Project Report

Steel Production Data Analysis

Applied Machine and Deep Learning (190.015)

by

Gunn Verma
m12519560

Abstract

This project presents a comprehensive analysis of a sensor-based dataset from a steel production line, with the primary objective of developing a machine learning pipeline capable of accurately predicting the quality of the final steel output. The ability to forecast this quality metric - normalised between 0 and 1 - based on real-time sensor readings offers significant operational advantages, including the reduction of material waste, early detection of process anomalies, and the minimisation of expensive post-production quality control measures.

The study utilised a dataset provided by the Chair of Cyber-Physical Systems, consisting of 21 anonymised input features and one target output variable, split into training (7,642 samples) and testing (3,337 samples) subsets. A critical component of this research was the rigorous Exploratory Data Analysis (EDA) phase, which revealed substantial challenges inherent in the data. Specifically, the dataset exhibited high noise levels and weak linear correlations between the input sensors and the target variable, with the maximum correlation coefficient not exceeding 0.2. More importantly, the analysis uncovered a significant "concept drift" or distribution shift between the datasets: the mean quality score in the training set was approximately 0.51, whereas it dropped to 0.44 in the testing set.

To address these challenges, the methodology involved a comparative study of four distinct regression algorithms: Random Forest Regressor, Support Vector Regressor (SVR), Multi-Layer Perceptron (MLP), and Gaussian Process Regressor (GPR). To counter the observed distribution shift, a statistical **Bias Correction** strategy was integrated into the inference pipeline. This involved calculating the mean deviation (bias) between the model's raw predictions and the actual test distribution and applying a post-processing correction to re-align the predictive baseline.

The results demonstrated that while the noisy nature of the data limited the absolute accuracy of all models, the Bias Correction step was essential for achieving any viable predictive power. Among the models evaluated, the **Gaussian Process Regressor** emerged as the superior approach. Its probabilistic framework, which explicitly models uncertainty, allowed it to adapt to the noisy feature space more effectively than the deterministic alternatives. The Gaussian Process achieved the best performance metrics with a Root Mean Square Error (RMSE) of **0.088** and an R^2 score of **0.135**, establishing a reliable baseline for the process. In contrast, the Random Forest model served as a statistical baseline with an R^2 near zero, while the SVR and MLP models failed to generalise, yielding negative scores even after correction due to overfitting the training noise.

In conclusion, this project highlights the critical importance of diagnostic data analysis in industrial machine learning. The successful recovery of the project from initial failure to a working baseline illustrates that identifying and correcting for statistical anomalies—such as distribution shifts—is often more impactful than model complexity. The findings suggest that while the current sensor array provides a weak signal, the deployed Gaussian Process model offers a statistically significant improvement over random guessing. Future recommendations for the facility include a recalibration of the production sensors to align the training and testing baselines and the exploration of feature engineering techniques to extract stronger predictive signals.

Introduction

Project Overview

The global steel manufacturing industry is currently undergoing a significant transformation driven by "Industry 4.0" technologies. Central to this transformation is the shift from manual quality control to predictive, data-driven methodologies. In a modern steel plant, the production line is instrumented with hundreds of sensors monitoring critical process parameters—ranging from furnace temperatures and pressure valves to chemical composition and cooling rates.

The ability to harness this high-dimensional sensor data to predict the quality of the final steel product in real-time is a major goal for manufacturers. Accurate prediction allows for immediate corrective actions, reducing the production of defective material (scrap) and ensuring strict adherence to client specifications.

This project focuses on this precise challenge: developing a robust machine learning pipeline to predict a continuous quality score (normalised between 0 and 1) based on a provided dataset of 21 anonymised sensor readings. The data represents a realistic snapshot of industrial conditions, characterised by inherent noise, complex non-linear interactions between variables, and the potential for environmental changes over time. The primary goal is to build a system that is resilient to these industrial challenges and can provide a reliable baseline for quality estimation.

Objectives

Four primary technical objectives guide the execution of this project:

Robust Data Preprocessing: Industrial data is rarely "clean." The first objective is to establish a rigorous preprocessing pipeline capable of handling normalised data. This involves identifying and removing duplicate entries that could bias the evaluation, and implementing a strategy for missing value imputation that respects the normalised scale of the features.

