Ancient Artifacts

Team: Weixi Chen, Shea Conaway, Shuyang Lin, Sydney Simmons

Team Lead: Mubarak Ganiyu

Agenda

1. Project Overview and Research Goal
2. Data Synopsis
3. Data Science Approach & Exploratory Data Analysis
4. **Key Findings and Takeaways**

1 Project Overview

Project Objective:

Identify the locations of Mayan stone tool manufacturing areas by characterizing the composition of soil samples from ancient villages

- Ancient stonemakers would remove large debris from sites, but small particles would remain, called lithic microdebitage
- Our team trained a model to distinguish between what particles are stone and soil
- Using our model, future particle samples can be analyzed to identify stone debris and predict manufacturing areas
- Identifying tool manufacturing sites can give valuable insight into past cultural activity in Mayan villages

2 Data Synopsis

Initial Datasets

- Initial data included two datasets (one for stone and one for soil samples) in order to train our model to recognize the differences between the particles and **identify stone**

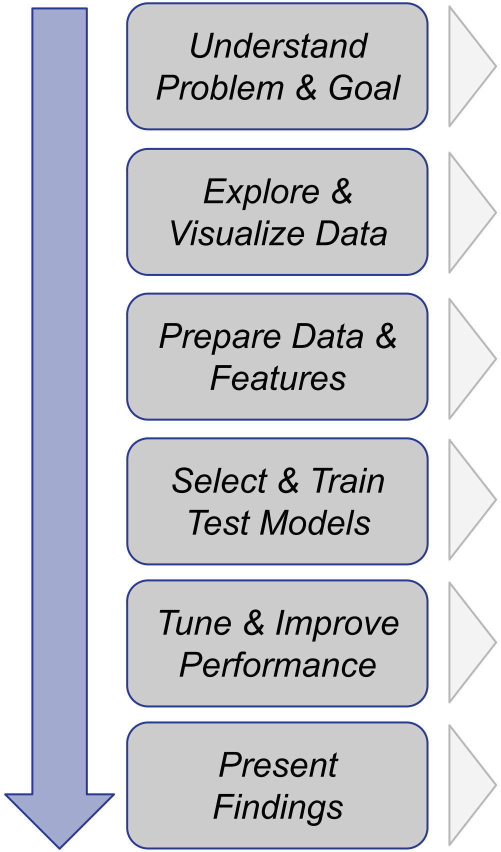
Data Attributes

- Combined dataset of both stone and soil included over **75,000 rows of data and 48 features**
- Features were numeric measurements and ratios to describe the particles, including transparency, volume, area, and thickness

Splitting Data

- After aggregation and data cleaning, the team split the data into data to use to train the model and data to use to test the performance of the model to ensure an **appropriate product for future use**

3 Data Science Approach



Understand Problem & Goal

- Model is not end goal, but part of larger project
- Check assumptions

Explore & Visualize Data

- Visualize univariate and bivariate response relationships
- Develop insights relevant to problem solution

Prepare Data & Features

- Transform provided measurements
- Engineer additional features for model

Select & Train Test Models

- Supervised learning classification problem
- Test applicable models and compare performance

Tune & Improve Performance

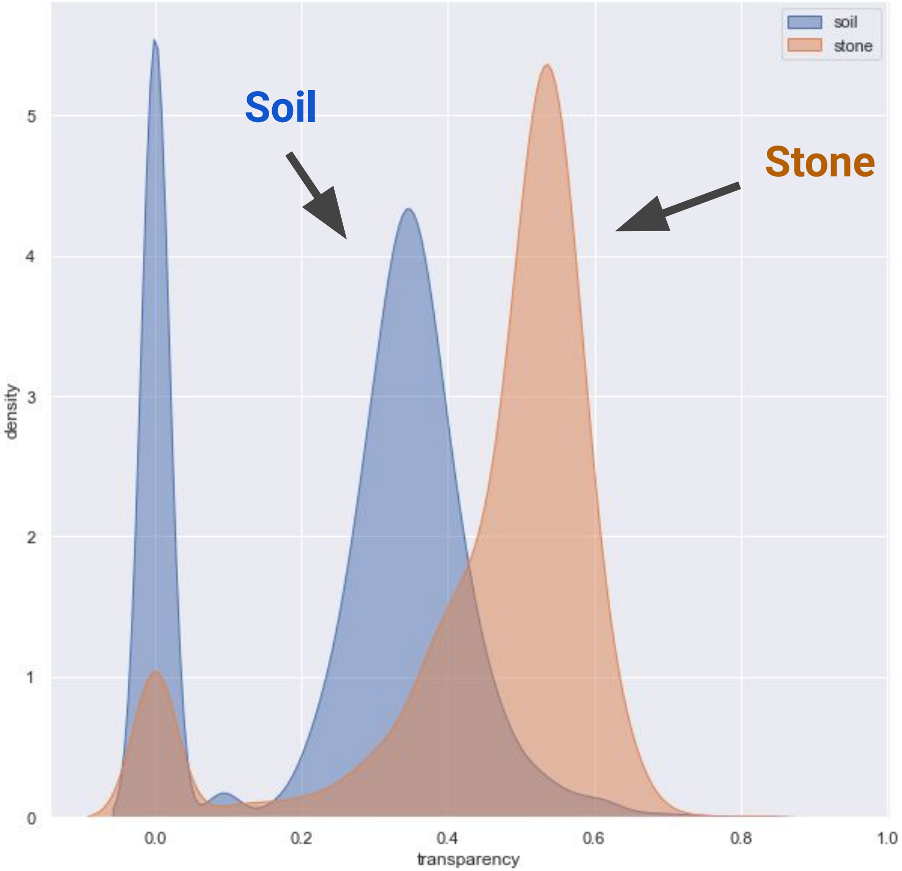
- Tune selected model
- Train on entire dataset

