

Mushrooms

Ina Ding

Introduction

This section includes an introduction to the project motivation, data, and research question.

The research question and motivation are clearly stated in the introduction, including citations.

The primary goal of this project is to best predict whether or not a mushroom is poisonous depending on various physical characteristics, rarity, and habitat of the fungus. The data set consists of 8124 hypothetical samples, constructed from the Audobon Society Field Guide. The samples correspond to 23 species of mushrooms from the Agaricus and Lepiota Families. Each mushroom is categorized as either poisonous or edible, with mushrooms ‘not recommended for eating’ or of unknown edibility counted as poisonous. Though the observations are hypothetical mushrooms, analyzing them can still provide beneficial results that can be applied to help identify the edibility of the near 14,000 existing species of mushrooms.

1. cap-shape:	bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
2. cap-surface:	fibrous=f,grooves=g,scaly=y,smooth=s
3. cap-color:	brown=n,buff=b,cinnamon=c,gray=g,green=r, pink=p,purple=u,red=e,white=w,yellow=y
4. bruises?:	bruises=t,no=f
5. odor:	almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s
6. gill-attachment:	attached=a,descending=d,free=f,notched=n
7. gill-spacing:	close=c,crowded=w,distant=d
8. gill-size:	broad=b,narrow=n
9. gill-color:	black=k,brown=n,buff=b,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e, white=w,yellow=y
10. stalk-shape:	enlarging=e,tapering=t

11. stalk-root:	bulbous=b,club=c,cup=u,equal=e, rhizomorphs=z,rooted=r,missing=
12. stalk-surface-above-ring:	fibrous=f,scaly=y,silky=k,smooth=s
13. stalk-surface-below-ring:	fibrous=f,scaly=y,silky=k,smooth=s
14. stalk-color-above-ring:	brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
15. stalk-color-below-ring:	brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
16. veil-type:	partial=p,universal=u
17. veil-color:	brown=n,orange=o,white=w,yellow=y
18. ring-number:	none=n,one=o,two=t
19. ring-type:	cobwebby=c,evanescent=e,flaring=f,large=l, none=n,pendant=p,sheathing=s,zone=z
20. spore-print-color:	black=k,brown=n,buff=b,chocolate=h,green=r, orange=o,purple=u,white=w,yellow=y
21. population:	abundant=a,clustered=c,numerous=n, scattered=s,several=v,solita
22. habitat:	grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d

Citations Mushroom. (1987). UCI Machine Learning Repository. <https://doi.org/10.24432/C5959T>.

Methodology

This section includes a brief description of your modeling process. Explain the reasoning for the type of model you're fitting, predictor variables considered for the model including any interactions. Additionally, show how you arrived at the final model by describing the model selection process, interactions considered, variable transformations (if needed), assessment of conditions and diagnostics, and any other relevant considerations that were part of the model fitting process.

Because this data set only contains categorical variables and no one attribute appeared immediately more important in categorizing a mushroom as poisonous or not, we were interested in including all the variables as predictors in our model. After some experimentation with LASSO and all subset variable selection models, there turned out to be so many variables and categories within each variable that it seemed better to just keep all variables in. Some light data cleaning was done, including removing the veil-type variable because all mushrooms in the set had the same veil-type ('partial').

The primary outcome of interest is whether or not a mushroom is poisonous or edible. Since this is a binary outcome (ambiguous or unknown edibilities were deemed poisonous in this data set) using a log-odds model seemed the most reasonable.

```
install.packages("leaps")
```

```

mushrooms <- read.csv("Mushrooms - Sheet1.csv")
no_veil <- subset(mushrooms, select = -c(veil))
library(tidymodels)
library(tidyverse)
library(leaps)
library(glmnet)

#sapply(lapply(no_veil, unique), length)

no_veil$poisonous[no_veil$poisonous == 'p'] <- 1
no_veil$poisonous[no_veil$poisonous == 'e'] <- 0

no_veil$poisonous <- as.numeric(as.character(no_veil$poisonous))

##m1 <- glm(poisonous ~ .,
            ## data = no_veil,
            ## family = "binomial")
##summary(m1)

#m_all <- regsubsets(poisonous ~ .,
                    # data = no_veil,
                    # nbest = 1, nvmax = 5, really.big=T)
#m_all

#cart ... look up for classification tree model

m1 <- lm(poisonous ~ ., data = no_veil)
summary(m1)

