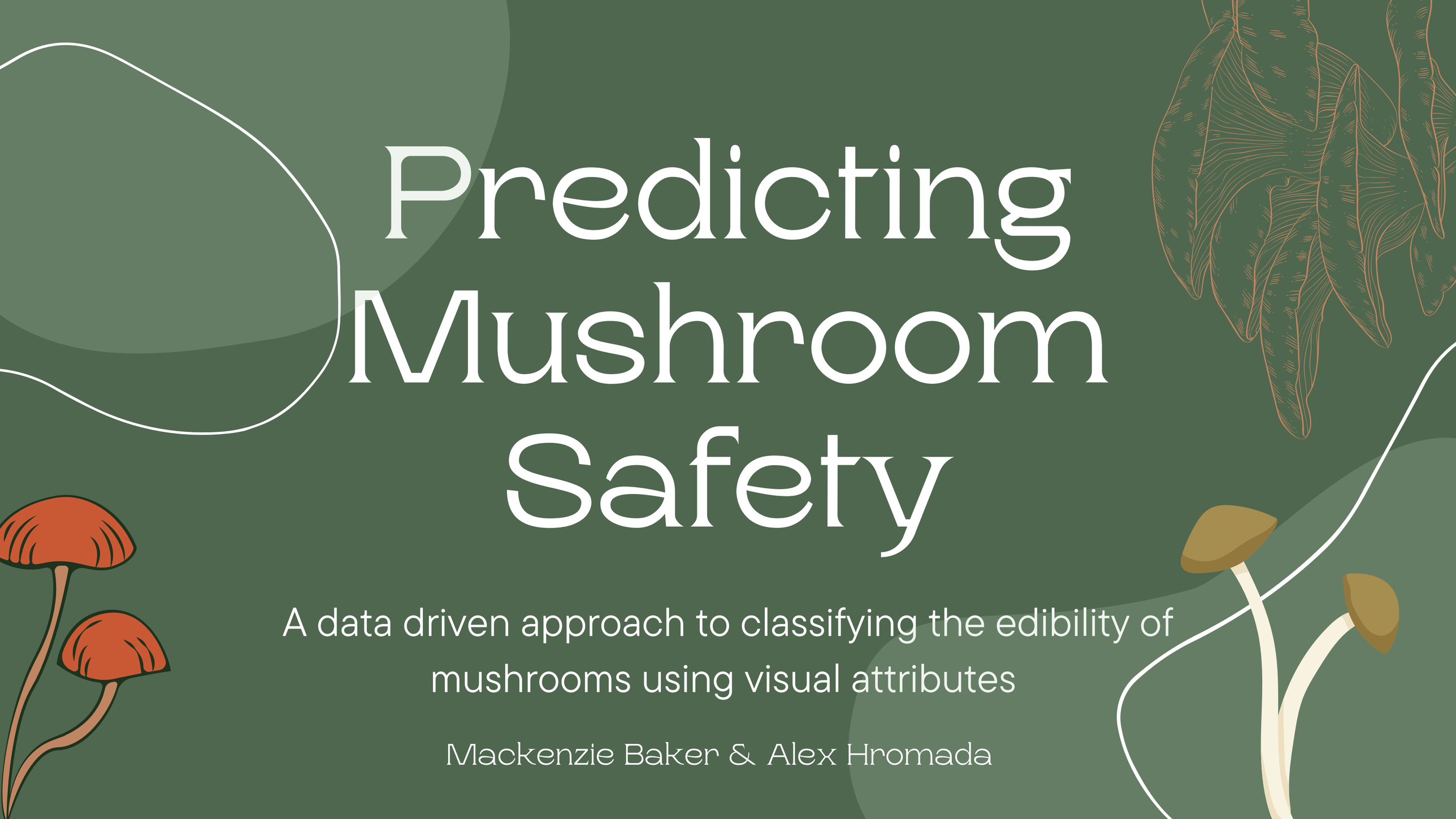

Predicting Mushroom Safety
A data driven approach to classifying the edibility of mushrooms using visual attributes
Mackenzie Baker & Alex Hromada

Agenda

Topics Covered

Summary of Pre-Processing

Data spending

Model Training

Top 5 Models

Predictions on Test set

Can we accurately predict the edibility of mushrooms based on visual attributes?