Present Findings

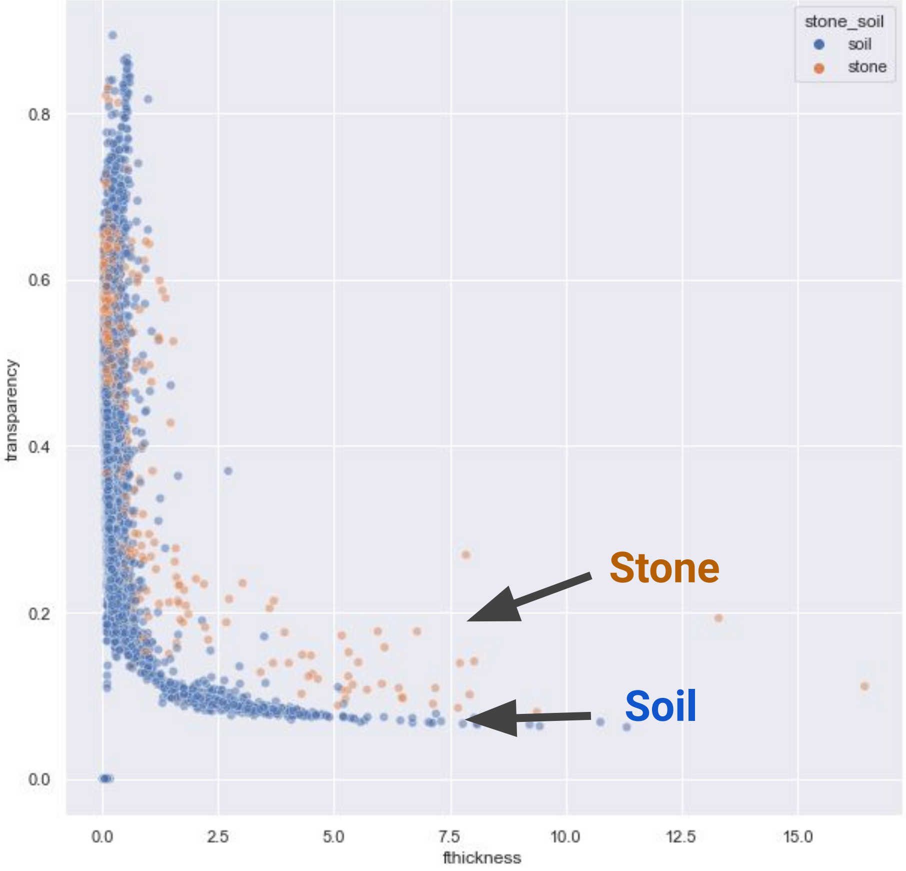
- Explain project findings
- Recommendations

3 Explore, Visualize, and Engineer Features

Transparency



Transparency x Thickness



3 Model Assessment & Selection

Initial Model Candidates	F1 Scores	
Logistic Regression	0.424	✗
Decision Tree	0.611	✓
Naive Bayes	0.139	✗
Support Vector Machine	0.206	✗

Random forest
F1: 0.688

Long training time

Extra trees
F1: 0.717

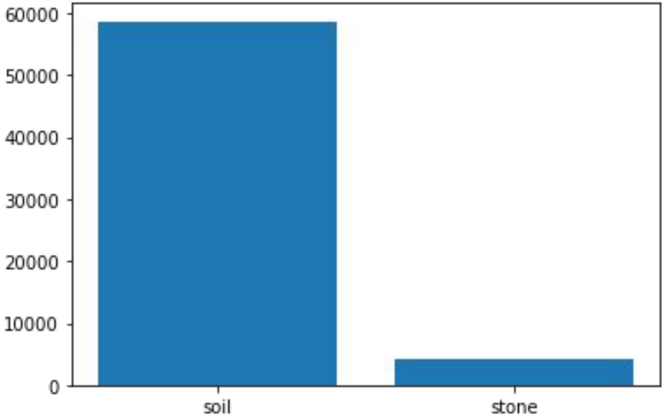
✓

4 Key Findings: Model Improvement (1/2)

- We selected **Extra Trees Classifier** as our final model model to limit overfitting and maintain a fast, effective model for the end user.
- In order to improve our model, we also:
  - **Added more stone particles to our training data**
  - Limited the used features using variable importance

- Reorganized the train and test size with respect to stone and soil.
- Used grid search cross validation to look for better hyper-parameters.
- Increased random state and applied bootstrap to improve out-of-bag performance.

