CNC Cost Estimation - Final Report

My Approach

To estimate CNC machine costs, I followed a structured approach:

1) Collected data from CNCZone, Practical Machinist, Fictiv, Xometry, Alibaba.

2) Cleaned data, removed outliers/duplicates, and handled missing values by either imputing or removing rows, depending on the context..

3) Engineered features: volume, cost/mm³, Working-area-to-volume ratio, density.

4) Visualized data using Python libraries and Tableau. Univariate analysis (e.g., distribution plots, box plots) and Bivariate analysis (e.g., heatmaps, pairplots) was conducted. These helped in understanding relationships between features and the target variable.

5) Trained various models:

   a) Linear Models: Linear Regression, Ridge

   b) * Ensemble Methods: Random Forest, Bagging, Extra Trees, AdaBoost, Gradient Boosting

   c) * Others: XGBoost, SVR, Decision Tree, KNN

6) Selected Extra Trees Regressor as final model based on evaluation metrics(RMSE, MAE, R^2, Adjusted R^2, and MAPE).

Summary of Results

Extra Trees Regressor performed best:

```
Training Performance:
           MAE           MSE          RMSE  R² Score
0  1542.830724  4.092685e+06  2023.038487  0.988128

Validation Performance:
          MAE           MSE          RMSE  R² Score
0  4606.04636  6.613165e+07  8132.136926  0.773388
```

Other models like Bagging showed signs of overfitting, and Ridge performed worse than a mean predictor. Similarly, the Linear Regression models underperform, with relatively lower R^2 scores (around 0.55 on training and 0.61–0.62 on validation), and very high MAPE values (86–97%), reflecting a lack of accuracy in percentage terms.

Overall, the Extra Trees Regressor not only provides the best balance between bias and variance but also ensures more reliable predictions, making it the most suitable model for this task. The parameters for the final model is {'bootstrap': False, 'max_depth': 8, 'max_features': 'log2', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 229} explaining about 77.34% variance in the data.

Final Analysis & Reflection

What worked well:

1) Extra Trees Regressor captured non-linear patterns effectively.

2) Feature engineering boosted model performance.

3) Ensemble methods were robust against noise.

Challenges faced:

- Small dataset (~100 samples) limited generalizability.
- Manual data collection from inconsistent sources was time-consuming.
- Some useful features (e.g., material type) were missing or unstructured.

Improvements with more data:

- Collect more labeled samples with consistent formats.
- Include domain-specific features like toolpath complexity, spindle speed.

Reflection on manufacturing data:

- Real-world manufacturing data is often sparse and messy.
- Requires domain knowledge to engineer meaningful features.

Additional features to improve accuracy:

- Machining time, tolerance specs, machine type, surface finish quality.
- Supplier location, labor cost, material procurement lead time.

Key Insights

- **Skewed Distributions:** Cost, weight, and volume are heavily right-skewed; outliers seen in working area, cost/mm³, area-to-volume ratio, and density. Log-transform or handle outliers before modeling.
- **Top Predictors:** Volume (corr ≈ 0.88) and weight (corr ≈ 0.83) strongly impact cost. Lead time (corr ≈ 0.49) has moderate influence. Other features show weak linear correlation.
- **Material & Working Area:** Wood, plastic, and aluminum occupy larger areas—cheap and easy to machine. Brass and iron used in smaller areas—costly and harder to process.
- **Area-to-Volume Ratio:** Brass shows high ratio—likely flat/thin parts (e.g., covers). Steel and iron show lower ratios—denser, blocky parts. High ratio → longer machining; low ratio → more tool wear.
- **Warranty Insights:** Leather, acrylic, and copper have longer warranties—likely premium parts. Steel and iron moderate—machining difficulty may affect durability.
- **Lead Time Variability:** Plastic and metal show high variance and outliers—may face complex machining or supply chain delays. Aluminum, brass, and iron show consistency—ideal for stable production.
- **Material Impact:** Metal and steel are heavy and voluminous—higher cost and cycle time. Plastic and aluminum offer better machinability and volume balance.
- **Cost Efficiency:** Metal is costliest per mm³ but favored for strength. Plastic is most cost-effective—ideal for prototyping and low-load parts.