61,069 Observations

20 Predictors

of Categorical Predictors: 15

# of Logical (Boolean) Predictors: 2

# of Continuous Predictors: 3

Categorical Response

2 Classes: Edible (E) or Poisonous (P)

Preprocessing Summary

Missing Values

Removed predictors with > 50% missing values & applied mode imputation for the rest

5 Predictors Removed

Degenerative Distributions

No strong correlations among numeric predictors
1 NZV predictor - Removed

1 Predictor Removed

Dummy Variables

Convert all remaining categorical predictors to dummy variables.

Dimensions: 61,069 x 88

Transformations Applied

- Applied BoxCox to correct skewness
- Applied Spatial Sign for Outliers
- Center and Scaled

Spending Data

61069 total samples
45% of mushroom samples are classified as edible
55% of mushrooms samples are classified as poisonous

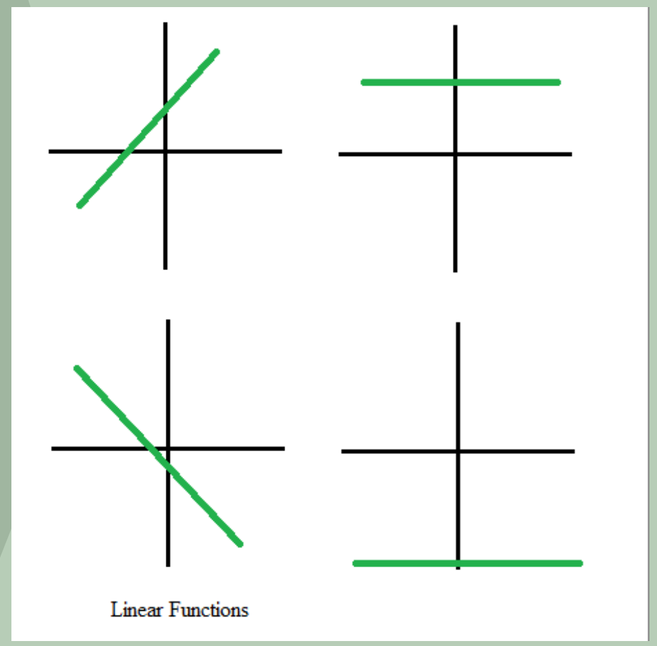
Based on these criteria, we chose to use a random split using 70% of our data as training data and 30% as testing data

Resampling

We have a data set with 61069 total samples and 92 predictors which is fairly large. Because of this we have decided to do a simple 5-fold Cross validation (No Repeats).

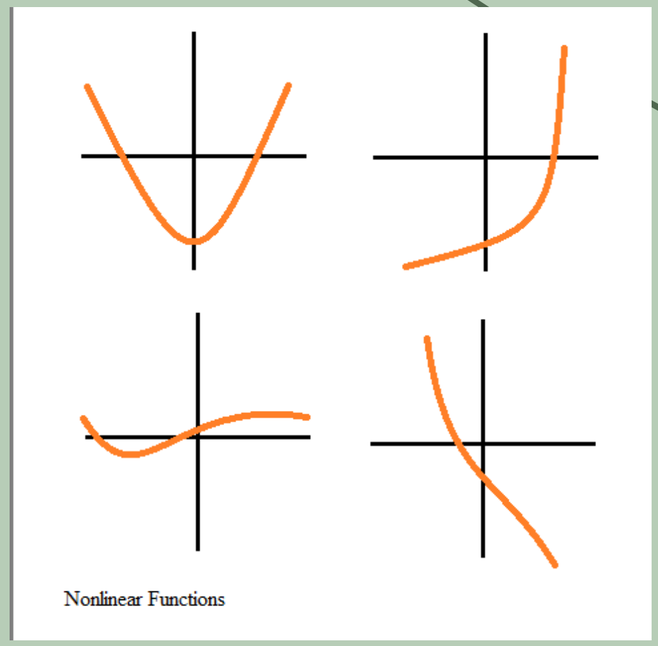
Originally attempted 10-fold but computation time was > 30 hours for some models

