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

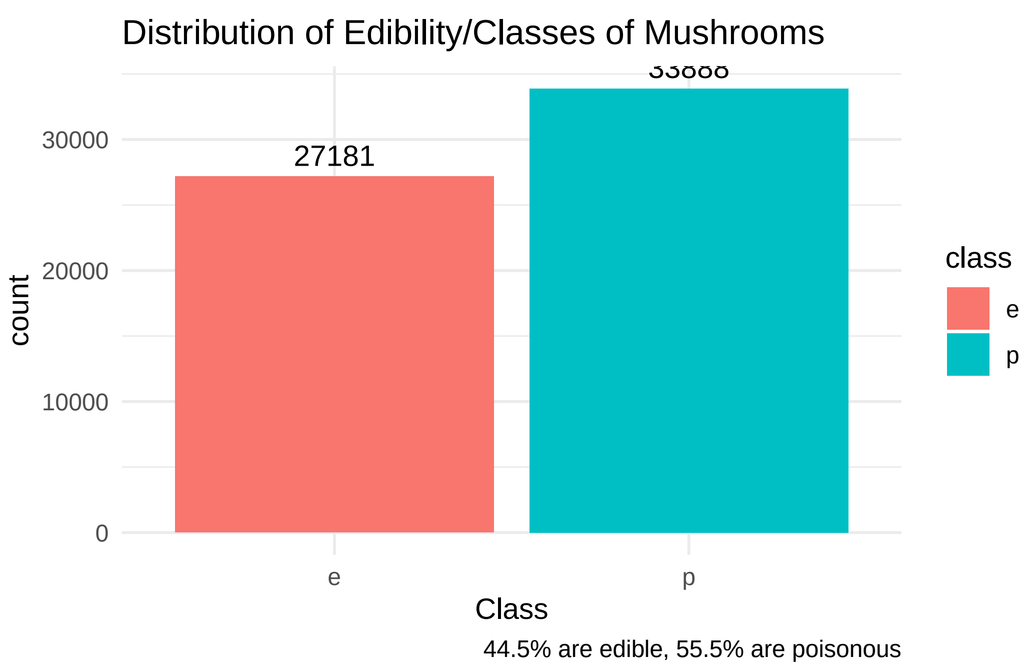
The `gill.color` and `stem.color` are:

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.

