# CRAB AGE PREDICTION

Group 7

# Agenda

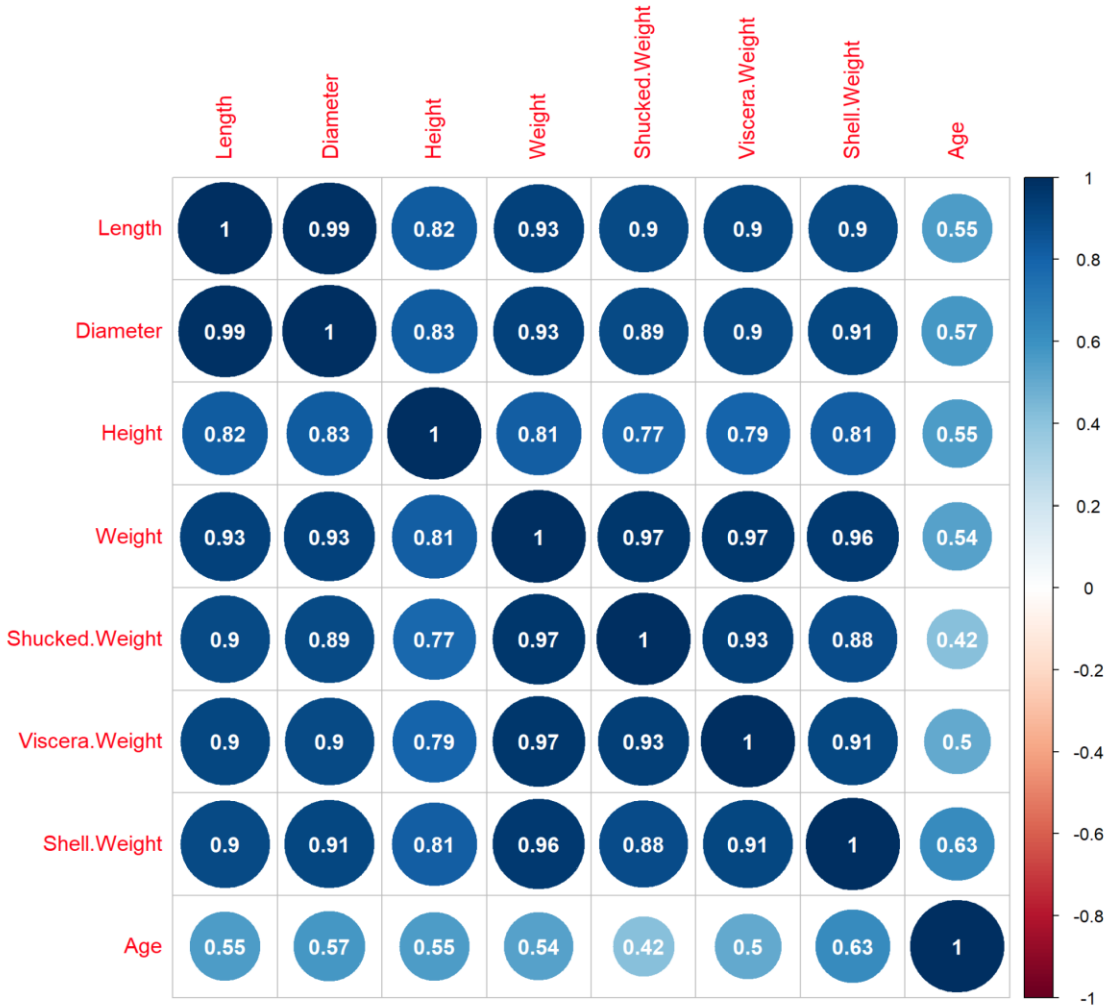
- Exploratory Data Analysis
- Feature Engineering
- Models
- Conclusion

# Exploratory Data Analysis

# Correlation Matrix

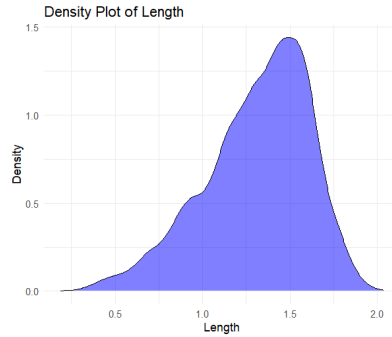
Plot to represent the linear relationship between various predictor variables

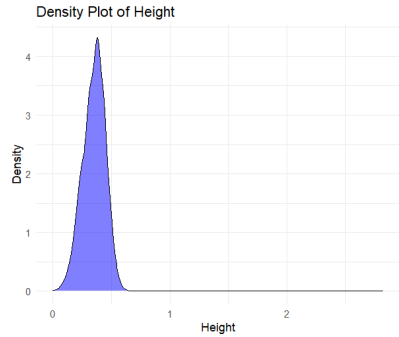
- Strong multicollinear relationship between all the weight variables
- All features except age are positively correlated with each other

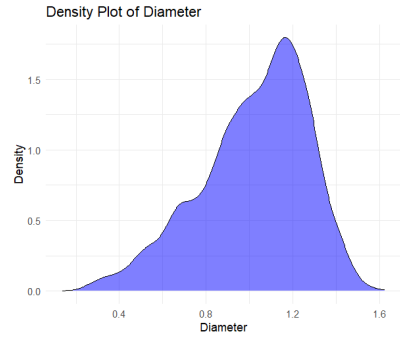


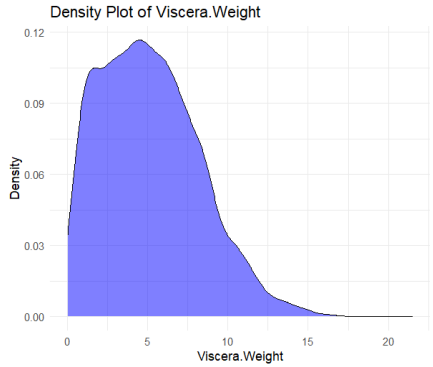
# Density Plots

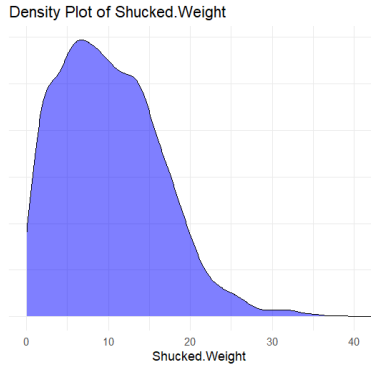
Understand the underlying shape and characteristics of the data

