Mushroom Edibility Analysis

Tofu-FC - Huiwen Wang, Rocky Zhang, Darrick Zhang

2024-10-28

Introduction

Project Motivation / Background:

Mushrooms are vital to the general wellness of the ecosystem, decomposing and recycling the nutrients in the soil. Mushrooms also provide a valuable food source full of nutrients for human beings and other important organisms. However, some mushroom species can also be poisonous and harmful.

The importance of this research has been highlighted in a multitude of studies. Take this quote, for example:

> The ingestion of wild and potentially toxic mushrooms is common in the United States, with poison centers logging cases in the National Poison Data System (NPDS) for over 30 years. From 1999 to 2016, there were 133,700 reported cases of mushroom exposure, mostly unintentional and involving children under six years old. While the majority of cases resulted in no or minor harm, there were 704 instances of major harm and 52 fatalities, primarily due to cyclopeptide-producing mushrooms ingested unintentionally by older adults. Misidentification of edible mushroom species is a common cause of poisoning and may be preventable through education (Brandenburg and Ward 2018).

As shown by studies and other similar studies, accurate classification of mushrooms is crucial for preventing poisoning incidents. Many toxic mushroom species closely resemble edible varieties, making it easy for foragers to misidentify them. Thus, our research will focus on what physical features and environmental factors of mushrooms humans can use to identify toxic/poisonous mushrooms in the wild. By conducting a research study on how to distinguish between safe and dangerous species, we can mitigate the incidence of mushroom poisoning and ensure safer foraging practices.

Research Question:

What environmental factors and/or physical features of mushrooms indicate that a wild mushroom is poisonous?

Hypothesis:

Mushrooms in the wild with obvious physical features like white gills, white rings, red caps, or red stems tend to be poisonous. These obvious physical traits are more likely to be spotted by animals, which would provide an evolutionary disadvantage unless they contain certain self-defense mechanisms, such as poison or toxins. Additionally, the habitat and season in which mushrooms are planted and grow may also affect whether they're poisonous. Different temperatures, humidity, and light can affect the production of toxins, which may also affect the edibility of mushrooms.

Data Description:

The data was curated on April 26, 1987, and submitted to the UCI by the National Audubon Society Field Guide. The National Audubon Society conducted extensive field research throughout North America, recording their observations on various aspects of mushrooms. Their research incorporate a wide range of physical characteristics, including size, shape, color, and texture of the mushrooms. Additionally, they documented environmental factors such as the type of habitat and seasonal variations. Importantly, the study also focused on the toxicity of the mushrooms, noting which species were poisonous. This comprehensive dataset provides valuable insights into the relationship between mushrooms and their environments, contributing significantly to the understanding of the factors influencing mushroom toxicity.

Our response variable is `class`, which is a qualitative variable labeled "e" for edible or "p" for poisonous.

Key quantitative predictor variables include `stem.height` and `stem.width`, the height (cm) and width (cm) of the stem of the mushroom. We are also interested in `cap.diameter`, the diameter of the mushroom cap (cm).

Key qualitative predictor variables include `cap.shape`, the shape of the mushroom cap; `gill.color`, the color of the fungi gills, `stem.color`, the color of the mushroom stem; `habitat`, the habitat that the mushroom is grown/found; and `season`, the season that the mushroom is grown/found (spring=s, summer=u, autumn=a, winter=w).

The `cap.shape` codes are:

b	c	x	f	s	p	o
bell	conical	convex	flat	sunken	spherical	others

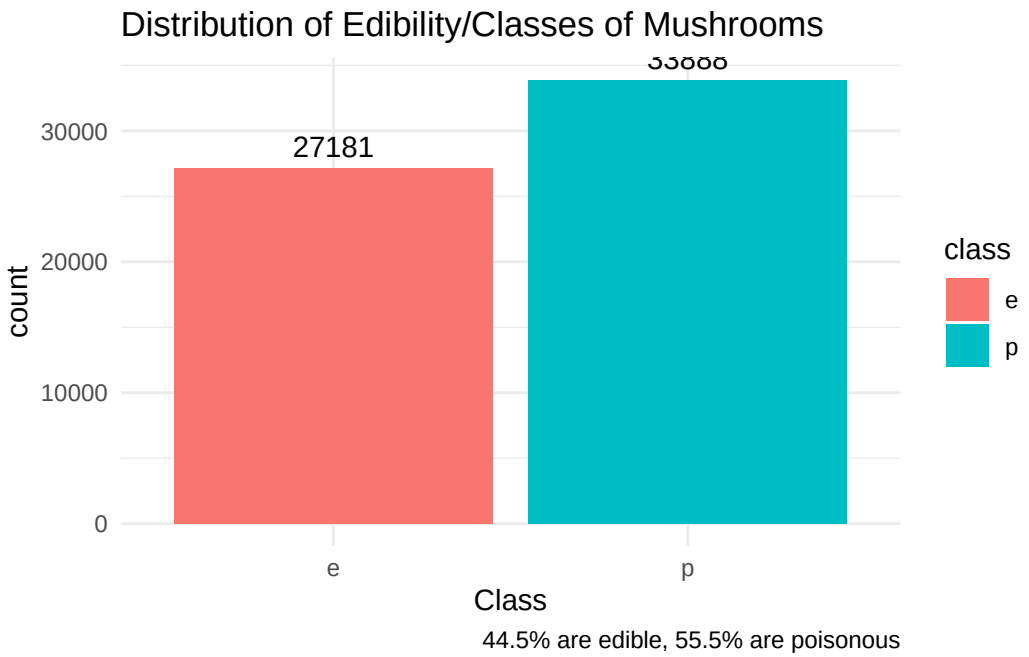
The `gill.color` and `stem.color` are:

n	b	g	r	p	u	e	w	y	l	o	k	f
brown	buff	gray	green	pink	purple	red	white	yellow	blue	orange	black	none

The `habitat` codes are:

g	l	m	p	h	u	w	d
grasses	leaves	meadows	paths	heaths	urban	waste	woods

Exploratory Data Analysis



Looking at the overall distribution of our response variable `class`, most of the mushrooms in our dataset seem to be poisonous ("p"). 33888 of the observations, or 55.5% of them are labeled poisonous, as opposed to 27181 (44.5%) of them as edible.

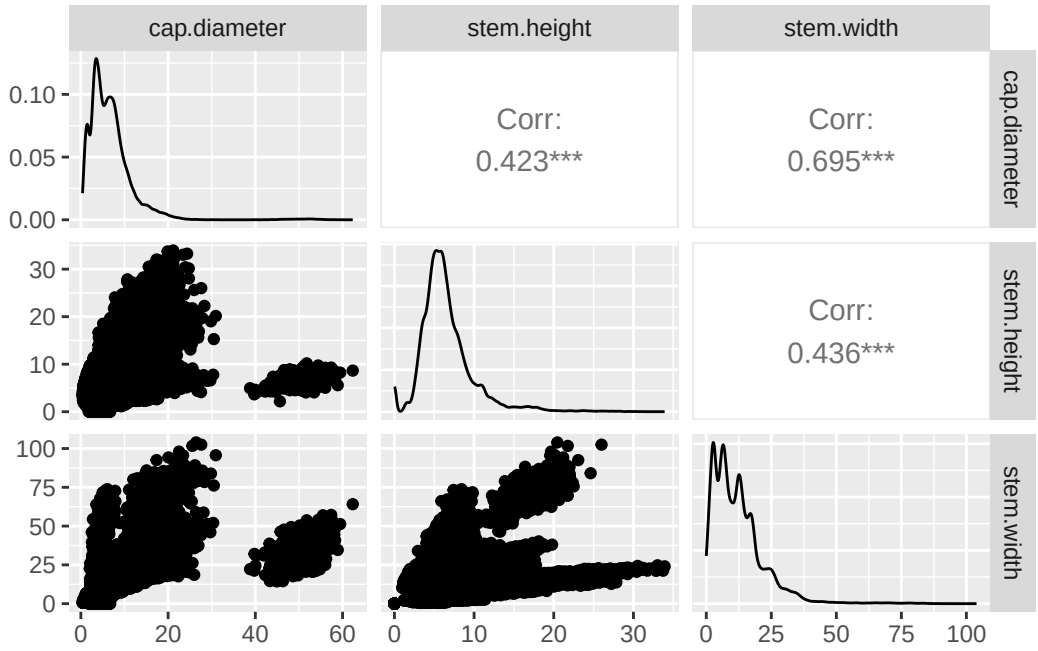


Table 4: cap diameter summary statistics

min	q1	median	q3	max	mean	sd
0.38	3.48	5.86	8.54	62.34	6.734	5.265

Table 5: stem height summary statistics

min	q1	median	q3	max	mean	sd
0	4.64	5.95	7.74	33.92	6.582	3.37

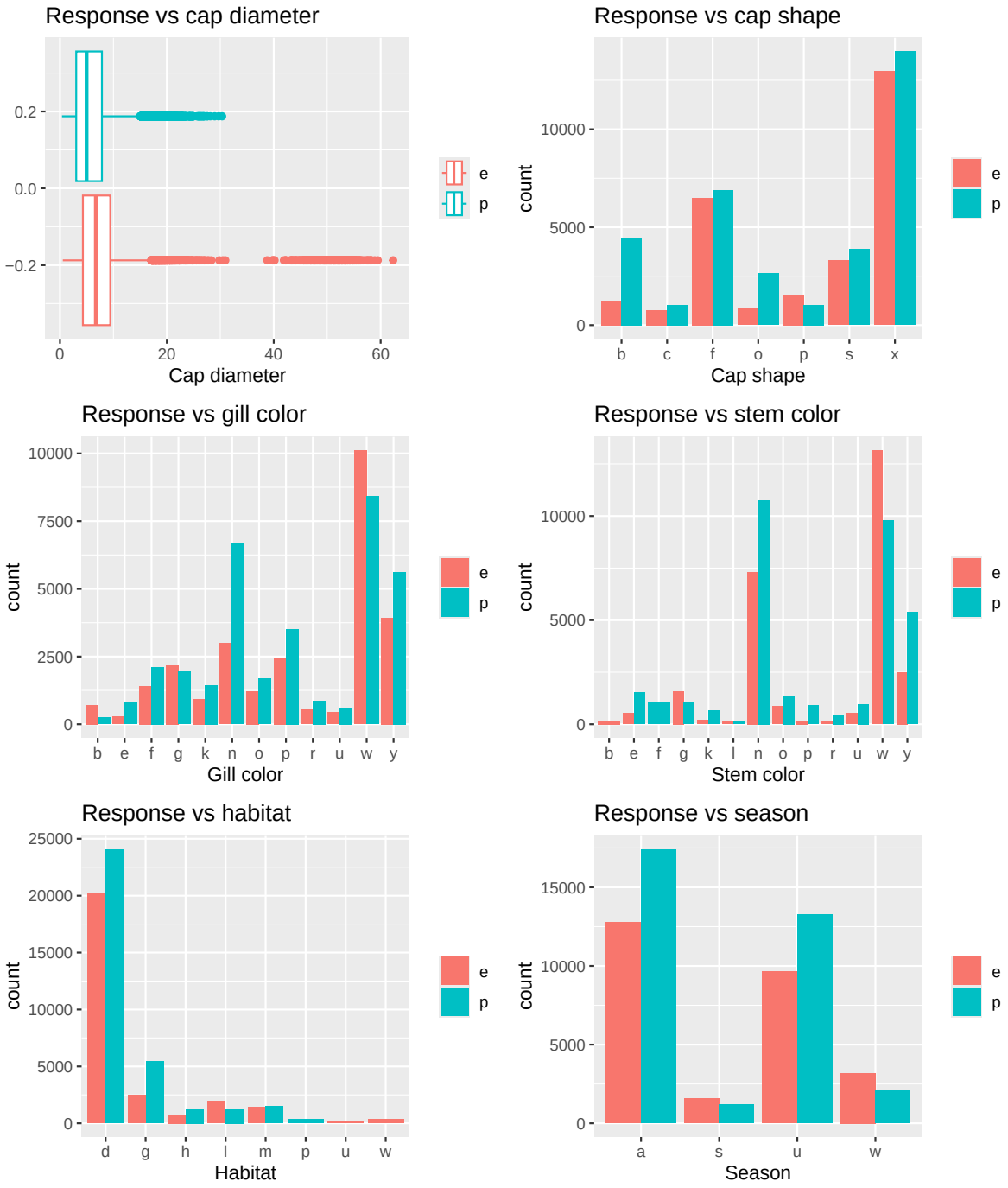
Table 6: stem width summary statistics

min	q1	median	q3	max	mean	sd
0	5.21	10.19	16.57	103.91	12.149	10.036

We were first interested in seeing the relationship between our continuous, quantitative variables. While not totally linear, with correlations of 0.423 and 0.436, they graphs seem to be somewhat linear, with some redundancy in their information. Thus, for the rest of the EDA we will focus mainly on visualizing `cap.diameter`. We may consider adding them back for the

final model, but were more interested in seeing some of the EDA with the categorical variables. The mean cap diameter is 6.734cm, with a SD of 5.265cm. The mean stem height is 6.582cm, with a SD of 3.37cm. The mean stem width is 12.149cm, with a SD of 10.036cm.

