Terminal vs non-terminal predictors of age in Blacklip Abalone (Haliotis rubra)

Barry Crouch

07/01/2020

Introduction

Blacklip abalone (H. rubra) are aquatic gastropods belonging to the genus Haliotis. Their shells, prized for their pearlescent colors are composed of alternating layers of calcium carbonate and protein. Each year a new protein / carbonate layer is deposited. The age of an abalone can therefore be determined by cutting a cross section through the shell, staining and examination under microscope. This proces is labor intensive and time consuming and so numerous efforts have been made to estimate the age of abalone using only rapidly obtainable metrics. However many parameters strongly predictive of age such as dry shell weight require that the animal is culled. This could present a significant ethical conundrum for researchers monitoring the wild population of endangered abalone species. Here we will first examine to what degree the age of an abaline can be determined using regression modelling based on parameters routinely collected by biologists. We shall also examine whether similar results can be obtained when only measurements that do not require culling of the animal are used.

Libraries

In addition to base R, This project utilized the tidyverse and caret packages. The knitr and gridExtra packages also utilized during creation of this report.

Methods

The Abalone dataset

This project made use of the publically available 'Abalone data set', a subset of data collected as part of a previously published population biology study Nash et al. (1978). The version used here is presently available via the University of California, Irvine's machine learning repository. A copy of the dataset (.csv format) was also uploaded to the edx platform alongside this report.

http://archive.ics.uci.edu/ml/datasets/Abalone

The following code can be used to load the abalone dataset to a dataframe with column names as specified in the documentation (see 'attribute information' in URL above). An additional binary column is also added at this stage (described below). Numeric variables were also scaled via division by 200 in the UCI hosted data set. This is reversed here to return data to their original (standard) units.

```r
#read data from file to dataframe
fname = 'abalone.csv'
alldata = data.frame(read.csv(file=fname, header=FALSE, sep=","))
#rename columns as described in documentation
alldata = alldata %>% rename(sex = V1, length = V2, diameter = V3, height = V4,
                             whole_weight = V5, shucked_weight = V6, viscera_weight = V7,
                             shell_weight = V8, rings = V9)
#add new binary variable to denote adult or infant
alldata = alldata %>% mutate(adult = !str_detect(sex,'I'))
alldata[,2:8] = alldata[,2:8]*200
```

The dataframe now consists of 10 variables (columns) recorded for 4177 animals (rows). The first 10 entries are displayed below.

sex	length	diameter	height	whole_weight	shucked_weight	viscera_weight	shell_weight	rings	adult
M	91	73	19	102.8	44.9	20.2	30	15	TRUE
M	70	53	18	45.1	19.9	9.7	14	7	TRUE
F	106	84	27	135.4	51.3	28.3	42	9	TRUE
M	88	73	25	103.2	43.1	22.8	31	10	TRUE
I	66	51	16	41.0	17.9	7.9	11	7	FALSE
I	85	60	19	70.3	28.2	15.5	24	8	FALSE
F	106	83	30	155.5	47.4	28.3	66	20	TRUE
F	109	85	25	153.6	58.8	29.9	52	16	TRUE
M	95	74	25	101.9	43.3	22.5	33	9	TRUE
F	110	88	30	178.9	62.9	30.2	64	19	TRUE

In brief these variables represent;