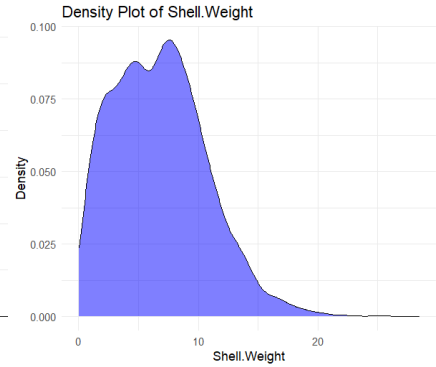


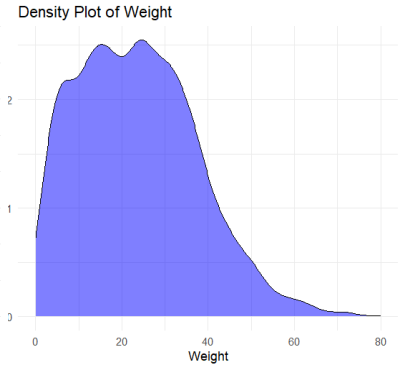



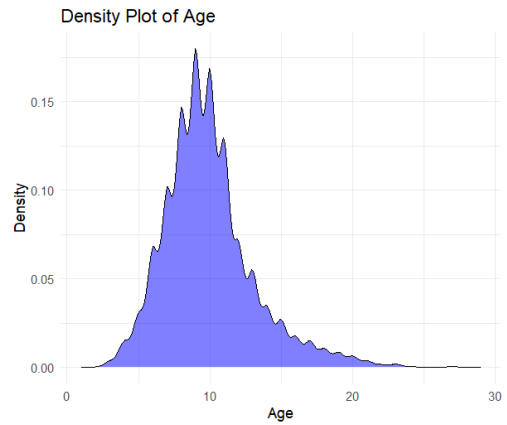



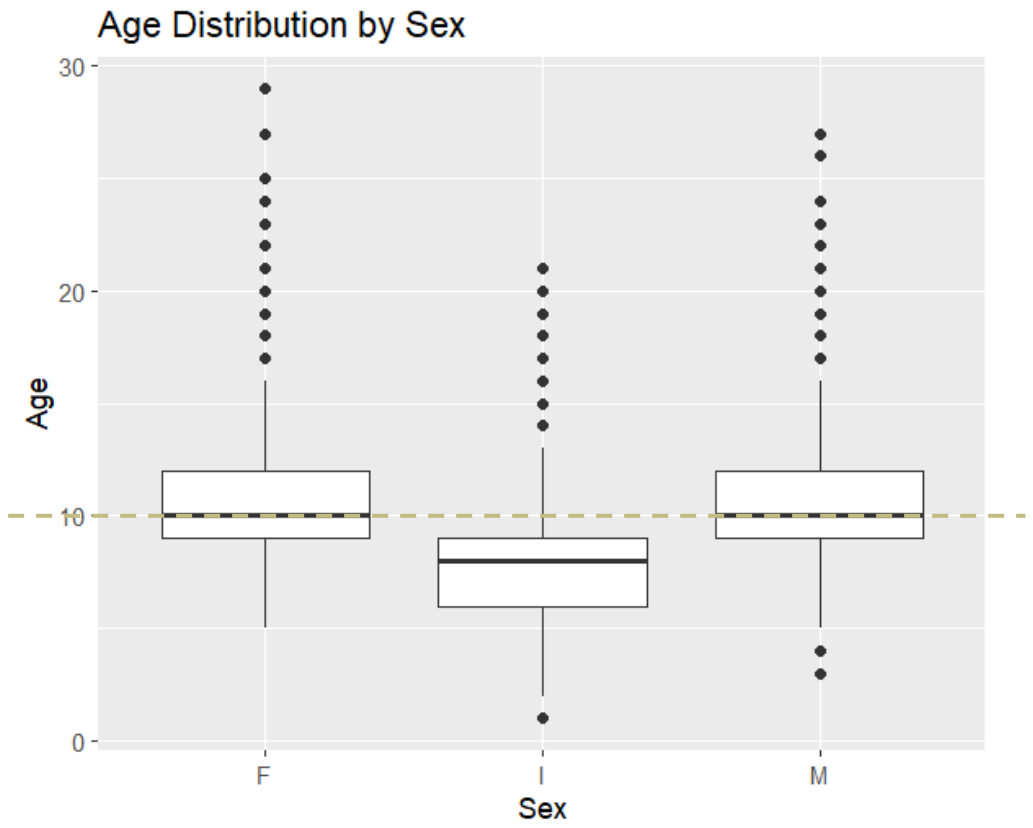



# How important is “Sex” variable to predict Age




Distribution of crab sex is nearly equal for all the categories

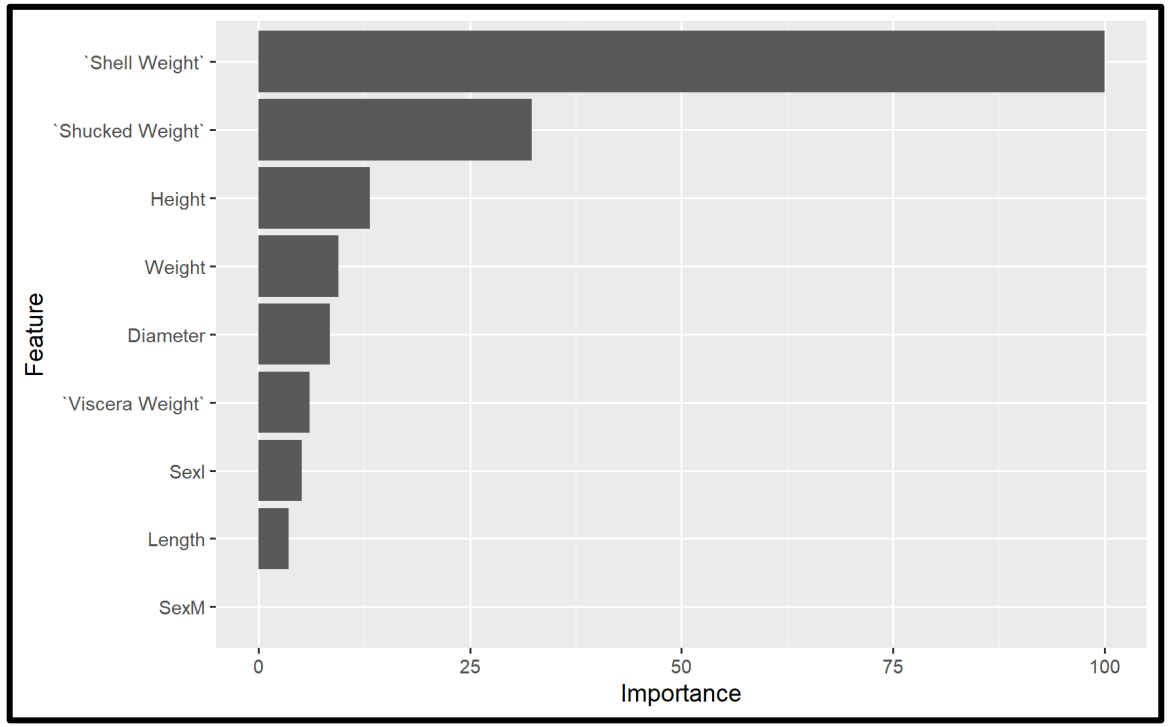
From the below box plot we can infer that for ‘Sex’==I, the average median age is less than the ones for M, F