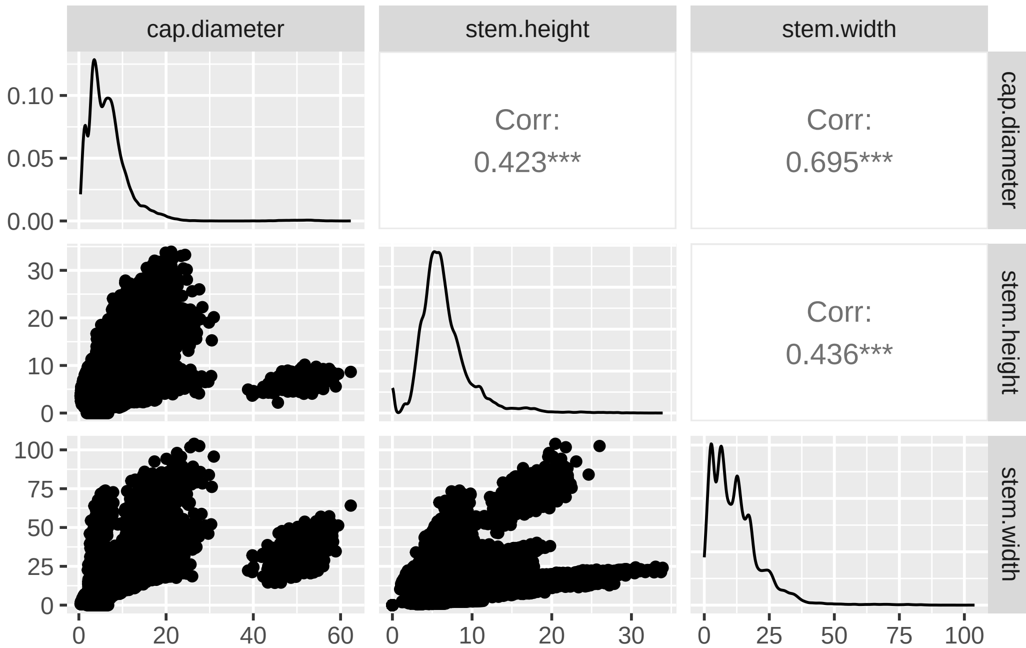


Table 4: cap diameter summary statistics

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

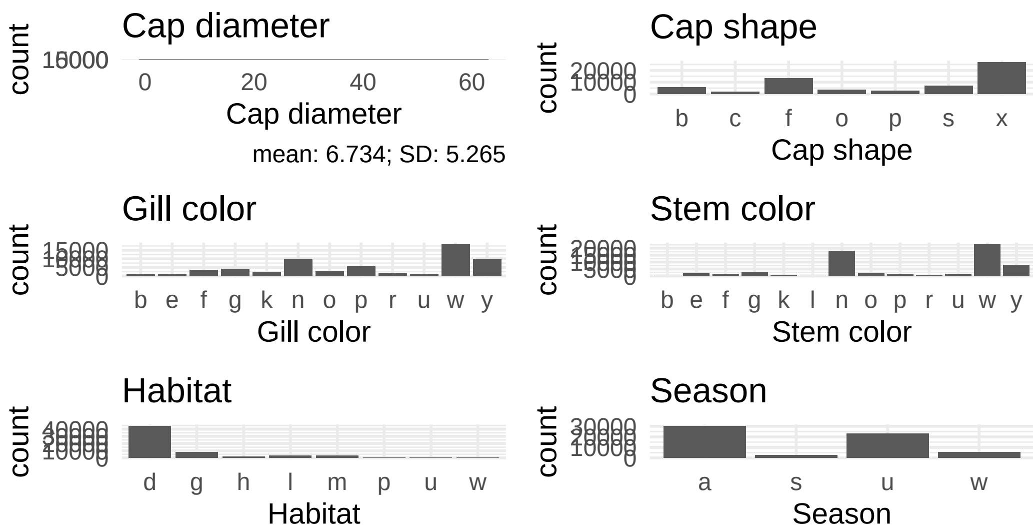
Table 5: stem height summary statistics

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

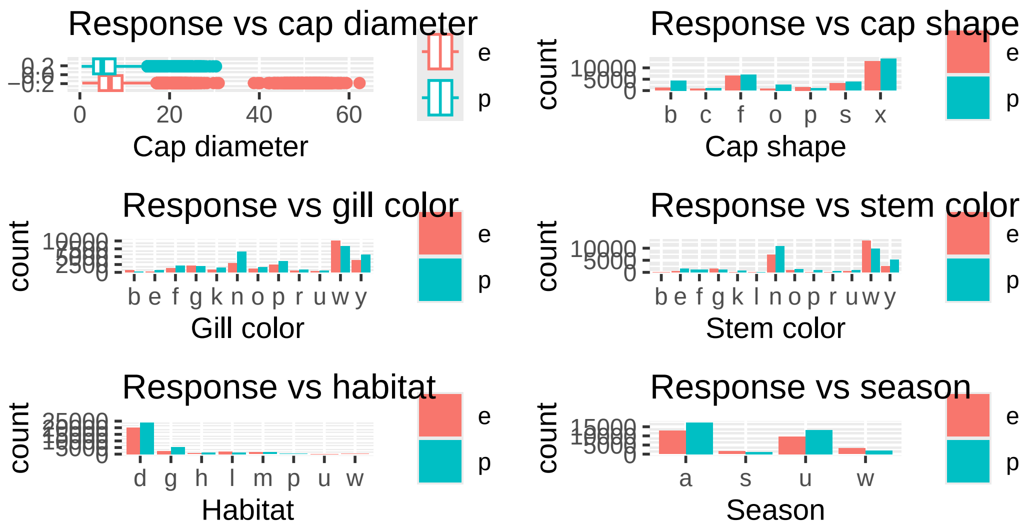
min	q1	median	q3	max	mean	sd
0	5.21	10.19	16.57	103.91	12.149	10.036

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.



