ASME IDETC-CIE 2023

International Design Engineering Technical Conferences
& Computers and Information in Engineering Conference

CONFERENCE AUGUST 20–23, 2023

ASME IDETC2023 Hackathon – Autodesk Problem

Automating Material Selection for Product Design

Lequn Chen

Contact: chen1470@e.ntu.edu.sg

PhD Candidate

Nanyang Technological University, Singapore

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

ASME IDETC-CIE 2023
International Design Engineering Technical Conferences & Computers and Information in Engineering Conference

ASME SETTING THE STANDARD
EWF
The American Society of Mechanical Engineers®
ASME®

CONTENTS

01 Dataset Analysis

02 Feature Engineering

03 ML Modelling

04 Evaluation and Discussion

05 Future Works

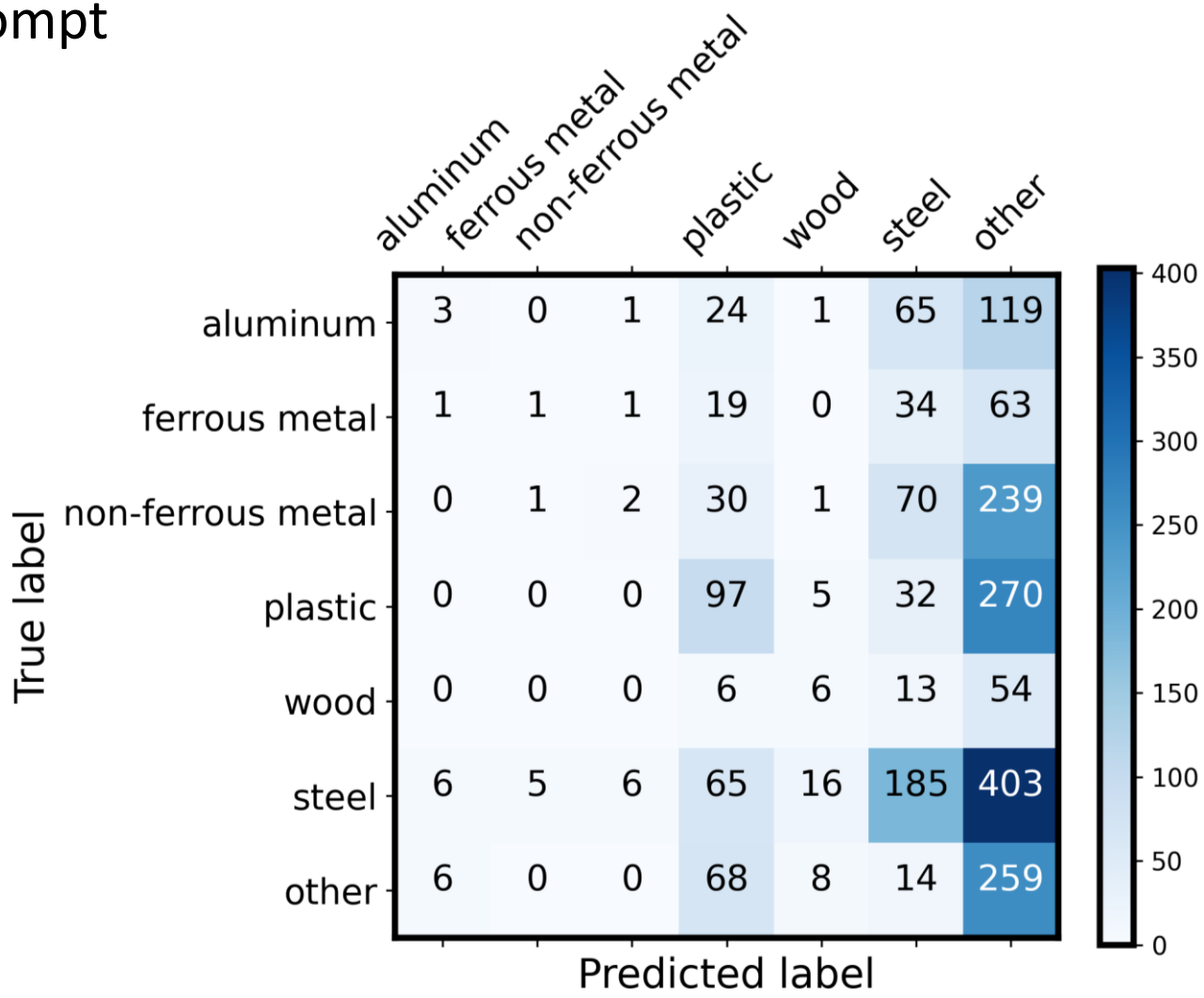
1. Dataset Analysis

Assembly JSON file hierarchical structure

- Keys: (['tree', 'root', 'occurrences', 'components', **'bodies'**, 'as_built_joints', 'contacts', 'holes', **'properties'**])
- Properties:
  - Assembly level information
  - *['name', 'bounding_box', 'vertex_count', 'edge_count', 'face_count', 'loop_count', 'shell_count', 'body_count', 'area', 'volume', 'density', 'mass', 'center_of_mass', 'principal_axes', 'xyz_moments_of_inertia', 'surface_types', 'vertex_valence', 'design_type', 'likes_count', 'comments_count', 'views_count', 'products', 'categories', 'industries']*)
- Bodies:
  - Individual Component level information
  - *{'name': 'Body7', 'type': 'BRepBody', 'png': 'e397626e-0608-11ec-a470-02052889e3af.png', 'smt': 'e397626e-0608-11ec-a470-02052889e3af.smt', 'step': 'e397626e-0608-11ec-a470-02052889e3af.step', 'obj': 'e397626e-0608-11ec-a470-02052889e3af.obj', 'physical_properties': {'center_of_mass': {'type': 'Point3D', 'x': 0.25315821192952176, 'y': 0.03333797923237883, 'z': 2.7849977545097624}, 'area': 19.067095914836415, 'volume': 1.12856425487885 2}, 'material_category': 'Plastic'}*

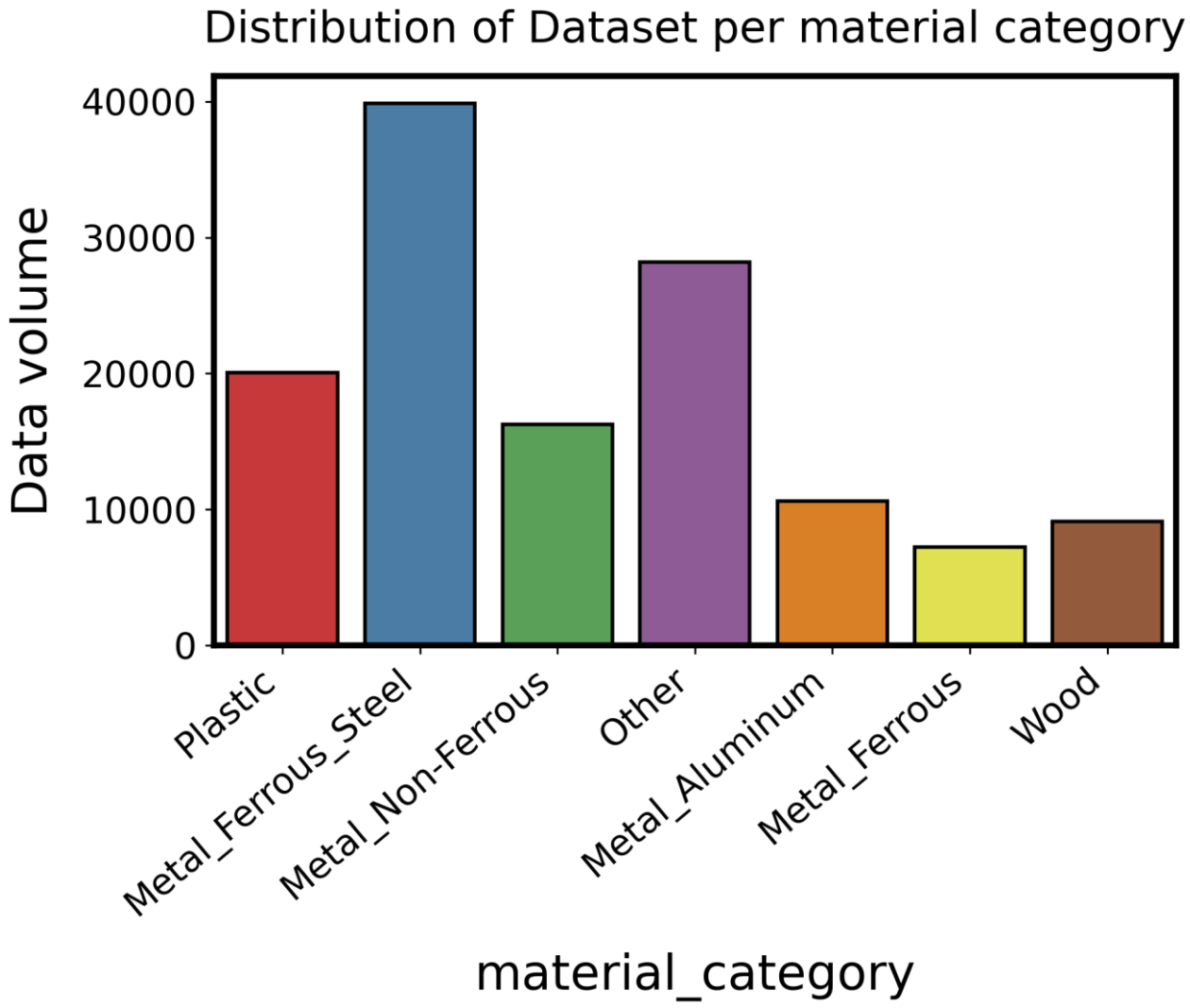
ASME IDETC-CIE 2023 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference
The American Society of Mechanical Engineers® ASME®

Baseline Model performance – GPT3.5 + only “name” in the prompt



ASME IDETC-CIE 2023
International Design Engineering Technical Conferences & Computers and Information in Engineering Conference
The American Society of Mechanical Engineers® ASME®

1. Dataset Analysis (Training set)



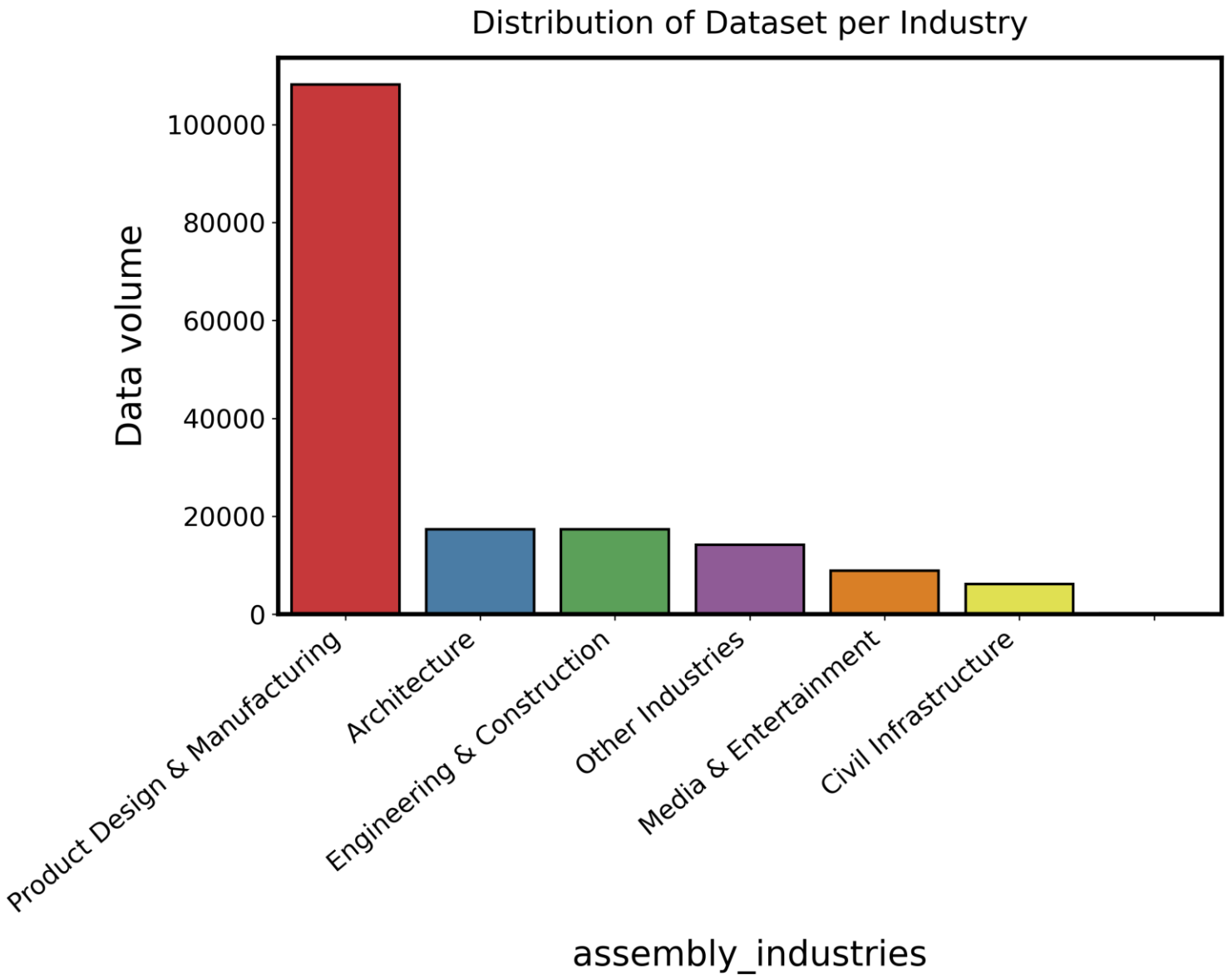
Notes:

- ❑ **Dataset is imbalanced**
- ❑ **Need to consider this factor during the modelling process.**

ASME IDETC-CIE 2023
International Design Engineering Technical Conferences & Computers and Information in Engineering Conference

The American Society of Mechanical Engineers ®
ASME®

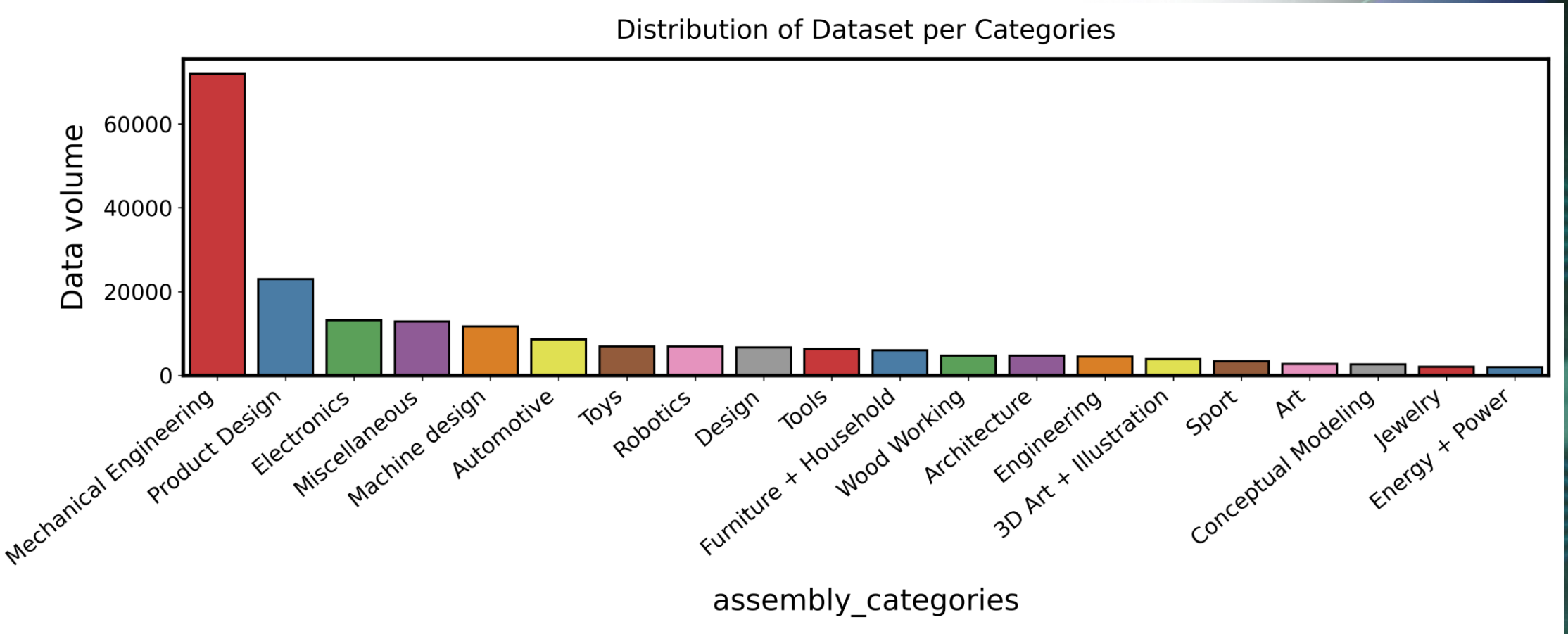
1. Dataset Analysis (Training set)



Notes:

- ❑ **Categorical Data.**
- ❑ **A part could belong to multiple industries.**
- ❑ **Multi-label Binarization for multiple categorical features**

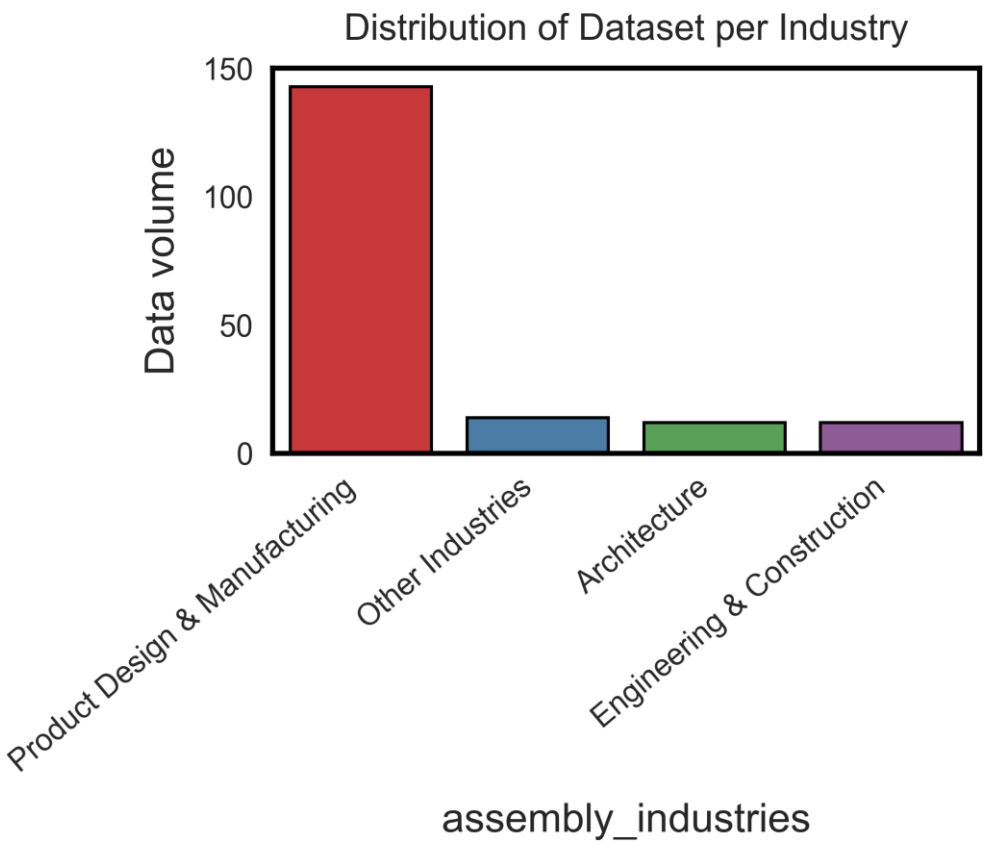
1. Dataset Analysis (Training set)

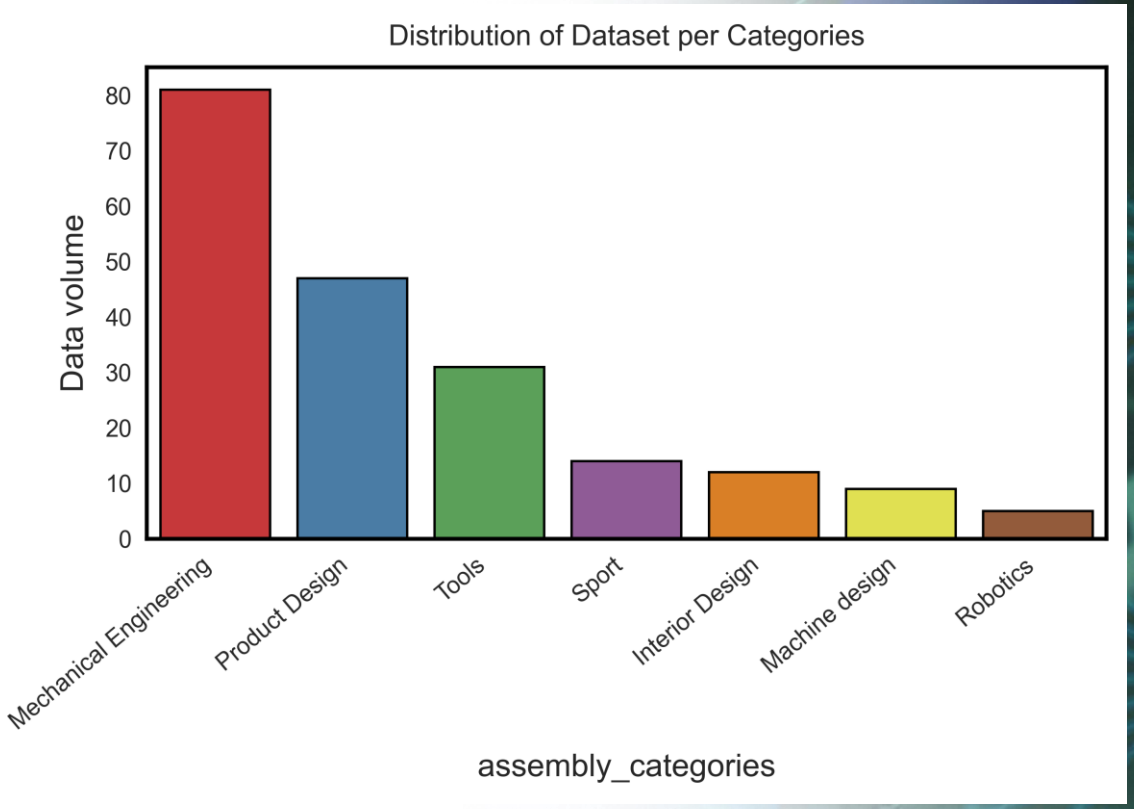


Notes:

- ❑ **Categorical Data.**
- ❑ **A part could belong to multiple categories.**
- ❑ **Multi-label Binarization for multiple categorical features**

1. Dataset Analysis (Test set)





- ❑ Test set has fewer industries, categories.
- ❑ Obviously, it is not the same distribution as in the training set. This means there could be a distribution shift, so we need to make our ML as robust as possible.

2. Feature Engineering

- **tree and root:** represent the *hierarchical structure* of the design. The depth of the tree (how many levels of hierarchy are there) might be indicative of the complexity of the assembly.
  - Potential feature: ***depth*** of the tree.

- **Occurrences:**
  - *Give instances of components. These can show how many times a component is repeated in a design.*
  - *Extracted: number of unique components, total number of components (including repetitions).*

- **Components**:
  - sub-assemblies
  - Potential features: number of components, average number of bodies per component.

- **joints**:
  - define the relative position and orientation between occurrences.
  - Potential features: number of joints, *distribution of different types* of joints (e.g., ball, revolute, slider), *average degrees of freedom* per joint.

2. Feature Engineering

- **as_built_joints:** joints maintain their position and don't get transformed
  - Potential features: number of as_built_joints.

- **Holes**:
  - Holes in the assembly can be crucial for certain materials (e.g., *metals might be preferred where there are holes for screws*).
  - **Features extracted (body-level feature)**:
    (1) average *size of holes* across a body
    (2) average *lengths of holes* across a body
    (3) Distribution of holes across each bodies: indicate which bodies have the highest number of holes and potentially indicate their significance in the assembly.
      (1) Total number of holes in each body,
      (2) distribution of *hole types* (e.g., RoundHoleWithBlindBottom, RoundBlindHoleWithSphericalBottom, etc.)

ASME IDETC-CIE 2023
International Design Engineering Technical Conferences & Computers and Information in Engineering Conference

ASME SETTING THE STANDARD
EWF
The American Society of Mechanical Engineers®
ASME®

2. Feature Engineering

- **Physical Properties (global assembly-level feature)**: Area, mass, density, and moments of inertia (xx, yy, zz, xy, yz, xz); Different assemblies will have distinct densities and mass distributions.

- **Geometry Information (individual bodies)**:
  - ***Bounding Box Dimensions***
  - ***Vertex*** *count,* ***Edge*** *count,* ***Face*** *count,* ***Loop*** *count,* ***Shell*** *count, Body count*: These counts can give an idea about the **complexity** of the geometry.
  - ***Surface Types***: The types of surfaces (e.g., cylindrical, planar, toroidal) and their counts can provide insights into the *manufacturing processes* suitable for the part, which in turn might hint at the material.
    - Surfaces such as ***planes*** might be used in structural parts of an assembly, demanding materials with *higher strength*. On the other hand, ***cylindrical*** or toroidal surfaces might indicate rotational parts, hinting at the need for materials with *low friction properties*.

2. Feature Engineering

- **Categories and Industries (assembly level feature)**: The industry and category the design is intended for can be a strong hint. For instance, designs in the "Aerospace" category might prefer lightweight and strong materials like titanium or composite materials.

- **Vertex Valence:** number of edges incident to a vertex
  - **High Valence**: many components or features are connected at that point, potential require more *durable and robust material*
  - ***Distribution***: A higher average vertex valence might suggest a densely interconnected assembly.

- **Derived Features**: Some features can be derived from the existing data to enhance its expressiveness. For instance:

  - *Aspect Ratios*: Derived from the *bounding box dimensions*, they can indicate if a part is *elongated, flat, or almost cubic*, which might influence material choice.

  - *Complexity Metrics*: Ratios like ***vertex-to-edge count***, ***face-to-loop count***, can be indicative of the overall **complexity** of the design.

  - *Volume fraction*: Body volume divided by the assembly volume. Large volume fraction could indicate *primary component* and might be made of a material suitable for *bearing loads* or ***stresses***. Conversely, small volume fractions might represent *accessory* or *supplementary* components.

