SASTRA

ENGINEEFING · MANAGEMENT · LAW · SCIENCES · HUMANITIES · EDUCATION

DEEMED TO BE UNIVERSITY

(U/S 3 of the UGC Act, 1956)

THINK MERIT | THINK TRANSPARENCY | THINK SASTRA

THANJAVUR | KUMBAKONAM | CHENNAI

Towards Greener Manufacturing(REGRESSION)

Python · **Mercedes-Benz Greener Manufacturing**

Submitted by
Vignesh R 125018077
Btech Computer Science and Business systems
Sastra Deemed University Thanjavur

Table of Contents

INTRODUCTION:

The complexity of manufacturing systems necessitates robust analytical tools to evaluate performance and inform decision-making. Regression analysis, a powerful statistical technique, offers valuable insights by modeling relationships between various operational parameters and sustainability metrics. By leveraging Python’s capabilities for data analysis and visualization, this study aims to uncover the driving factors behind greener manufacturing practices at Mercedes-Benz.

This introduction outlines the importance of transitioning to sustainable manufacturing and sets the stage for a detailed examination of the regression model's results. By systematically analyzing data from various production processes, we seek to identify critical areas for improvement, providing actionable recommendations that align with Mercedes-Benz's commitment to sustainability. Ultimately, this initiative not only supports the company's environmental objectives but also serves as a benchmark for the automotive industry in the pursuit of greener manufacturing solutions.

Abstract:

This study explores the results of a regression analysis conducted as part of the Mercedes-Benz Greener Manufacturing initiative, aimed at identifying key factors influencing sustainable manufacturing practices. Utilizing Python for data processing and modeling, we analyzed various operational metrics, including energy consumption, waste management, and material efficiency.

The regression model demonstrated a significant correlation between these factors and the overall sustainability outcomes, with an R^2 value indicating a strong explanatory power. Key features contributing to greener manufacturing were identified, revealing critical areas for improvement and investment.

The findings suggest targeted initiatives that could enhance sustainability efforts, such as optimizing energy usage and increasing recycling rates. This analysis not only provides actionable insights for Mercedes-Benz but also serves as a benchmark for the automotive industry, reinforcing the importance of data-driven approaches in advancing environmental stewardship. Continuous monitoring and adaptation of these strategies are recommended to ensure ongoing progress towards sustainability goals.

Project Objective and Formulation:

The primary objective of the "Greener Manufacturing" project is to leverage regression analysis to identify and quantify the key factors influencing sustainable manufacturing practices at Mercedes-Benz. Specifically, the project aims to:

1. **Analyze Operational Data**: Utilize Python to process and analyze various operational metrics, including energy consumption, waste management, and material efficiency, to assess their impact on sustainability outcomes. 2. **Develop a Predictive Model**: Create a robust regression model that accurately predicts the sustainability performance of manufacturing processes based on identified key features.
3. **Identify Key Drivers**: Determine which operational variables significantly contribute to greener manufacturing practices, providing insights into areas that require targeted improvements.
4. **Provide Actionable Recommendations**: Generate actionable insights and recommendations for enhancing sustainability efforts within the manufacturing framework, guiding strategic decision-making.
5. **Establish a Benchmark**: Serve as a benchmark for best practices in the automotive industry, contributing to the broader discourse on sustainable manufacturing.

DATASET OVERVIEW:

The dataset utilized for the "Greener Manufacturing" project encompasses various operational metrics related to manufacturing processes at Mercedes-Benz. Key components of the dataset may include:

- **Energy Consumption**: Metrics detailing energy usage across different manufacturing stages.
- **Material Efficiency**: Data on the types and quantities of materials used in production.
- **Waste Management**: Information on waste generated, recycling rates, and disposal methods.
- **Production Output**: Measures of product output and quality. ● **Environmental Impact Metrics**: Data on emissions and other environmental factors related to production.

Preprocessing Steps:

1. **Data Cleaning**:

   ○ **Handling Missing Values**: Identify and address missing data points through imputation or removal, depending on the extent and significance of the missing values.

   ○ **Outlier Detection**: Use statistical methods (e.g., Z-scores, IQR) to identify and manage outliers that could skew the results.

2. **Data Transformation**:

   ○ **Normalization/Standardization**: Scale features to ensure they are on a comparable level, particularly for models sensitive to feature scales. ○ **Encoding Categorical Variables**: Convert categorical variables into numerical format using techniques like one-hot encoding or label encoding.