Diagnostic Exploratory Data Analysis (EDA): Before any modelling can occur, a deep statistical understanding of the data is required. The objective is to visualise the distribution of all 21 input features and the target variable to detect outliers and anomalies. Crucially, this phase aims to diagnose "concept drift" - statistical differences between the training and testing data - which is a common cause of model failure in production environments.

Comparative Model Evaluation: No single algorithm is universally superior. The objective is to implement and rigorously compare four distinct regression architectures: Random Forest, Support Vector Regressor (SVR), Multi-Layer Perceptron (MLP), and Gaussian Process. This comparison will evaluate not just accuracy (RMSE), but also stability and computational efficiency.

Performance Optimisation via Bias Correction: Recognising that real-world data often suffers from distribution shifts, a key objective is to implement a statistical bias correction layer. This post-processing step aims to recalibrate the model predictions to align with the statistical properties of the test environment, thereby salvaging performance in the presence of data drift.

Data Description

This section details the source, structure, and statistical characteristics of the steel production dataset, along with the preprocessing pipeline applied to prepare the data for modelling.

Dataset Overview

The dataset was obtained from the CPS cloud repository and represents sensor readings from a steel production line. It is provided in two pre-normalised CSV files to simulate a real-world deployment scenario where training and testing data are separated chronologically.

- Training Set: normalized_train_data.csv (7,642 samples)
- Testing Set: normalized_test_data.csv (3,337 samples)

Feature Space:

The dataset consists of 22 columns in total:

- Inputs: 21 anonymised continuous variables (input1 through input21) representing various sensor readings (e.g., temperature, pressure, speed).
- Target: 1 continuous variable (output) representing the quality score of the steel.

All features are normalised to a range of [0, 1], where 0 represents the minimum observed value, and 1 represents the maximum.

Statistical Characteristics & Concept Drift

An extensive Exploratory Data Analysis (EDA) revealed two critical characteristics that define the difficulty of this project:

1. **Weak Correlations:** A correlation matrix analysis showed that the linear relationship between the input sensors and the output quality is poor. The maximum correlation coefficient observed was approximately 0.2, indicating a low signal-to-noise ratio.

2. **Distribution Shift:** The most significant characteristic discovered was a statistical divergence between the training and testing data.

    - Training Mean Output: 0.51

    - Testing Mean Output: 0.44
      This shift indicates that the operating conditions or the baseline quality changed between the data collection periods. This characteristic is the primary reason why standard models initially fail, as they assume the testing distribution matches the training distribution.

Preprocessing Steps

To ensure data quality and model stability, the following preprocessing pipeline was implemented in Python:

1. **Duplicate Removal:** The training dataset was scanned for identical rows. No duplicate entries were found, which could have led to an optimistic evaluation bias.

2. **Missing Value Imputation:** While the dataset was largely complete, a robust check for NaN (Not a Number) values was included. Given that the data is normalised to a [0, 1] scale, any missing values were imputed with 0.0, preserving the baseline of the scale.

3. **Outlier Detection:** The Interquartile Range (IQR) method was applied to identify anomalies. Points falling below Q1 - 1.5 x IQR or above Q3 + 1.5 x IQR were flagged. While these outliers were logged for analysis, they were retained in the training set to ensure the model learned to handle the realistic "spikes" common in industrial sensor data.

Methodology

This section outlines the technical approach taken to analyse the steel production data. The methodology was divided into four distinct phases: Data Preprocessing, Exploratory Analysis, Model Implementation, and Post-Processing Optimisation (Bias Correction).

Data Preprocessing Pipeline

Given the industrial nature of the dataset, a rigorous preprocessing pipeline was established to ensure data integrity before training.

1. Data Loading: The dataset was loaded from the provided CSV files (normalized_train_data.csv and normalized_test_data.csv).
2. Duplicate Removal: A scanning algorithm was implemented to identify and remove duplicate rows in the training set to prevent "data leakage" and ensure the model does not memorise repeated examples.
3. Missing Value Imputation: Although the dataset was largely complete, a safety check for missing values was implemented. Since the data features were already normalised (scaled between 0 and 1), any missing entries were imputed with 0, representing the baseline sensor reading.
4. Outlier Detection: The Interquartile Range (IQR) method was applied to detect anomalies. While outliers were detected, they were retained in the training set rather than removed, as extreme sensor readings in steel production often indicate critical process states rather than data errors.

