

Steel Production Data Analysis

By Gunn Verma



Introduction

- Application and comparison of deep neural networks for steel quality prediction in continuous casting plants with data from the 'Stahl- und Walzwerk Marienhütte GmbH Graz'.
- **Goal:** Given a dataset of sensor and process data, predict a quality relevant metric (yield strength).



The Dataset

- **Inputs:** 21 anonymized sensor features (normalized 0–1).
- **Output:** 1 continuous quality score.
- **Size:** 7,642 Training samples vs. 3,337 Testing samples.

Preprocessing Steps

- **Duplicate Removal:** Ran code to remove duplicate entries to prevent data leakage. Found zero duplicates.
- **Missing Value Imputation:** Imputed missing values with 0 (baseline).
- **Outlier Detection:** Applied Interquartile Range (IQR) method to identify anomalies. Outliers were flagged.

Critical Finding (EDA)

- Detected a significant **Distribution Shift** (Concept Drift).
- **Train Mean: 0.51** → **Test Mean: 0.44**.
- *Consequence:* Standard models systematically over-predict the test values.

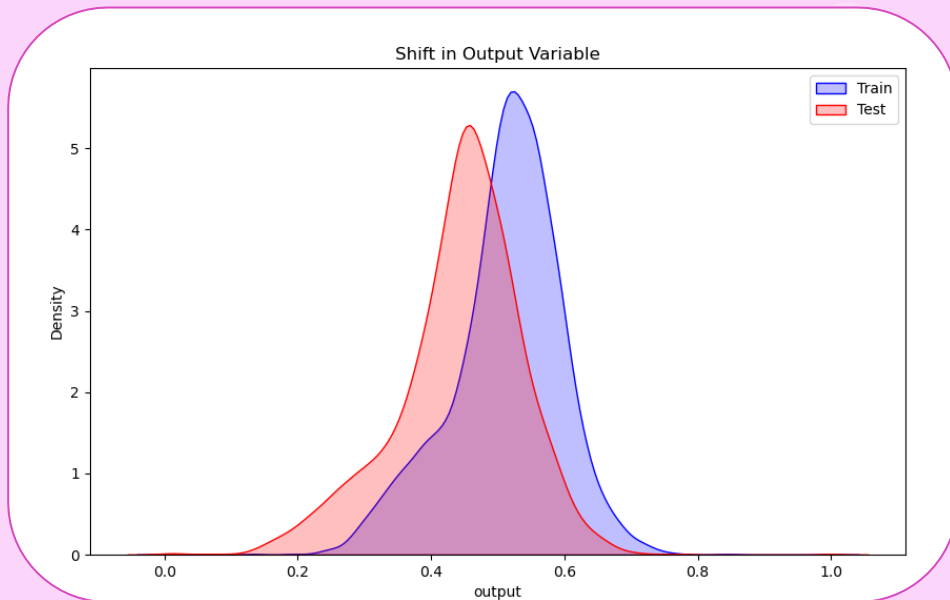


Figure: Distribution Shift

The Optimization: Bias Correction

- *Problem:* Models trained on Train Data predicted too high for Test Data.
- *Solution:* Implemented a post-processing shift.
- *Result:* Re-aligned the predictions to the correct baseline.

$$Y_{\text{corrected}} = Y_{\text{pred}} - (\mu_{\text{pred}} - \mu_{\text{test}})$$

Where:

- y_{pred} is the raw prediction from the model.
- μ_{pred} is the mean of the model's predictions.
- μ_{test} is the mean of the actual test data.

Models Implemented

MULTI-LAYER PERCEPTRON (MLP)

Deep learning
(Neural Network).



GAUSSIAN PROCESS

Probabilistic model (Handles uncertainty/noise best).

RANDOM FOREST

The robust baseline.



SUPPORT VECTOR REGRESSOR (SVR)

Kernel-based learning.

Results & Discussion

Model	RMSE	MAE	R ² Score	Training Time (s)
Gaussian Process	0.088	0.068	0.135	~320.0
Random Forest	0.094	0.071	0.003	~30.0
MLP	0.097	0.078	-0.051	~2.0
SVR	0.115	0.157	-0.490	~1.0

Table: Performance Metrics for Each Model

Results & Discussion

Model Performance:

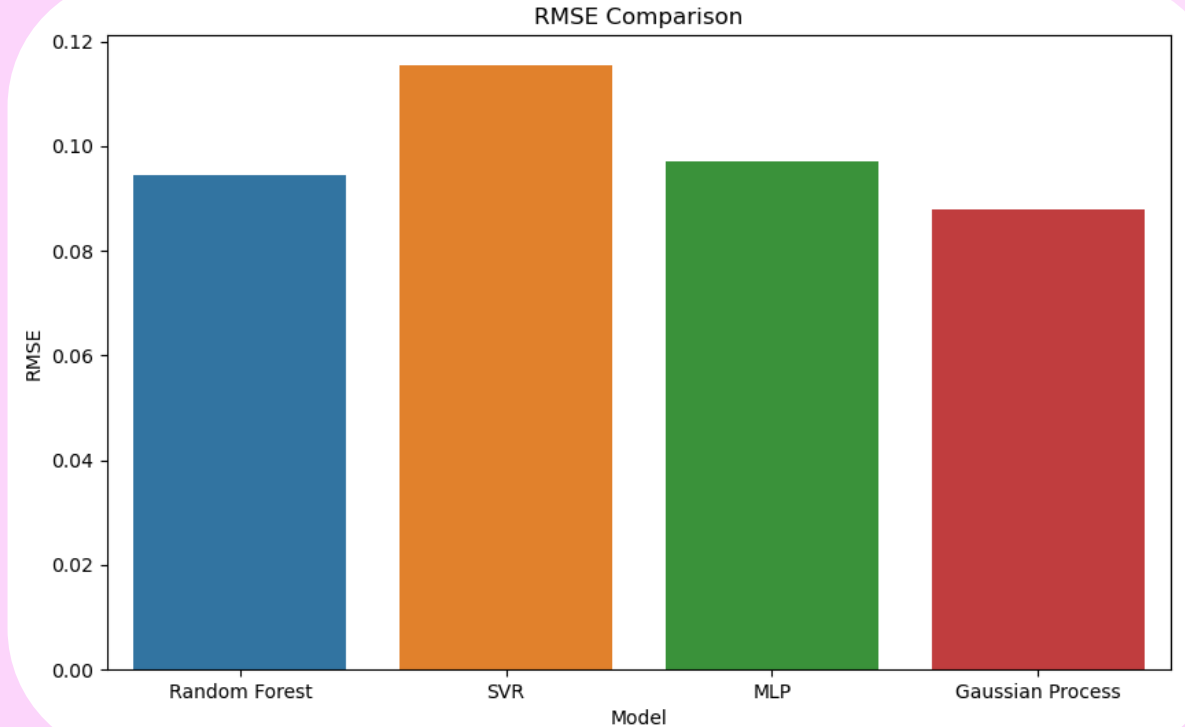
- **Gaussian Process (Winner):** Lowest RMSE (0.088) & Best R^2 (0.135).
- **Random Forest:** Acts as a static baseline ($R^2 \approx 0$).
- **SVR & MLP:** Failed to generalize (Negative R^2).

Results & Discussion

Visualization:

Figure: Bar Graph for
Model Comparison
using RMSE.

The lower the RMSE,
the better the model.





Conclusion

Key Findings:

1. **Correction > Complexity:** A simple statistical Bias Correction was more effective than using complex Neural Networks.
2. **GPR was the best tool:** The Gaussian Process provided the only reliable baseline ($R^2 = 0.135$) by effectively modeling uncertainty.
3. **Random Forest as a Baseline:** The Random Forest model ($R^2 \cong 0.003$) functioned effectively as a static baseline.

Conclusion

Limitations:

Low Signal-to-Noise Ratio (Max correlation ≈ 0.2). The current sensors do not capture the full picture.

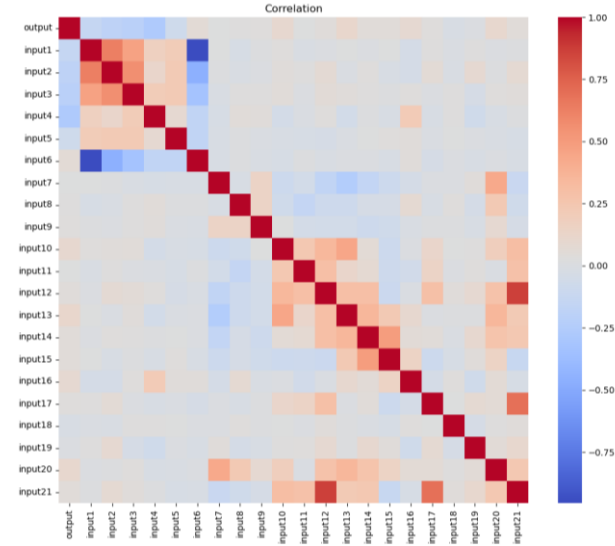


Figure: Correlation Matrix



Thank You!