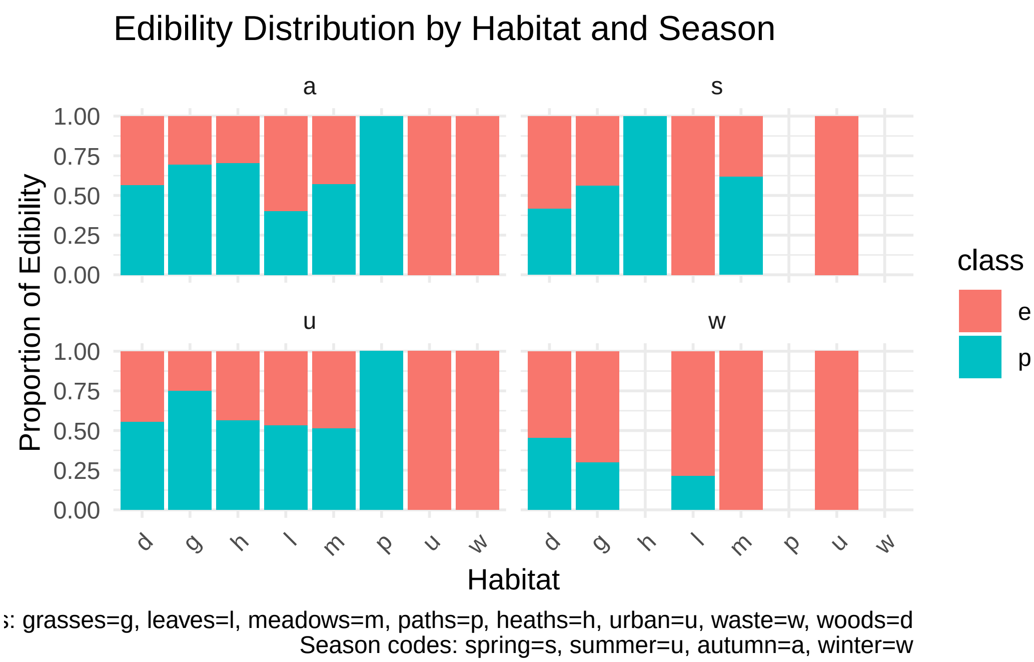
cap shape: bell=b, conical=c, convex=x, flat=f, sunken=s, spherical=p, others=o
colors: brown=n, buff=b, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y, blue=l, orange=o, black=k
habitat: grasses=g, leaves=l, meadows=m, paths=p, heaths=h, urban=u, waste=w, woods=d
season: spring=s, summer=u,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

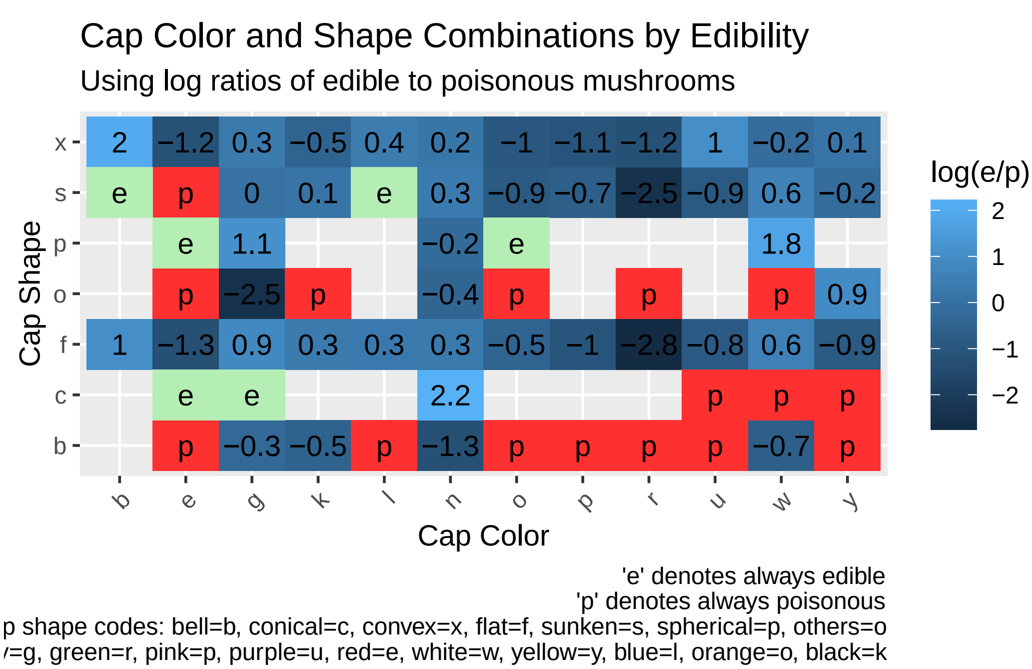
To visualize the distribution of some of our predictor variables, we use a histogram for our continuous variable `cap.diameter` and bar graphs for our categorical variables. The distribution of cap diameter seems to be unimodal, skewed right. From the For qualitative variables, there appears to be more 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.