```

Call:

```
lm(formula = poisonous ~ ., data = no_veil)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.359e-13	-3.800e-16	0.000e+00	3.800e-16	8.127e-13

Coefficients: (10 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.627e-14	8.119e-15	2.004e+00	0.045076 *
cap.shapec	-2.171e-14	5.420e-15	-4.005e+00	6.27e-05 ***

cap.shapef	5.169e-15	6.161e-16	8.390e+00	< 2e-16	***
cap.shapek	3.035e-15	6.698e-16	4.531e+00	5.95e-06	***
cap.shapes	2.996e-15	2.112e-15	1.419e+00	0.155920	
cap.shapex	1.984e-15	5.902e-16	3.362e+00	0.000778	***
cap.surfaceg	-5.259e-15	6.626e-15	-7.940e-01	0.427454	
cap.surfaces	2.631e-15	3.527e-16	7.461e+00	9.49e-14	***
cap.surfacey	1.007e-15	2.956e-16	3.408e+00	0.000657	***
cap.colorc	-4.406e-15	2.047e-15	-2.152e+00	0.031405	*
cap.colore	-1.825e-15	9.529e-16	-1.916e+00	0.055441	.
cap.colorg	-1.915e-15	9.130e-16	-2.098e+00	0.035975	*
cap.colorn	-2.709e-15	9.316e-16	-2.907e+00	0.003655	**
cap.colorp	-3.426e-15	1.172e-15	-2.922e+00	0.003491	**
cap.colorr	-4.787e-15	3.435e-15	-1.394e+00	0.163495	
cap.coloru	-3.425e-15	3.435e-15	-9.970e-01	0.318813	
cap.colorw	-3.186e-15	9.186e-16	-3.468e+00	0.000527	***
cap.colory	-1.132e-15	9.776e-16	-1.158e+00	0.247060	
bruise	5.000e-01	3.147e-15	1.589e+14	< 2e-16	***
odorc	5.000e-01	3.765e-15	1.328e+14	< 2e-16	***
odorf	-5.000e-01	9.425e-15	-5.305e+13	< 2e-16	***
odorl	1.591e-15	6.620e-16	2.404e+00	0.016259	*
odorm	2.500e+00	1.482e-14	1.686e+14	< 2e-16	***
odorn	-1.500e+00	9.092e-15	-1.650e+14	< 2e-16	***
odorp	-1.000e+00	1.132e-14	-8.834e+13	< 2e-16	***
odors	-5.000e-01	9.441e-15	-5.296e+13	< 2e-16	***
odory	-5.000e-01	9.441e-15	-5.296e+13	< 2e-16	***
gill.attachmentf	-3.190e-16	3.121e-15	-1.020e-01	0.918592	
gill.spacingw	4.868e-16	1.351e-15	3.600e-01	0.718662	
gill.sizen	-1.500e+00	8.521e-15	-1.760e+14	< 2e-16	***
gill.colore	-5.000e-01	5.644e-15	-8.859e+13	< 2e-16	***
gill.colorg	-5.000e-01	5.589e-15	-8.946e+13	< 2e-16	***
gill.colorh	-5.000e-01	5.578e-15	-8.963e+13	< 2e-16	***
gill.colork	-5.000e-01	5.596e-15	-8.935e+13	< 2e-16	***
gill.colorn	-5.000e-01	5.585e-15	-8.952e+13	< 2e-16	***
gill.coloro	-5.000e-01	5.805e-15	-8.613e+13	< 2e-16	***
gill.colorp	-5.000e-01	5.574e-15	-8.970e+13	< 2e-16	***
gill.colorr	-5.000e-01	6.041e-15	-8.277e+13	< 2e-16	***
gill.coloru	-5.000e-01	5.601e-15	-8.927e+13	< 2e-16	***
gill.colorw	-5.000e-01	5.562e-15	-8.989e+13	< 2e-16	***
gill.colory	-5.000e-01	5.756e-15	-8.687e+13	< 2e-16	***
stalk.shapet	-1.000e+00	5.551e-15	-1.802e+14	< 2e-16	***
stalk.rootb	-5.000e-01	3.675e-15	-1.361e+14	< 2e-16	***
stalk.rootc	-3.000e+00	1.721e-14	-1.743e+14	< 2e-16	***
stalk.roote	5.000e-01	3.382e-15	1.478e+14	< 2e-16	***

stalk.rootr	-3.500e+00	1.482e-14	-2.361e+14	< 2e-16	***
ss.abovek	-7.934e-17	7.010e-16	-1.130e-01	0.909892	
ss.aboves	-9.433e-17	5.635e-16	-1.670e-01	0.867064	
ss.abovey	2.220e-15	7.700e-15	2.880e-01	0.773121	
ss.belowk	-4.879e-17	7.010e-16	-7.000e-02	0.944519	
ss.belows	-4.761e-17	5.635e-16	-8.400e-02	0.932676	
ss.belowy	5.000e-01	4.705e-15	1.063e+14	< 2e-16	***
sc.abovec	NA	NA	NA	NA	
sc.abovee	7.428e-17	1.534e-15	4.800e-02	0.961383	
sc.aboveg	-2.195e-17	8.151e-16	-2.700e-02	0.978519	
sc.aboven	-4.625e-18	6.370e-16	-7.000e-03	0.994207	
sc.aboveo	-1.000e+00	7.102e-15	-1.408e+14	< 2e-16	***
sc.abovep	-1.847e-17	6.370e-16	-2.900e-02	0.976875	
sc.abovew	-3.027e-17	7.263e-16	-4.200e-02	0.966758	
sc.abovey	3.000e+00	1.425e-14	2.105e+14	< 2e-16	***
sc.belowc	NA	NA	NA	NA	
sc.belowe	-3.398e-17	1.534e-15	-2.200e-02	0.982331	
sc.belowg	-2.784e-17	8.151e-16	-3.400e-02	0.972751	
sc.belown	-2.112e-17	6.370e-16	-3.300e-02	0.973550	
sc.belowo	NA	NA	NA	NA	
sc.belowp	-2.149e-17	6.370e-16	-3.400e-02	0.973083	
sc.beloww	-5.477e-17	7.263e-16	-7.500e-02	0.939889	
sc.belowy	-9.600e-18	3.371e-15	-3.000e-03	0.997728	
veil.coloro	8.452e-18	1.351e-15	6.000e-03	0.995009	
veil.colorw	NA	NA	NA	NA	
veil.colory	NA	NA	NA	NA	
ring.numo	2.500e+00	1.364e-14	1.833e+14	< 2e-16	***
ring.numt	NA	NA	NA	NA	
ring.typef	1.000e+00	6.825e-15	1.465e+14	< 2e-16	***
ring.typel	NA	NA	NA	NA	
ring.typhen	NA	NA	NA	NA	
ring.typep	5.000e-01	2.862e-15	1.747e+14	< 2e-16	***
spore.colorh	NA	NA	NA	NA	
spore.colork	4.510e-16	1.935e-15	2.330e-01	0.815756	
spore.colorn	-6.587e-17	1.911e-15	-3.400e-02	0.972506	
spore.coloro	-3.252e-17	1.911e-15	-1.700e-02	0.986423	
spore.colorr	2.500e+00	9.127e-15	2.739e+14	< 2e-16	***
spore.coloru	-6.215e-16	2.703e-15	-2.300e-01	0.818125	
spore.colorw	1.500e+00	8.755e-15	1.713e+14	< 2e-16	***
spore.colory	1.777e-27	1.911e-15	0.000e+00	1.000000	
populationc	-1.470e-15	1.632e-15	-9.010e-01	0.367781	
populationn	-1.162e-16	9.459e-16	-1.230e-01	0.902192	
populations	2.558e-17	6.756e-16	3.800e-02	0.969802	

