84949  0.5406475
lasso lasso 3.993510 26.62851  0.5444282
```

Another advantage of our model is its natural robustness to irrelevant or multicollinear variables. It discards “Adiposity” automatically, which doesn’t help in splitting the data.

However, as a trade-off of simplicity, our model might be under-fitted since we only used 2 out of the 15 variables. The model’s predictions could also be biased due to the loss of information from the rest of the variables.

Conclusion

Our Decision Tree model accurately estimates body fat percentage using chest and abdomen circumferences. While it excels in simplicity and robustness, using only two out of 15 variables, it may sacrifice some complexity. Further enhancements are possible.

Contributions	Xilin Qiao	Vaishnavi Borwankar	Yuchen Dou	Daniel Ye
Presentation	Responsible for slides 1-3	Responsible for slides 4-5, 7-8 and 12.	Page 10-11	Responsible for slides 6 and 9
Summary	Responsible for the Variable and Model Selection part Reviewed and edited final model statement and conclusion	Reviewed and edited the summary.	Finish the draft of the summary	Responsible for the Variable and Model Selection, and model strength/weakness part
Code	Responsible for Github Repository creation, updates and README Reviewed and edited data cleaning part	Reviewed and provided feedback on the code.	Adjust the code when writing the draft of the summary and some visualization	Responsible for all the code
Shiny App	Reviewed and provided feedback on Shiny App	Responsible for the Shiny app.	Review	Reviewed