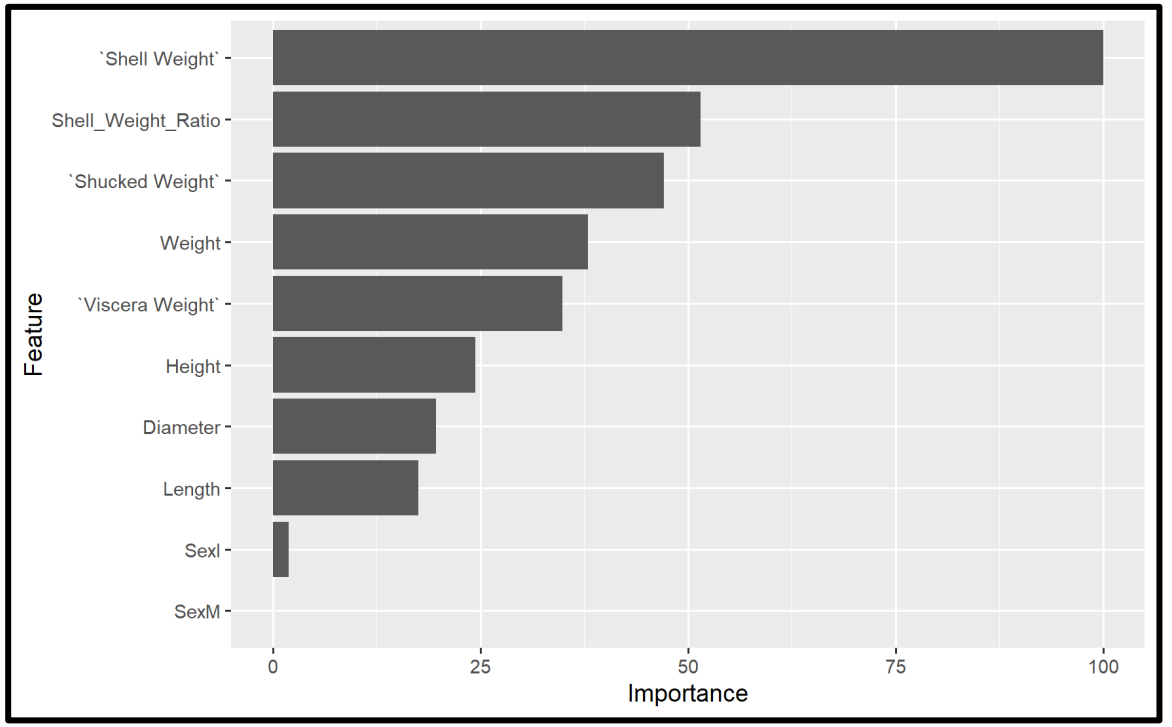
Perhaps, its harder to tell the sex of the crab when it’s younger!

