Defect Rate Prediction Documentation

1. Data Preparation

I. Data Understanding (Business Context)

The dataset originates from a manufacturing firm aiming to identify and reduce defects in their product line. Multiple datasets represent different stages of the testing process, namely: Defect Rates, Step1_Mount_terminal_Resin, Step1_Mount_Terminals, Step2_Wind_Wire, Step3_Peel_Wire, Step4_Check_Alignment and classifying if the product is defect or not.

The business goal is to predict defect rates based on test results to improve production quality, reduce wastage, and optimize the manufacturing process. These datasets were provided with timestamps and metrics from their respective testing stages.

Key Considerations:

Data integration requires consistent timestamps for accurate analysis. Identifying potential issues or anomalies in the dataset is critical before proceeding to model building.

II. Descriptive Analytics

Dataset Size: A total of 6 data sets with 10,000 rows were available .

Key Observations:

Three out of six datasets had dates listed as 28-08-2024, while others were dated 29-08-2024. The time format across datasets was inconsistent. Most datasets used MM:SS, but one had the format HH:MM:SS. There are no Missing values were present

III. Assumptions

Date Adjustment: Given the small interval between dates (1 day), the three datasets with earlier dates were adjusted forward by one day to ensure uniformity. This assumption was validated through exploratory analysis, confirming the logical continuity between dates.

Time Standardization: Converted the HH:MM:SS format to match MM:SS by removing the hour component, as it was redundant for the analysis.

Consistency Check: Merged datasets were cross-validated using timestamps to ensure alignment of testing stages.

IV. Code Flow

Data Loading:

Loaded six datasets using pandas, Inspected data types, missing values, and inconsistencies.

Data Cleaning:

Adjusted date inconsistencies and standardized time formats. And combined Time and Date columns and formed a Timestamp column and converted to datetime format .

Data Merging:

Merged all datasets on DateTime and Time as primary keys and joined each dataset using 'inner' join. Created a unified dataframe representing all test stages for each timestamp.

Feature Engineering:

Generated additional features like rolling means and standard deviations for temporal patterns. Derived lag features to incorporate dependencies across timestamps.

2. Model Building

I. Data Splitting

Data was split into: Training Set (80%): Used for model training. Testing Set (20%):

Used for model evaluation. The splitting maintained sequential integrity, ensuring no data leakage across timeframes.

II. Preprocessing

Scaling: Numerical features were scaled using RobustScaler, which is robust to outliers.

Encoding: Categorical features were one-hot encoded to convert them into numerical representations suitable for machine learning models.

III. Model Selection

Algorithm: Random Forest Regressor

Justification: Handles non-linear relationships, robust to overfitting with proper hyperparameter tuning, and provides feature importance metrics.

Hyperparameters:

n_estimators=100: Number of trees in the forest.

Default values for other parameters were used.

IV. Model Evaluation

Metrics Used:

Mean Absolute Error (MAE): Measures the average magnitude of errors. The MAE obtained is 0.01253 which explains that the error is very small.

R2 Score: Indicates how well the model explains the variance in the target variable. And the R2 score obtained is 0.9963103.

Adjusted R2: Accounts for the number of predictors to prevent overestimation of model performance. Adjusted R2 score obtained is: 0.99629417.
From the evaluation it is clear that model performed well in training without underfitting or overfitting.

3. Defect Cause Analysis

Feature Importance

Objective: Identify key factors contributing to defect rates.

Approach: Feature importance was extracted from the Random Forest model, which ranks features based on their predictive power.

Insights:

Top 10 features were identified, revealing significant contributors like Trg1PitchUpper, MeasurementCount LowerRight_ProductCenter_IrradiationDistanceX

These insights can guide manufacturing process adjustments to reduce defect rates.

4. Prediction Intervals

Sequential Prediction

Forecasted defect rates for a 7-day period using a sequential approach. Utilized the model's outputs as inputs for subsequent predictions to simulate real-time prediction scenarios.