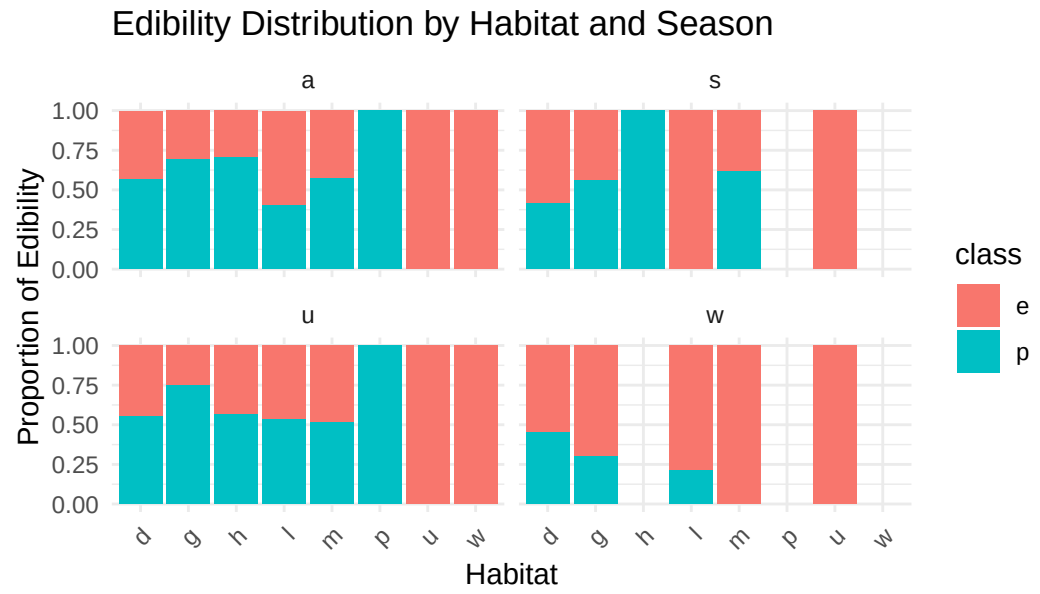
To visualize the distribution of some of our predictor variables, we use a histogram for our continuous variable `cap.diameter` and bar graphs for our categorical variables. The distribution of cap diameter seems to be unimodal, skewed right. From the For qualitative variables, there appears to be more common physical and environmental characteristics. For cap shape, flat and convex tends to be the most common; for stem root the most common was "missing data" (which likely suggests that this may be a variable to remove); for stem color the most common is white, yellow, and brown; for habitat, woods is the most common (see above for code meanings); for season, autumn and summer tends to be the most common.



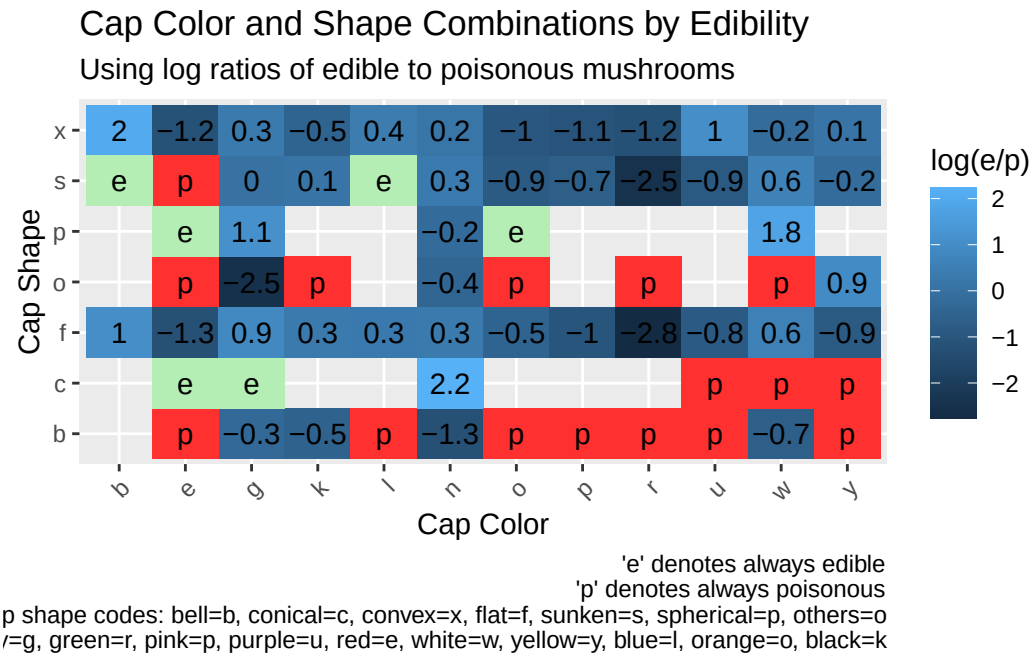
cap shape: bell=b, conical=c, convex=x, flat=f, sunken=s, spherical=p, others=o
colors: brown=n, buff=b, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y, blue=l, orange=o, black=k
habitat: grasses=g, leaves=l, meadows=m, paths=p, heaths=h, urban=u, waste=w, woods=d
season: spring=s, summer=u, autumn=a, winter=w

The bivariate exploratory data analysis also shows some interesting findings for predictors.