# Feature Engineering

# Incorporating “Shell Weight Ratio”

Shell.Weight.Ratio = Shell.Weight / Weight



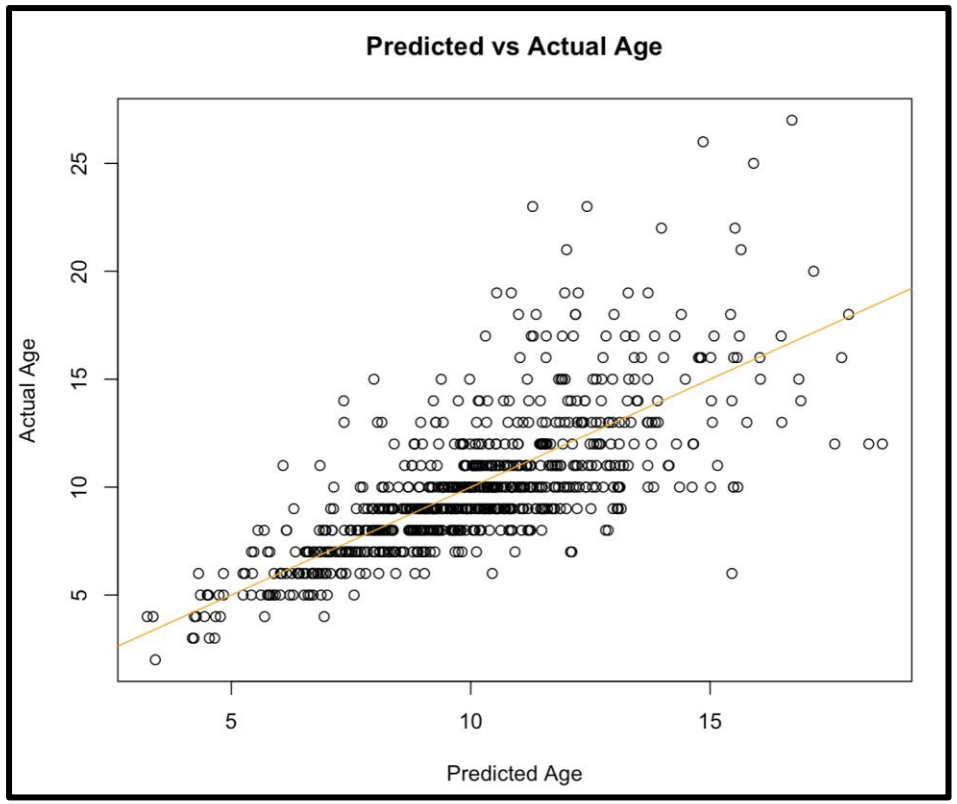
# Models

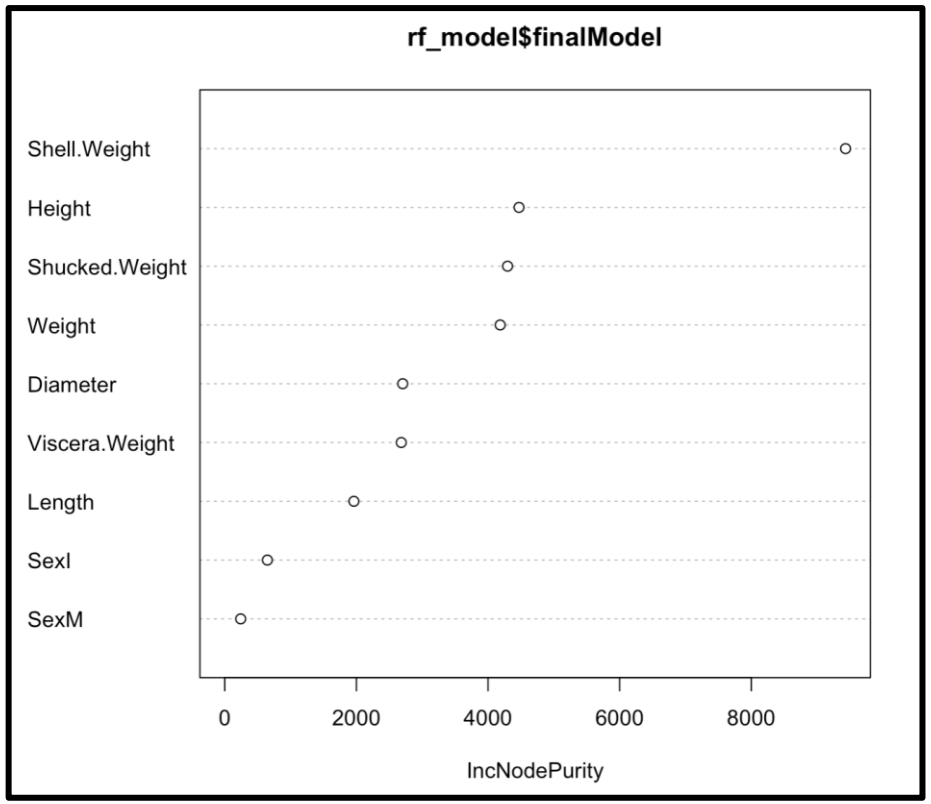
# Model Summary

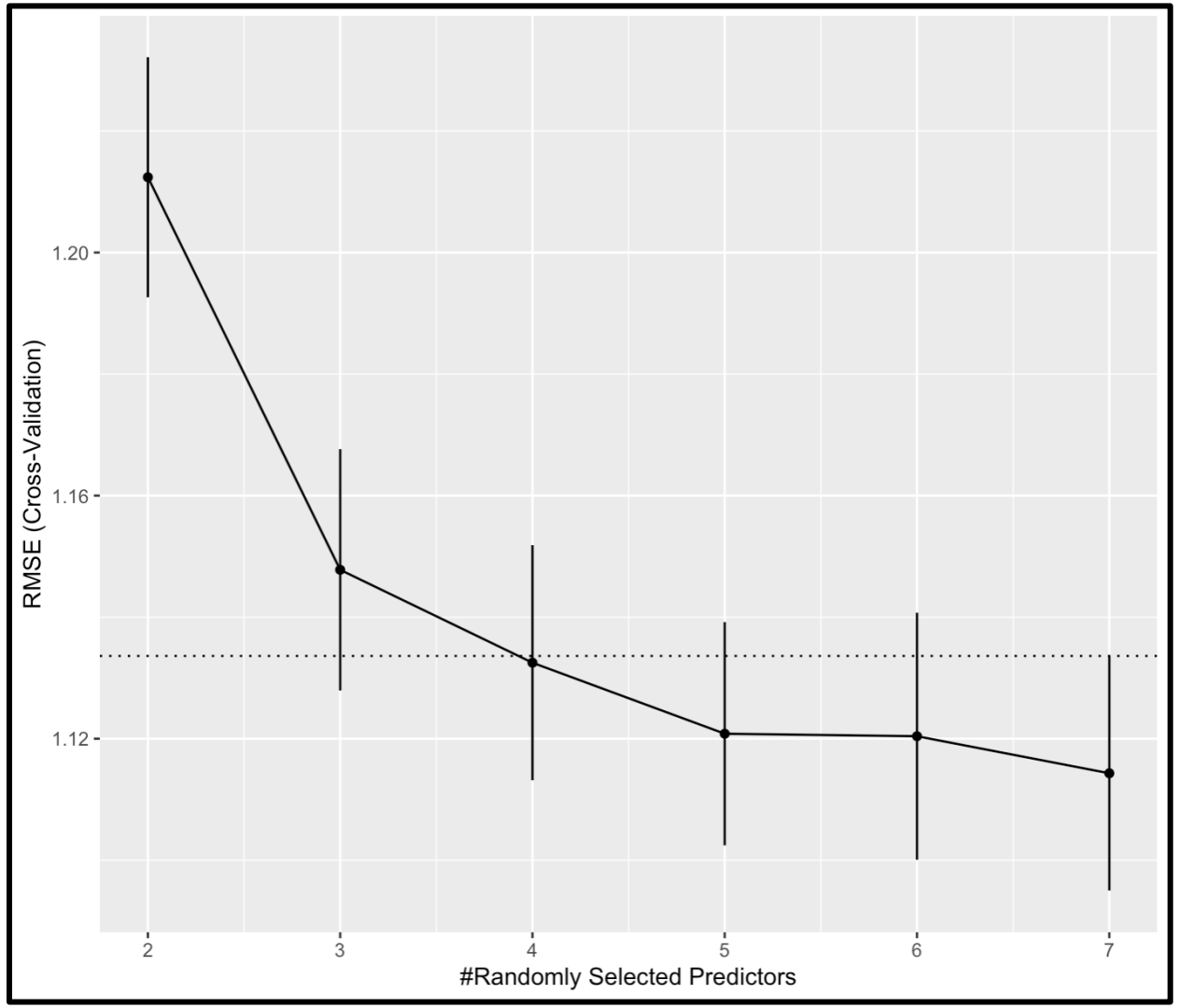
| Model Name | RMSE Before Shell Weight Ratio | RMSE After Shell Weight Ratio | Improvement |
|---|---|---|---|
| Random Forest | 2.262 | 2.096 | 7.34% |
| Multiple Linear Regression | 2.303 | 2.127 | 7.64% |
| Bagging | 2.315 | 2.127 | 8.12% |
| Lasso Regression | 2.154 | 2.141 | 0.60% |
| Single Tree | 2.265 | 2.223 | 1.85% |
| Ridge Regression | 2.266 | 2.259 | 0.31% |
| Boosting | 2.268 | 2.263 | 0.22% |
| KNN | 2.265 | 2.265 | 0% |
| Pruned Tree | 2.468 | 2.279 | 7.66% |

