NaturalistAI: Models for Mushroom Identification

Jacob Serfaty

What's the Difference?


Edible

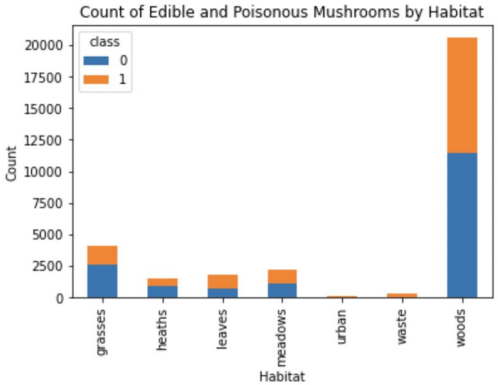

Poisonous

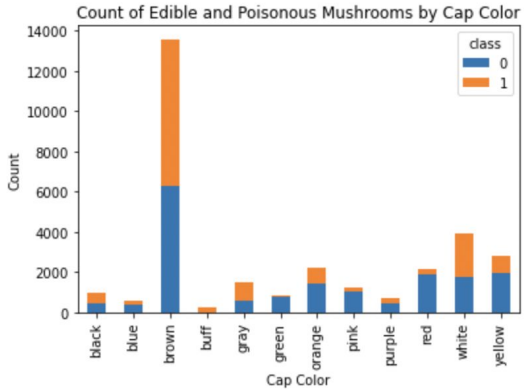
NaturalistAI: Company Goals

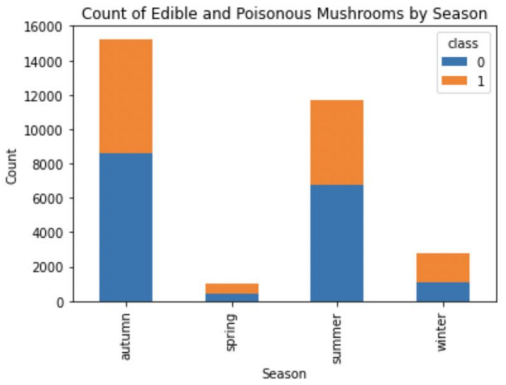
- Create a database that can catalogue mushrooms based on their morphology, habitat, season, and location
- Pair the database with a classifier that can accurately identify edible mushrooms from poisonous ones
- Create an application where users can identify mushrooms in the wild based off of this classifier and the features found in the database

Mushroom Data Review

- Mushrooms are usually brown, found in the woods, and grow in the Summer and Autumn

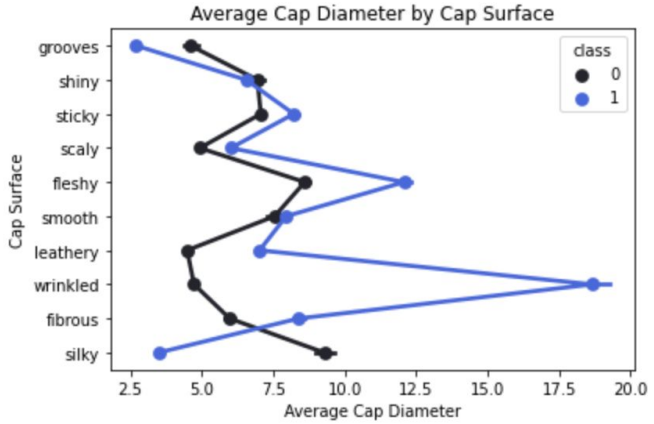


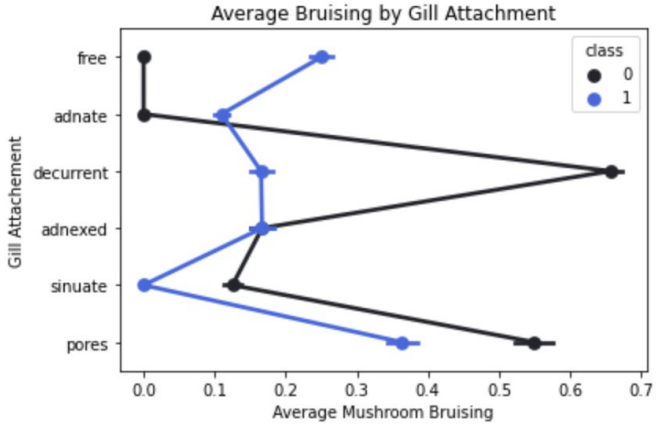




Mushroom Data Review

- Relevant features which distinguish poisonous from edible mushrooms come from the morphology of the mushroom and more importantly the combination of morphological features.





Classification Models: Assessment

- The metric used to assess the success of the model was precision, which weighs the model in favor of false negatives.
- This is necessary for this model as false negative account for poisonous mushrooms identified as edible