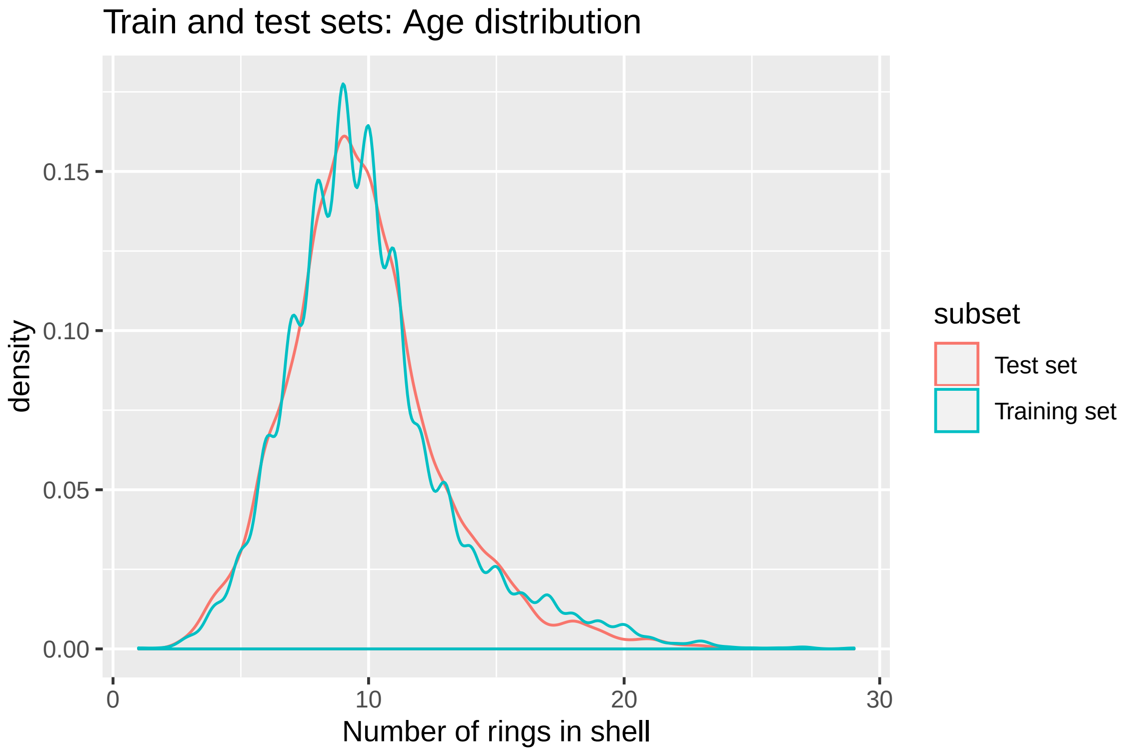
- Sex (non-terminal) - The sex of the animal (M = male, F = female, I = infant)
- Length (non-terminal) - Length of shell at longest point (mm)
- Diameter (non-terminal) - Measurement of shell perpendicular to axis of length measurement (mm)
- Height (non-terminal) - Height measured with animal in shell (mm)
- Whole weight (non-terminal) - Weight of whole animal (g)
- Shucked weight (terminal) - Weight of body after removal from shell (g)
- viscera weight (terminal) - Weight of gut after bleeding (g)
- Shell weight (terminal) - Weight of shell after being dried post removal of animal (g)
- Rings - Rings counted after sectioning and staining of shell (+1.5 converts n rings to age in years)
- Adult - Boolean simply indicating whether or not animal has reached sexual maturity

Training and testing partitions

Before exploring the data further it is nessecary to draw off a fraction of the data to used for final validation of models. The original dataframe 'fulldata' is divided into a training set (80% of data) and a test set (remaining 20%). These are allocated to the variable names 'trainset' and 'testset' rerspectively.