Interval Calculation

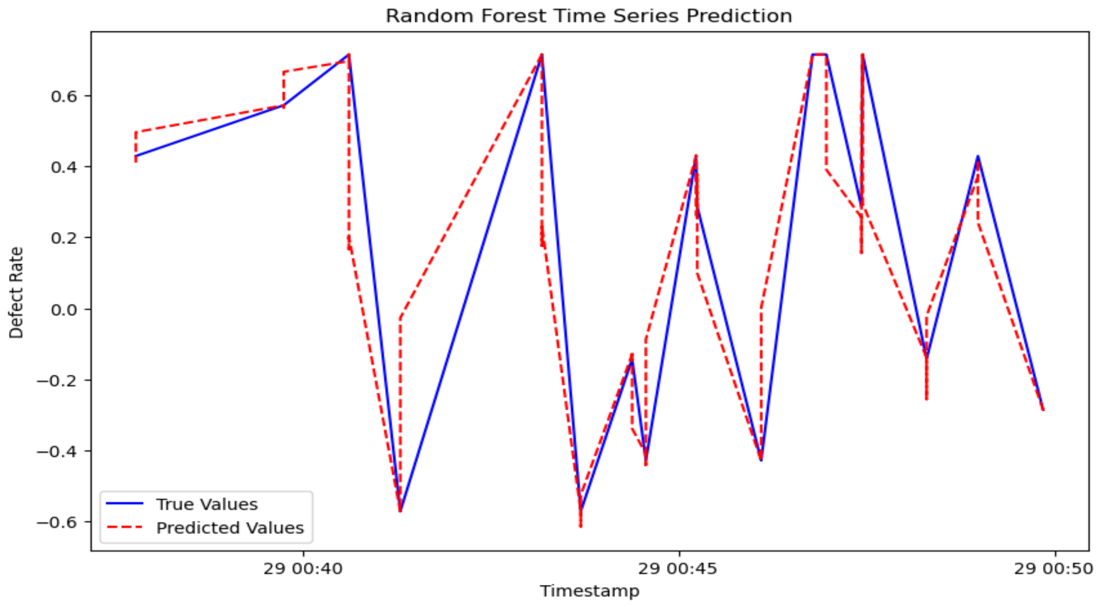
Calculated prediction intervals using the model's standard deviation of predictions. This provides a confidence range for predicted defect rates, helping decision-makers understand uncertainty.

5. Visualization Insights

I. Defect Rate Prediction

Description: Line chart comparing predicted defect rates to actual observed values.

Insights: Evaluates model accuracy and highlights periods with significant prediction errors.