# Random Forest

| RMSE | 2.263 |
|---|---|



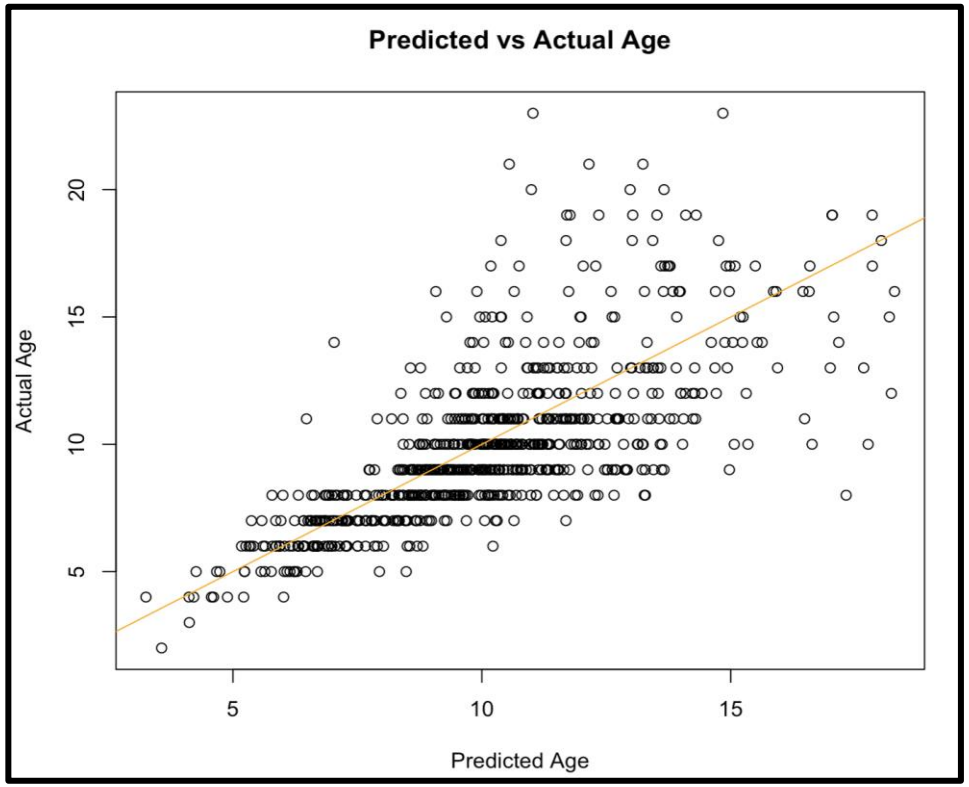



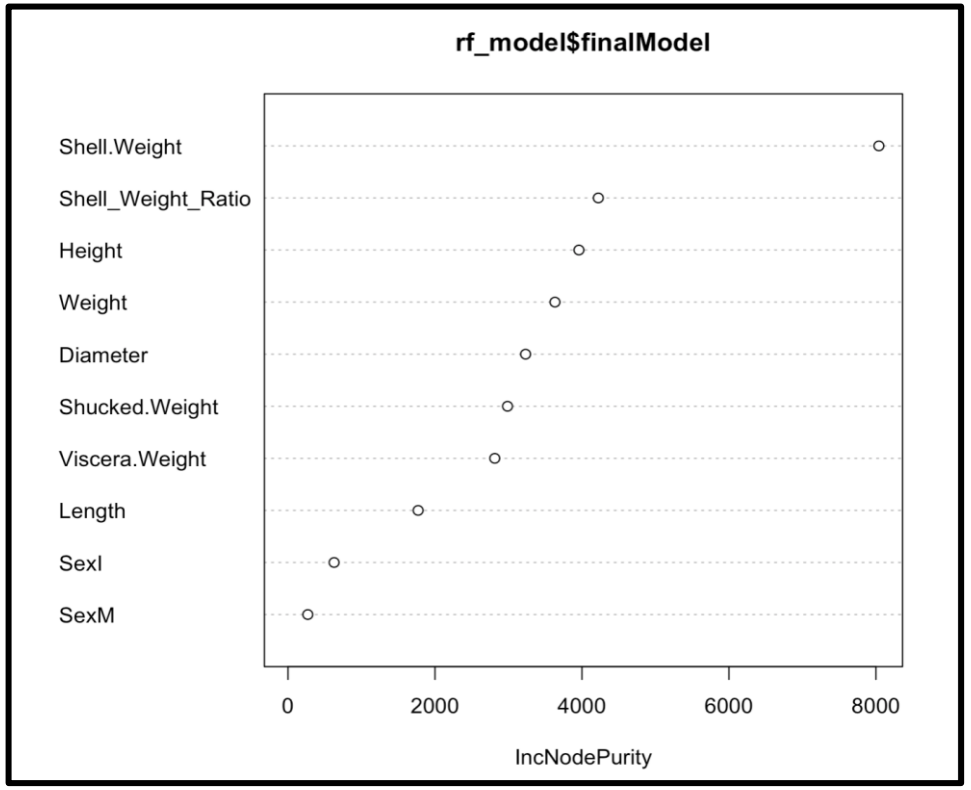


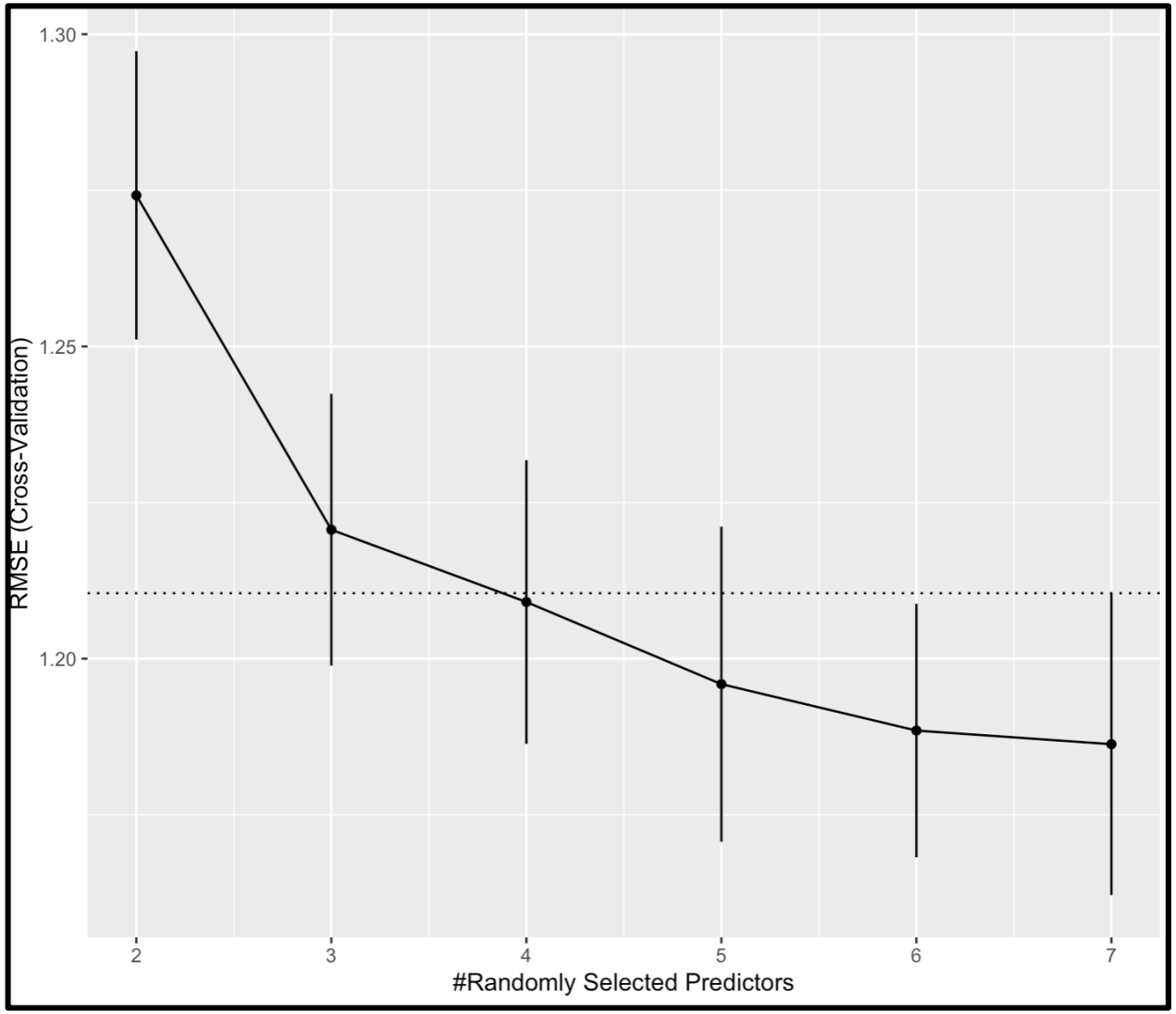

# Optimal Random Variables

# Random Forest (with Shell Weight Ratio)

| RMSE | 2.096 |
|---|---|

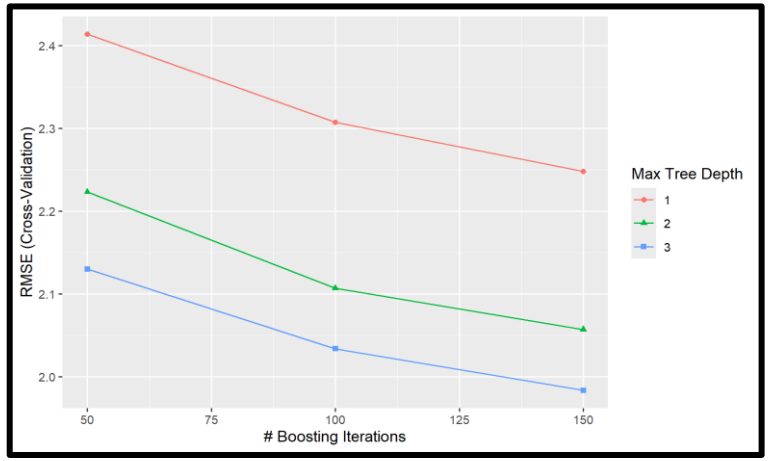