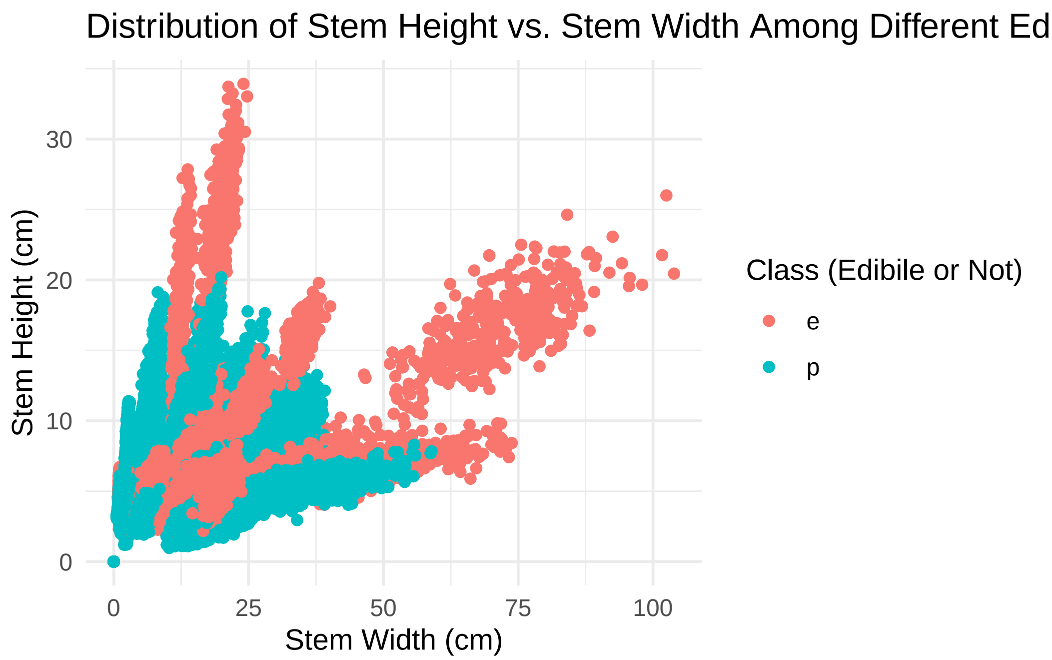




We believe that these predictors may have potential interactive effects that could help us with our model to predict edibility of mushrooms due to the fact that certain more specific characteristics tends to be poisonous.

Habitat × 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 × 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:

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

Predictor terms were choosen 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 a likelihood ratio test was performed.

```
[1] -41960.76
```

```
[1] -33818.07
```

```
[1] 16285.38
```

```
[1] 0
```

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 + habitat	61017	67636.13	51	16285.39	0

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

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.

#what if we remove cap.diameter and kept stem height and width, and had an interaction term of the two?

#Why is df so high?

#High p-values in tidy table of full_model? Is that a point of concern?

term	df.residual	residual.deviance	df	deviance	p.value
class_binary ~ cap.diameter + season + cap.shape + cap.color + gill.color + stem.color + habitat	61017	67636.13	NA	NA	NA
class_binary ~ cap.diameter + season + cap.shape + cap.color + gill.color + stem.color + habitat + habitat * season + cap.shape * cap.color	60953	62378.33	64	5257.798	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] 67740.13
```