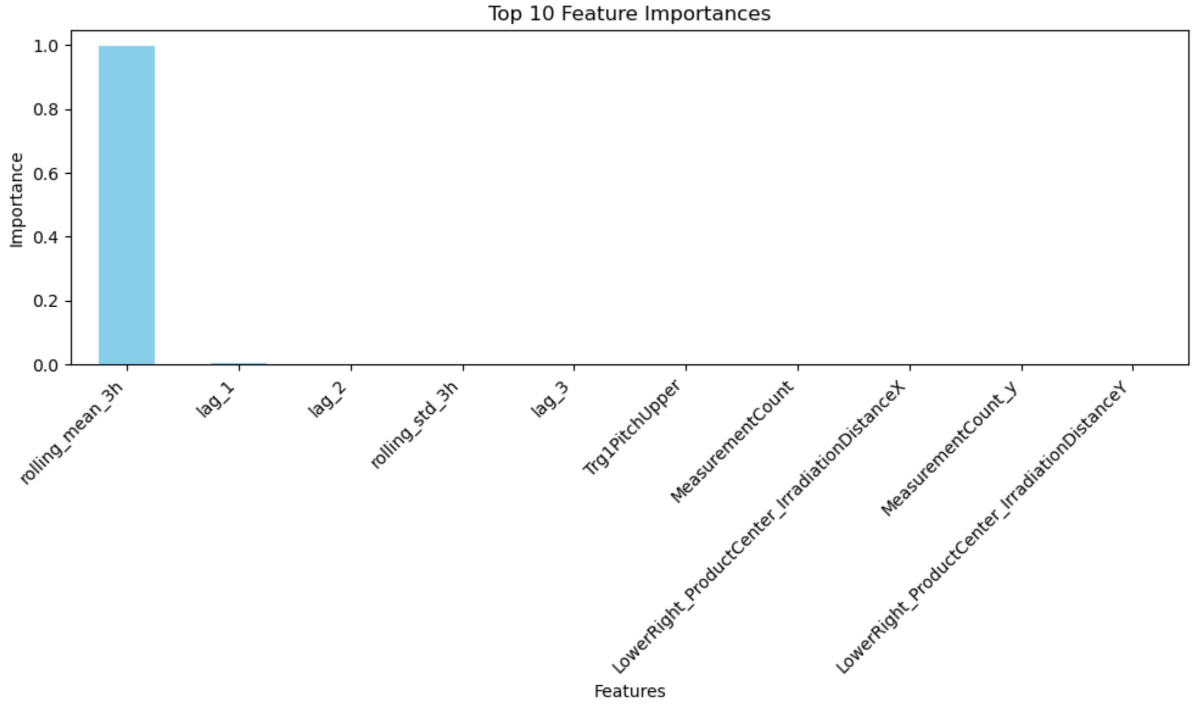


II. Feature Importance

Description: Bar chart visualizing the top 10 features affecting defect rates.

Insights: Identifies critical factors influencing defects, enabling targeted process improvements.

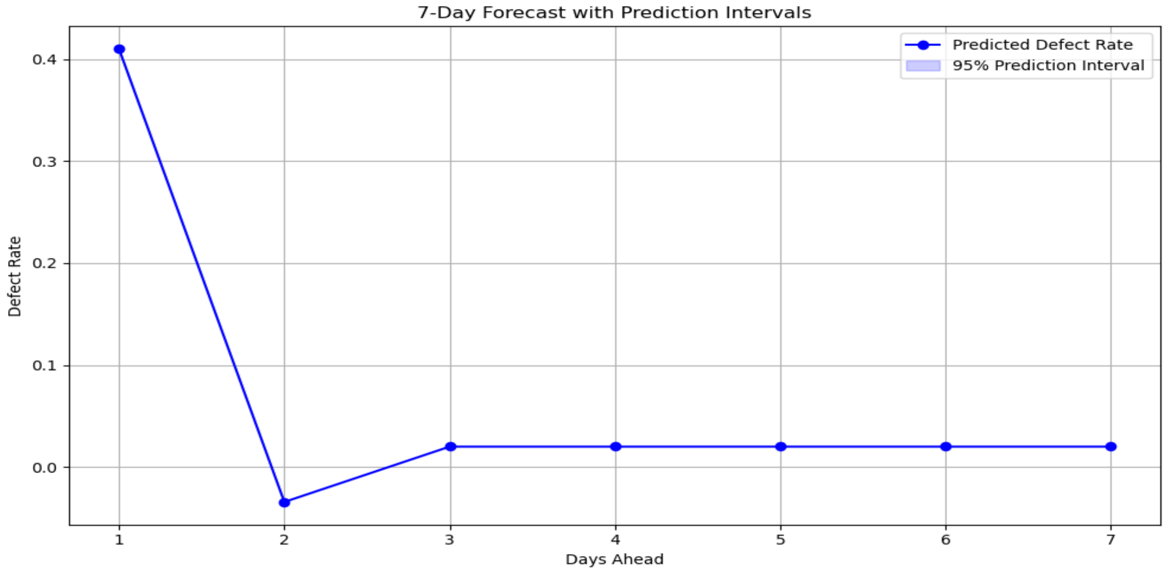
(Placeholder for chart)



III. Prediction Intervals

Description: Line chart with shaded confidence intervals for defect rate predictions.

Insights: Demonstrates prediction reliability and highlights potential risk periods.


7-Day Forecast with Prediction Intervals
Predicted Defect Rate
95% Prediction Interval
Defect Rate
Days Ahead