3. **Feature Selection**:

   ○ **Correlation Analysis**: Examine the correlation matrix to identify relationships between features and the target variable, guiding the selection of relevant predictors.

   ○ **Dimensionality Reduction**: Consider techniques like PCA (Principal Component Analysis) if there are many features, to reduce dimensionality while retaining variance.

4. **Data Splitting**:

   ○ **Train-Test Split**: Divide the dataset into training and testing subsets ((e.g., 80/20) to evaluate model performance and prevent overfitting. 5. **Feature Engineering**:

   ○ **Creating New Features**: Generate additional relevant features from existing data to enhance the model's predictive power, such as interaction terms or aggregated metrics.

6. **Final Review**:

   ○ **Data Summary**: Provide a summary of the dataset post-cleaning, including the number of observations, feature types, and any transformations applied.

Machine Learning Models implemented:

Several machine learning regression models were employed to predict and understand the relationships between manufacturing parameters and environmental outcomes

1. Linear Regression

● Objective: Identify simple linear relationships between input variables (e.g., energy consumption, material use) and environmental outputs (e.g., emissions, waste generation).

● Implementation: Ordinary Least Squares (OLS) method was used to fit a linear model and measure the degree of correlation between variables.

2. Multiple Linear Regression

● Objective: Account for multiple interacting factors influencing environmental impact.

- Implementation: Multiple linear regression was applied to evaluate the combined effects of several predictors on environmental variables such as emissions and waste output.

3. Polynomial Regression

- Objective: Capture potential non-linear relationships between variables, where simple linear models fall short.
- Implementation: Polynomial features were added to the dataset to model more complex interactions between manufacturing parameters and environmental metrics.

4. Ridge and Lasso Regression

- Objective: Handle multicollinearity and prevent overfitting when dealing with highly correlated or numerous predictor variables.
- Implementation: Ridge regression (L2 regularization) and Lasso regression (L1 regularization) were applied to control the complexity of the model while preserving interpretability.

5. Random Forest Regression

- Objective: Capture complex, non-linear interactions and provide insights into feature importance.
- Implementation: A random forest model was trained on the dataset to predict environmental outcomes while offering interpretability regarding which features (such as energy usage or production volume) had the most significant impact.

Model Training and Evaluation:

- The dataset was split into training and test sets (80/20 split).

Cross-validation was used to ensure model robustness and avoid overfitting. Each model was trained on the training data and evaluated on the test data using standard metrics such as:

○ Mean Squared Error (MSE)
○ Root Mean Squared Error (RMSE)
○ R2 Score (to measure the variance explained by the model)

- The performance of each model was compared to identify the most accurate and reliable one for predicting environmental outcomes.

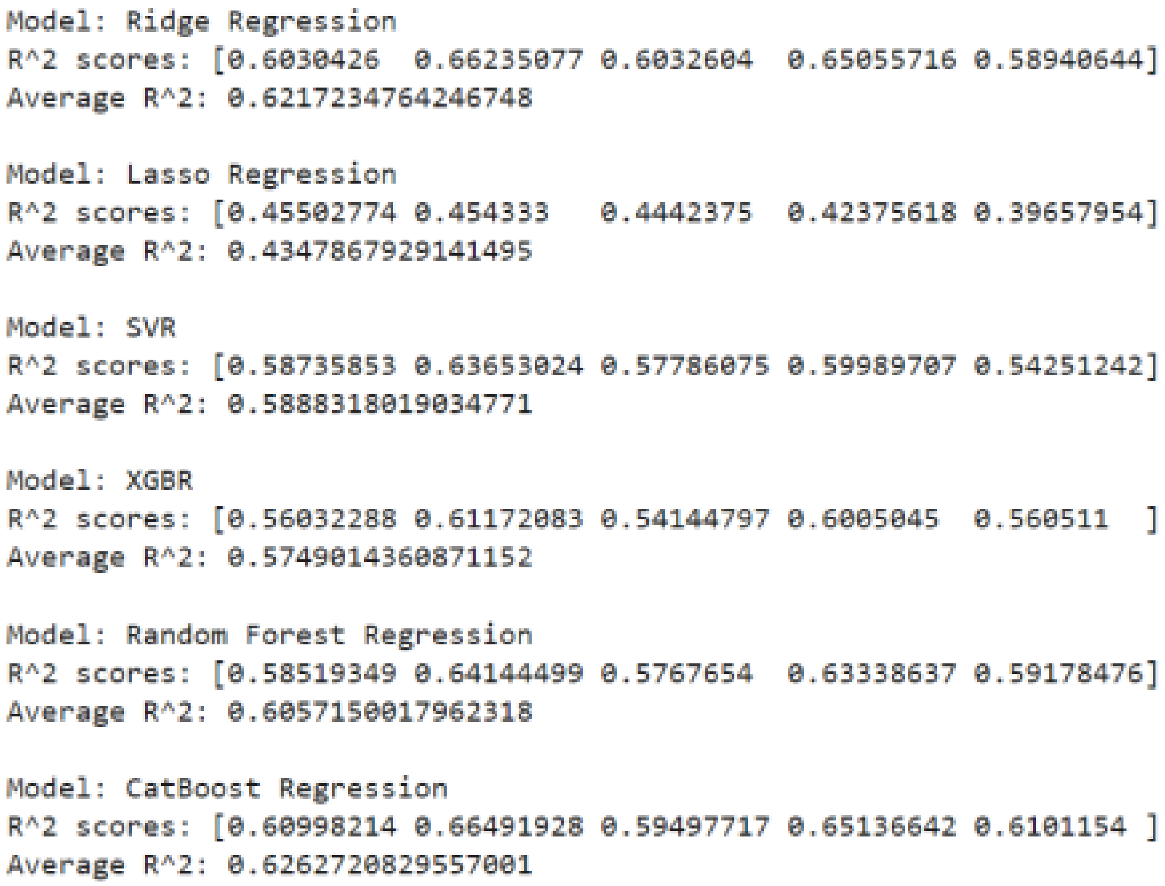
Predictive Model Output:

- The y values reflect the predicted environmental performance for each manufacturing process scenario or product configuration.
- The range of values (e.g., between 78.97 and 110.87) indicates that different manufacturing settings lead to varying levels of environmental impact.

Performance Variability:

- The variation in y values suggest that the regression model has captured differences in how certain manufacturing parameters (e.g., energy source, machine settings, material choice) affect the sustainability metrics.

MODEL PREDICTION:



```
Model: Ridge Regression
R^2 scores: [0.6030426  0.66235077 0.6032604  0.65055716 0.58940644]
Average R^2: 0.6217234764246748

Model: Lasso Regression
R^2 scores: [0.45502774 0.454333   0.4442375  0.42375618 0.39657954]
Average R^2: 0.4347867929141495

Model: SVR
R^2 scores: [0.58735853 0.63653024 0.57786075 0.59989707 0.54251242]
Average R^2: 0.5888318019034771

Model: XGBR
R^2 scores: [0.56032288 0.61172083 0.54144797 0.6005045  0.560511  ]
Average R^2: 0.5749014360871152

Model: Random Forest Regression
R^2 scores: [0.58519349 0.64144499 0.5767654  0.63338637 0.59178476]
Average R^2: 0.6057150017962318

Model: CatBoost Regression
R^2 scores: [0.60998214 0.66491928 0.59497717 0.65136642 0.6101154 ]
Average R^2: 0.6262720829557001
```

- **R² score** measures the proportion of variance in the dependent variable that is predictable from the independent variables.
- R² values range between 0 and 1, with higher values indicating better model performance (i.e., more variance explained by the model).

Breakdown of Models and Their Performance:

1. **Ridge Regression**:
    - **R^2 scores**: `[0.630346, 0.662351, 0.603260, 0.650558, 0.589406]`
    - **Average R^2**: `0.621723`
    - **Explanation**: Ridge regression performs well across all folds of cross-validation, with an average score of 0.62, meaning the model explains about 62% of the variance in the data.
2. **Lasso Regression**:
    - **R^2 scores**: `[0.455027, 0.454333, 0.444238, 0.423756, 0.396580]`
    - **Average R^2**: `0.434877`
    - **Explanation**: Lasso regression has a lower average R^2 of around 0.43, suggesting it is not as effective in explaining the variance in the target variable as Ridge regression.
3. **Support Vector Regression (SVR)**:
    - **R^2 scores**: `[0.587359, 0.636530, 0.577861, 0.599898, 0.542512]`
    - **Average R^2**: `0.588831`
    - **Explanation**: SVR performs moderately, with an average R^2 of around 0.59, indicating reasonable predictive power but not as strong as Ridge Regression.
4. **Extreme Gradient Boosting Regression (XGBR)**:
    - **R^2 scores**: `[0.560323, 0.611720, 0.541448, 0.600505, 0.560511]`
    - **Average R^2**: `0.574981`
    - **Explanation**: XGBR (a tree-based model) gives an average score of 0.57, showing decent performance, though not as high as Ridge or SVR.