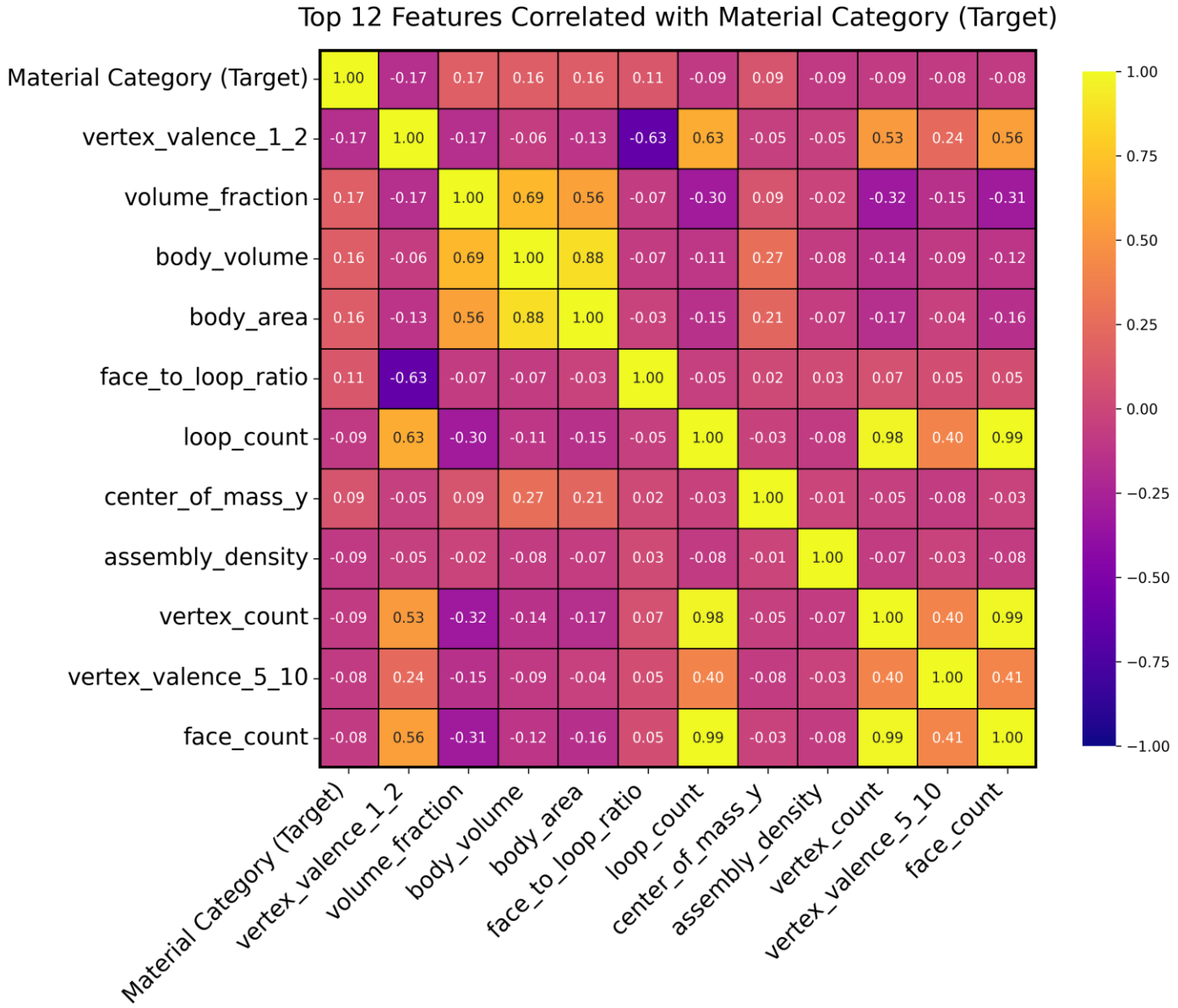
2. Feature Engineering

- ***Derived feature***:
  - *Center of mass distance*: the ***Euclidean distance*** between the body's center of mass (COM) and the assembly's COM. → how the material is distributed within the assembly; gives an indication of how central or peripheral a body is within the assembly.

- **Online Metrics**: While these metrics (**likes, comments, views**) might not directly influence the material, they might *correlate with certain material choices (**popularity**)* if **popular** designs tend to use specific materials.

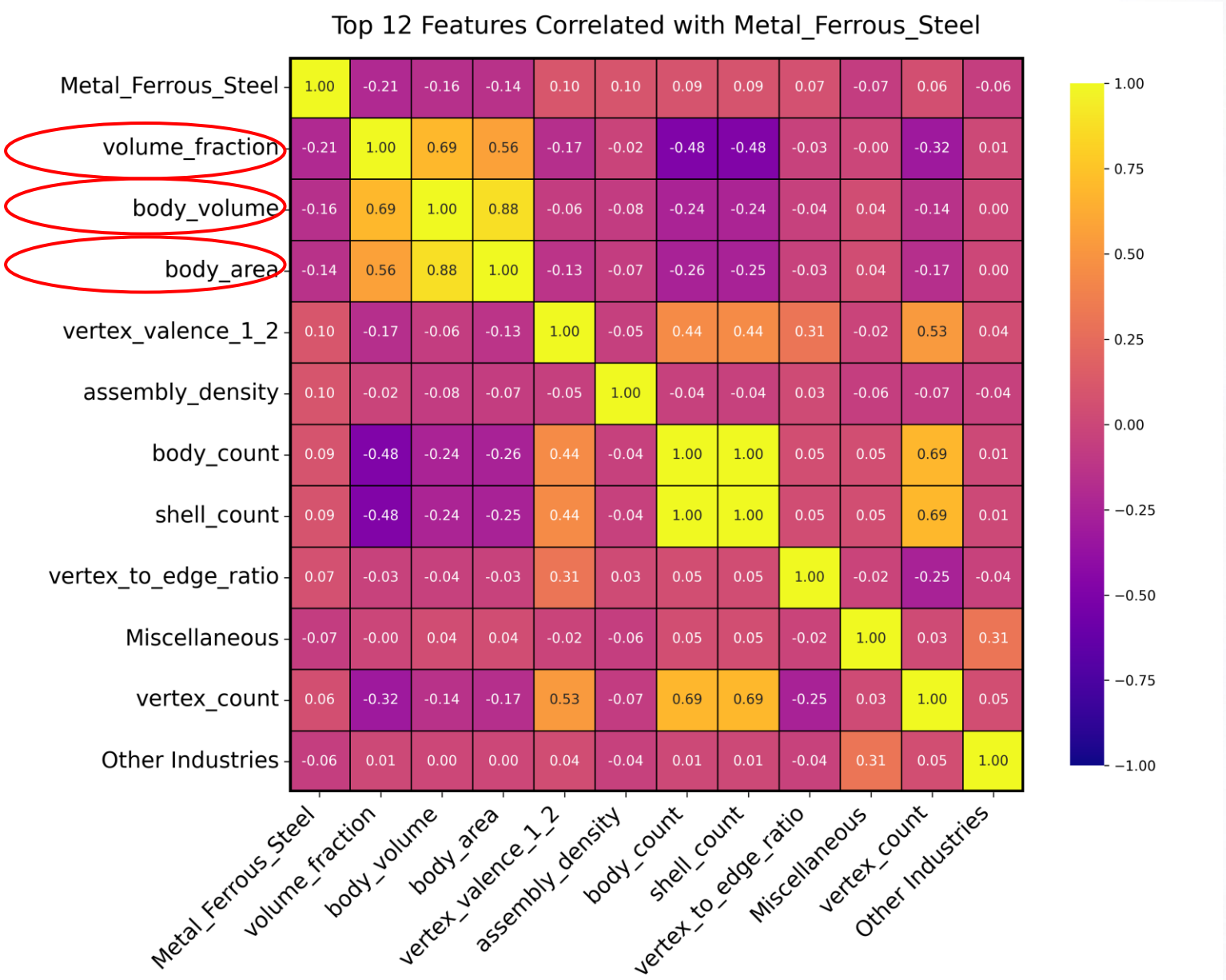
- **Components and Hierarchical Structure (future work)**: The relationships between components can be meaningful. If certain sub-components are made of a specific material, it might influence the material choice of the main body. However, this requires a more complex modeling approach to take **hierarchy** into account. → potentially need to use **knowledge graph** to model the relationship