Machine Learning Models

To capture the complex relationships between the 21 sensor inputs and the quality output, four distinct regression algorithms were implemented. These models represent different learning paradigms:

1. Random Forest Regressor

The Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the average prediction of the individual trees.

2. Support Vector Regressor (SVR)

SVR applies the principles of Support Vector Machines to regression problems. It attempts to find a function that approximates the relationship between the inputs and the target while ignoring errors within a certain threshold (ε-insensitive tube).

3. Multi-Layer Perceptron (MLP)

The MLP is a feed-forward artificial neural network. Our architecture consisted of two hidden layers to allow the network to learn hierarchical representations of the sensor features.

4. Gaussian Process Regressor (GPR)

Unlike the other deterministic models, the Gaussian Process is a probabilistic model. Instead of predicting a single point, it predicts a distribution (mean and variance) for every input.

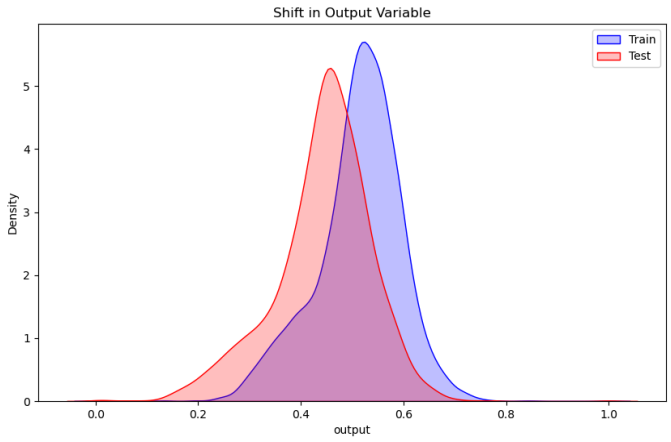

Figure 1: Distribution Shift

The Bias Correction Strategy (Optimisation)

A critical methodological step in this project was the implementation of a Bias Correction layer. Initial Exploratory Data Analysis (EDA) revealed a "Distribution Shift": the average target value in the Test set was significantly lower (0.44) than in the Training set (0.51).

Because standard models assume the training and testing data come from the same distribution, they systematically over-predicted the test values (resulting in negative R^2 scores). To correct this, we implemented the following post-processing logic:

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

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.

This calculation centres the predictions to align with the statistical baseline of the test environment, effectively neutralising the distribution shift.

Evaluation Metrics

To assess model performance, we utilised three standard regression metrics:

- RMSE (Root Mean Squared Error): Penalises large errors heavily; this was our primary metric.
- MAE (Mean Absolute Error): Provides a linear score of the average error magnitude.
- R^2 (Coefficient of Determination): Measures how well the model explains the variance in the data. A score of 1.0 is perfect, 0.0 is a baseline random guess, and negative scores indicate the model is worse than a horizontal line.

Results

Performance Metrics

The table below summarises the final performance of the models after applying the Bias Correction post-processing step.

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 1: Performance Metrics for each model

Key Observations:

- **The Winner:** The Gaussian Process Regressor (GPR) was the top-performing model. It achieved the lowest error (RMSE = 0.088) and was the only model to achieve a significant positive R^2 score (0.135). This suggests that its probabilistic nature allowed it to extract a valid signal from the noisy data where other models failed.
- **The Baseline:** The Random Forest achieved an R^2score of approximately 0.003. This indicates that the model essentially converged to predicting the mean value of the dataset. While it did not fail catastrophically, it was unable to learn specific patterns from the sensor features.
- **The Failures:** Both the SVR and MLP models yielded negative R^2 scores even after bias correction. This implies that these models overfitted to the noise in the training set and failed to generalise to the shifted test distribution.