```
[1] 62610.33
```

```
[1] 68209.16
```

```
[1] 63656.62
```

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

Model Results

term	estimate	std.error	statistic	p.value
(Intercept)	-1.593	522.797	-0.003	0.998
cap.diameter	-0.056	0.003	-21.264	0.000
seasons	-1.432	0.073	-19.607	0.000
seasonu	-0.011	0.024	-0.461	0.645
seasonw	-0.630	0.043	-14.768	0.000
cap.shapec	0.996	361.900	0.003	0.998
cap.shapef	-17.211	277.170	-0.062	0.950
cap.shapeo	-19.185	277.170	-0.069	0.945
cap.shapep	-2.010	0.146	-13.802	0.000
cap.shapes	-34.275	863.450	-0.040	0.968
cap.shapex	-18.610	277.170	-0.067	0.946
cap.colore	1.499	654.918	0.002	0.998
cap.colorg	-16.194	277.170	-0.058	0.953
cap.colork	-16.494	277.170	-0.060	0.953
cap.colorl	1.688	719.028	0.002	0.998
cap.colorn	-15.718	277.170	-0.057	0.955
cap.coloro	0.745	559.226	0.001	0.999
cap.colorp	1.645	649.071	0.003	0.998
cap.colorr	1.351	587.712	0.002	0.998

term	estimate	std.error	statistic	p.value
cap.coloru	1.152	636.556	0.002	0.999
cap.colorw	-16.024	277.170	-0.058	0.954
cap.colory	1.223	0.167	7.305	0.000
gill.colore	1.375	0.132	10.391	0.000
gill.colorf	-0.744	0.125	-5.938	0.000
gill.colorg	0.545	0.112	4.847	0.000
gill.colork	1.152	0.117	9.821	0.000
gill.colorn	1.439	0.108	13.284	0.000
gill.coloro	1.018	0.114	8.929	0.000
gill.colorp	0.958	0.111	8.664	0.000
gill.colorr	-0.104	0.131	-0.790	0.430
gill.coloru	1.461	0.137	10.640	0.000
gill.colorw	0.681	0.106	6.453	0.000
gill.colory	0.947	0.107	8.823	0.000
stem.colore	18.571	443.275	0.042	0.967
stem.colorf	36.896	470.466	0.078	0.937
stem.colorg	17.267	443.275	0.039	0.969
stem.colork	20.751	443.275	0.047	0.963
stem.colorl	17.179	443.275	0.039	0.969
stem.colorn	18.493	443.275	0.042	0.967
stem.coloro	18.190	443.275	0.041	0.967
stem.colorp	20.005	443.275	0.045	0.964
stem.colorr	19.079	443.275	0.043	0.966
stem.coloru	18.430	443.275	0.042	0.967
stem.colorw	17.615	443.275	0.040	0.968
stem.colory	18.764	443.275	0.042	0.966
habitatg	0.413	0.045	9.137	0.000
habitath	0.397	0.074	5.381	0.000
habitatl	-0.474	0.064	-7.384	0.000
habitatm	-0.432	0.073	-5.926	0.000
habitatp	16.057	372.899	0.043	0.966
habitatu	-17.594	1214.547	-0.014	0.988
habitatw	-18.083	493.434	-0.037	0.971