In particular, categories that have a disparity between the two different classes could offer potential modeling power in classification of the classes (toxicity). Larger, more extreme cap diameter is often linked to edibility. Cap shape of convex, bell, and others is more likely to be poisonous/toxic than edible. We also see that gill color of brown and yellow tends to be poisonous. For stem color, yellow and brown tends to be more poisonous than not. And lastly the habitat of woods and the season of autumn and summer also observes a similar observation.



s: grasses=g, leaves=l, meadows=m, paths=p, heaths=h, urban=u, waste=w, woods=d
Season codes: spring=s, summer=u, autumn=a, winter=w

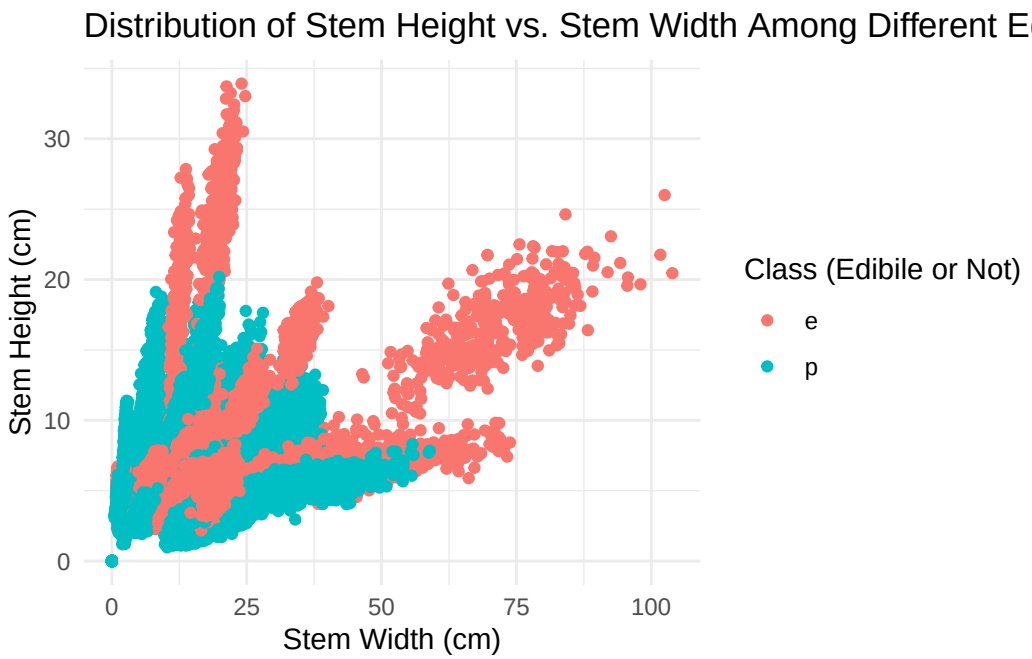


'e' denotes always edible
'p' denotes always poisonous
p shape codes: bell=b, conical=c, convex=x, flat=f, sunken=s, spherical=p, others=o
/=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y, blue=l, orange=o, black=k

We believe that these predictors may have potential interactive effects that could help us with our model to predict edibility of mushrooms due to the fact that certain more specific characteristics tends to be poisonous.

Habitat $\times$ **Season**: Our EDA reveals that mushrooms in certain habitats might only be edible during specific seasons. For example, mushrooms in the meadows are edible in the winter, but may be poisonous in other seasons.

Cap color $\times$ **Cap shape**: Certain combinations of cap color and cap shape are always edible or poisonous. Additionally, the log ratios across combinations of cap colors and cap sizes are varied with no pattern – for a mushroom with a sunken cap shape, it could be always edible (if the color is buff) to always poisonous (if the color is red). Similarly, if a mushroom is brown, it could be high likely it is edible (if the cap shape is conical) or likely it is poisonous (if the cap shape is bell).



Finally, we look at multivariate data analysis including 2 predictors and our response variable. Here, we visualize the effect of both stem width and stem height on the response variable, `class`. Interestingly, it seems like mushrooms with either high stem width or stem height seem to be edible. This suggests there may be some potential interaction effects between stem height and stem width – the low value of one alone does not seem to predict if the mushroom is poisonous, but requires the low value of both.