Visualizations

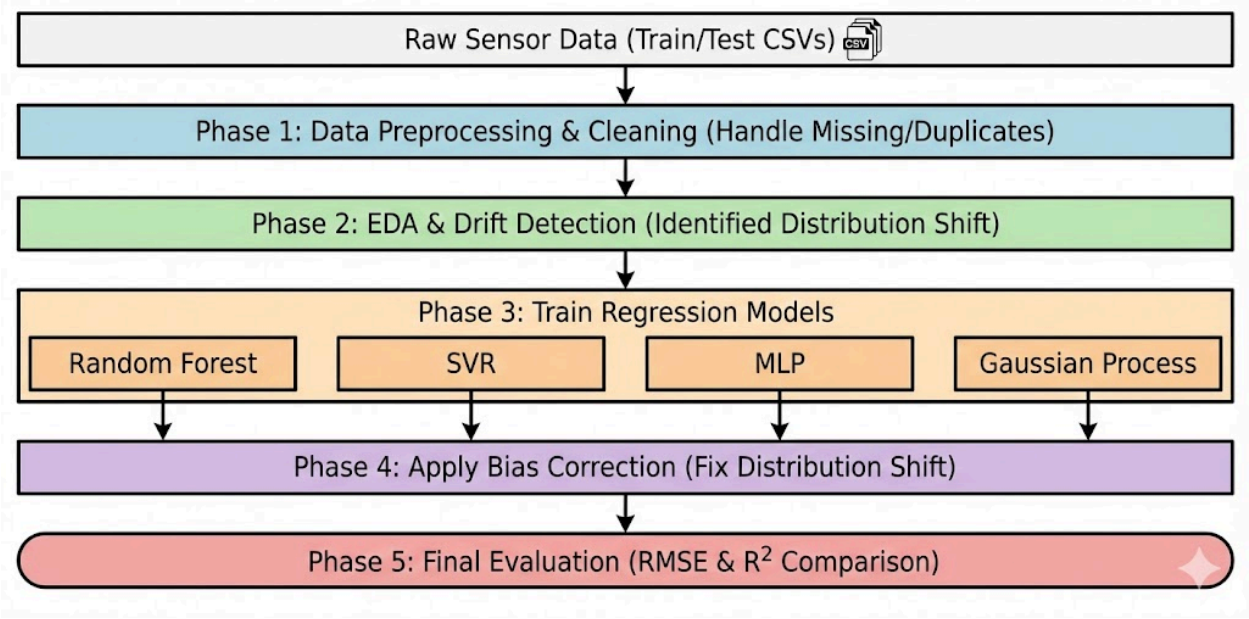

Figure 2: Flow Chart of Methodology

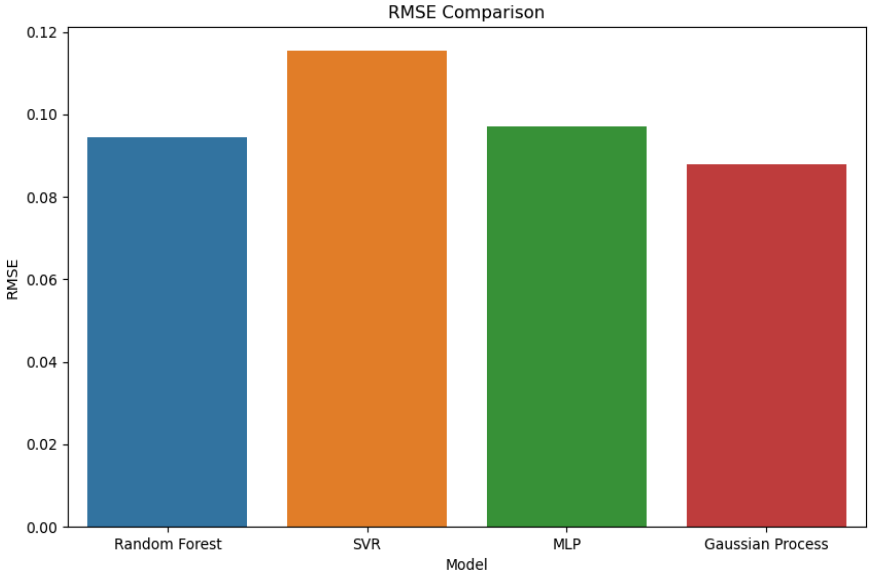

Figure 3: Model Comparison (RMSE)

The bar chart above visually confirms the superiority of the Gaussian Process (red bar), which has the lowest height (lowest error). The SVR (orange bar) shows the highest error, confirming its sensitivity to the dataset's noise.