2. Feature Selection – Spearman's Rank Correlation matrix

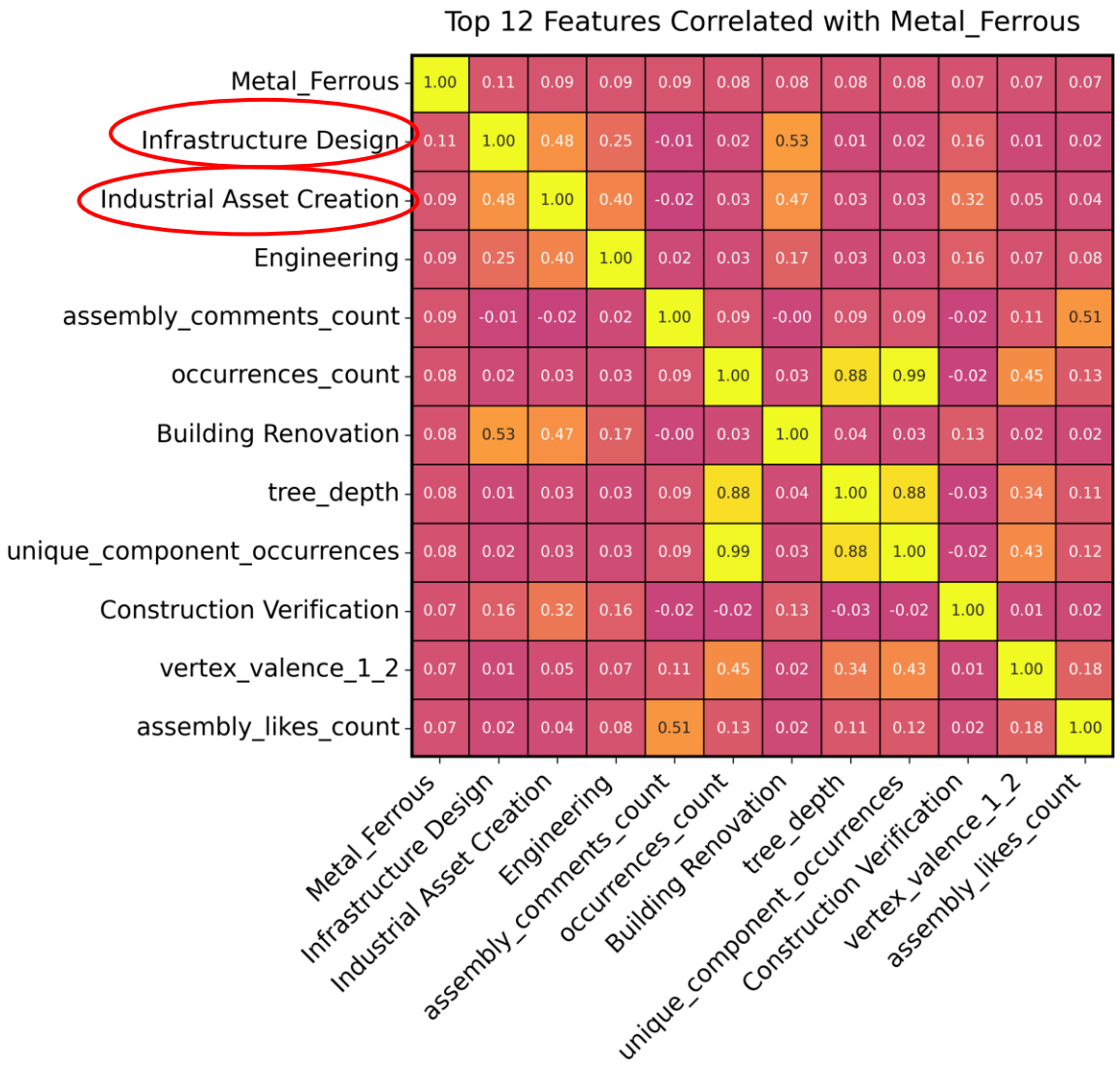


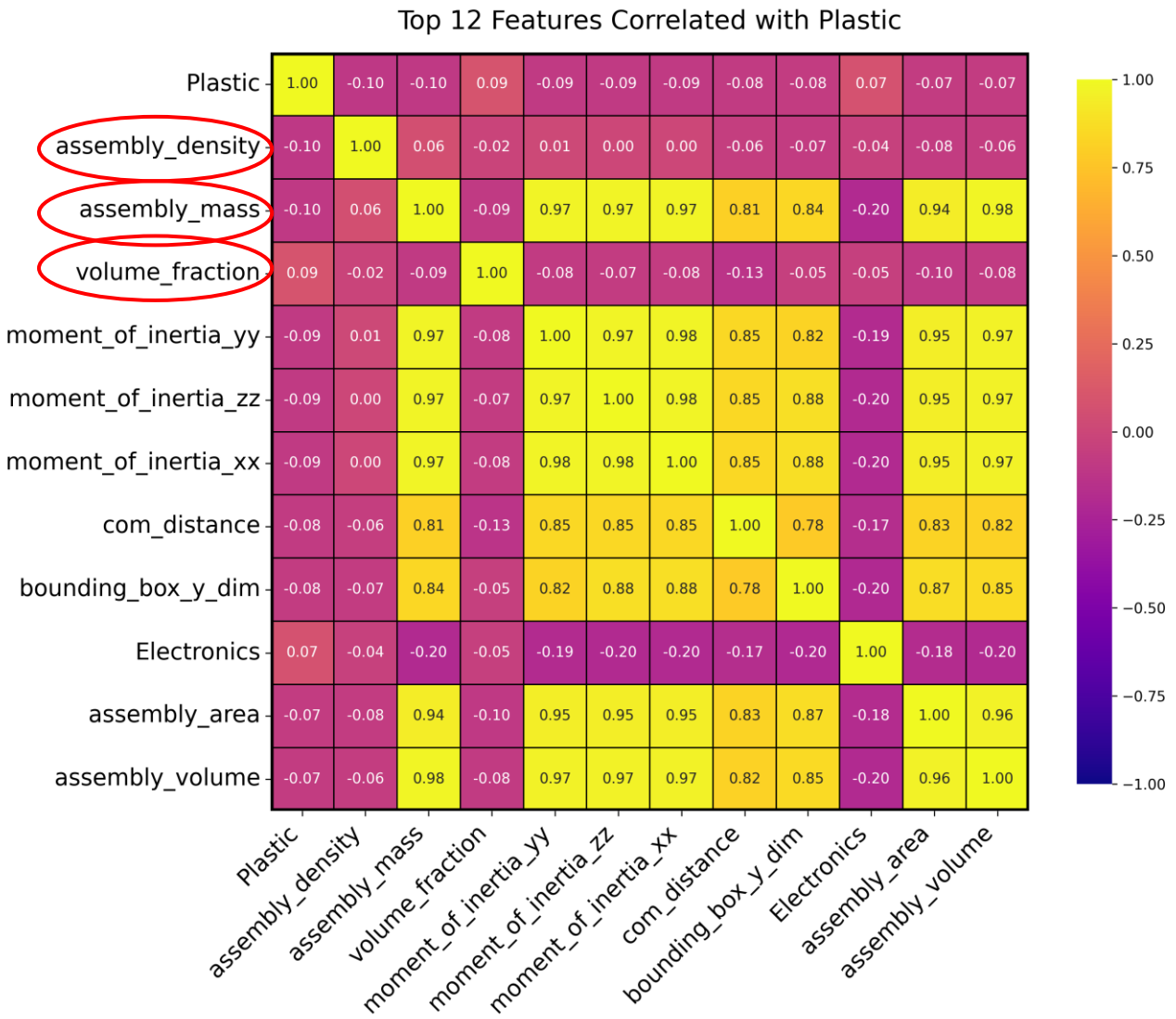
- The relationship between variables might *not be strictly linear*, **Spearman's** matrix is chosen to evaluate the feature monotonic relationships
- ***Multicollinearity***: highly correlated independent variables can lead to multicollinearity, which might make the model unstable. → need to identify
- ***Feature Importance***: By looking at the correlation of each feature with the *target variable*, we can get the importance or influence of each feature on the target.

2. Feature Selection – Spearman's Rank Correlation matrix



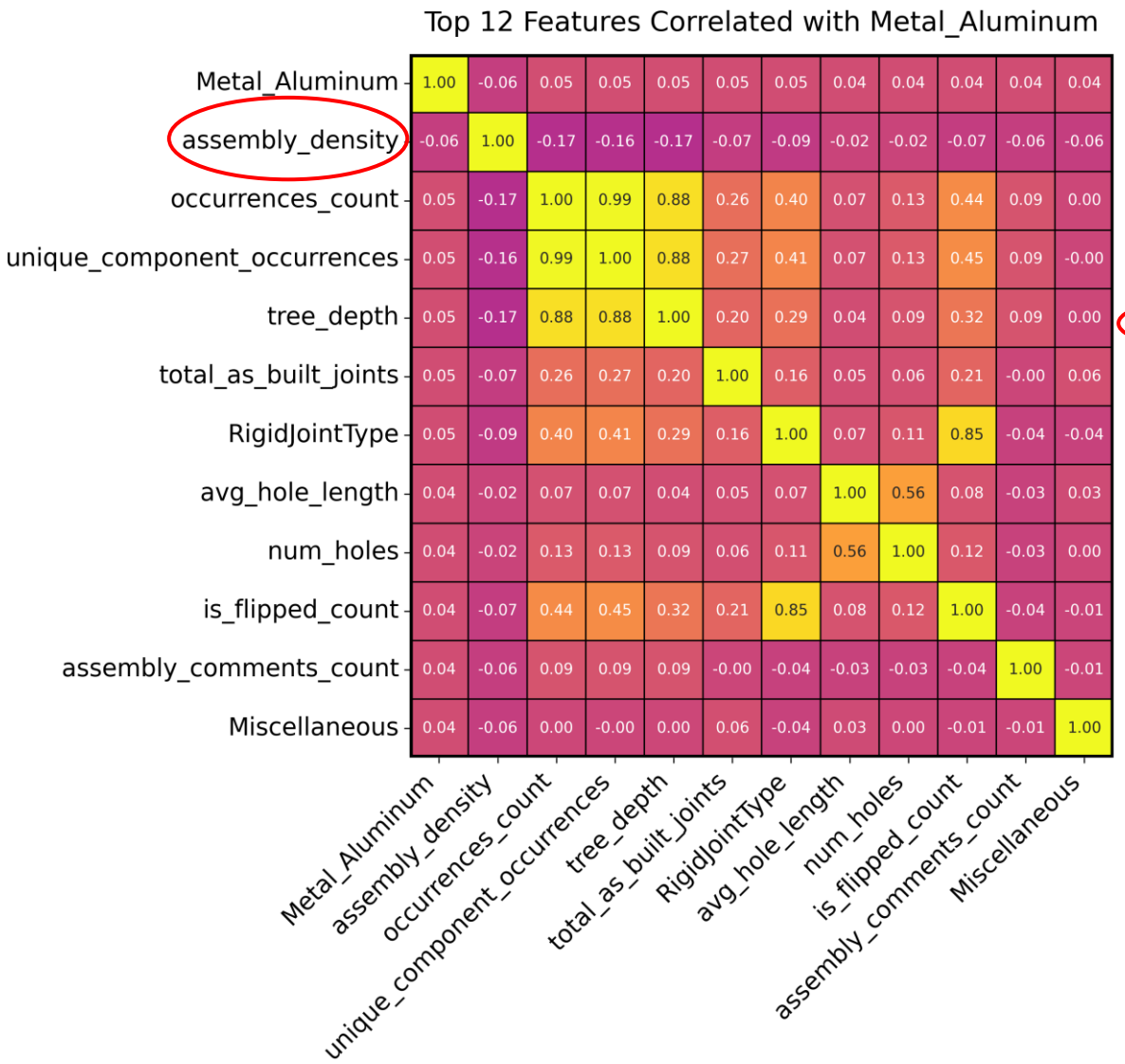
2. Feature Selection – Spearman's Rank Correlation matrix with *individual material* category

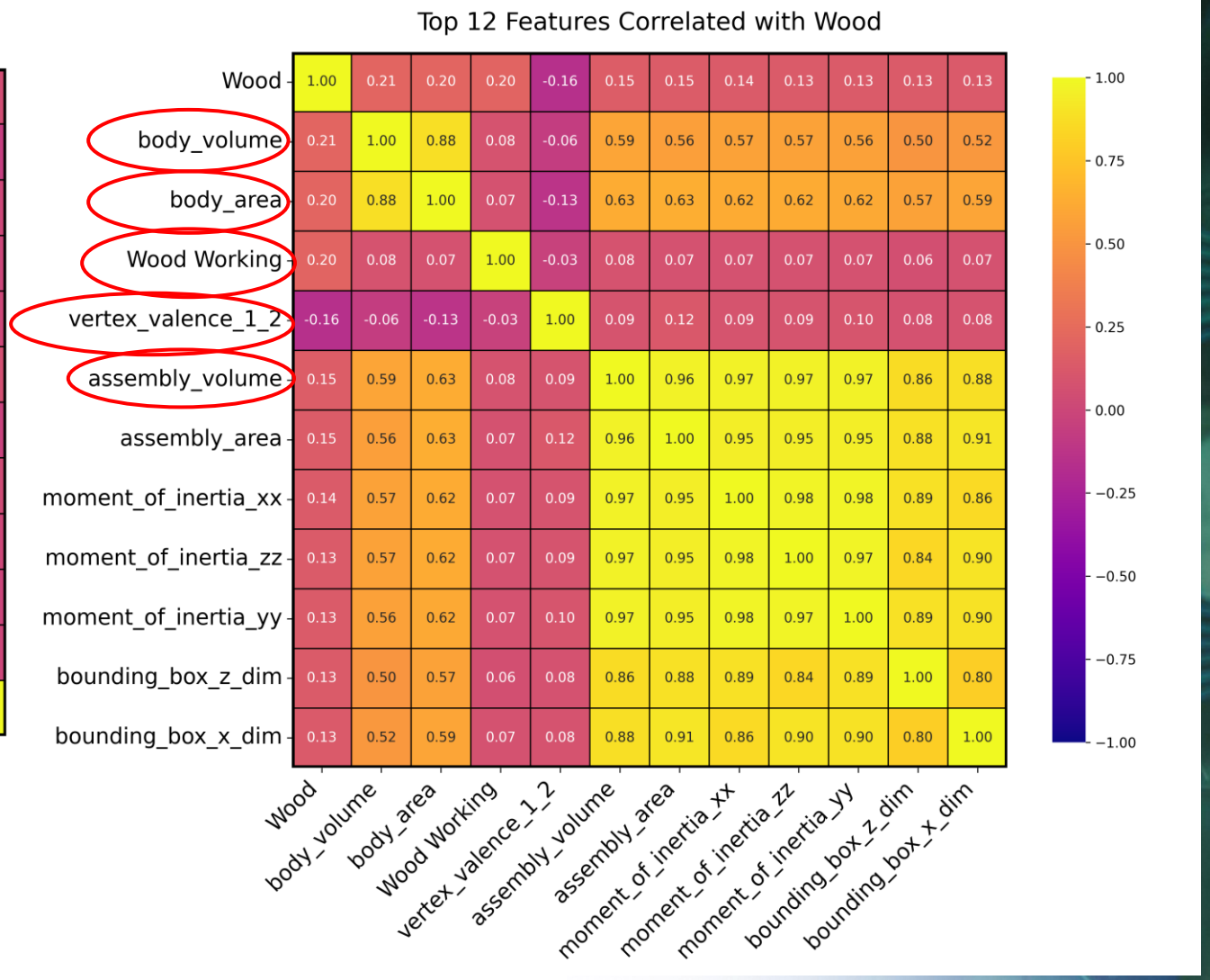




ASME SETTING THE STANDARD
EWF
The American Society of Mechanical Engineers ®
ASME®

2. Feature Engineering – Spearman's Rank Correlation matrix with *individual material* category





2. Feature Engineering – Spearman's Rank Correlation matrix with *individual material* category

Top 12 Features Correlated with Metal_Non-Ferrous

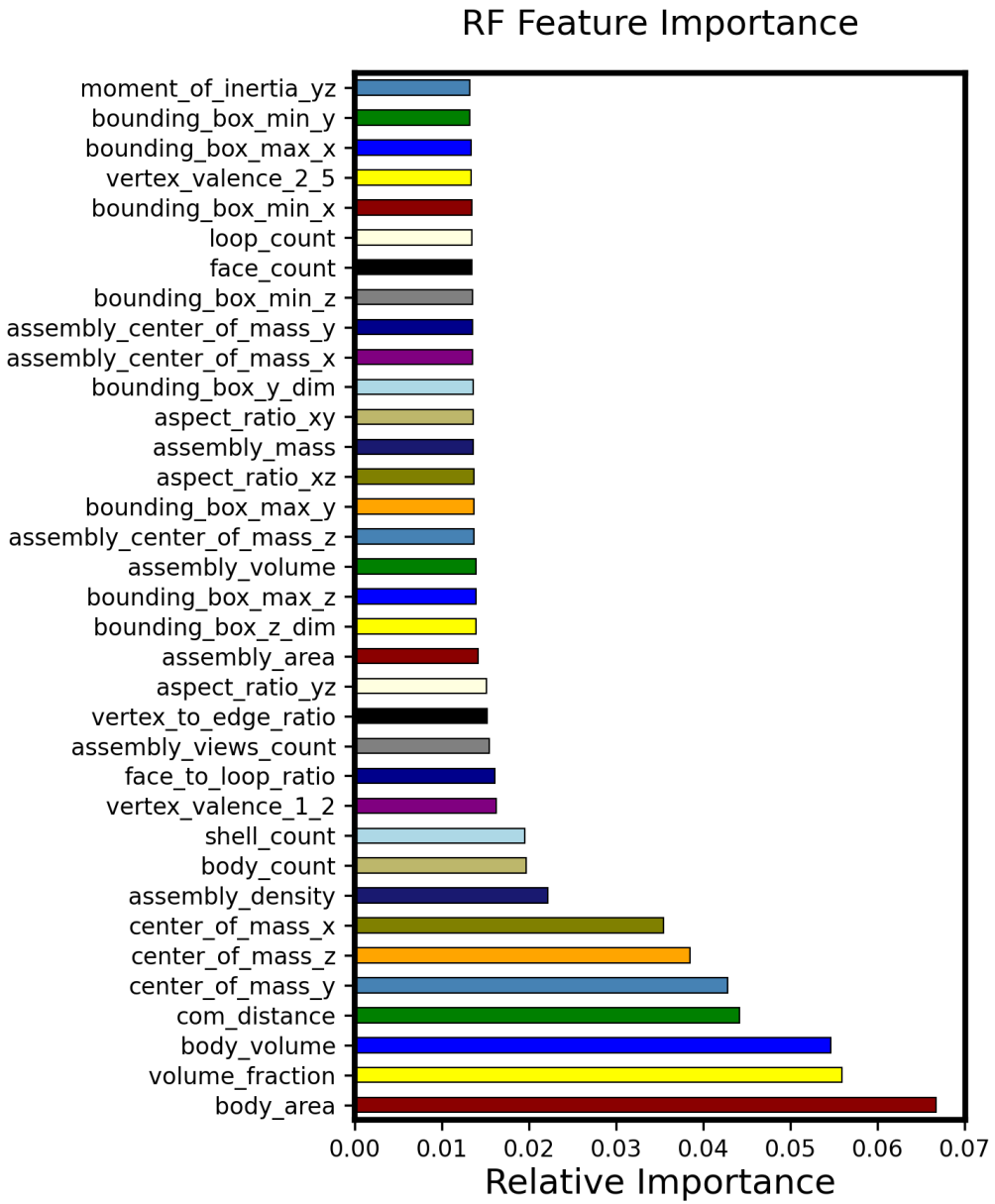
	Metal_Non-Ferrous	body_area	Jewelry	assembly_area	assembly_volume	bounding_box_x_dim	bounding_box_max_y	body_volume	moment_of_inertia_zz	bounding_box_y_dim	assembly_density	bounding_box_max_x
Metal_Non-Ferrous	1.00	-0.10	0.10	-0.09	-0.09	-0.08	-0.08	-0.08	-0.08	-0.08	0.08	-0.08
body_area	-0.10	1.00	-0.09	0.63	0.63	0.59	0.55	0.88	0.62	0.61	-0.07	0.53
Jewelry	0.10	-0.09	1.00	-0.11	-0.11	-0.06	-0.08	-0.08	-0.11	-0.09	-0.01	-0.05
assembly_area	-0.09	0.63	-0.11	1.00	0.96	0.91	0.78	0.56	0.95	0.87	-0.08	0.79
assembly_volume	-0.09	0.63	-0.11	0.96	1.00	0.88	0.77	0.59	0.97	0.85	-0.06	0.77
bounding_box_x_dim	-0.08	0.59	-0.06	0.91	0.88	1.00	0.70	0.52	0.90	0.78	-0.06	0.87
bounding_box_max_y	-0.08	0.55	-0.08	0.78	0.77	0.70	1.00	0.48	0.80	0.89	-0.06	0.63
body_volume	-0.08	0.88	-0.08	0.56	0.59	0.52	0.48	1.00	0.57	0.53	-0.08	0.46
moment_of_inertia_zz	-0.08	0.62	-0.11	0.95	0.97	0.90	0.80	0.57	1.00	0.88	0.00	0.79
bounding_box_y_dim	-0.08	0.61	-0.09	0.87	0.85	0.78	0.89	0.53	0.88	1.00	-0.07	0.69
assembly_density	0.08	-0.07	-0.01	-0.08	-0.06	-0.06	-0.06	-0.08	0.00	-0.07	1.00	-0.06
bounding_box_max_x	-0.08	0.53	-0.05	0.79	0.77	0.87	0.63	0.46	0.79	0.69	-0.06	1.00