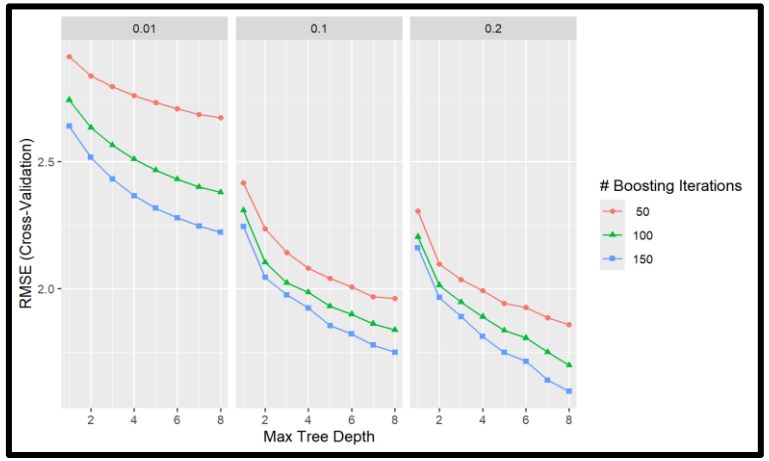



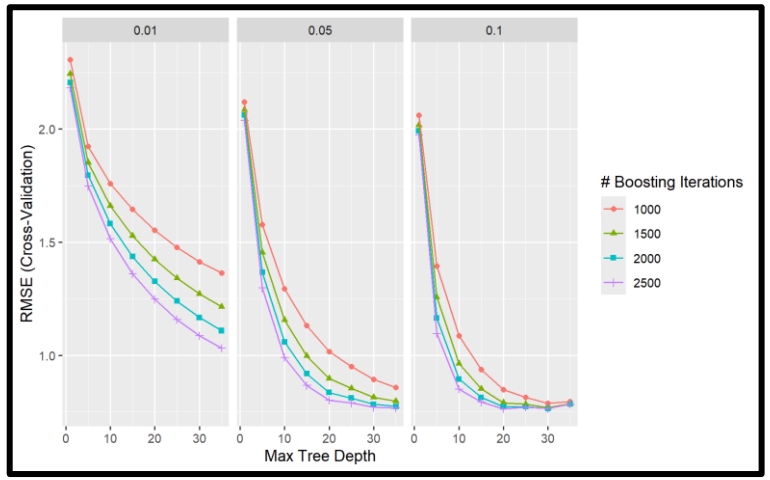


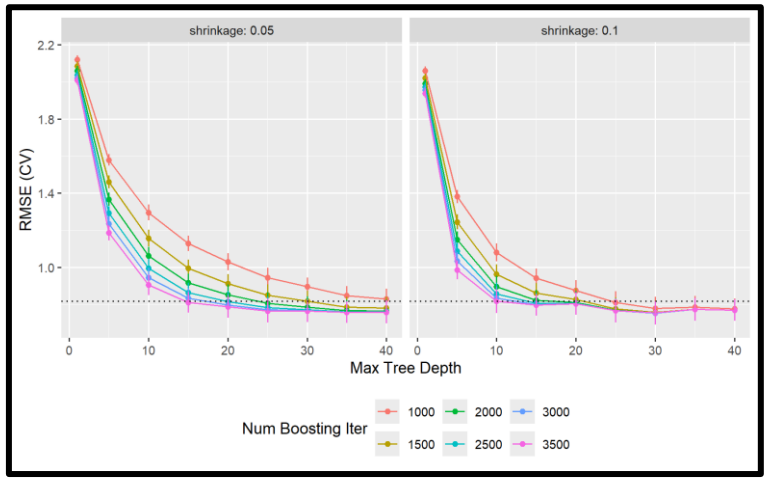

# Optimal Random Variables

# Boosting – Tuning



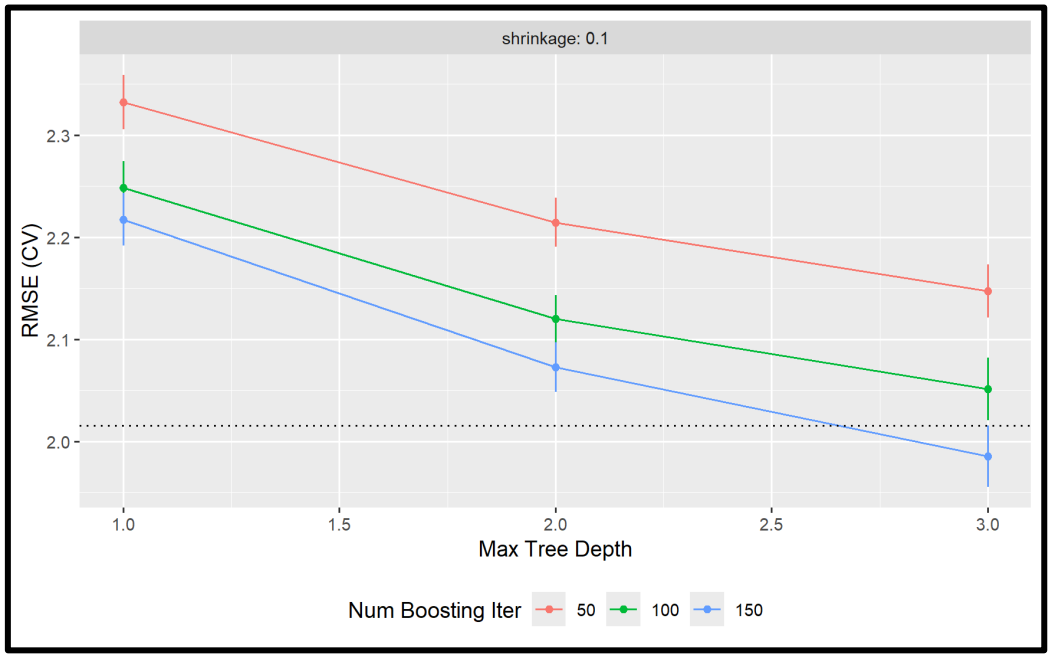



# Boosting - Model

| RMSE | 2.263 |
|---|---|

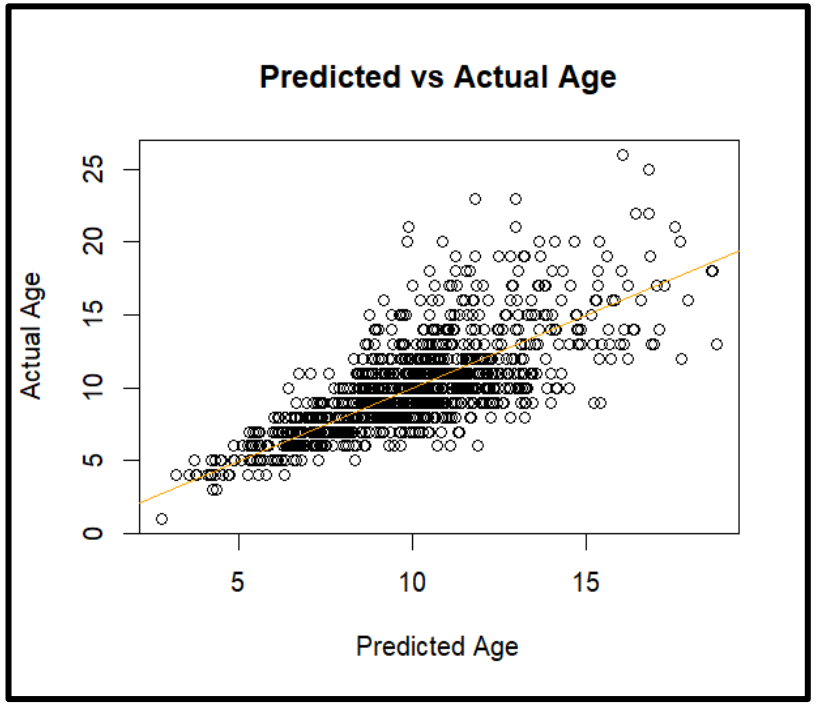