```

populationv      -1.470e-15  9.158e-16 -1.605e+00  0.108445
populationy      -1.392e-15  9.487e-16 -1.467e+00  0.142350
habitatg         -1.030e-17  5.622e-16 -1.800e-02  0.985384
habitatl         1.701e-18  5.201e-16  3.000e-03  0.997391
habitatm        -1.124e-15  9.573e-16 -1.174e+00  0.240468
habitatp         3.402e-18  4.112e-16  8.000e-03  0.993400
habitatw         2.723e-15  9.801e-16  2.778e+00  0.005478 **
habitatw         NA          NA          NA          NA

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.362e-15 on 8038 degrees of freedom

Multiple R-squared: 1, Adjusted R-squared: 1

F-statistic: 2.723e+29 on 85 and 8038 DF, p-value: < 2.2e-16

```
install.packages("glmnet")
```

```
library(glmnet)
```

```
y <- no_veil$poisonous
```

```
x <- model.matrix(poisonous ~ .,
                  data = no_veil)
```

```
m_lasso_cv <- cv.glmnet(x, y, alpha = 1)
```

```
best_lambda <- m_lasso_cv$lambda.min
```

```
best_lambda
```

```
[1] 0.0002523176
```

```
m_best <- glmnet(x, y, alpha = 1, lambda = best_lambda)
```

```
m_best$beta
```