seasons:habitatg	0.200	0.176	1.142	0.254
seasonu:habitatg	0.303	0.064	4.758	0.000
seasonw:habitatg	-1.047	0.123	-8.518	0.000
seasons:habitath	17.064	961.695	0.018	0.986
seasonu:habitath	-0.486	0.109	-4.469	0.000
seasonw:habitath	NA	NA	NA	NA
seasons:habitatl	-16.934	402.574	-0.042	0.966
seasonu:habitatl	0.333	0.103	3.239	0.001

term	estimate	std.error	statistic	p.value
seasonw:habitatl	0.213	0.137	1.549	0.121
seasons:habitatm	1.476	0.193	7.631	0.000
seasonu:habitatm	-0.290	0.106	-2.750	0.006
seasonw:habitatm	-17.949	401.116	-0.045	0.964
seasons:habitatp	NA	NA	NA	NA
seasonu:habitatp	0.188	546.689	0.000	1.000
seasonw:habitatp	NA	NA	NA	NA
seasons:habitatu	1.415	1714.595	0.001	0.999
seasonu:habitatu	0.218	1754.954	0.000	1.000
seasonw:habitatu	0.802	1607.339	0.000	1.000
seasons:habitatw	NA	NA	NA	NA
seasonu:habitatw	0.015	689.407	0.000	1.000
seasonw:habitatw	NA	NA	NA	NA
cap.shapec:cap.colore	-38.756	1084.634	-0.036	0.971
cap.shapef:cap.colore	0.229	654.918	0.000	1.000
cap.shapeo:cap.colore	21.486	796.959	0.027	0.978
cap.shapep:cap.colore	-34.132	1015.384	-0.034	0.973
cap.shapes:cap.colore	34.058	1100.224	0.031	0.975
cap.shapex:cap.colore	1.168	654.918	0.002	0.999
cap.shapec:cap.colorg	-16.985	852.116	-0.020	0.984
cap.shapef:cap.colorg	16.180	277.170	0.058	0.953
cap.shapeo:cap.colorg	22.096	277.171	0.080	0.936
cap.shapep:cap.colorg	1.092	0.220	4.964	0.000
cap.shapes:cap.colorg	33.906	863.450	0.039	0.969
cap.shapex:cap.colorg	17.871	277.170	0.064	0.949
cap.shapec:cap.colork	NA	NA	NA	NA
cap.shapef:cap.colork	16.943	277.170	0.061	0.951
cap.shapeo:cap.colork	38.969	545.368	0.071	0.943
cap.shapep:cap.colork	NA	NA	NA	NA
cap.shapes:cap.colork	34.317	863.450	0.040	0.968
cap.shapex:cap.colork	18.870	277.170	0.068	0.946
cap.shapec:cap.colorl	NA	NA	NA	NA
cap.shapef:cap.colorl	-1.155	719.028	-0.002	0.999
cap.shapeo:cap.colorl	NA	NA	NA	NA
cap.shapep:cap.colorl	NA	NA	NA	NA
cap.shapes:cap.colorl	-1.904	1384.885	-0.001	0.999
cap.shapex:cap.colorl	-0.018	719.028	0.000	1.000
cap.shapec:cap.colorn	-4.495	361.900	-0.012	0.990
cap.shapef:cap.colorn	15.594	277.170	0.056	0.955
cap.shapeo:cap.colorn	18.614	277.170	0.067	0.946