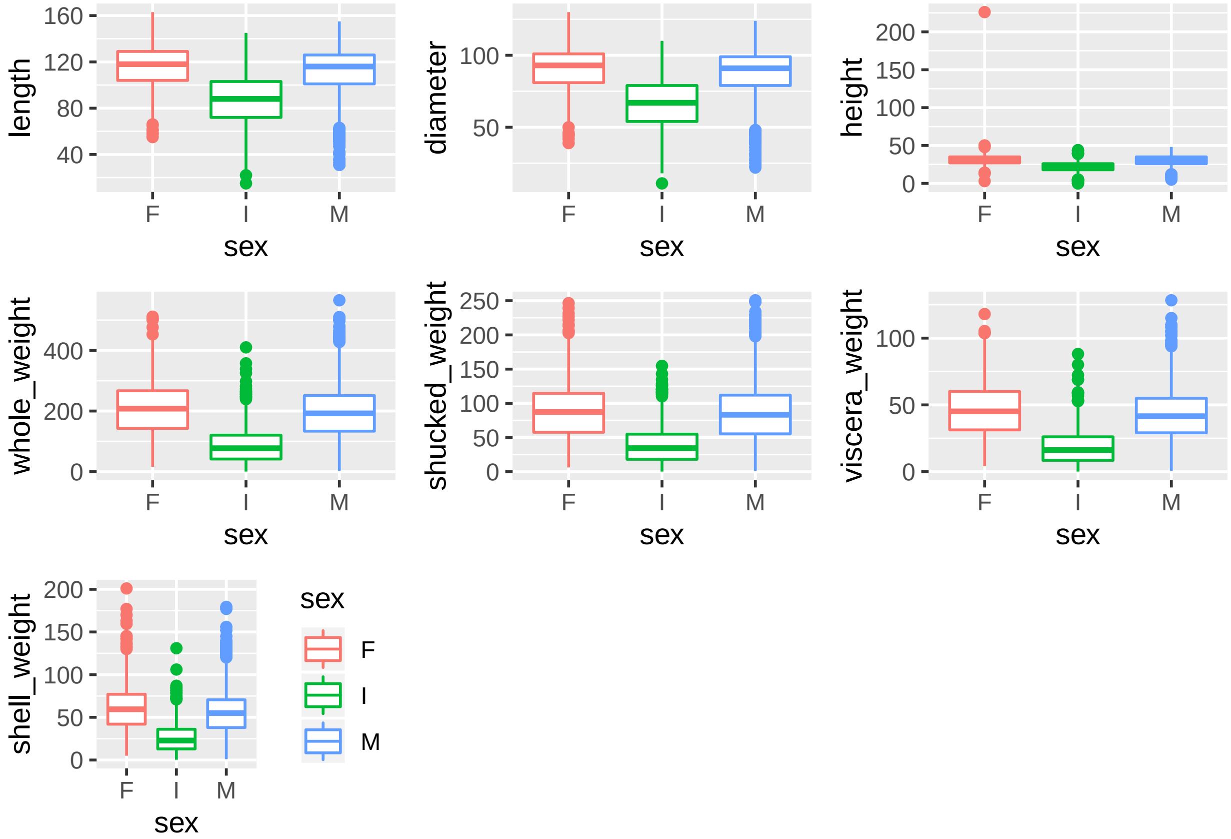
```
#draw off validation set (20%)
set.seed(1988,sample.kind = 'Rounding')
inds = createDataPartition(alldata$rings, p = 0.2, times = 1, list = FALSE)
trainset = alldata[-inds,]
testset = alldata[inds,]
```

The training and testibg subset are approximately equivalent in their representation of subject age (ring count). The test set is therefore representative of the training data. The (now redundant) dataframe 'fulldata' can be erased to recover memory.