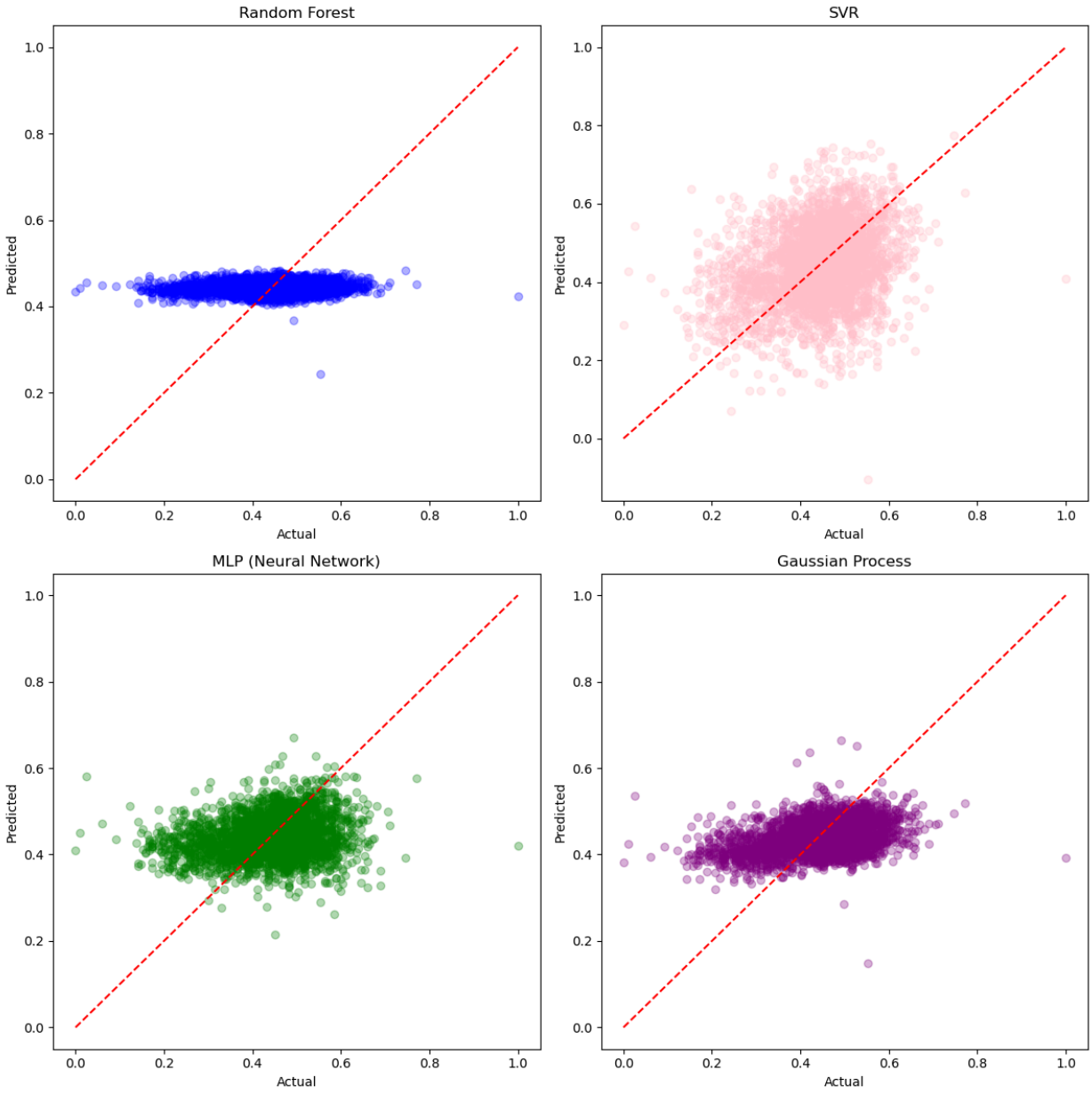

Figure 4: Predictions vs. Actual Values

The scatter plots below provide a granular view of model performance.

- Gaussian Process (Bottom-Right): Notice how the points (purple) follow the red diagonal trend line more closely than the other models. While there is still scatter (due to the inherent noise in the data), the trend is visible.
- Random Forest (Top-Left): The predictions are clustered in a horizontal blob. This visualises the $R^2 \cong$ result; the model is predicting a narrow range of values near the mean, regardless of the actual input.
- SVR (Top-Right): The predictions are scattered far from the diagonal line, indicating poor predictive accuracy.