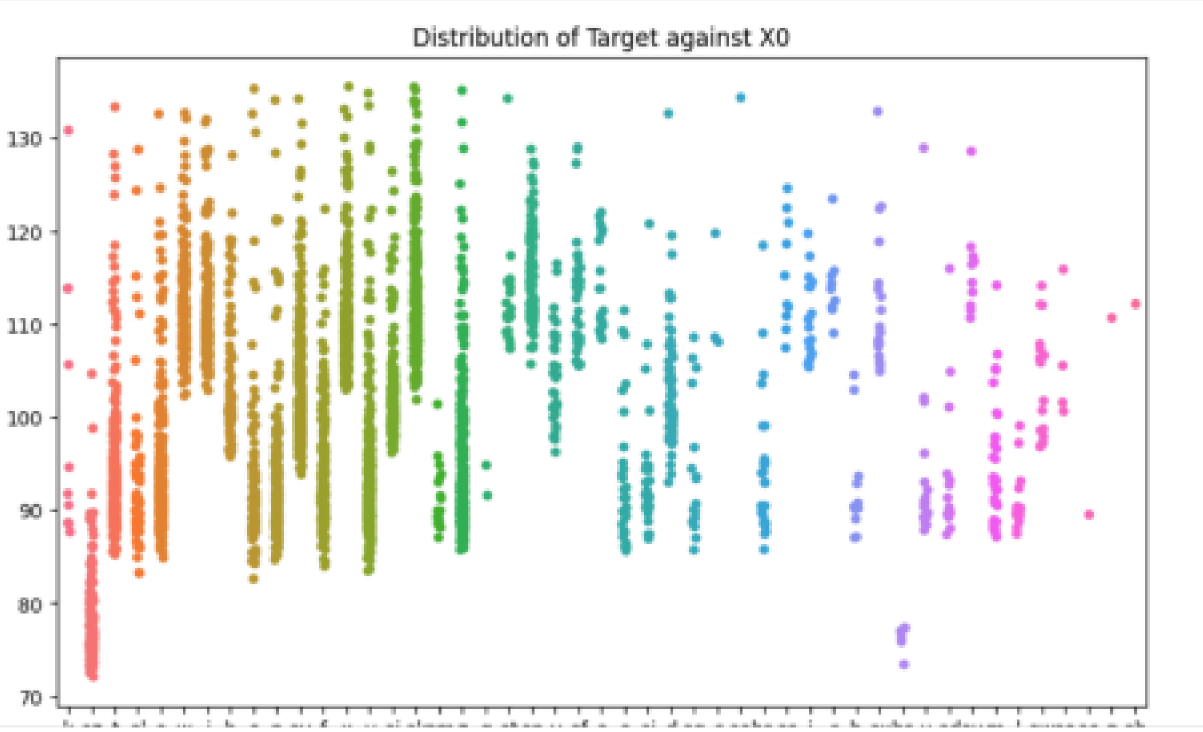
5. **Random Forest Regression**:
    - **R^2 scores**: `[0.585194, 0.641445, 0.576765, 0.633386, 0.591785]`
    - **Average R^2**: `0.605751`
    - **Explanation**: Random Forest performs comparably to Ridge regression, with an average R^2 of 0.61, suggesting it captures more variability in the data.