Training Classification Models



Linear Models

- Logistic Regression
- Linear Discriminant Analysis
- PLS Discriminant Analysis
- Penalized Model
- Nearest Shrunken Centroid



Non-Linear Models

- Neural Network
- Flexible Discriminant Analysis
- Support Vector Machine
- KNN
- Naïve Bayes

Linear Models' Results (Training Set)

MODEL	SPECIFICITY	SENSITIVITY	ROC	ACCURACY	KAPPA
Logistic Regression	0.7826	0.7325	0.8406871	0.7604	0.5149
Linear Discriminant Analysis	0.7801	0.7402	0.8393885	0.7624	0.5195
PLS Discriminant Analysis	0.7708	0.7229	0.8394725	0.7495	0.4933
Penalized Model (GLMNet)	0.8693	0.4541	0.8414859	0.6847	0.3362
Nearest Shrunken Centroid	0.7613	0.6047	0.7551841	0.6917	0.3696

Non-Linear Models' Results (Training Set)

MODEL	SPECIFICITY	SENSITIVITY	ROC	ACCURACY	KAPPA
Neural Network	0.9464	0.9588	0.9999838	0.9519	0.9029
Flexible Discriminant Analysis	0.7466	0.5832	0.7932640	0.6739	0.3331
Support Vector Machine	0.9653	0.9615	0.9997658	0.9636	0.9264
K Nearest Neighbors	0.9987	0.9957	0.9985157	0.9974	0.9947
Naïve Bayes	0.4066	0.8299	0.8008022	0.5948	0.2237

Best 5 Models (Test Set)

MODEL	SPECIFICITY	SENSITIVITY	ROC	ACCURACY	KAPPA
K Nearest Neighbors	0.9987	0.9971	0.9997	0.998	0.9959
Logistic Regression	0.7804	0.7329	0.8333	0.7592	0.513
Neural Network	0.9988	0.9969	1	0.998	0.9959
Penalized Model	0.7873	0.7361	0.8389	0.7644	0.5234
Support Vector Machine	0.9989	0.9892	0.9998	0.9946	0.9891

Best Chosen Model

MODEL	SPECIFICITY	SENSITIVITY	ROC	ACCURACY	KAPPA
Neural Network	0.9988	0.9969	1	0.998	0.9959

Model with the highest ROC Value

Model with the highest Sensitivity

Model with 2nd Highest Specificity (0.001 difference)

In addition to our metric for selection (ROC) we are most interested in specificity because It is far more important to ensure we accurately predict poisonous mushrooms rather than edible ones

Most Important Predictors

nnet variable importance

only 20 most important variables shown (out of 84)

	Overall
gill-attachment_p	100.00
stem-width	95.74
stem-color_w	87.86
cap-surface_y	86.86
stem-color_y	75.37
does-bruise-or-bleed_TRUE	74.16
gill-attachment_x	70.89
gill-spacing_d	68.83
gill-attachment_d	67.03
gill-attachment_s	65.50
has-ring_TRUE	64.97
does-bruise-or-bleed_FALSE	59.94
gill-color_w	58.80
has-ring_FALSE	58.64
cap-surface_g	57.98
cap-surface_d	53.10
gill-color_p	51.82
gill-spacing_c	49.96
gill-attachment_e	49.63
gill-attachment_a	49.58

The most important 5 predictors for this model are gill-attachment_p, stem-width, stem-color_w, cap-surface_y, and stem-color_y


Thank you!