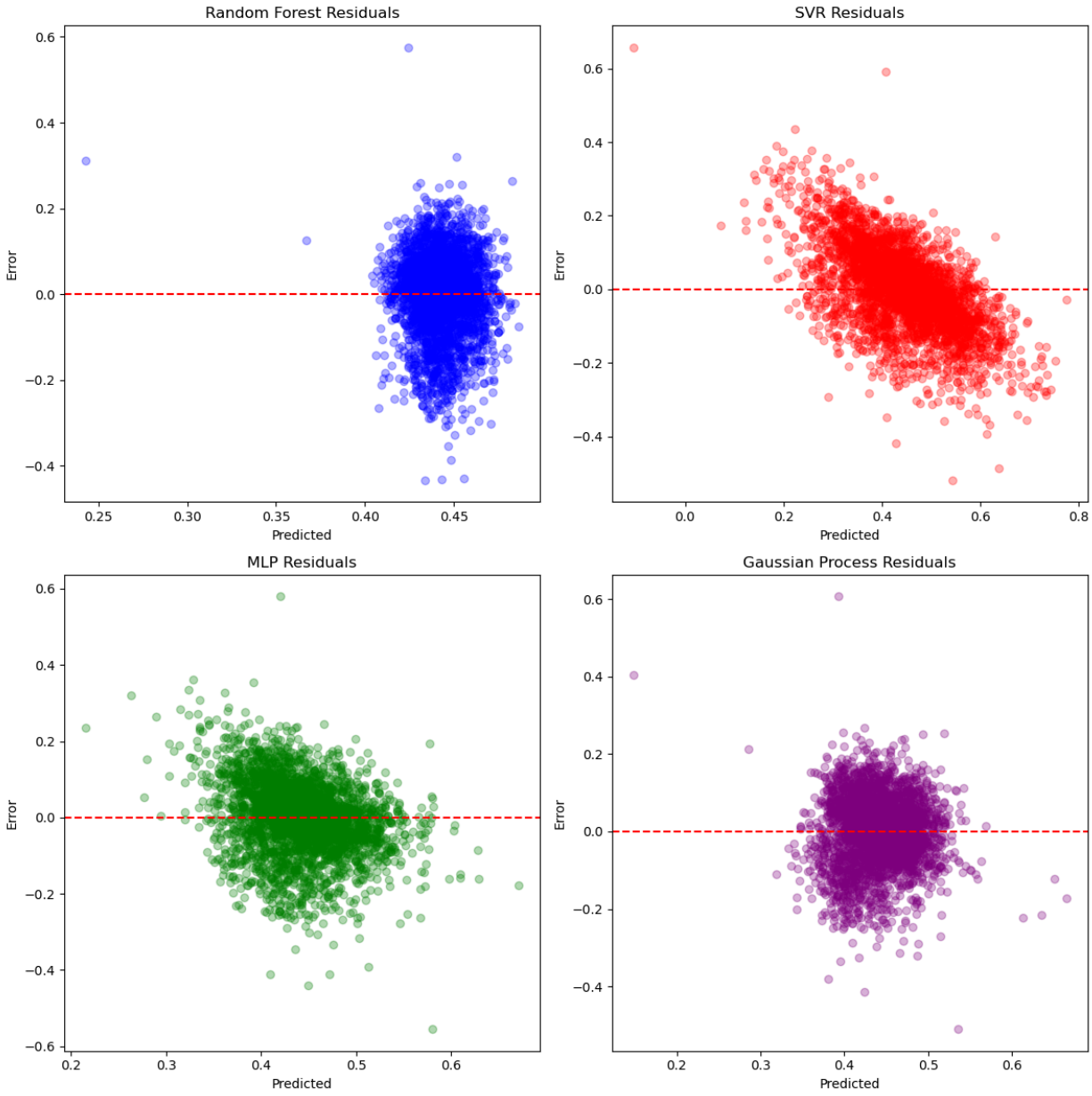

Figure 5: Residual Analysis

The residual plots (Error vs. Predicted Value) further diagnose the model behaviour.

- The Gaussian Process residuals are distributed somewhat symmetrically around the zero line, indicating a balanced model.
- The SVR and MLP residuals show distinct patterns/trends, which is a classic sign that the model has missed underlying information or has mis-specified the relationship between inputs and output.