Exploratory analysis 1: Effect of sex / maturity

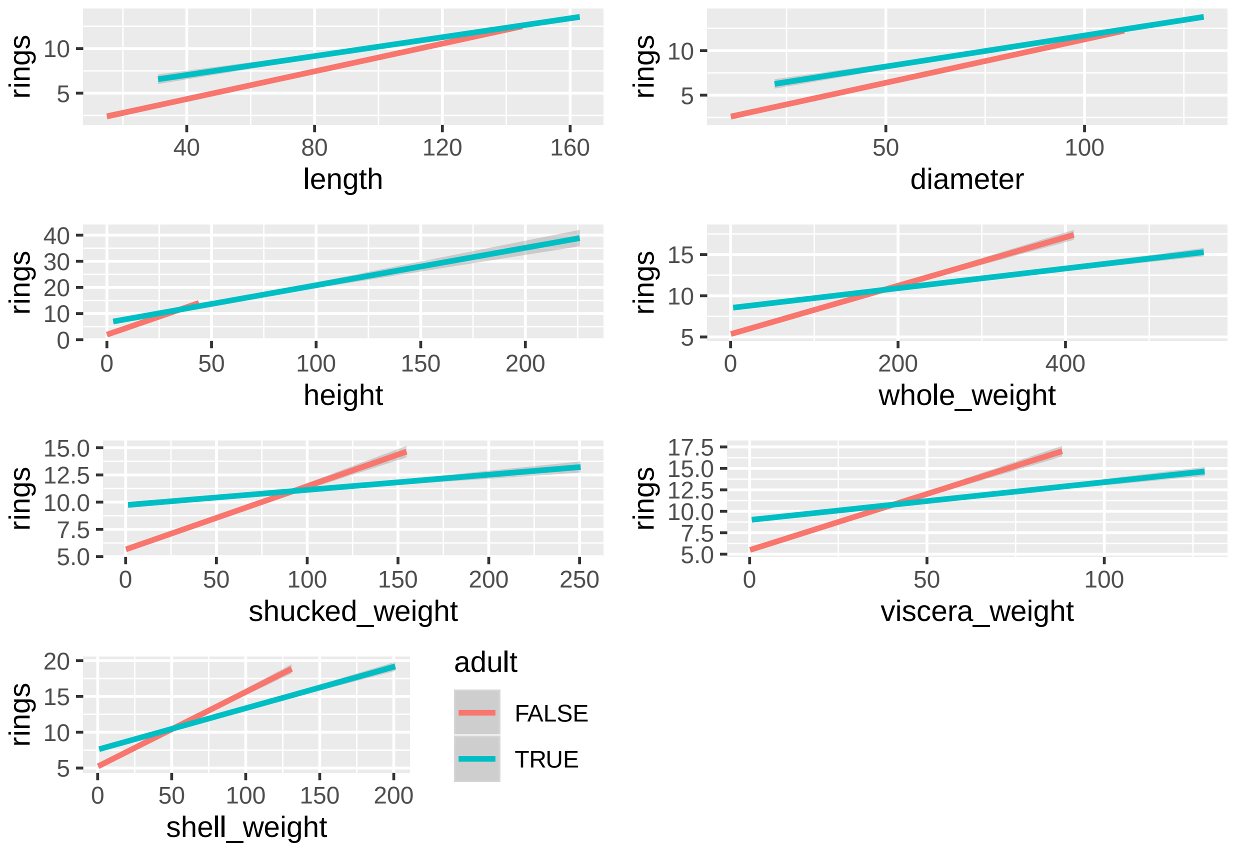
Abalone species do not exhibit sexual dimorphism Ali et al. (2009) therefore it is unlikely that predictors contained in the training set will vary substantially by sex. Howeve it is highly likely that they will vary by whether or not the animal has reached sexual maturity. A preliminary visualization is therefore required to confirm these assumptions.



The results indicate that there is little point in discriminating between male and female abalone but it may be adviasble to stratify models according to whether the animal has reached maturity or not.

Exploratory analysis 2: Relationship between predictors and ring count in adult vs infant abalone

Most organisms do not grow continuously throught their lifespan. It is therefore intuitive that the physical dimensions of infant abalone will be more predictive of age than for adults whos physical growth has plateaud. If the relationship of predictive variables to age does in fact depend upon life stage, this would imply that adault and infant abalone are best modelled seperately. This can be explored by plotting ring count against each predictor for adults and infants individually using geom_smooth with method set to general linear model.



The results indicate that the relationship between ring count and each predictor is indeed steeper among infant abalone (adult = FALSE) than for adults (adult = TRUE). Therefore it is adviasable that separate models are applied to predict the age of infant and adult abalone.