Analysis

The base model:

$$\log\left(\frac{\text{P(class = poisonous)}}{\text{P(class = edible)}}\right) = \beta_0 + \beta_1 \cdot \text{cap.diameter} + \beta_2 \cdot \text{season} + \beta_3 \cdot \text{cap.shape} + \beta_4 \cdot \text{cap.color} + \beta_5 \cdot \text{gill.color} + \beta_6 \cdot \text{ste}$$

Predictor terms were chosen from exploratory data analysis and general physical or environmental factors that are generally understood and easy to evaluate by everyone. To determine the usefulness of these selected predictor terms we first decided to look at the coefficients from the tidy function.

term	estimate	std.error	statistic	p.value
(Intercept)	-15.328	172.082	-0.089	0.929

term	estimate	std.error	statistic	p.value
cap.diameter	-0.076	0.002	-34.334	0.000
seasons	-1.089	0.052	-21.099	0.000
seasonu	0.003	0.020	0.156	0.876
seasonw	-0.837	0.037	-22.799	0.000
cap.shapec	-0.813	0.068	-11.889	0.000
cap.shapef	-1.244	0.044	-28.509	0.000
cap.shapeo	0.173	0.077	2.253	0.024
cap.shapep	-1.041	0.057	-18.110	0.000
cap.shapes	-1.241	0.049	-25.426	0.000
cap.shapex	-1.473	0.041	-36.105	0.000
cap.colore	1.974	0.089	22.141	0.000
cap.colorg	0.689	0.087	7.875	0.000
cap.colork	1.286	0.103	12.483	0.000
cap.colorl	0.853	0.113	7.551	0.000
cap.colorn	0.427	0.081	5.296	0.000
cap.coloro	1.415	0.090	15.652	0.000
cap.colorp	1.065	0.105	10.115	0.000
cap.colorr	3.003	0.112	26.844	0.000
cap.coloru	1.094	0.099	11.106	0.000
cap.colorw	0.803	0.084	9.552	0.000
cap.colory	0.804	0.084	9.540	0.000
gill.colore	1.058	0.121	8.715	0.000
gill.colorf	-0.397	0.109	-3.637	0.000
gill.colorg	0.226	0.101	2.227	0.026
gill.colork	0.784	0.107	7.351	0.000
gill.colorn	1.292	0.097	13.255	0.000
gill.coloro	0.670	0.102	6.538	0.000
gill.colorp	0.778	0.099	7.824	0.000
gill.colorr	0.554	0.115	4.827	0.000
gill.coloru	0.990	0.120	8.218	0.000
gill.colorw	0.434	0.095	4.580	0.000
gill.colory	0.771	0.097	7.969	0.000
stem.colore	16.299	172.082	0.095	0.925
stem.colorf	31.777	185.273	0.172	0.864
stem.colorg	14.813	172.082	0.086	0.931
stem.colork	17.654	172.082	0.103	0.918
stem.colorl	14.725	172.082	0.086	0.932
stem.colorn	16.180	172.082	0.094	0.925
stem.coloro	15.797	172.082	0.092	0.927
stem.colorp	17.598	172.082	0.102	0.919
stem.colorr	16.734	172.082	0.097	0.923

term	estimate	std.error	statistic	p.value
stem.coloru	16.325	172.082	0.095	0.924
stem.colorw	15.400	172.082	0.089	0.929
stem.colory	16.390	172.082	0.095	0.924
habitatg	0.473	0.031	15.074	0.000
habitath	0.053	0.054	0.987	0.324
habitatl	-0.476	0.045	-10.493	0.000
habitatm	-0.428	0.048	-8.862	0.000
habitatp	15.871	121.020	0.131	0.896
habitatu	-15.782	215.397	-0.073	0.942
habitatw	-16.260	126.274	-0.129	0.898

The Wald's Significance Tests for multiple coefficients of the same predictor variables reveals that for certain categories there may be limited data (also seen through EDA) and such there may be tons of coefficients that are not helpful in prediction. For these variables we aim to transform them into fewer categories for the simplicity of our model. Stem.color of "w", "y"m and "n" were kept while the other observations were assigned "Other". For habitat, "d" and "g" were kept.