Top 12 Features Correlated with Other

	Other	tree_depth	occurrences_count	unique_component_occurrences	body_area	vertex_valence_1_2	face_to_loop_ratio	body_volume	Archeology	bounding_box_max_z	bounding_box_z_dim	total_as_built_joints
Other	1.00	-0.12	-0.12	-0.12	0.11	-0.09	0.09	0.08	0.08	0.07	0.06	-0.06
tree_depth	-0.12	1.00	0.88	0.88	-0.02	0.34	-0.20	-0.00	-0.01	0.04	0.05	0.20
occurrences_count	-0.12	0.88	1.00	0.99	0.01	0.45	-0.25	0.05	0.01	0.08	0.11	0.26
unique_component_occurrences	-0.12	0.88	0.99	1.00	0.02	0.43	-0.24	0.05	0.02	0.08	0.11	0.27
body_area	0.11	-0.02	0.01	0.02	1.00	-0.13	-0.03	0.88	-0.03	0.47	0.57	0.03
vertex_valence_1_2	-0.09	0.34	0.45	0.43	-0.13	1.00	-0.63	-0.06	-0.03	0.07	0.08	0.11
face_to_loop_ratio	0.09	-0.20	-0.25	-0.24	-0.03	-0.63	1.00	-0.07	0.04	-0.00	-0.01	-0.07
body_volume	0.08	-0.00	0.05	0.05	0.88	-0.06	-0.07	1.00	-0.04	0.40	0.50	0.05
Archeology	0.08	-0.01	0.01	0.02	-0.03	-0.03	0.04	-0.04	1.00	-0.01	-0.01	-0.02
bounding_box_max_z	0.07	0.04	0.08	0.08	0.47	0.07	-0.00	0.40	-0.01	1.00	0.81	0.04
bounding_box_z_dim	0.06	0.05	0.11	0.11	0.57	0.08	-0.01	0.50	-0.01	0.81	1.00	0.04
total_as_built_joints	-0.06	0.20	0.26	0.27	0.03	0.11	-0.07	0.05	-0.02	0.04	0.04	1.00

ASME SETTING THE STANDARD
EWF
The American Society of Mechanical Engineers® ASME®

2. Feature Engineering – RF feature importance ranking



- The features that are most influential in making predictions.
- The highest importance in our case is around 0.065, which is considered low.
- It means even a small importance score can be significant.
- Features are selected from the correlation analysis and feature importance analysis

2. Feature Engineering – Correlation matrix

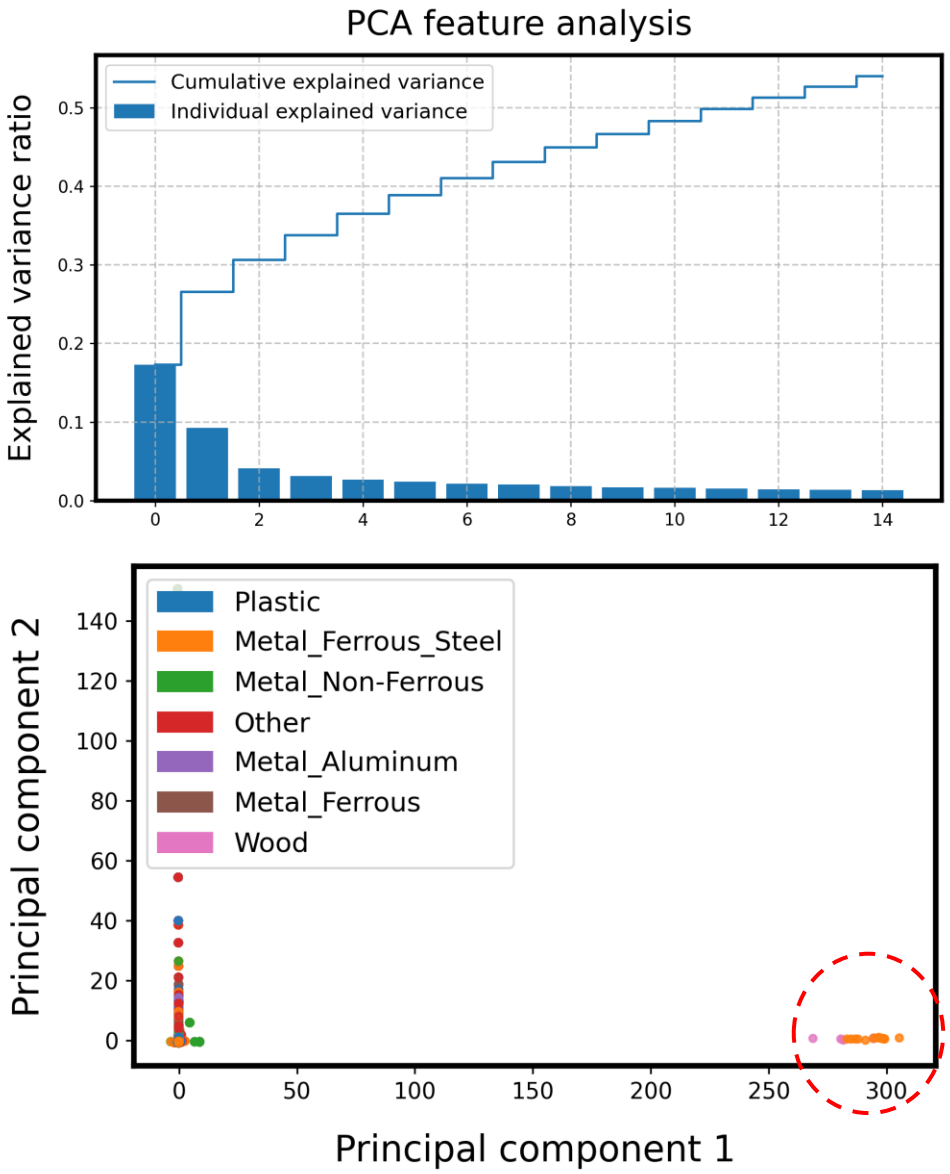
Observations:

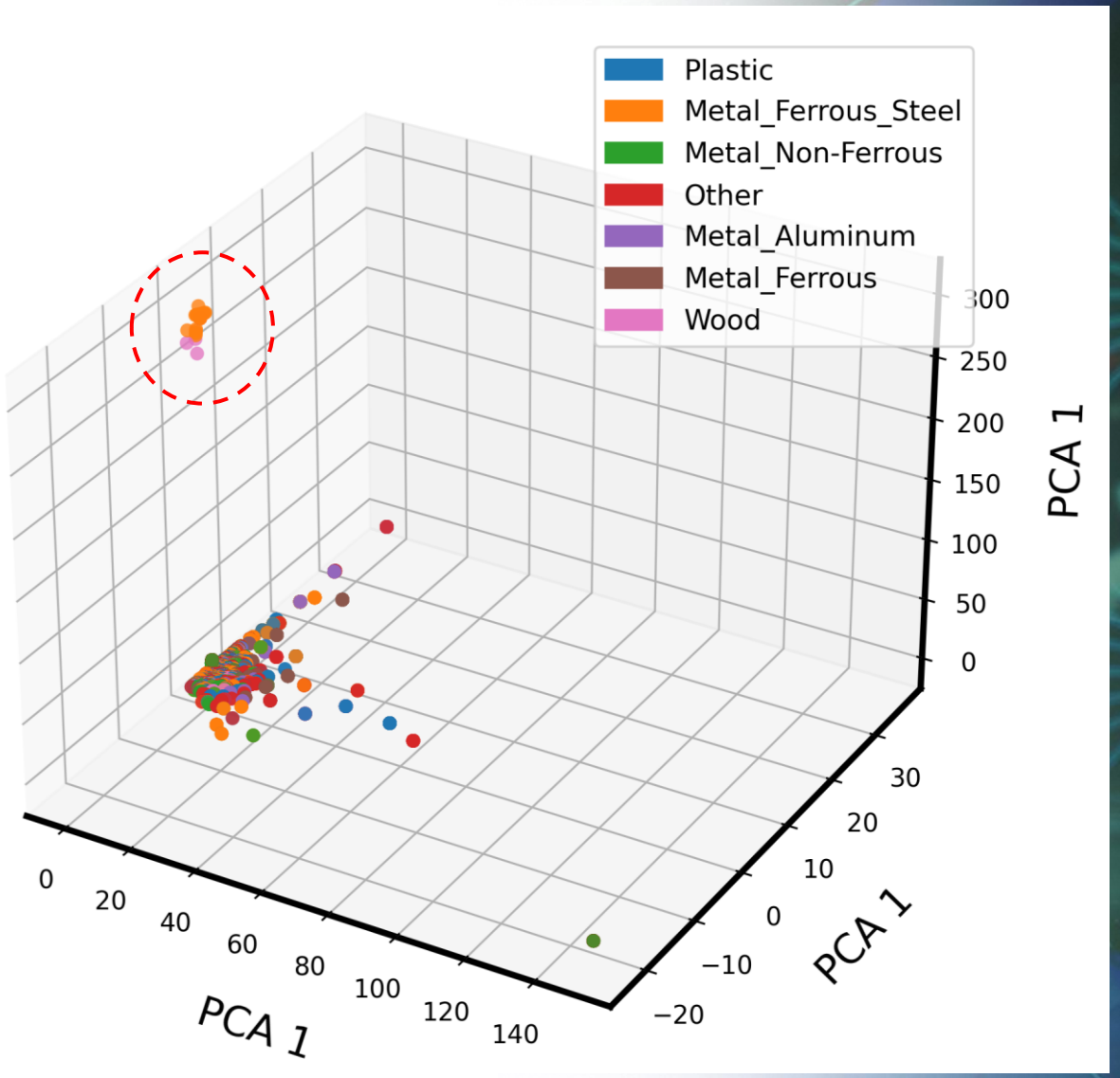
- **Weak Correlation**: This suggests that the relationship between these features and the target might be *complex* and highly *non-linear*.
  - more complex models that can capture non-linearities (e.g., decision trees, random forests, or neural networks) might be effective.
  - there's room for more *feature engineering* or to derive features or external features could be beneficial for prediction.
  - Inherent Complexity

ASME IDETC-CIE 2023
International Design Engineering Technical Conferences & Computers and Information in Engineering Conference

ASME SETTING THE STANDARD
EWF
The American Society of Mechanical Engineers® ASME®

3. Visualization and dimensionality reduction





3. ML modelling

General Procedure:

- ❑ Train-validation split = 8:2. Data Normalization (StandardScaler).
- ❑ Selected models (7): **["KNN", "DT", "AdaBoost", "RF", "BalancedRandomForest", 'LightGBM', "XGBoost"]**
- ❑ 5-fold cross-validation for hyperparameter tuning
  - Initial tuning with **RandomSearchCV**
  - Fine-tuning: **GridSearchCV**
- ❑ Best hyperparameter is used to train the model on the training set, and validate on validation set
- ❑ Ensemble learning: ***Weighted Voting***
- ❑ The final model with best validation score is used to make inference on the testing set.

```
Best parameters for KNN:  {'n_neighbors': 11, 'weights': 'uniform'}
Best cross-validation score for KNN: 0.69

Best parameters for DT:  {'ccp_alpha': 3.893829205071642e-05, 'max_depth': 21, 'min_samples_leaf': 1, 'min_samples_split': 1
3}
Best cross-validation score for DT: 0.58

Best parameters for Ada:  {'algorithm': 'SAMME.R', 'learning_rate': 0.5345936375621236, 'n_estimators': 475}
Best cross-validation score for Ada: 0.38

Best parameters for RF:  {'bootstrap': False, 'max_depth': 30, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_s
plit': 2, 'n_estimators': 460}
Best cross-validation score for RF: 0.81

Best parameters for BalancedRandomForestClassifier:  {'bootstrap': False, 'max_depth': 30, 'max_features': 'sqrt', 'min_samp
les_leaf': 2, 'min_samples_split': 2, 'n_estimators': 460}
Best cross-validation score for BalancedRandomForestClassifier: 0.78

Best parameters for LightGBM:  {'colsample_bytree': 0.6558555380447055, 'learning_rate': 0.2700340105889054, 'max_depth': 4
2, 'min_child_weight': 4, 'n_estimators': 359, 'num_leaves': 63, 'reg_alpha': 1.8789978831283782, 'reg_lambda': 1.7896547008
552977, 'subsample': 0.7989499894055425}
Best cross-validation score for LightGBM: 0.83
```

3. ML modelling: Regularization and pruning (prevent overfitting)

Hyperparameter tuning:

1. KNN:

- **Complexity Control**: *k* - the number of neighbours. A small *k* can lead to a noisy model (overfitting), while a large *k* can lead to a too smooth model (underfitting).
- **Distance Metric**: Euclidean, Manhattan. *Weighted* KNN, where closer neighbours have a bigger influence, can also be considered.