Classification Models: Initial Model

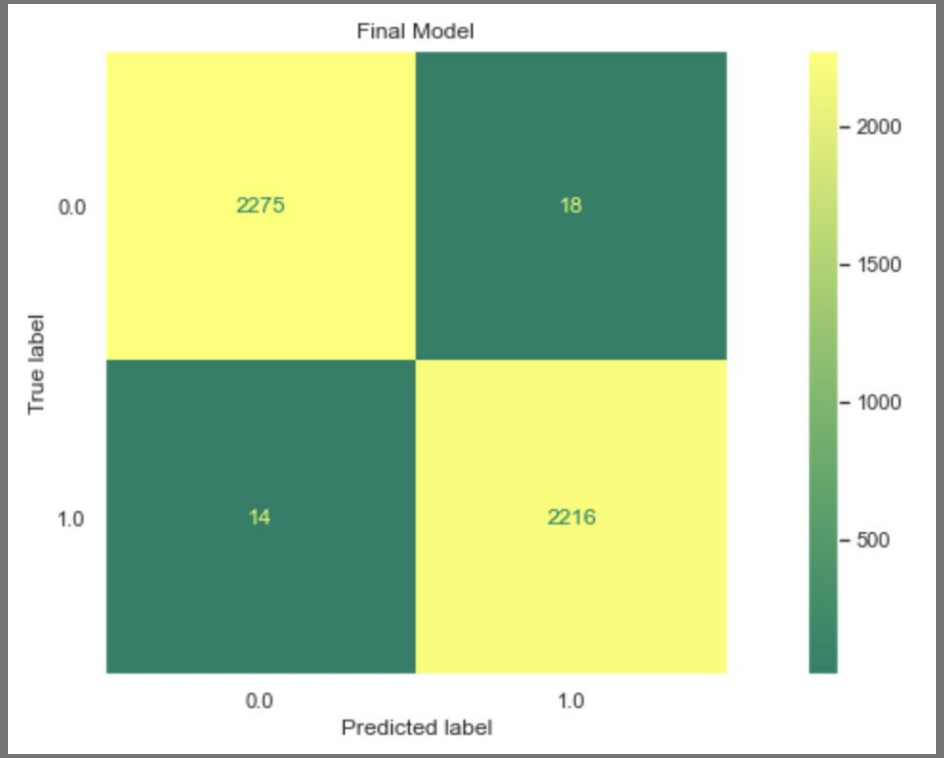
- The initial logistic regression model, without any tuning, had a precision of 0.84
- While this is a decent precision score there was room for improvement in the model.

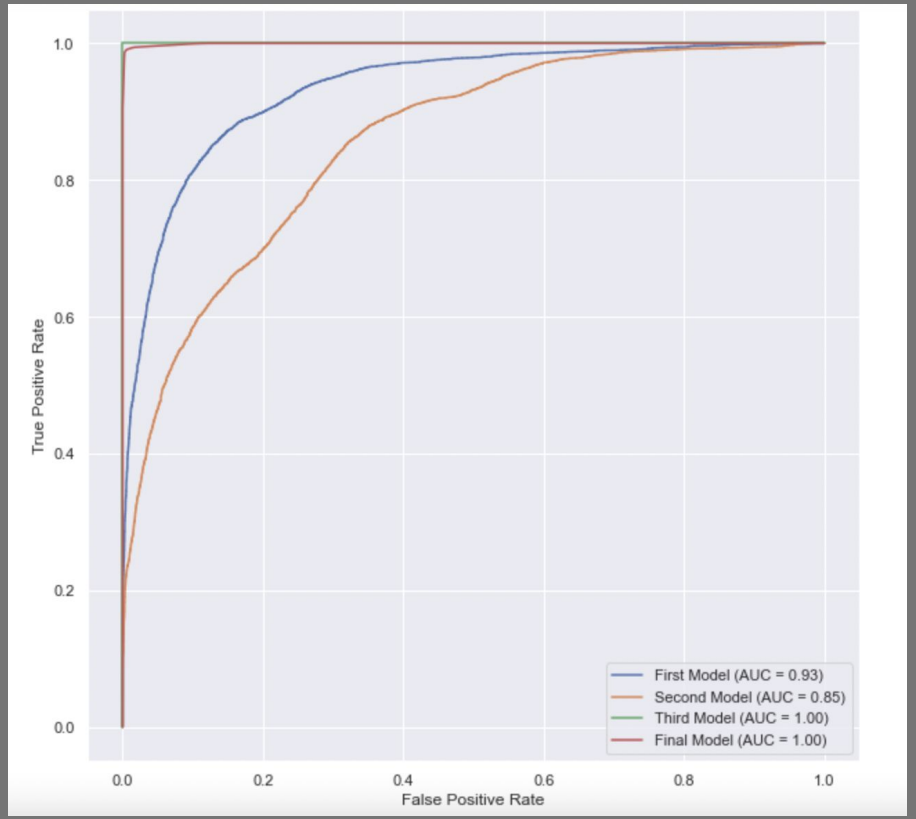
Second Model: Feature Selection

- The second model was the same as the first model, but features were selected based on their high levels of multicollinearity
- Examples of features removed from the model were
  - ring type - none
  - stem color - brown
  - stem color - white
  - cap diameter
- Surprisingly, this model did worse than my original model

Final Model: Decision Tree

- The final model was created by a process of searching through all of the tuning hyperparameters to find the best model, which landed on a decision tree model
- Features within the model that were identified as important features were cap shape, cap surface, cap color, and gill attachment
- This model gave me a precision of 0.99 for both my training and test data


Final Model
2275
18
14
2216
0.0
1.0
True label
Predicted label
2000
1500
1000
500


True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
First Model (AUC = 0.93)
Second Model (AUC = 0.85)
Third Model (AUC = 1.00)
Final Model (AUC = 1.00)

Business Recommendations

- Create an application that identifies mushrooms based on features a user can put into the application
- Create a catalog of mushrooms with their features so that users can search for those mushrooms based on those features
- Allow for users to add images and features of those mushrooms so that the database the model is built off of can continually improve itself

Future Improvements

- Combine the feature classification model with an image classification model, such as a convolutional neural network
- Include more specific features for each mushroom, such as its specific habitat, which trees they grow under, weather they grow on dead trees or alive trees, weather they grow on the ground or on the side of trees, whether they cluster or grow individually, etc.

Jacob Serfaty

Data Scientist | Bioinformatics Specialist

LinkedIn: www.linkedin.com/in/jacob-serfaty
GitHub: www.github.com/jacobserfaty