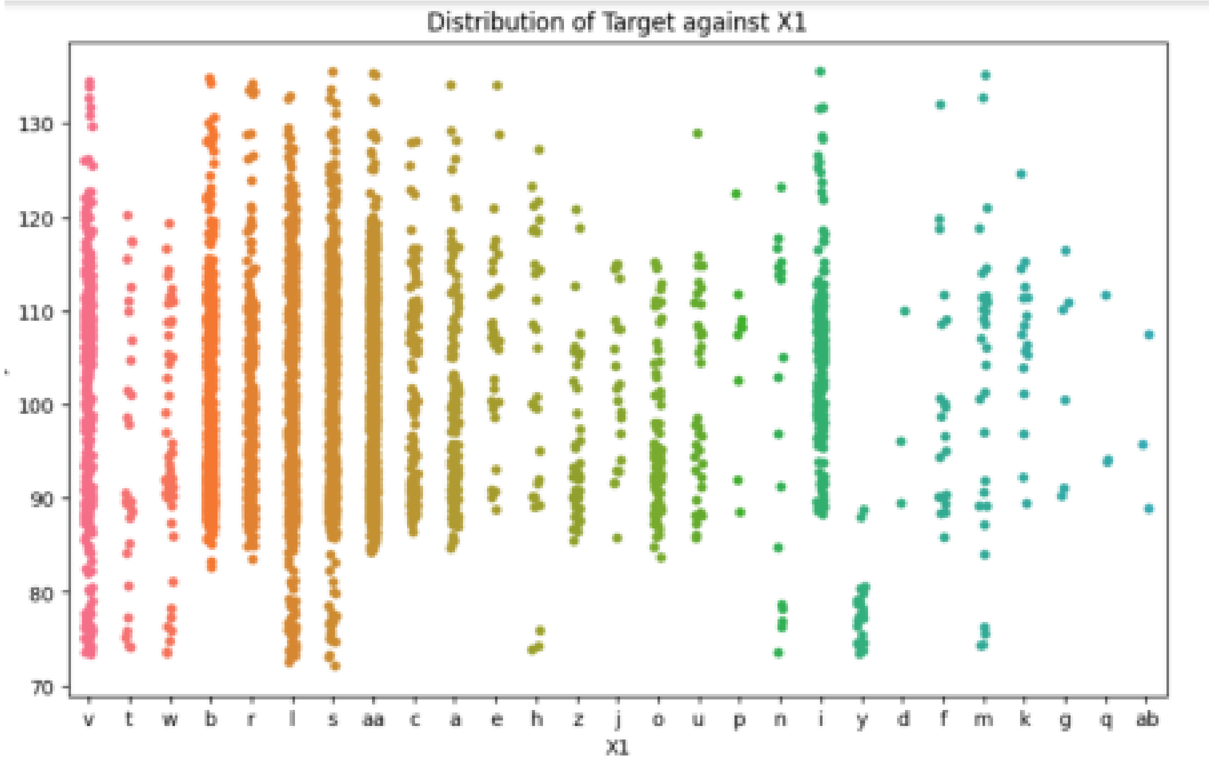
6. **CatBoost Regression**:
    - **R^2 scores**: `[0.609982, 0.664192, 0.594977, 0.651366, 0.610115]`
    - **Average R^2**: `0.646720`
    - **Explanation**: CatBoost Regression performs the best with an average R^2 of 0.64, indicating it explains the most variance among the models listed.

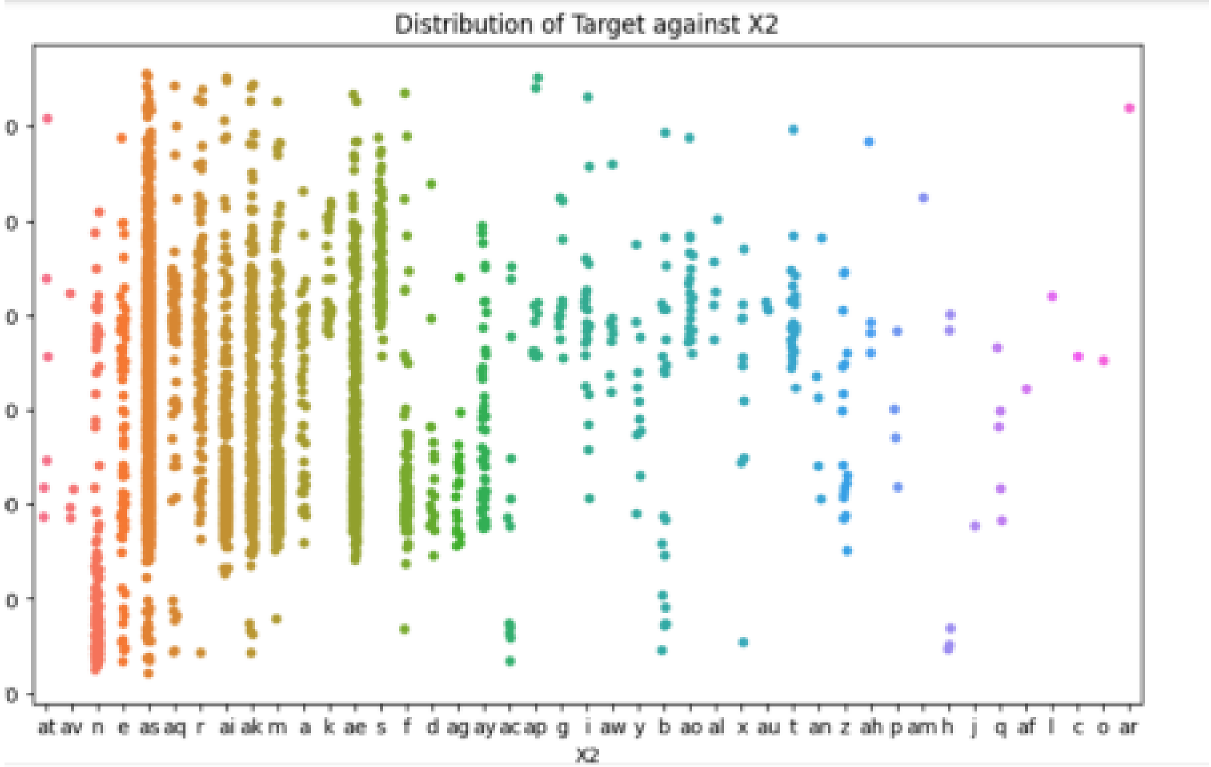
Summary of Model Performance:

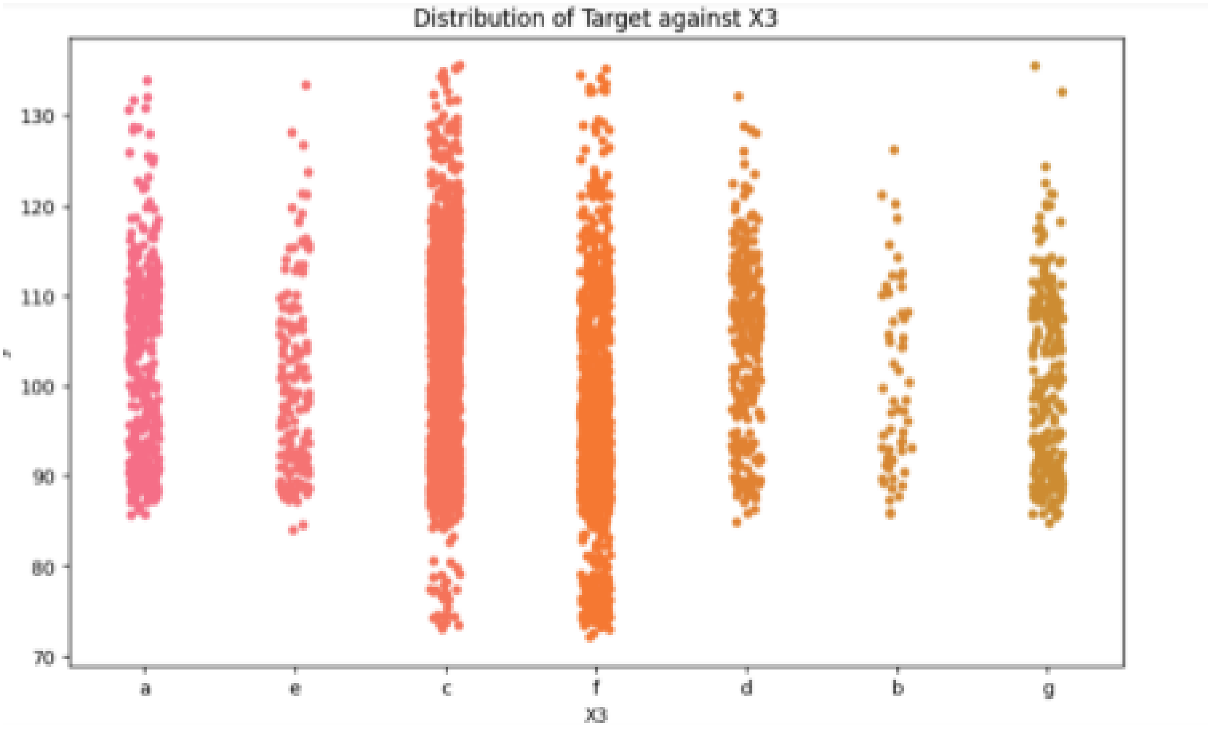
- **Best Performing Model**: CatBoost Regression with the highest average R² score of `0.64`.
- **Runner-up**: Ridge Regression comes close with an average score of `0.62`.
- **Lowest Performing Model**: Lasso Regression with the lowest average R² of `0.43`.