cap.shapep:cap.colorn	1.354	0.165	8.225	0.000

term	estimate	std.error	statistic	p.value
cap.shapes:cap.colorn	32.701	863.450	0.038	0.970
cap.shapex:cap.colorn	16.965	277.170	0.061	0.951
cap.shapec:cap.coloro	NA	NA	NA	NA
cap.shapef:cap.coloro	-0.256	559.226	0.000	1.000
cap.shapeo:cap.coloro	18.773	700.495	0.027	0.979
cap.shapep:cap.coloro	-33.458	940.720	-0.036	0.972
cap.shapes:cap.coloro	17.544	990.684	0.018	0.986
cap.shapex:cap.coloro	1.581	559.226	0.003	0.998
cap.shapec:cap.colorp	NA	NA	NA	NA
cap.shapef:cap.colorp	-0.671	649.071	-0.001	0.999
cap.shapeo:cap.colorp	NA	NA	NA	NA
cap.shapep:cap.colorp	NA	NA	NA	NA
cap.shapes:cap.colorp	15.352	1044.038	0.015	0.988
cap.shapex:cap.colorp	0.443	649.071	0.001	0.999
cap.shapec:cap.colorr	NA	NA	NA	NA
cap.shapef:cap.colorr	1.965	587.712	0.003	0.997
cap.shapeo:cap.colorr	18.373	733.441	0.025	0.980
cap.shapep:cap.colorr	NA	NA	NA	NA
cap.shapes:cap.colorr	18.695	1007.039	0.019	0.985
cap.shapex:cap.colorr	2.096	587.712	0.004	0.997
cap.shapec:cap.coloru	-1.695	903.306	-0.002	0.999
cap.shapef:cap.coloru	-0.174	636.556	0.000	1.000
cap.shapeo:cap.coloru	NA	NA	NA	NA
cap.shapep:cap.coloru	NA	NA	NA	NA
cap.shapes:cap.coloru	17.111	1036.304	0.017	0.987
cap.shapex:cap.coloru	-0.570	636.556	-0.001	0.999
cap.shapec:cap.colorw	17.244	506.515	0.034	0.973
cap.shapef:cap.colorw	16.189	277.170	0.058	0.953
cap.shapeo:cap.colorw	37.091	351.249	0.106	0.916
cap.shapep:cap.colorw	NA	NA	NA	NA
cap.shapes:cap.colorw	33.121	863.450	0.038	0.969
cap.shapex:cap.colorw	18.048	277.170	0.065	0.948
cap.shapec:cap.colory	NA	NA	NA	NA
cap.shapef:cap.colory	-0.171	0.200	-0.857	0.392
cap.shapeo:cap.colory	NA	NA	NA	NA
cap.shapep:cap.colory	NA	NA	NA	NA
cap.shapes:cap.colory	16.595	817.754	0.020	0.984
cap.shapex:cap.colory	NA	NA	NA	NA

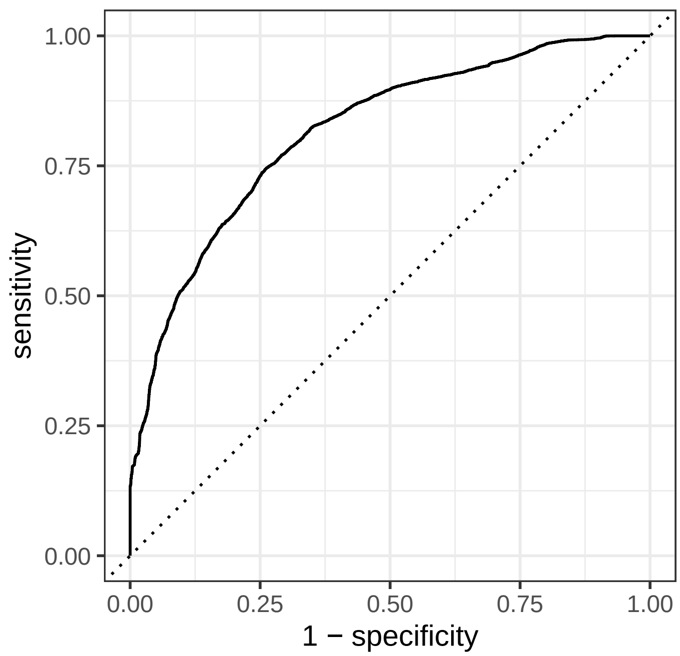
Final Model

$$

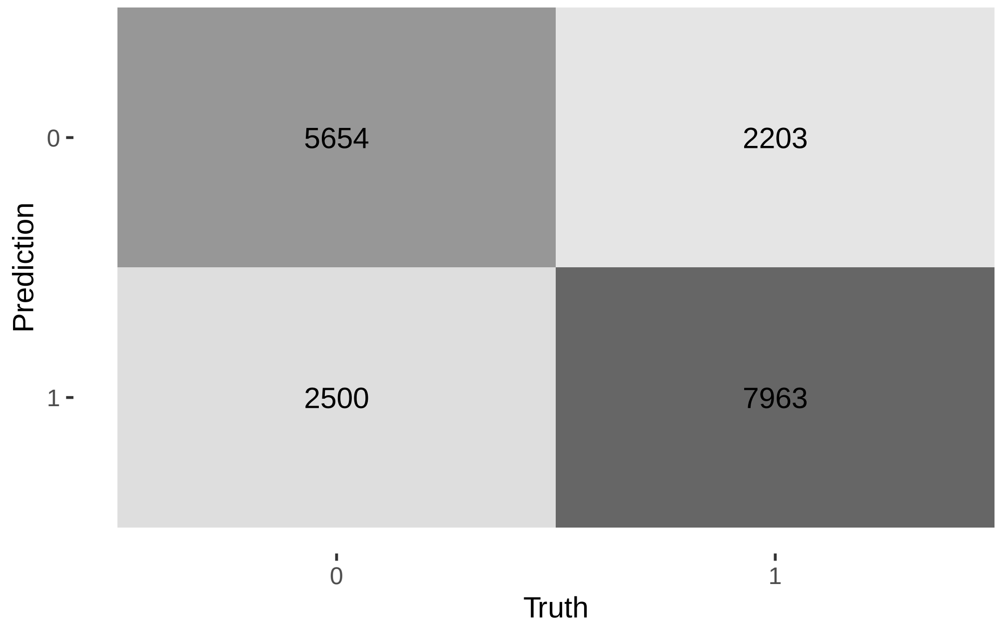
log(Odds(class_binary = p)) = _0 + _1 cap.diameter + _2 season + _3 cap.shape + _4 cap.color \+ _5 gill.color + _6 stem.color + _7 habitat + _8 (habitat season) + _9 (cap.shape cap.color)

$$ ### 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         0.816
```



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.