Project Proposal

Jake Kaihewalu (4012753) Jasmine Kellogg (4266367) Andrew Nguyen (5495395)
Masaki Osato (4222154)

February 20, 2020

Research Question

Abalone age is conventionally determined by counting its rings, but it is a tedious and invsave process.

Our research question is: Can abalone age be determined by other variables other than its rings, and what model does this best?

Data

Our response variable is the recorded number of rings. There are 8 predictors, one being categorical and the rest being numeric.

Analysis Plan

A quick analysis at descriptive statistics would include frequency on 'sex' and mean and median calculations on the rest of the predictors. Some exploratory graphics we plan to include are a scatterplot matrix to look at individual interactions and distribution plots for each variable. The simple model fit for this proposal was a logistic regression model based on a new varible that indicated whether the number of rings are higher or lower than the median. To answer our question we will explore different regression models along with regression trees to predict number of rings. In addition, we will also group the number of rings into a few classes in order to use K-NN to see if this kind of classification is more suitable to acheiving our goal. We will use a mixture of AIC/BIC on our regression models and cross validation on our regression tree and KNN models. We plan to use 70% of our dataset for training and 30% for testing/training since we have a moderate number of observations.

References

Abalone Data Set. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Data Overview

```
library(plyr)
## read in data
abalone_data_full <- read.csv("abalone_data.csv", header=FALSE)
abalone_data_full <- rename(abalone_data_full, c('V1' = 'sex', 'V2' = 'length',
                       'V3' = 'diameter','V4' = 'Height',
                       'V5'='whole_weight', 'V6' = 'shucked_weight',
                       'V7' = 'viscera_weight', 'V8' = 'shell_weight',
                       'V9' = 'rings'))
```

```
library(magrittr)
##classify rings
abalone_data = abalone_data_full %>%
mutate(old=as.factor(ifelse(rings <= median(rings), 1, 0)))

## missingness
apply(is.na(abalone_data), 2, sum)
```

```
##            sex         length       diameter         Height   whole_weight
##              0              0              0              0              0
## shucked_weight viscera_weight   shell_weight          rings            old
##              0              0              0              0              0
```

“‘

```
## Split the data
set.seed(3)
id <- sample(1:nrow(abalone_data), 0.75*nrow(abalone_data))
abalone_data.train <- abalone_data[id,]
abalone_data.test <- abalone_data[-id,]
dim(abalone_data.train)
```

```
## [1] 3132   10
```

```
dim(abalone_data.test)
```

```
## [1] 1045   10
```

```
#sumarize response
summary(abalone_data.train$old)
```

```
##    0    1
## 1535 1597
```

```
#basic model
glm.fit <- glm(old ~ whole_weight + length + Height, data = abalone_data.train, family = binomial)
summary(glm.fit)
```

```
##
## Call:
## glm(formula = old ~ whole_weight + length + Height, family = binomial,
##     data = abalone_data.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3974  -0.8607   0.2796   0.8416   6.7391
```

```
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    4.4175     0.4668   9.463  < 2e-16 ***
## whole_weight  -1.2673     0.3229  -3.925 8.67e-05 ***
## length        -0.1728     1.2886  -0.134    0.893
## Height       -23.2685     2.8894  -8.053 8.07e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4340.6  on 3131  degrees of freedom
## Residual deviance: 3284.8  on 3128  degrees of freedom
## AIC: 3292.8
##
## Number of Fisher Scoring iterations: 5
```