Infant abalone

The training and test datasets were filtered to produce train and test data containing only infant abalone. It is crucial to note that this filtration was performed using the 'adult' boolean variable, derived from the sex categorical variable.

```
infant_train = filter(trainset,!adult)
infant_test = filter(testset,!adult)
```

Naive benchmark

A benchmark for model performance was calculated by determining the root mean squared error (RMSE) between actual ring counts in the training set and a vector of the same length where every entry is set to the average ring count. The result is approximately 2.46 rings. Age estimates based on this model alone can therefore be expected to be off by 2 and a half years on average.

```
#RMSE calculation function
rmse = function(yhat,y){
  e = y-yhat
  se = e^2
  mse = mean(se)
  sqrt(mse)}
mu = mean(infant_train$rings)
y = infant_train$rings
yhat = (y*0)+mu
#naine baseline
rmse(y,yhat)
```

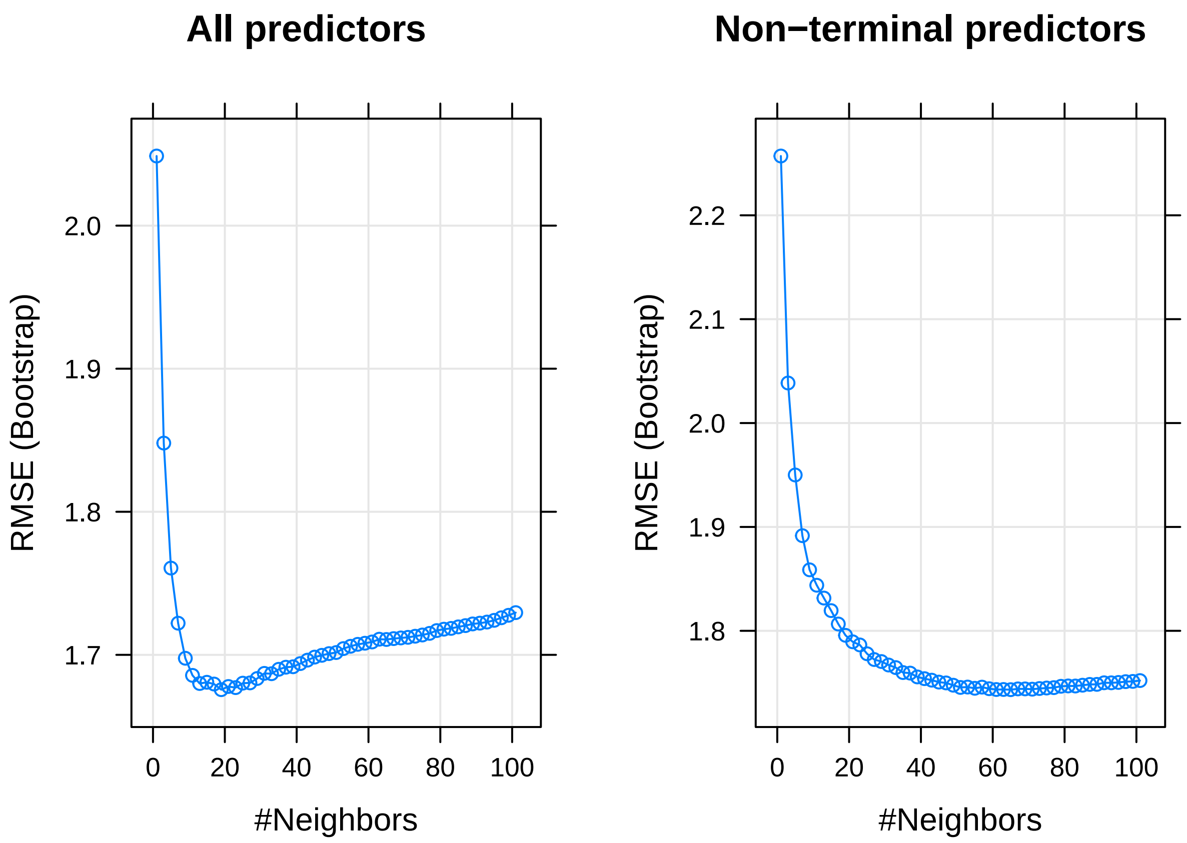
```
## [1] 2.458858
```

Model training

1 linear model, 1 general linear model and 1 k-nearest neighbours (KNN) model were trained on the infant tarining set using all available predictors. A further 3 corresponding models were trained using only non-terminal predictors. KNN models were tuned to minimal RMSE (bootstrap) using k values from 1 to 101. Trained models were then tested using the infant test set. Performance on the test set was quantified by RMSE between actual (y) and predicted (yhat) values.

```
#Model training / tuning
KNN_tunegrid = data.frame(k = seq(1,101,2))
LM_terminal = train(rings~length+diameter+height+whole_weight+shucked_weight+
                      viscera_weight+shell_weight, method = 'lm',data=infant_train)
GLM_terminal = train(rings~length+diameter+height+whole_weight+shucked_weight+
                       viscera_weight+shell_weight,method = 'glm',data=infant_train)
KNN_terminal = train(rings~length+diameter+height+whole_weight+shucked_weight+
                       viscera_weight+shell_weight,method = 'knn',data=infant_train,
                     tuneGrid = KNN_tunegrid)
LM_nonterminal = train(rings~length+diameter+height+whole_weight,
           method = 'lm',data=infant_train)
GLM_nonterminal = train(rings~length+diameter+height+whole_weight,