Now lets look at the tidy table again and run a likelihood ratio test to evaluate the overall significance of the coefficents of the new model with modified categorical variables.

term	df.residual	residual.deviance	df	deviance	p.value
class_binary ~ 1	61068	83921.51	NA	NA	NA
class_binary ~ cap.diameter + season + cap.shape + cap.color + gill.color + stem.color.modified + habitat.modified	61031	70569.07	37	13352.44	0

term	estimate	std.error	statistic	p.value
(Intercept)	0.590	0.111	5.333	0.000
cap.diameter	-0.074	0.002	-36.247	0.000
seasons	-0.975	0.049	-20.006	0.000
seasonu	-0.005	0.020	-0.243	0.808
seasonw	-0.766	0.035	-21.870	0.000
cap.shapec	-0.881	0.064	-13.674	0.000
cap.shapef	-1.155	0.042	-27.246	0.000
cap.shapeo	1.027	0.074	13.920	0.000
cap.shapep	-1.003	0.056	-17.825	0.000
cap.shapes	-1.082	0.047	-22.843	0.000

term	estimate	std.error	statistic	p.value
cap.shapex	-1.335	0.040	-33.690	0.000
cap.colore	1.989	0.087	22.762	0.000
cap.colorg	0.557	0.085	6.515	0.000
cap.colork	0.887	0.102	8.674	0.000
cap.colorl	0.650	0.109	5.971	0.000
cap.colorn	0.358	0.079	4.512	0.000
cap.coloro	1.359	0.088	15.402	0.000
cap.colorp	1.614	0.099	16.296	0.000
cap.colorr	2.886	0.109	26.419	0.000
cap.coloru	1.207	0.094	12.784	0.000
cap.colorw	0.821	0.082	9.975	0.000
cap.colory	0.771	0.083	9.311	0.000
gill.colore	1.262	0.117	10.767	0.000
gill.colorf	-0.247	0.104	-2.378	0.017
gill.colorg	0.222	0.097	2.302	0.021
gill.colork	0.936	0.102	9.165	0.000
gill.colorn	1.435	0.093	15.498	0.000
gill.coloro	0.720	0.098	7.350	0.000
gill.colorp	0.983	0.094	10.476	0.000
gill.colorr	0.696	0.110	6.354	0.000
gill.coloru	0.867	0.113	7.701	0.000
gill.colorw	0.548	0.090	6.096	0.000
gill.colory	0.971	0.092	10.562	0.000
stem.color.modifiedOther	-0.128	0.030	-4.339	0.000
stem.color.modifiedw	-0.744	0.025	-29.649	0.000
stem.color.modifiedy	0.224	0.035	6.391	0.000
habitat.modifiedg	0.471	0.031	15.277	0.000
habitat.modifiedOther	-0.341	0.027	-12.391	0.000

$$H_0 : \beta_j = 0 H_a : \beta_j \neq 0 \ for\ at\ least\ 1\ j$$

Since the p-value is small, and less than $\alpha = 0.05$, we reject the H_0. The data provide sufficent evidence of at least one non-zero coefficient in the model.

To determine adding interaction terms, a drop in deviance test was performed with the added interaction terms of habitat * season and cap.shape * cap.color.

#High p-values in tidy table of full_model? Is that a point of concern? The likelihood test is not picking up on stem color? Are the p-values for 1 term the wald's test? Intercept p value

is super high, should we exclude it? #Testing anova with the full dataset not just the train dataset

#EDA univariate or bivariate to make it shorter and more readable?

Make height of the bivariate taller, and describes trends, don't really need univariate EDA. Simplify categories of some predictors to change with p-values. All the modeling should be done with the training set. Prediction- AUC, BIC as priority. Add more to the tables so we can remove univariate eda.

term	df.residual	residual.deviance	df	deviance	p.value
class_binary ~ cap.diameter + season + cap.shape + cap.color + gill.color + stem.color.modified + habitat.modified	61031	70569.07	NA	NA	NA
class_binary ~ cap.diameter + season + cap.shape + cap.color + gill.color + stem.color.modified + habitat.modified + habitat.modified * season + cap.shape * cap.color	60977	65170.76	54	5398.315	0

Since the p-value is low, below $\alpha = 0.05$, we decide to include these interaction terms as there is convincing evidence that at least one of these interactive term coefficients are not 0 and thus helpful in the model.

Additionally, we looked at BIC and BIC to evaluate the base model in comparison with the interactive model.

```
[1] 70645.07
```

```
[1] 65354.76
```

```
[1] 70987.83
```

```
[1] 66184.58
```

For both AIC and BIC, the model with the interaction effect does much better in comparison to the main model.