2. Decision Trees (DT):

- **Pruning**: setting a limit on the *maximum depth, minimum samples* at a leaf, or *minimum samples* to split a node. → prevents the tree from growing too deep, which can lead to overfitting.
- **Regularization**: The `*ccp_alpha*` parameter used for *Minimal Cost-Complexity Pruning*. A higher `ccp_alpha` prunes more of the tree.

ASME IDETC-CIE 2023
International Design Engineering Technical Conferences & Computers and Information in Engineering Conference

ASME SETTING THE STANDARD
EWF
The American Society of Mechanical Engineers®
ASME®

3. ML modelling: Regularization and pruning

Hyperparameter tuning:

3. Random Forest (RF):

- *Tree Depth and Leaf Samples*: control the maximum depth of the trees in the forest or set a minimum number of samples required to be at a leaf node.
- *Number of Trees*: More trees increase the robustness of the model at the cost of computational efficiency. But after a certain point, adding more trees doesn't improve performance significantly.
- *Features per Split*: setting the `max_features` parameter to control how many features are considered for each split. This introduces randomness and can help in reducing overfitting.
- *Out-of-Bag (OOB) Score*: The OOB score can be used as an approximation to the validation error without a separate validation set. If the OOB error is much lower than the test error, it might indicate overfitting.

3. ML modelling: Dealing with imbalanced dataset

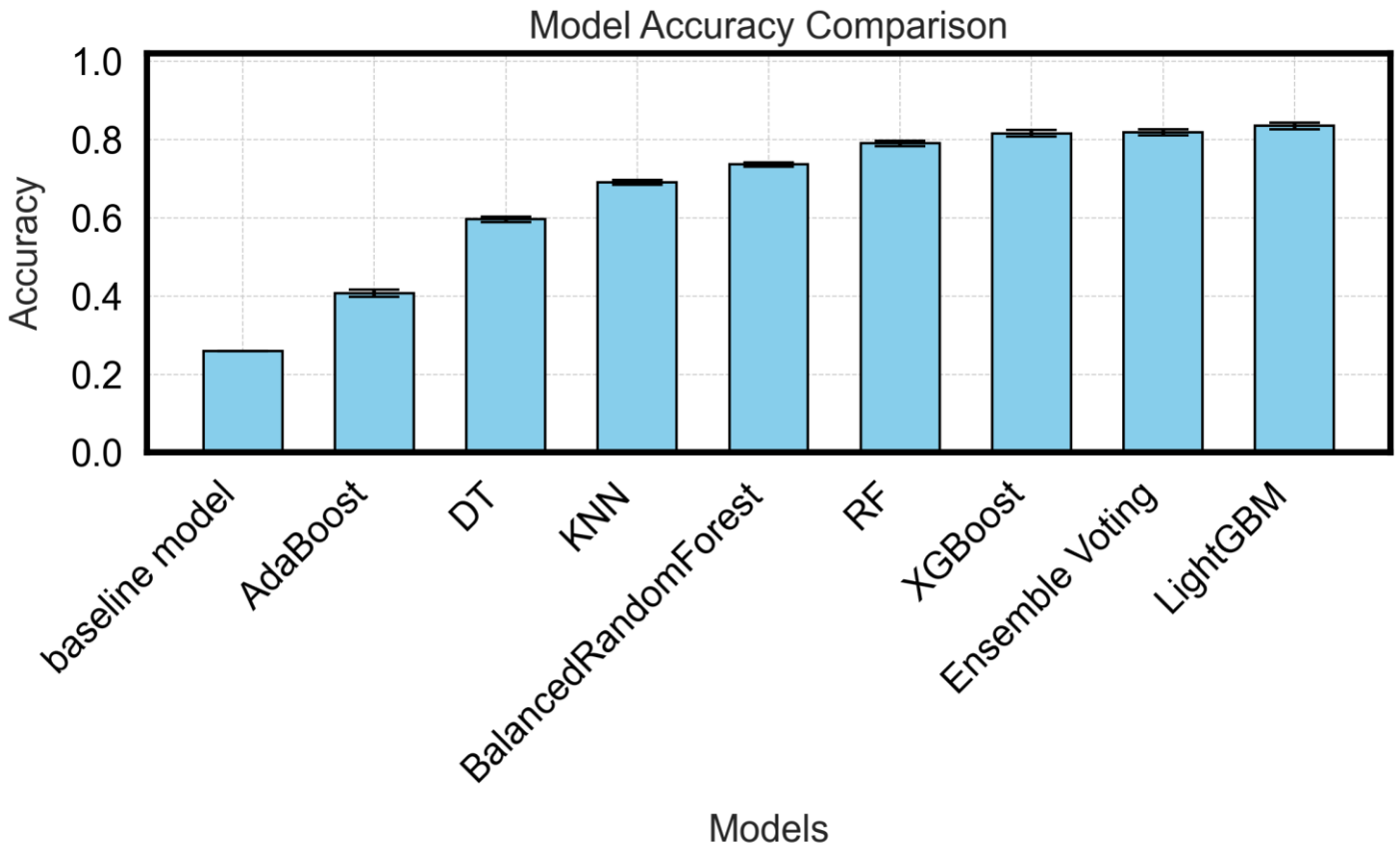
- **Stratified Split**: ensures that the train and validation sets have approximately the same percentage of samples of each target class as the complete dataset.
- Using **class_weight**: during the hyperparameter tuning phase (i.e., within RandomizedSearchCV or GridSearchCV) and during final training. This will ensure that during both phases, the classifier is taking into account the *imbalances* in the class distribution.
- **Evaluation Metrics** during Hyperparameter Tuning: choose F1 score (F1 weighted) instead of accuracy.
  - "f1_macro": Calculate metrics for each label and find their unweighted mean.
  - "*f1_weighted*": Calculate metrics for each label, and find their average weighted by support (the number of true instances for each label).

3. ML modelling: Enhancing model robustness

- **Ensemble learning (weighted voting classifer)**
- **Feature Jittering**
  - Feature jittering involves adding *small Gaussian noise* to the features.

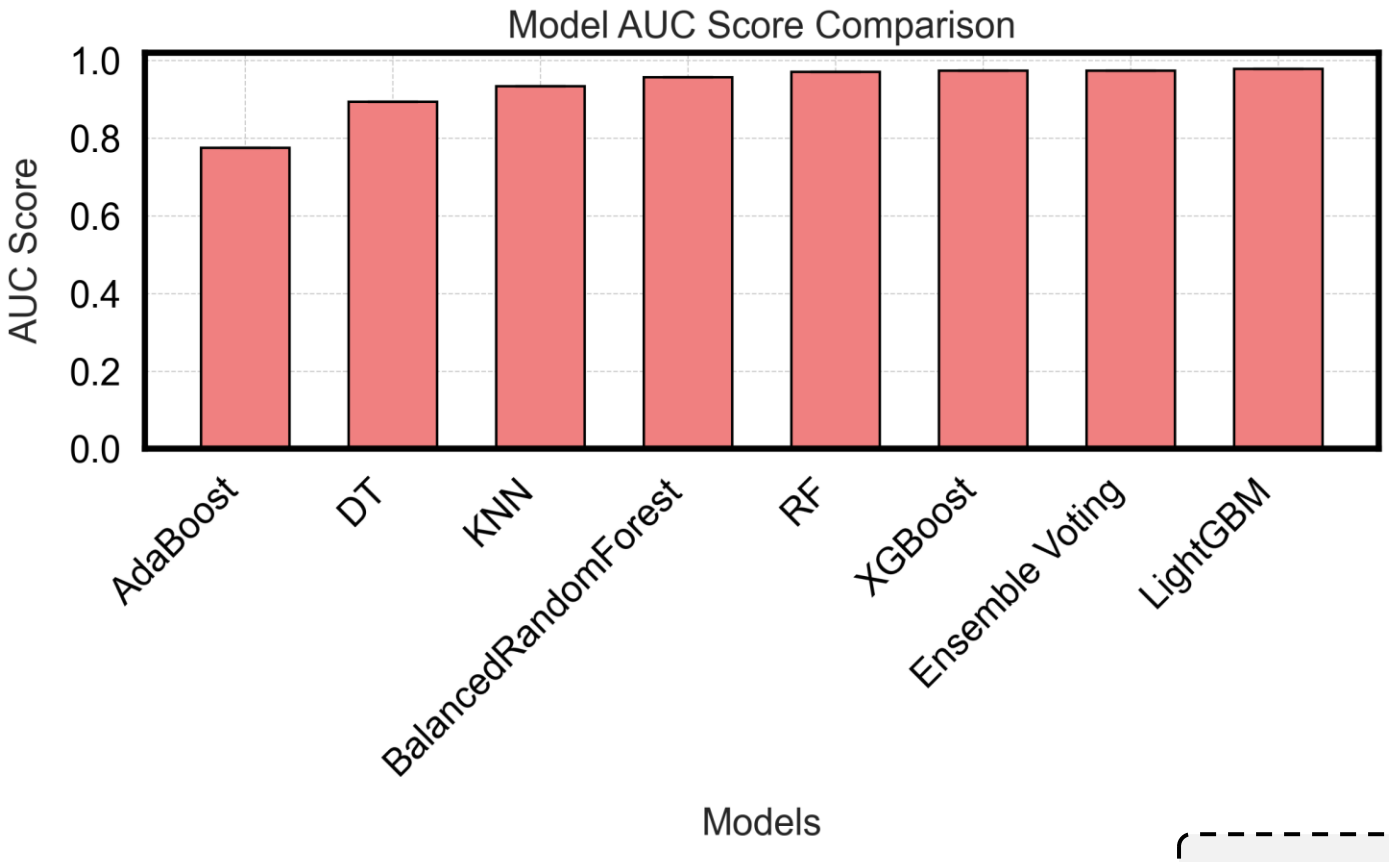
- **Feature Dropout**
  - A simple way to implement feature dropout is to randomly set a portion of the features to *zero* for each batch during training. This encourages the model to not overly rely on any single feature.

4. Results and Discussions



- Best performance on the validation set was achieved by the **LightGBM** *(light gradient-boosting machine)* Classifier (~83.4%).
- Followed by similar accuracy by RF, XGBoost.
- All models with the selected features obtained higher accuracy than the baseline model