           method = 'glm',data=infant_train)
KNN_nonterminal = train(rings~length+diameter+height+whole_weight,
           method = 'knn',data=infant_train,tuneGrid = KNN_tunegrid)

#KNN model tuning curves
p1 = plot(KNN_terminal,main='All predictors')
p2 = plot(KNN_nonterminal,main='Non-terminal predictors')
grid.arrange(p1, p2, ncol = 2)
```



```r
#Model testing
lm_t_perf = rmse(predict(LM_terminal,infant_test),t(infant_test$rings))
lm_n_perf = rmse(predict(LM_nonterminal,infant_test),t(infant_test$rings))
glm_t_perf = rmse(predict(GLM_terminal,infant_test),t(infant_test$rings))
glm_n_perf = rmse(predict(GLM_nonterminal,infant_test),t(infant_test$rings))
knn_t_perf = rmse(predict(KNN_terminal,infant_test),t(infant_test$rings))
knn_n_perf = rmse(predict(KNN_nonterminal,infant_test),t(infant_test$rings))
```

Results 1 - Performance summary

The below table provides a summary of the predicted (bootstrap) RMSE for each model as well as the 'actual' RMSE obtained from testing on the infant test set. The difference between the actual RMSE of models using all parameters vs those using only non-terminal parameters in the loss column (non-terminal RMSE - all parameter RMSE). The results indicate first that all models out-performed the naive benchmark. While model methods were broadly equivalent simple linear models achieved the highest accuracy. Critically models using terminal predictors consistently outperformed those using terminal parameters. In real terms however the difference in performance is negligable (~0.15 years). As age is calculated as an integer this discrepancy is likely of little consequence indicating that, at least for infants, terminal methods do not offer a significant benefit over those which do not require cull.

Method	terminal_predicted	terminal_actual	nonterminal_predicted	nonterminal_actual	loss
Linear model	1.643261	1.715064	1.707083	1.856488	0.1414246
General linear model	1.652449	1.715064	1.697728	1.856488	0.1414246
K-nearest neighbours	1.675652	1.773379	1.743365	1.879986	0.1066071

Adult abalone

In this section the precise steps used for the infant datasets are repeated for the adult datasets. For the sake of brevity, and to avoid repetition only the corresponding code is presented.

```
adult_train = filter(trainset,adult)
adult_test = filter(testset,adult)
```