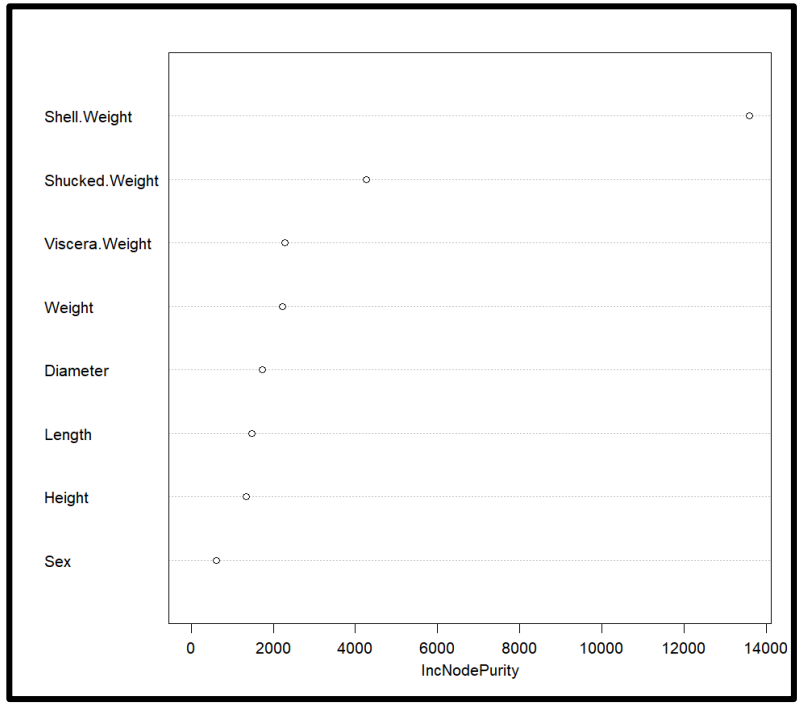

| # Trees | Depth | Shrinkage |
|---|---|---|
| 150 | 3 | .1 |

# Bagging

| RMSE | 2.127 |
|---|---|

Bagging is simply a special case of a random forest with m = p. Therefore, we use the function randomForest() to perform both random forests and bagging