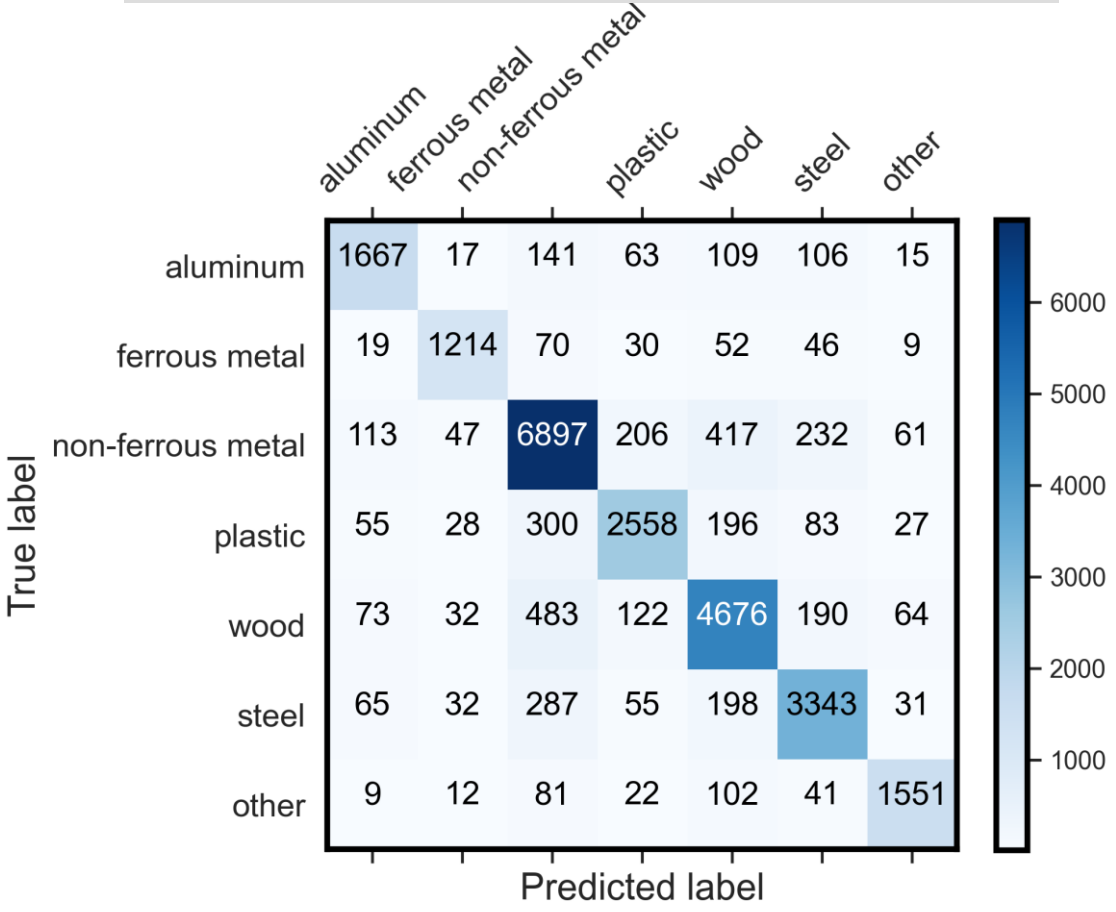
4. Results and Discussions

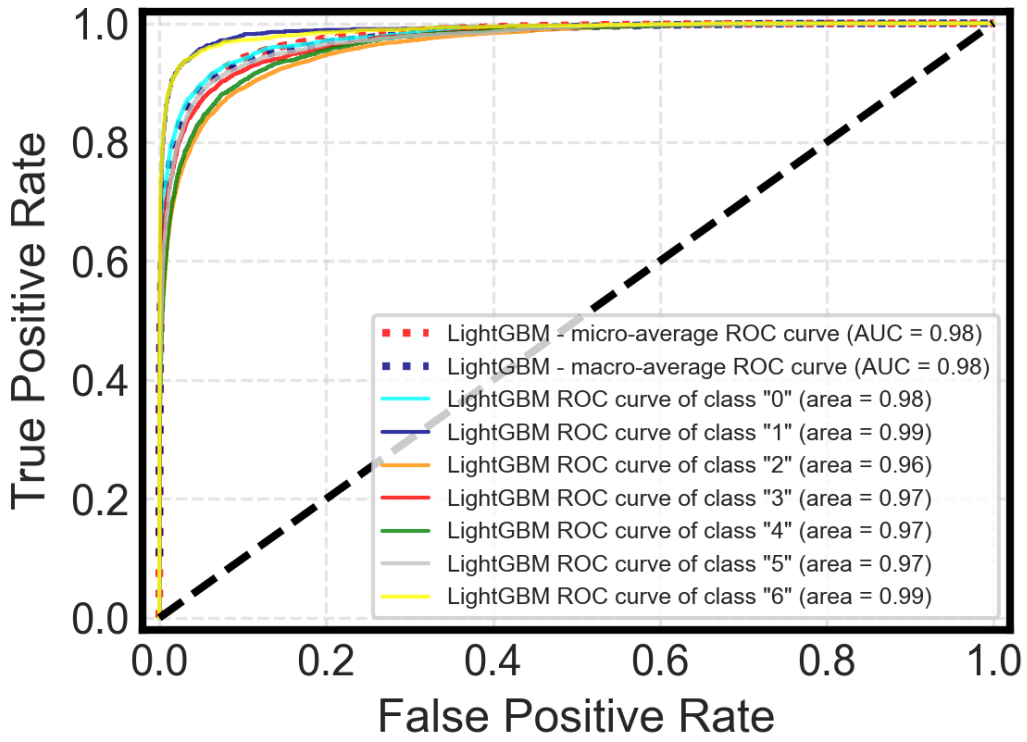


- Top 4 classifiers can achieve Reasonably good AUC score

4. Results and Discussions

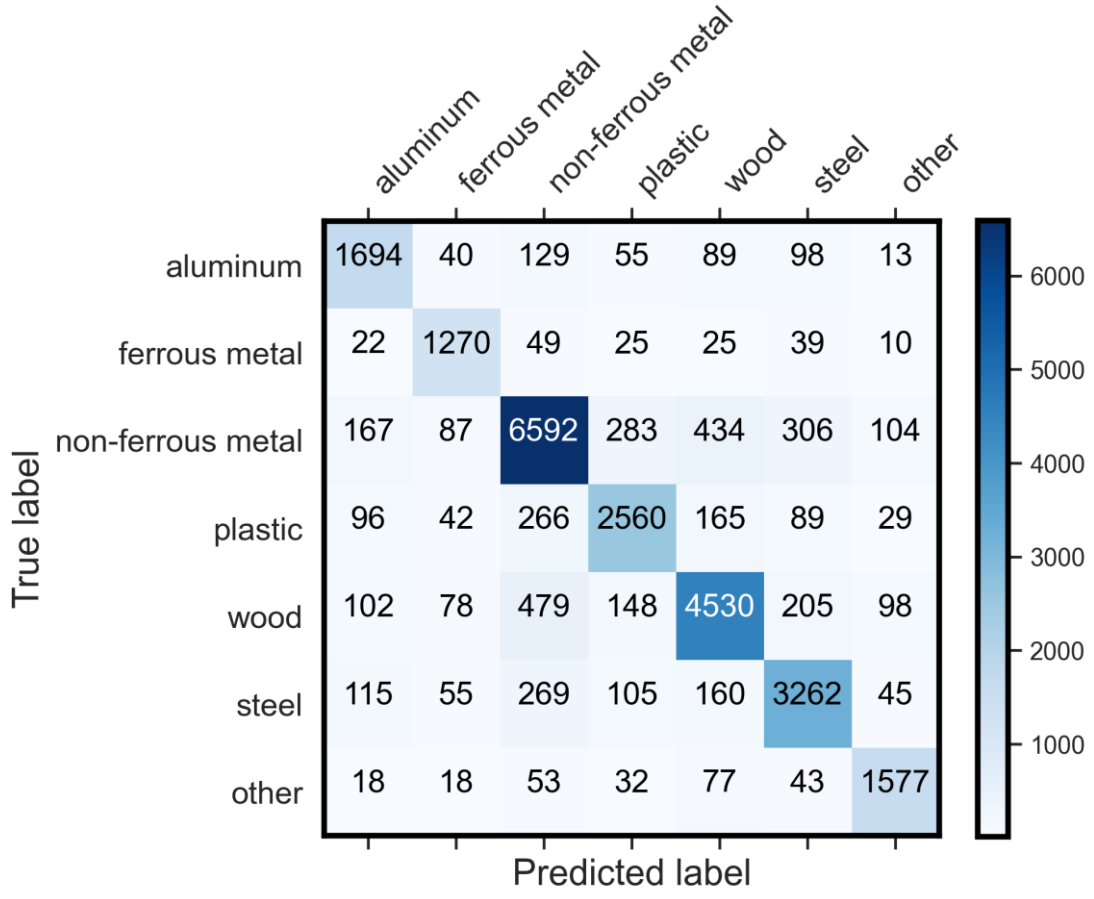
LightGBM (Best performance) – Validation accuracy : 83%