Naive benchmark

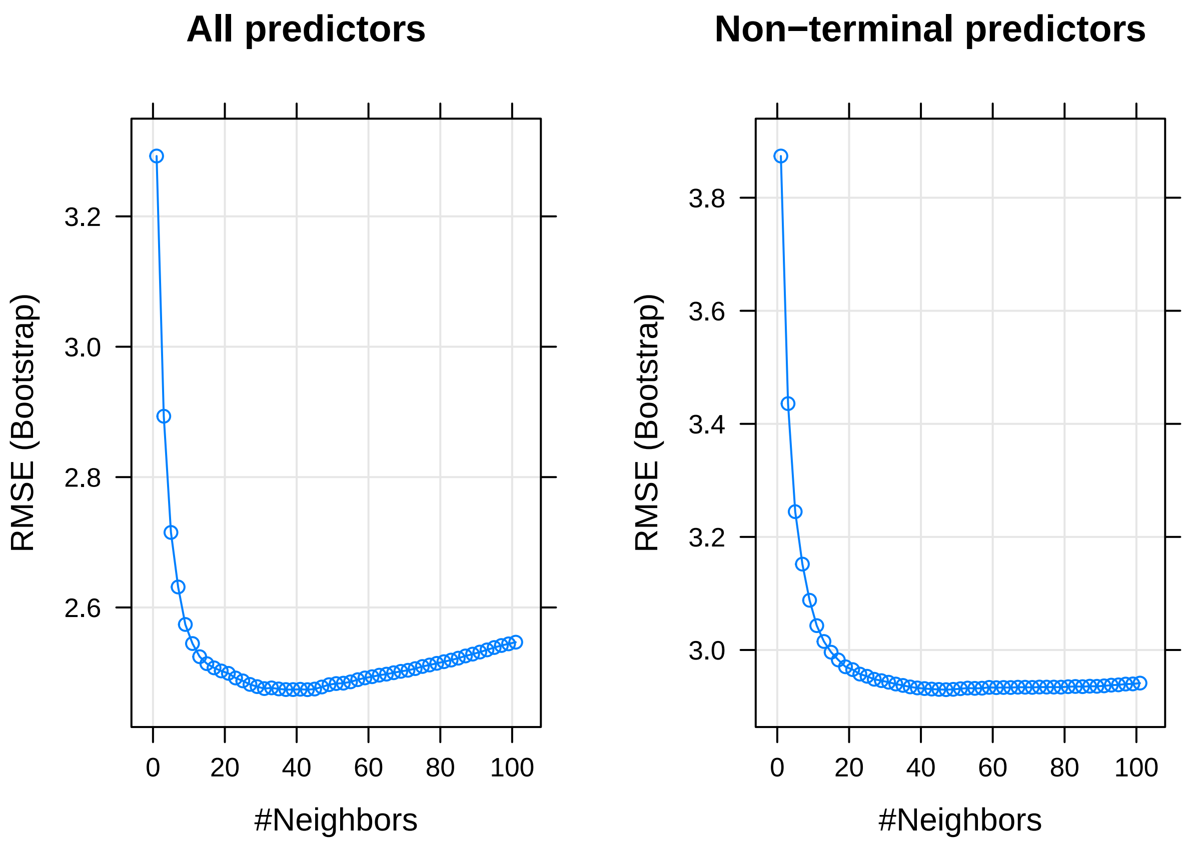
```
mu = mean(adult_train$rings)
y = adult_train$rings
yhat = (y*0)+mu
#naine baseline
rmse(y,yhat)
```

```
## [1] 3.151345
```

