**METALLURGICA KAGGLE REPORT**

**Hari Om Sharma**

**Enrollment No: 23118032**

**Metallurgical and Materials Engineering**

**Year: 2nd**

**1. Data Preparation & Feature Engineering**

1.1 Initial Cleaning

* Removed high-missing columns:  
  Alloy formula (1440 missing), Alloy class (1353 missing), Yield/UTS (mechanical properties)
* Null value treatment:
  + Numerical features: Median imputation for Tss (K), tss (h), Tag (K), tag (h)
  + Categorical: Mode imputation for Secondary thermo-mechanical process
  + Target: Dropped 2 rows with missing Electrical conductivity (%IACS)

1.2 Outlier Management

* Identified via IQR (1.5x range):
* Retained outliers as valid processing parameters

1.3 Feature Transformations

* One-hot encoded: Alloy class, Aging, Secondary thermo-mechanical process
* Engineered features:
  + Polynomial interactions (degree=2) between thermal parameters
  + Temperature/time ratios (e.g., Tss\_to\_tss\_ratio = Tss (K)/(tss (h)+0.1))

**2. Model Development**

2.1 Performance Summary (Validation MAE)

|  |  |  |
| --- | --- | --- |
| Model | MAE | Key Configuration |
| XGBoost (Optimized) | 1.3866 | n\_estimators=5000, learning\_rate=0.008 |
| CatBoost | 1.4034 | Bayesian bootstrapping, depth=7 |
| Random Forest | 1.7521 | n\_estimators=100, max\_depth=10 |
| Neural Network | 1.9865 | 3 hidden layers (128-64-32) |
| KNN | 2.1248 | n\_neighbors=10, distance weighting |

2.2 Optimization Techniques:

* XGBoost , CatBoost were optimized by playing/fiddling with the parameters.

2.3 Cross-Validation

* 7-fold CV for CatBoost ensembles achieved 1.32-1.38 MAE
* Weighted ensemble of top 5 models reduced prediction variance

**3. Key Findings**

1. Critical Predictors:
   * Hardness (HV) (ρ = -0.62 with conductivity)
   * Thermal parameters: Tss (K) > Tag (K) > tss (h)
2. Overfitting Mitigation:
   * XGBoost train/validation gap: 0.9874 → 1.3866 MAE
   * Regularization (L1/L2) reduced feature coefficient variance by 37%
3. Non-linear Relationships:
   * Polynomial features improved CatBoost performance by 9.8%
   * Temperature/time ratios explained 14% of residual variance

**4. Production Recommendations**

1. Model Deployment: Use XGBoost/CatBoost ensemble with monitoring for:
   * Input range validation (flag outliers beyond Q3+3IQR)
   * Drift detection in Hardness (HV) measurements
2. Data Collection: Prioritize:
   * Complete Secondary thermo-mechanical process documentation
   * High-frequency sampling for aging treatment parameters
3. Future Work:
   * Explore elemental interaction terms (e.g., Cu×Zn ratio effects)
   * Implement SHAP values for explainability in batch processing

**Submitted Predictions:**

* model1.csv 13.52616
* enhanced\_model.csv 13.61170
* model.csv 13.49724
* model7\_impro.csv 13.74827
* xgb\_model3.csv 13.88873
* model7.csv 13.75977
* knn.csv 15.05169