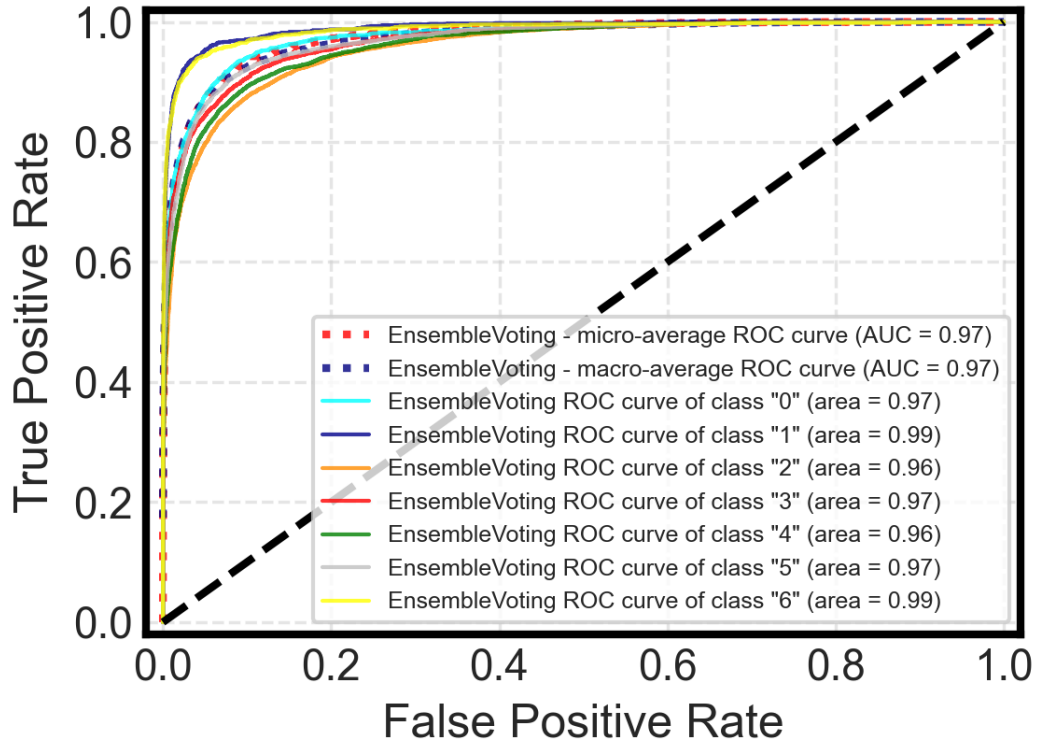




4. Results and Discussions

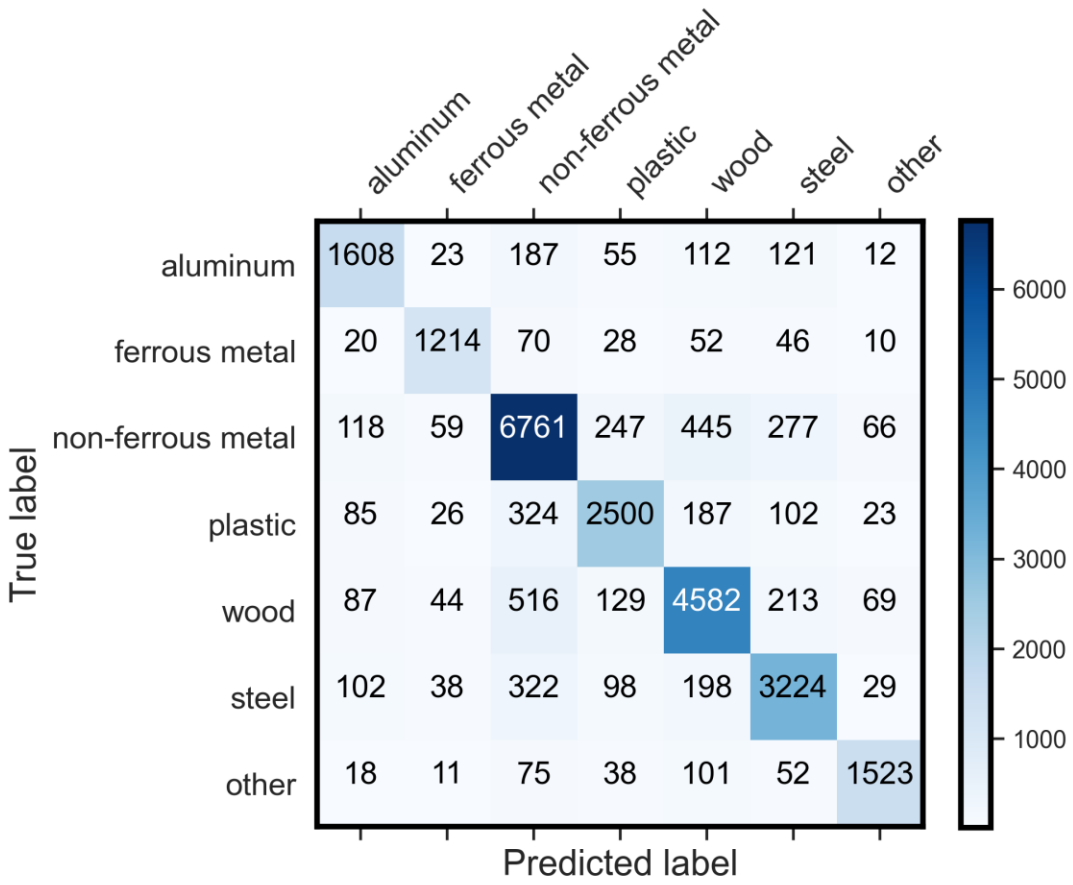
Ensemble Voting – Validation score: 82%

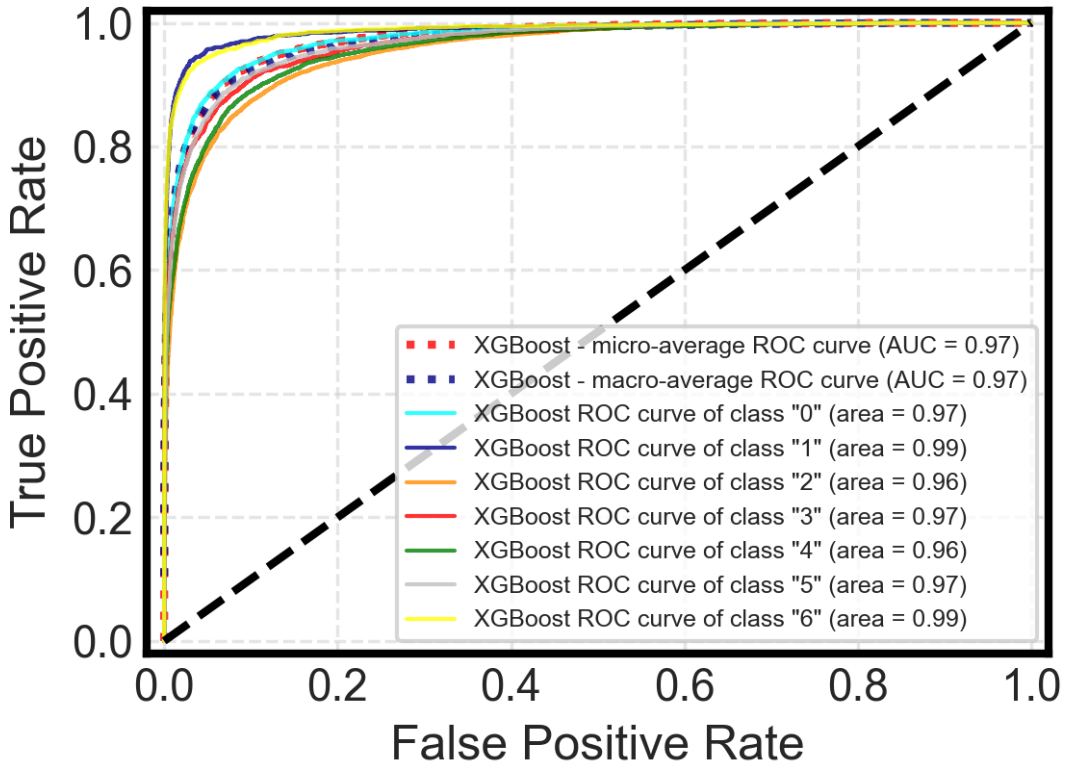




4. Results and Discussions

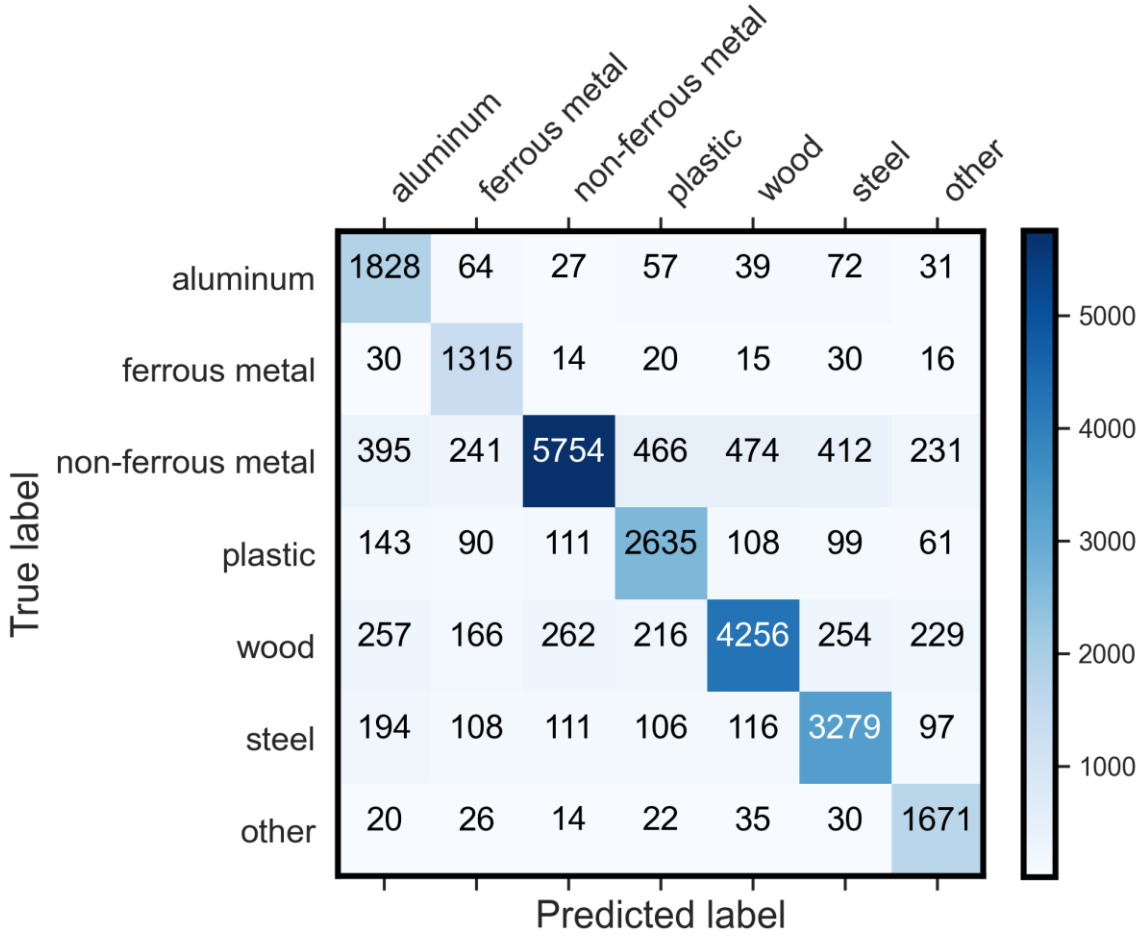
XGBoosting – Validation score: 80 %

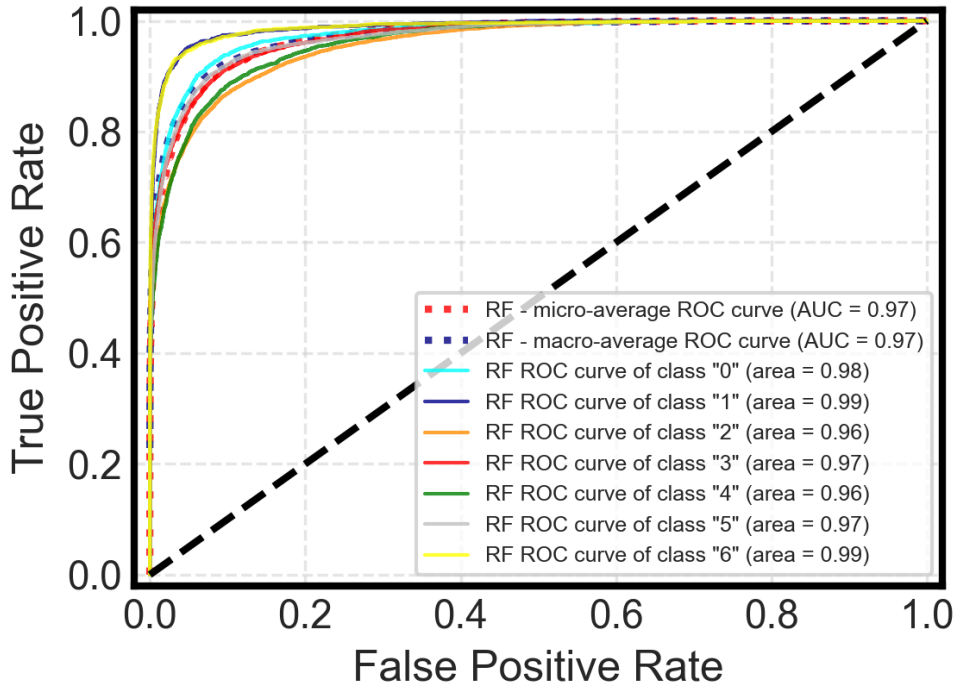




4. Results and Discussions

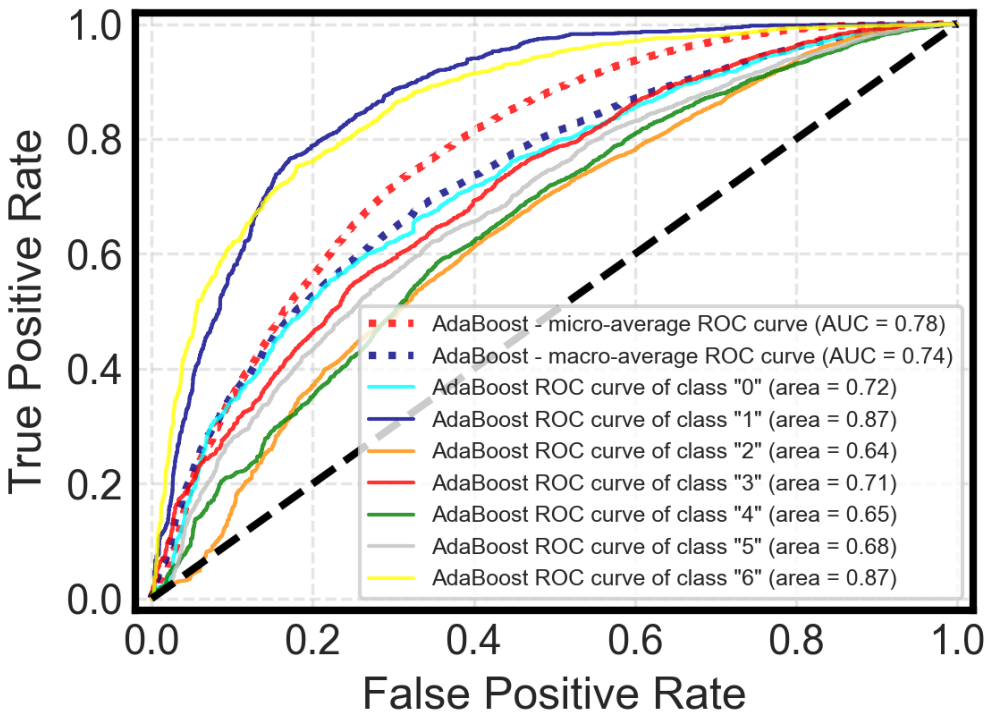
RF – Validation score: 79 %



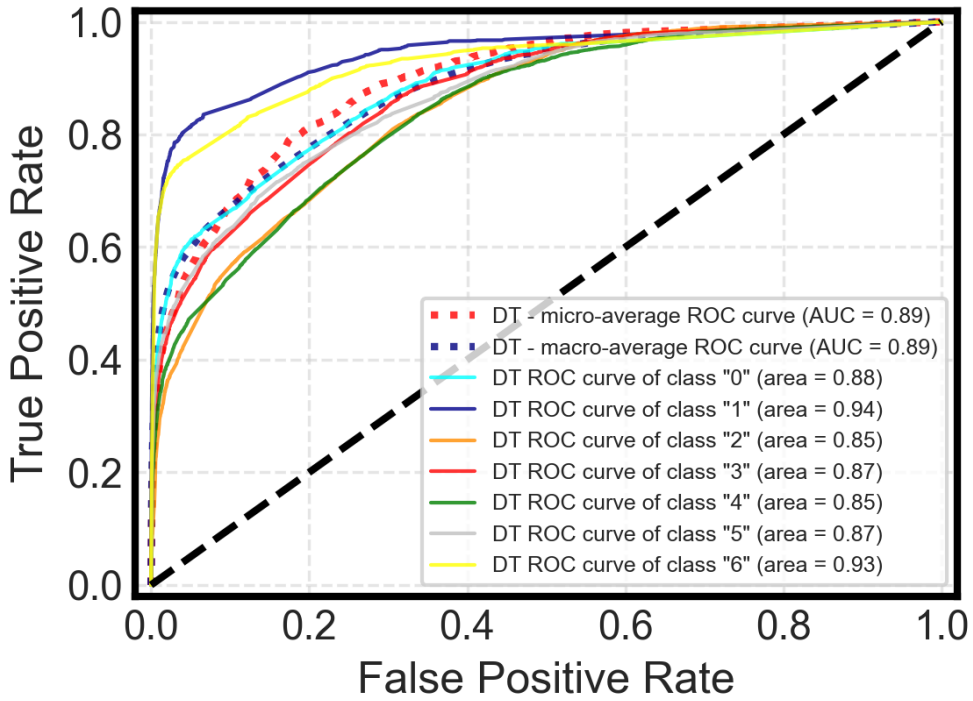


4. Results and Discussions (ROC-AUC curve)

Adaboost – 40%



Decision Tree – Validation score: 59%



4. **Discussion:** Model confidence or uncertainty estimation

Decision Function / Predict Probability (for Scikit-learn models):

These can provide measures of **confidence** in predictions.

- predict_proba gives class probabilities similar to the softmax output.
- decision_function gives a distance to the decision boundary. The magnitude (absolute value) can be interpreted as confidence.

5. Future works

1. **Enhanced Feature extraction**
   - ❑ **Sematic features:** Embed the semantics of component names and their occurrences.
   - ❑ **Advanced geometric features:** Leverage *3D point cloud processing* and deep learning to discover:
     - Segmentation: Understand spatial relationships between components.
     - Thickness and Normal Vectors: Utilize the normal vectors associated with surfaces for feature recognition. This could identify potential manufacturing processes and, subsequently, potential material categories.

2. **Ensemble Techniques**:
   - ❑ ***Stacking***: Train a *meta-model* that takes the predictions of each individual model as input and predicts the final class.
   - ❑ ***Blending***: Take a majority vote based on validation set performance.

3. **Graph Neural Network Modelling**
   - ❑ Leverage GNNs to delve deeper into *hierarchical relationships*, like contacts between components and the relationships between sub-components.

5. Future works

4. Domain Adaptation:

- **Adversarial Training:** Implement Adversarial Domain Adaptation techniques to achieve feature representations that remain *robust to domain shifts*, ensuring that feature representations from both training and test domains are indistinguishable.

5. Knowledge Graph:

- Extract and analyze *hidden relationships* between individual components and the overall assembly.
- Integrate knowledge graphs with machine learning to facilitate more interpretable inferences.

6. Language Logic Models (LLM):

- **Direct Interpretation**: Some body *names* inherently suggest potential material categories. Combining the insights from LLMs with other algorithms can be fruitful.
- **LLM Refinement**: Enhance LLM's predictive capabilities through prompt engineering and fine-tuning.

Thank you!
Q & A