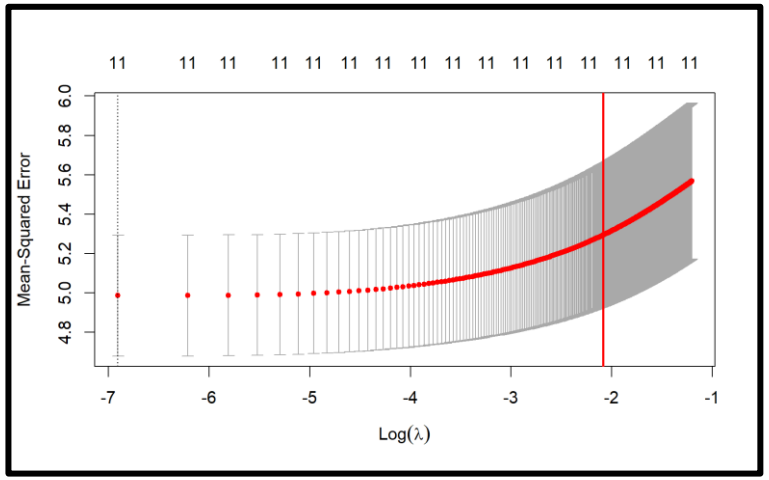


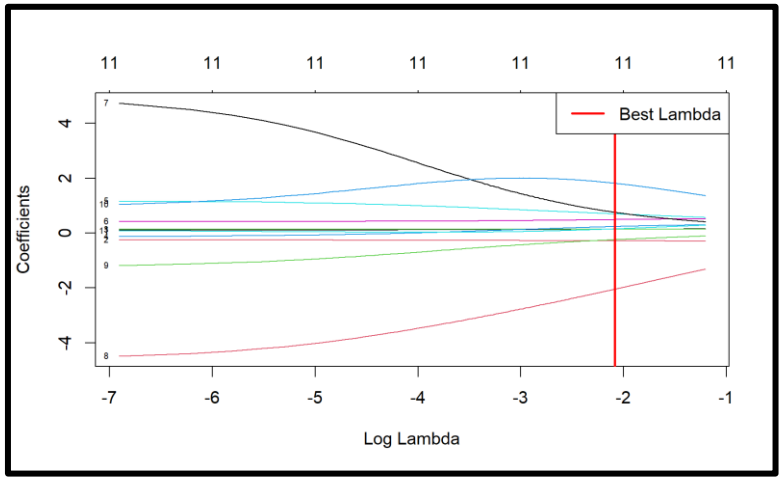




| mtry | ntree |
|---|---|
| 8 | 200 |

# Ridge Regression

| RMSE | 2.259 |
|---|---|



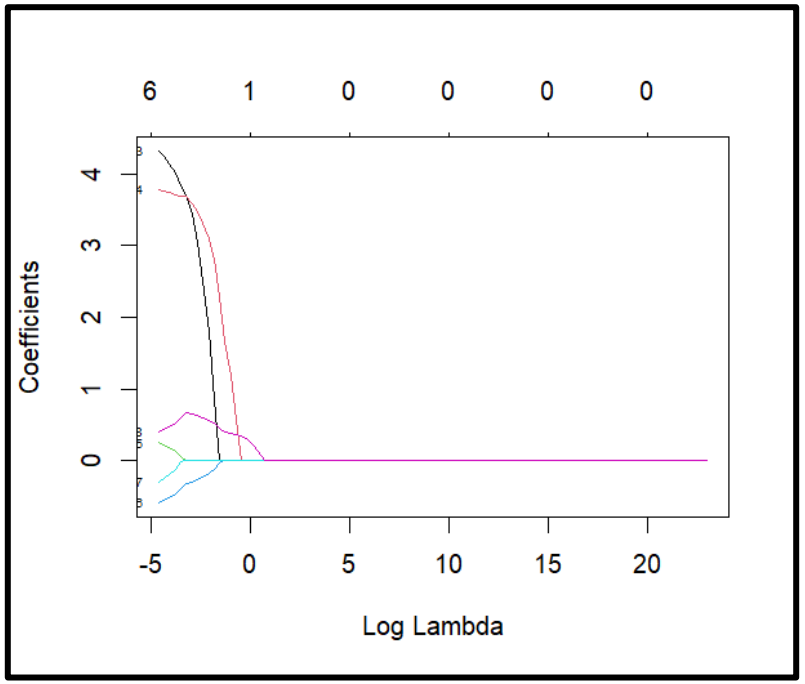



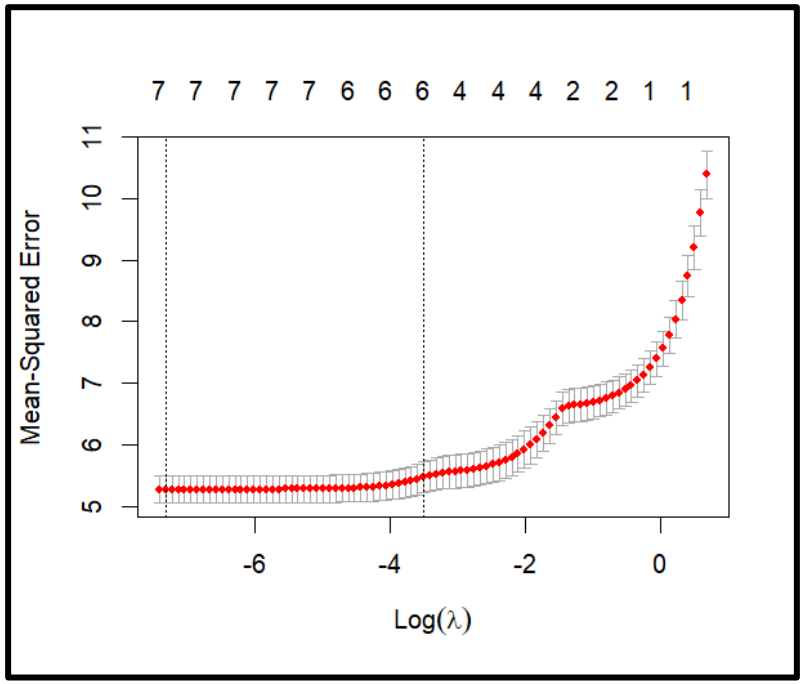

| | s0 |
|---|---|
| (Intercept) | 9.9502727 |
| SexF | 0.1389801 |
| SexI | -0.2823239 |
| SexM | 0.1399685 |
| Length | 0.2310161 |
| Diameter | 0.6998818 |
| Height | 0.4907809 |
| Weight | 0.7744060 |
| `Shucked Weight` | -2.0483138 |
| `Viscera Weight` | -0.2473548 |
| `Shell Weight` | 1.8036650 |
| Shell_Weight_Ratio | 0.1525858 |

| Lambda |
|---|
| .124 |

# Lasso Regression

| RMSE | 2.141 |
|---|---|




| | |
|---|---|
| (Intercept) | 2.6195140 |
| Sex | . |
| Length | . |
| Diameter | 4.6424868 |
| Height | 4.4815973 |
| Weight | 0.2372040 |
| Shucked.Weight | -0.6158695 |
| Viscera.Weight | -0.2060053 |
| Shell.Weight | 0.3602727 |
| Shell.Weight.Ratio | 1.1353637 |

# Multiple Linear Regression

RMSE before engineering: **2.303**

RMSE with shell weight ratio: **2.127**

**Important predictors**:

- Sexl: intersex likely to be younger
  - Could be the case that they cannot determine sexes in younger crabs, or that it is difficult
- Diameter (ft):
  - For every foot of increase in diameter we would expect an increase in age of **five years**
- Height (ft):
  - For every foot of increase in height we would expect an increase in age of **3.66 years**
- Weight - all forms (lbs.):
  - Shell weight: for every pound we would expect an increase in age of **0.34 years** (roughly four months)

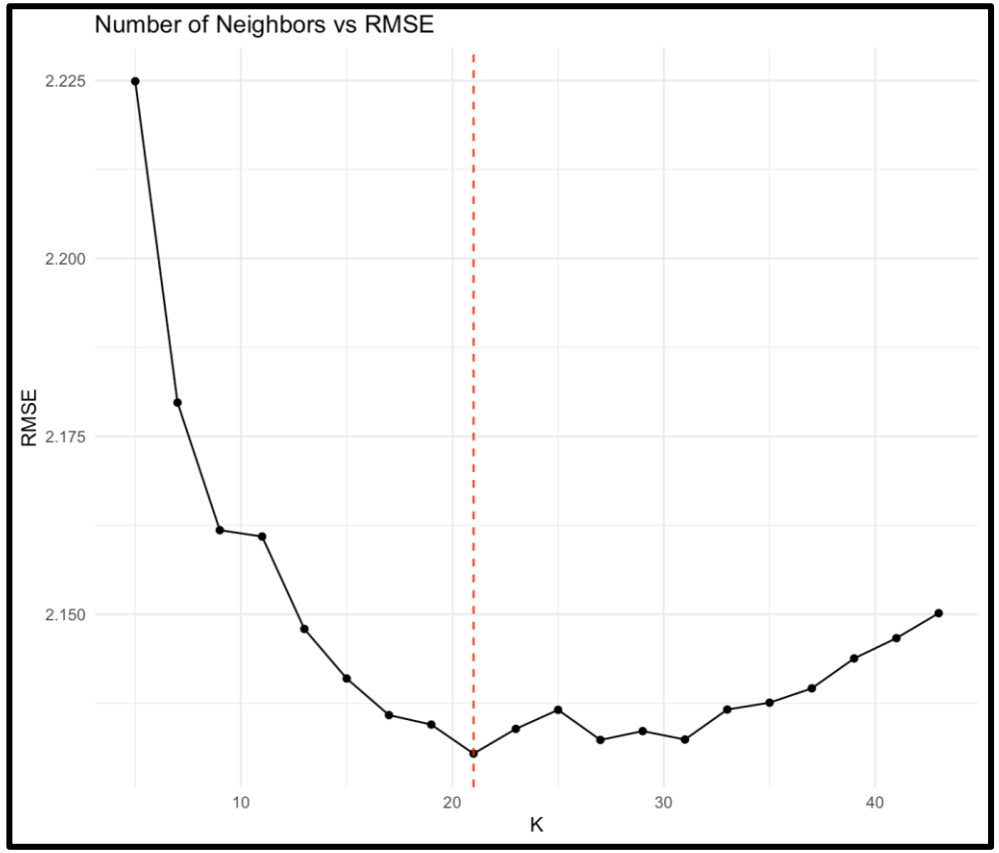
```
Residuals:
    Min      1Q  Median      3Q     Max
-8.7915 -1.3049 -0.3103  0.8600 11.8823

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.74020    0.33428  11.189  < 2e-16 ***
SexI            -0.78682    0.11801  -6.668 3.07e-11 ***
SexM             0.03690    0.09558   0.386    0.699
Length          -0.43112    0.82377  -0.523    0.601
Diameter         5.19253    1.02405   5.071 4.20e-07 ***
Height           3.66096    0.64931   5.638 1.87e-08 ***
Weight           0.29734    0.02896  10.266  < 2e-16 ***
Shucked.Weight -0.70339    0.03279 -21.451  < 2e-16 ***
Viscera.Weight -0.32988    0.05217  -6.323 2.93e-10 ***
Shell.Weight     0.34063    0.04548   7.489 8.96e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.18 on 3104 degrees of freedom
Multiple R-squared:  0.5441,    Adjusted R-squared:  0.5428
F-statistic: 411.6 on 9 and 3104 DF,  p-value: < 2.2e-16
```

| RMSE | 2.127 |
|---|---|

# K-Nearest Neighbors (KNN)

| RMSE | 2.265 |
|---|---|

- The K-Nearest Neighbors (KNN) had an out of sample RMSE of 2.265, putting in the bottom half of performers.
- The figure to the right shows the k neighbors with their corresponding RMSE values (in sample).
- The lowest RMSE can be found at **k=21**.
- From k=15 to k=37 the RMSE values hovered around 2.13 (in sample)



# Conclusion

## Goal

Predict crab **age** based on their **physical traits**

## Results

- **Random Forest & MLR** yielded the **best fit**
- Across all the models, **Shell Weight** turned out to be of the **most importance**
- The addition of **Shell Weight Ratio reduced out of sample RMSE** across the board
- **Bagging** had the **largest improvement** after including Shell Weight Ratio

# Questions?