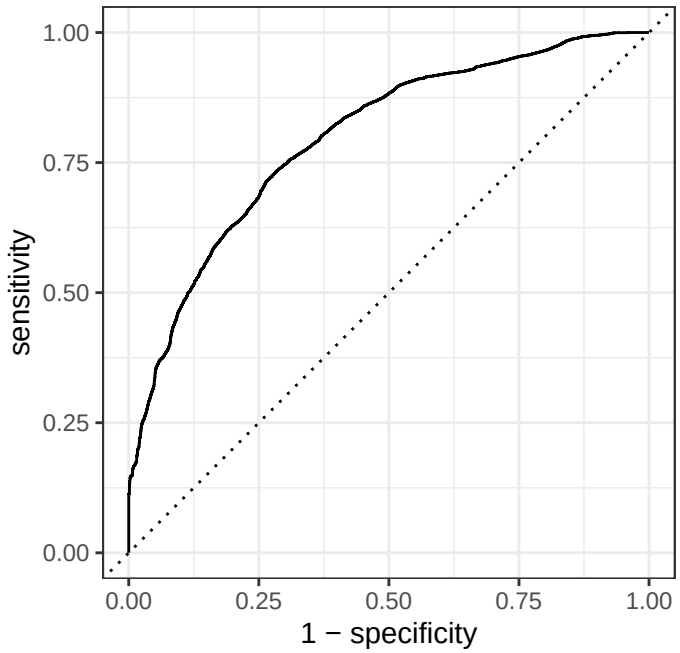
Model Results

Final Model

$\log(\text{Odds}(\text{class} = \text{poisonous}) = \beta_0 + \beta_1 \cdot \text{cap.diameter} + \beta_2 \cdot \text{season} + \beta_3 \cdot \text{cap.shape} + \beta_4 \cdot \text{cap.color} + \beta_5 \cdot \text{gill.color} + \beta_6 \cdot \text{ste}$

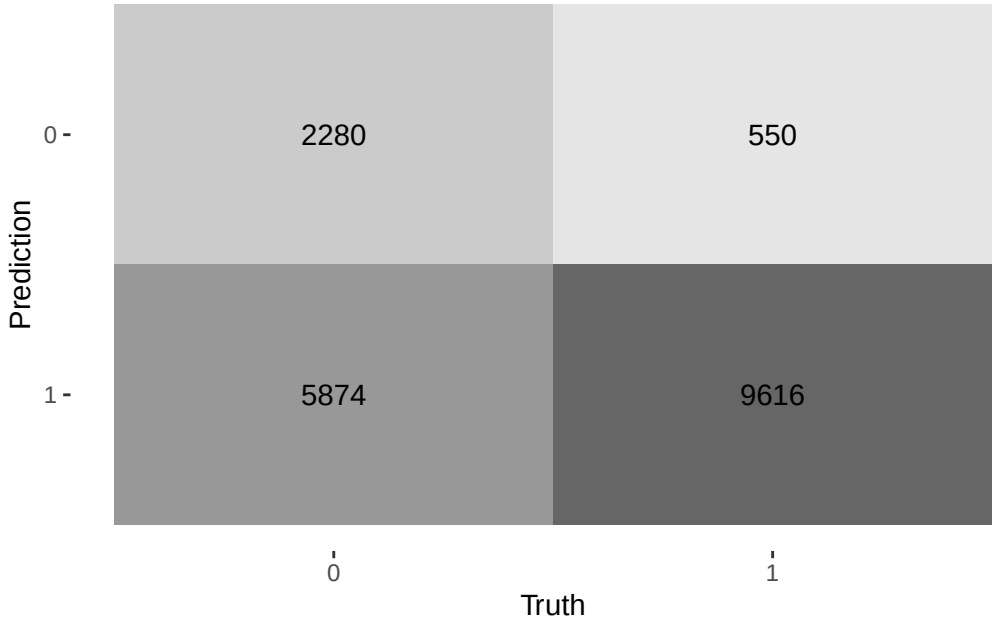
Confusion Matrix and ROC

class binary: edible = 0, poisonous = 1




```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.797
```

```
# A tibble: 6 x 3
  .threshold specificity sensitivity
       <dbl>       <dbl>       <dbl>
1      0.247       0.264       0.950
2      0.247       0.264       0.950
3      0.247       0.264       0.950
4      0.248       0.264       0.950
5      0.248       0.264       0.950
6      0.248       0.264       0.950
```



The model is decent as the AUC is 0.816 which is closer to 1 than 0.5. In a specified model, the threshold is determined to be p = ___ because the false positives are more "expensive" than the false negatives as eating a poisonous mushroom can be detrimental to one's health. We are better to be overly cautious and let our model predict more false negatives for a trade off of less false positives. #Need code to determine the p value

References

Brandenburg, William E., and Karlee J. Ward. 2018. "Mushroom Poisoning Epidemiology in the United States." *Mycologia* 110 (4): 637–41. https://doi.org/10.1080/00275514.2018.1479561.