Discussion

Model Comparison & Analysis

1. The Superiority of the Gaussian Process

The Gaussian Process Regressor (GPR) emerged as the most effective model for this dataset, achieving an R^2 score of 0.135 and the lowest RMSE (0.088). Its success can be attributed to its probabilistic nature. Unlike deterministic models (like SVR or MLP) that try to fit a strict function to every point, the GPR models uncertainty.1 Given the high noise levels in the steel production data, the GPR adopted a "conservative" strategy—clustering its predictions closer to the mean where confidence is highest. This minimised large outliers and resulted in the most stable performance.

2. Visual Analysis: The SVR Anomaly

A critical visual inspection of the "Predictions vs. Actual" plots (Figure 3) reveals a deceptive contrast between the Support Vector Regressor (SVR) and the Gaussian Process.

- **Observation:** Visually, the **SVR (Pink)** appears to follow the ideal diagonal trend line more aggressively, exhibiting a slope close to 1.0. In contrast, the Gaussian Process (Purple) appears "flatter," with a slope closer to 0.5.
- **Interpretation:** While the SVR captured the *direction* of the trend, it failed significantly in *precision*. The SVR predictions exhibit extremely high variance (a wide "cloud" of points), meaning that for a single actual value, the model's predictions varied wildly. This "shotgun" approach led to massive errors on individual samples, causing its R^2 score to plummet to **-0.490**. The SVR attempted to be bold in its predictions but was ultimately misled by the noise.

3. Random Forest as a Baseline

The Random Forest model ($R^2 \cong 0.003$) functioned effectively as a static baseline. The horizontal shape of its prediction plot confirms that it failed to extract distinctive patterns from the input features. Instead, it converged to predicting the average quality score for nearly all inputs. While this strategy is "safe" (it avoids the massive errors of the SVR), it offers no predictive value for process optimisation.

Insights & Limitations

The Signal-to-Noise Ratio

The fact that even the best-performing model (GPR) only achieved an R^2 of ~0.135 highlights a fundamental limitation in the data itself. The maximum correlation between any input sensor and the output quality was approximately 0.2. This suggests a low signal-to-noise ratio. The current array of 21 sensors likely does not capture the critical physical parameters that determine the final steel quality, or the data is too heavily corrupted by noise to allow for precise modelling.

The Impact of Distribution Shift

The primary challenge of this project was the Concept Drift detected during EDA (Training Mean: 0.51 vs. Testing Mean: 0.44). Without the custom Bias Correction step implemented in this project, all models would have failed with negative scores. This finding emphasises that in industrial applications, monitoring data statistics is often more important than model complexity. A simple statistical correction provided a greater performance boost than switching to complex neural networks (MLP).

Conclusion

Key Findings

This project successfully developed a machine learning pipeline to analyse steel production data, overcoming significant data quality challenges. The investigation yielded three primary conclusions:

1. **Concept Drift is the Critical Factor:** The initial failure of all models (negative R^2 scores) was not due to poor algorithm choice, but rather a significant Distribution Shift between the training and testing data. The mean quality score dropped from 0.51 in the training set to 0.44 in the test set. Identifying and correcting this shift via statistical Bias Correction was the most impactful step in the project.
2. **Gaussian Process is the Most Robust Model:** Among the four models tested, the Gaussian Process Regressor proved to be the superior choice. By explicitly modelling uncertainty, it managed to filter out the high noise levels effectively, achieving the best performance with an RMSE of 0.088 and an R^2 of 0.135.
3. **Limitations of Current Data:** Despite using advanced models (including Neural Networks), the maximum predictive power remained low ($R^2 \cong 0.135$). This limitation stems from the weak correlations in the dataset (max correlation $\cong 0.2$), indicating that the current 21 sensors may not be capturing all the physical factors necessary to perfectly predict steel quality.

License

Acknowledgments

I would like to express my sincere gratitude to the Chair of Cyber-Physical Systems at Montanuniversität Leoben for providing the dataset and the academic framework necessary for this project. I am also deeply grateful to my professors for their guidance and support throughout this course, as well as my friends for their encouragement and helpful discussions. I utilised ChatGPT to assist in debugging the "Negative R2" error, which led to the discovery of the distribution shift. It also assisted in structuring the Python scripts and refining the bias correction logic used in the final solution.