Model training

1 linear model, 1 general linear model and 1 k-nearest neighbours (KNN) model were trained on the infant tarining set using all available predictors. A further 3 corresponding models were trained using only non-terminal predictors. KNN models were tuned to minimal RMSE (bootstrap) using k values from 1 to 101. Trained models were then tested using the infant test set. Performance on the test set was quantified by RMSE between actual (y) and predicted (yhat) values.

```
#Model training / tuning
KNN_tunegrid = data.frame(k = seq(1,101,2))
LM_terminal = train(rings~length+diameter+height+whole_weight+shucked_weight+
                      viscera_weight+shell_weight, method = 'lm',data=adult_train)
GLM_terminal = train(rings~length+diameter+height+whole_weight+shucked_weight+
                       viscera_weight+shell_weight,method = 'glm',data=adult_train)
KNN_terminal = train(rings~length+diameter+height+whole_weight+shucked_weight+
                       viscera_weight+shell_weight,method = 'knn',data=adult_train,
                     tuneGrid = KNN_tunegrid)
LM_nonterminal = train(rings~length+diameter+height+whole_weight,
           method = 'lm',data=adult_train)
GLM_nonterminal = train(rings~length+diameter+height+whole_weight,
           method = 'glm',data=adult_train)
KNN_nonterminal = train(rings~length+diameter+height+whole_weight,
           method = 'knn',data=adult_train,tuneGrid = KNN_tunegrid)
#KNN model tuning curves
p1 = plot(KNN_terminal,main='All predictors')
p2 = plot(KNN_nonterminal,main='Non-terminal predictors')
grid.arrange(p1, p2, ncol = 2)
```



```
#Model testing
lm_t_perf = rmse(predict(LM_terminal,adult_test),t(adult_test$rings))
lm_n_perf = rmse(predict(LM_nonterminal,adult_test),t(adult_test$rings))
glm_t_perf = rmse(predict(GLM_terminal,adult_test),t(adult_test$rings))
glm_n_perf = rmse(predict(GLM_nonterminal,adult_test),t(adult_test$rings))
knn_t_perf = rmse(predict(KNN_terminal,adult_test),t(adult_test$rings))
knn_n_perf = rmse(predict(KNN_nonterminal,adult_test),t(adult_test$rings))
```

Results 2 - Performance summary

In line with pedictions the RMSE values for all models including the naive model were larger for adult abalone than for infants. Again all models outperformed the naive benchmark however in this case, the k-nearest neighbors method proved more accurate. Again excluding terminal parameters decreased the accuracy of all methods by ~0.42 - 0.45 rings.

Method	terminal_predicted	terminal_actual	nonterminal_predicted	nonterminal_actual	loss
Linear model	2.485357	2.191981	3.100068	2.612229	0.4202487
General linear model	2.504568	2.191981	3.008042	2.612229	0.4202487
K-nearest neighbours	2.473899	2.112861	2.929803	2.562885	0.4500237

Conclusion

The result of this short project indicate that (at least for the models deployed here) higher accuracy is obtained in the prediction of age in both adult and infant abalone when parameters that require culling of the animal are utilized. However, in the case of infant abalone the reduction in accuracy is particularly small and may be considered negligable. This is not to say that the culling of abalone to definitively establish age is unwaranted or ethically unjustified. However a conceptually similar approach may be worthy of consideration by (for example) conservation ecologists to establish whether an acceptably accurate measurement of age can be obtained while preserving population numbers.

References

Ali, A., A. Basmidi, M. Aideed, and Al-Quffail Saeed. 2009. "First Remarks on Abalone Biology (Haliotis Pustulata) on the Northern Coast of Aden Gulf, Yemen." *Journal of Fisheries and Aquatic Science* 4 (May): 210–27. https://doi.org/10.3923/jfas.2009.210.227.

Nash, W. J., T. L. Sellers, S. R. Talbot, A. J. Cawthorn, and W. B. Ford. 1978. *The Population Biology of Abalone (Haliotis Species) in Tasmania. I. Blacklip Abalone (H Rubra) from the North Coast and the Islands of Bass Strait.* Technical Report (Tasmania. Marine Laboratories). Sea Fisheries Division, Marine Research Laboratories - Taroona, Department of Primary Industry and Fisheries, Tasmania. https://books.google.co.uk/books?id=uLDOQwAACAAJ.