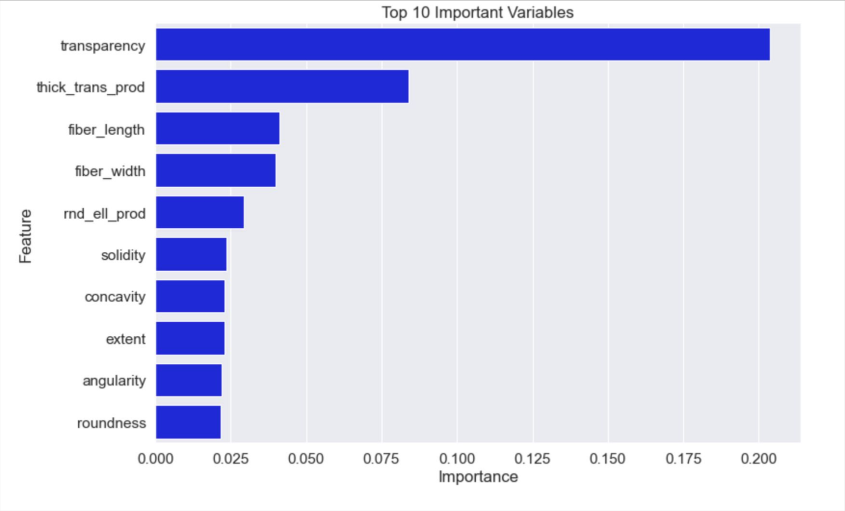
Histogram of Stone or Soil



4 Key Findings: Model Improvement (2/2)

- We selected **Extra Trees Classifier** as our final model model to limit overfitting and maintain a fast, effective model for the end user.
- In order to improve our model, we also:
  - Added more stone particles to our training data
  - **Limited the used features using variable importance**

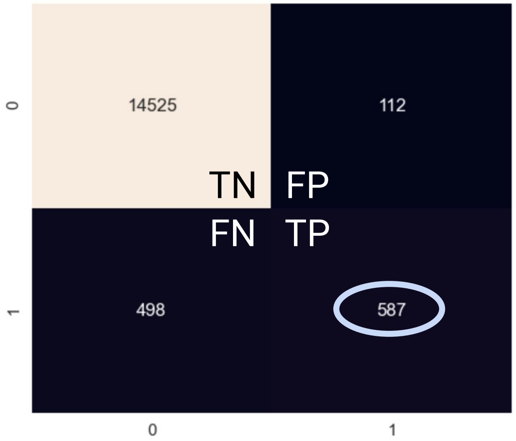
- Exported the variable importance table from the initial model
- Figured out the importance of size and transparency
- Attempted to do a sub-model on samples with a larger size



4 Key Findings: Model Performance

- In improving our model, we focused on increasing **recall and F1 score**, limiting false positives and negatives and increasing true positives (when the model predicted stone and the particle was stone)

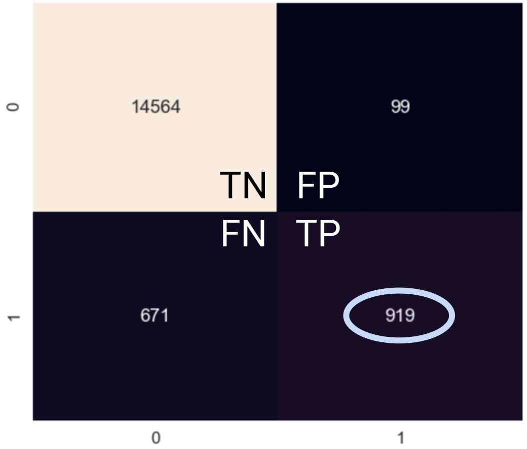
Original Extra Trees Model



	0	1
0	14525	112
1	498	587

Accuracy:	**0.96**
Precision:	**0.84**
Recall:	**0.54**
F1:	**0.66**

Improved Extra Trees Model



	0	1
0	14564	99
1	671	919

Accuracy:	**0.95**
Precision:	**0.91**
Recall:	**0.59**
F1:	**0.72**

4 Next Steps and Recommendations

- Sample sizes need to be large enough to have different particle sizes
- Soil / stone determination most accurate on big particles with large perimeters
- Constant model refinement and use to locate stone sites
  - Produce more samples for training
  - Evaluate real-world confusion manually for feedback and improve model
  - Monitor particle analyzer performance and continue to decrease false negative results