96 x 1 sparse Matrix of class "dgCMatrix"

s0

```

(Intercept)      .
cap.shapec       2.469270e-01
cap.shapef       .
cap.shapek       .
cap.shapes      -4.780328e-03

```

cap.shapex	.
cap.surfaceg	7.481027e-01
cap.surfaces	.
cap.surfacey	7.228459e-04
cap.colorc	-6.613558e-02
cap.colore	.
cap.colorg	.
cap.colorn	-9.789276e-04
cap.colorp	.
cap.colorr	.
cap.coloru	.
cap.colorw	2.720506e-03
cap.colory	.
bruise	6.459105e-04
odorc	9.665179e-01
odorf	7.966668e-01
odorl	-4.046078e-03
odorm	5.659035e-02
odorn	-4.517302e-02
odorp	9.423022e-01
odors	7.963789e-01
odory	7.964045e-01
gill.attachmentf	.
gill.spacingw	-1.719982e-02
gill.sizen	.
gill.colore	.
gill.colorg	.
gill.colorh	.
gill.colork	.
gill.colorn	.
gill.coloro	.
gill.colorp	.
gill.colorr	.
gill.coloru	.
gill.colorw	-1.959178e-04
gill.colory	.
stalk.shapet	6.117153e-03
stalk.rootb	-7.871970e-03
stalk.rootc	-4.573025e-02
stalk.roote	7.148989e-03
stalk.rootr	-8.618155e-01
ss.abovek	2.321158e-03
ss.aboves	.

ss.abovey	-8.035657e-01
ss.belowk	6.740357e-04
ss.belows	.
ss.belowy	8.175232e-01
sc.abovec	3.588411e-02
sc.abovee	.
sc.aboveg	.
sc.aboven	.
sc.aboveo	-7.569320e-03
sc.abovep	1.198385e-04
sc.abovew	.
sc.abovey	4.736608e-01
sc.belowc	2.978803e-06
sc.belowe	.
sc.belowg	.
sc.belown	-9.342302e-03
sc.belowo	-1.576968e-02
sc.belowp	.
sc.beloww	1.350767e-03
sc.belowy	2.835038e-02
veil.coloro	.
veil.colorw	.
veil.colory	2.771655e-01
ring.numo	9.290282e-02
ring.numt	-4.116992e-02
ring.typef	-1.562118e-01
ring.typel	8.587265e-03
ring.typen	.
ring.typep	.
spore.colorh	1.555109e-01
spore.colork	.
spore.colorn	-6.014691e-04
spore.coloro	.
spore.colorr	1.135084e+00
spore.coloru	-2.520651e-02
spore.colorw	1.481708e-01
spore.colory	.
populationc	5.019038e-02
populationn	-7.732554e-04
populations	.
populationv	.
populationy	.
habitatg	.


```

habitatl      .
habitatm      .
habitatp      8.385491e-05
habitatu      .
habitatw      -6.500746e-02

```

```

m2 <- glm(poisonous ~ .,
          data = no_veil,
          family = "binomial")
m2_aug <- augment(m2)

m2_aug <- m2_aug %>%
  mutate(prob = exp(.fitted)/(1 + exp(.fitted)),
         pred_pois = ifelse(prob > 0.5, "Poisonous", "Edible")) %>%
  select(.fitted, prob, pred_pois, poisonous)

table(m2_aug$pred_pois, m2_aug$poisonous)

```

	0	1
Edible	4208	0
Poisonous	0	3916

```
install.packages("rpart.plot")
```

Installing package into '/home/guest/R/x86_64-pc-linux-gnu-library/4.2'
(as 'lib' is unspecified)

```
install.packages("ISLR")
```

Installing package into '/home/guest/R/x86_64-pc-linux-gnu-library/4.2'
(as 'lib' is unspecified)

```

library(rpart)
library(rpart.plot)
library(ISLR)

```

```
no_veil$poisonous = as.factor(no_veil$poisonous)

set.seed(234) train = sample(1:nrow(Carseats.H), 200) Carseats.train=Carseats.H[train,]
Carseats.test=Carseats.H[-train,]

High.test=High[-train]
```