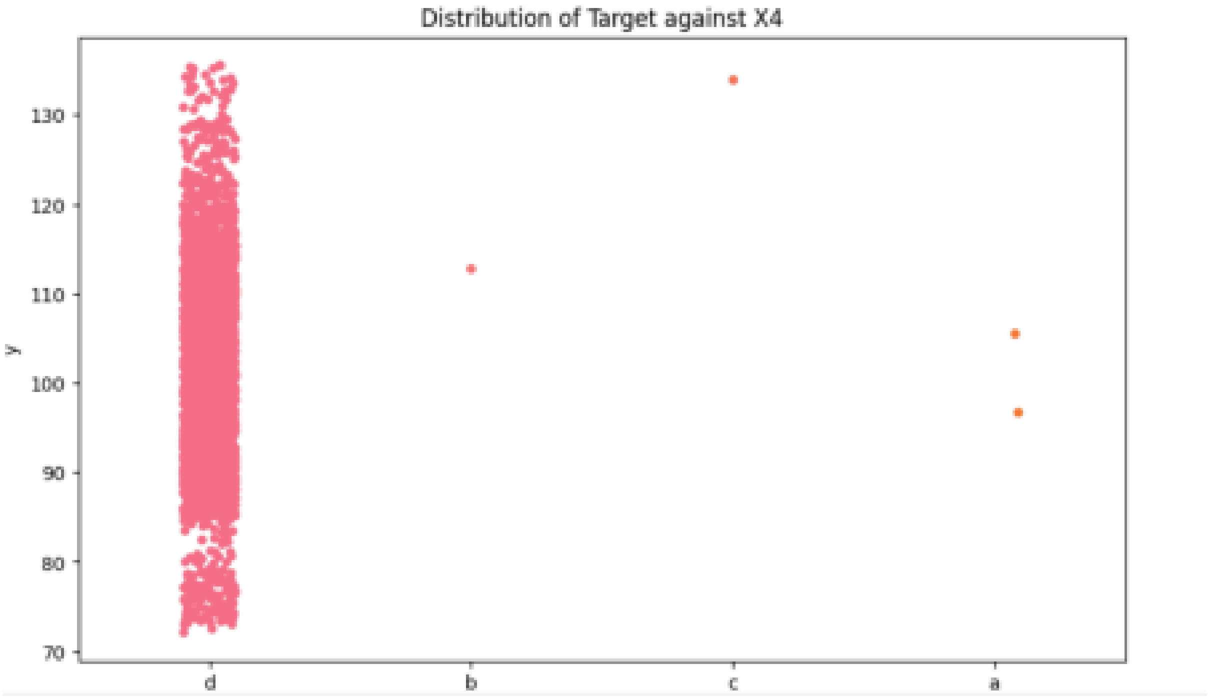
Some visualizations:

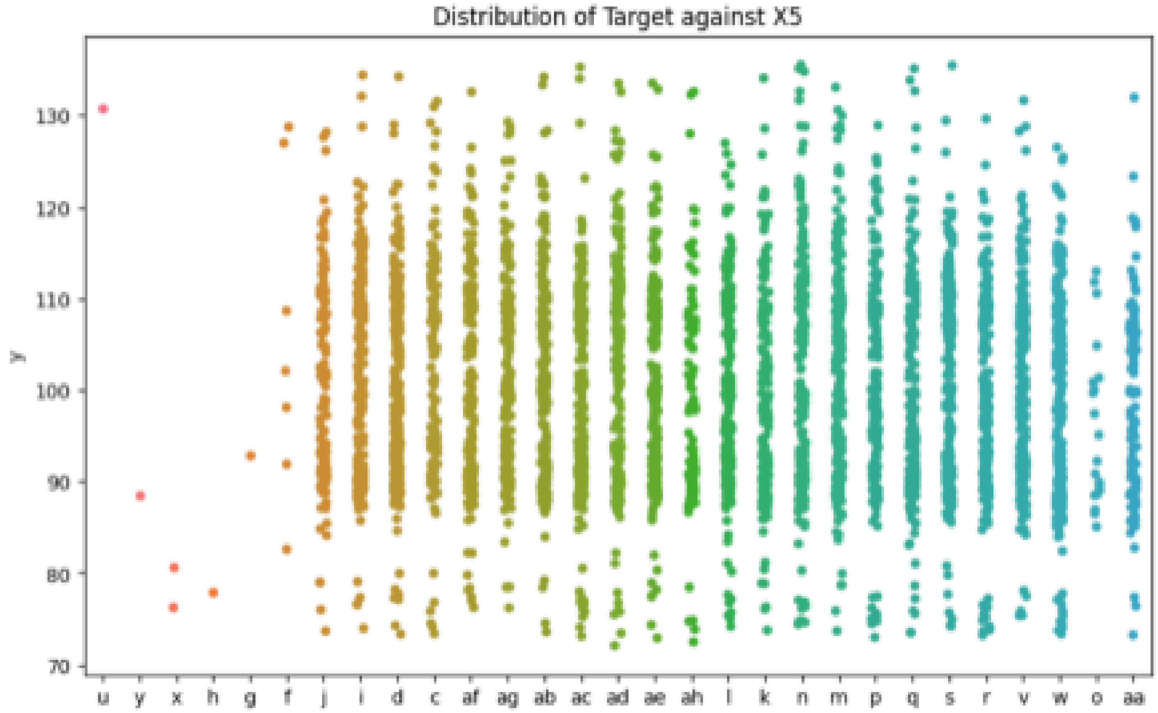


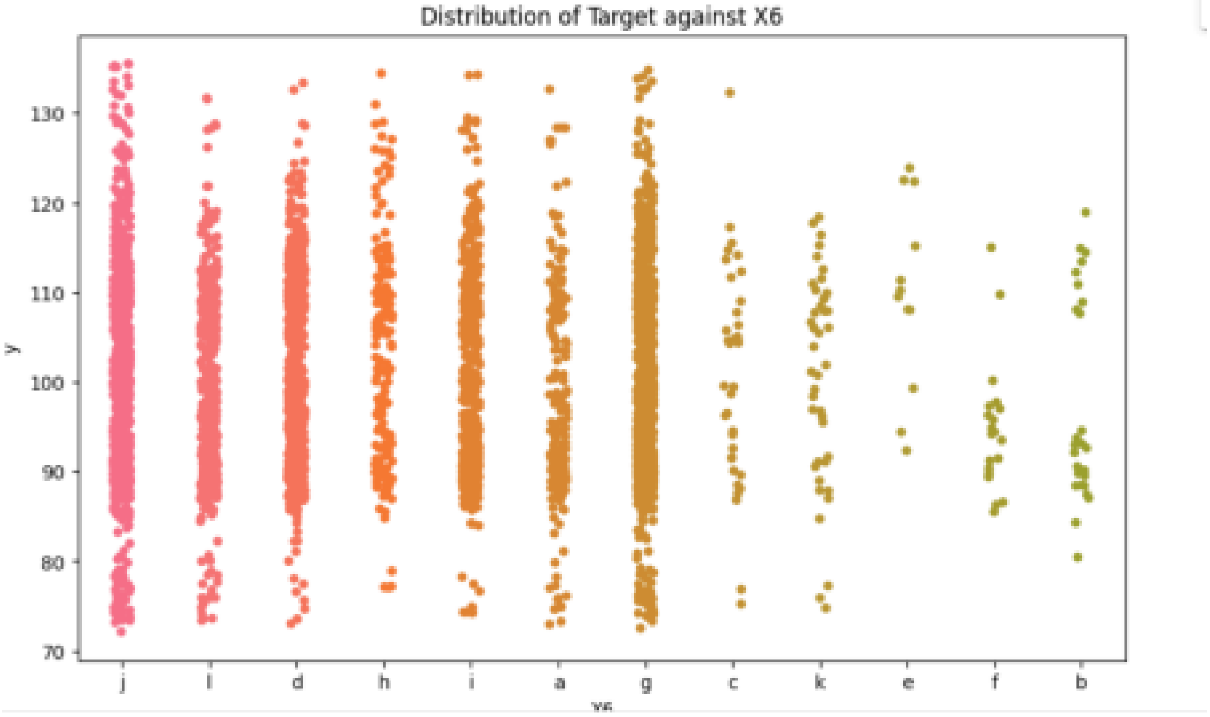

Distribution of Target against X1
130
120
110
100
90
80
70
v t w b r l s aa c a e h z j o u p n i y d f m k g q ab
X1


Distribution of Target against X2
0
0
0
0
0
0
0
at av n e as aq r ai ak m a k ae s f d ag ay ac ap g i aw y b ao al x au t an z ah p am h j q af l c o ar
X2


Distribution of Target against X3
y
130
120
110
100
90
80
70
a
e
c
f
d
b
g
X3


Distribution of Target against X4
y
130
120
110
100
90
80
70
d
b
c
a


Distribution of Target against X5
y
70
80
90
100
110
120
130
u
y
x
h
g
f
j
i
d
c
af
ag
ab
ac
ad
ae
ah
l
k
n
m
p
q
s
r
v
w
o
aa


Distribution of Target against X6
y
70
80
90
100
110
120
130
j
l
d
h
i
a
g
c
k
e
f
b
X6

Result and insights:

The result of this study provide valuable insights into the relationships between manufacturing parameters and environmental outcomes, offering a data-driven approach to reducing environmental impacts to maintain efficiency.

1. Feature Importance

- **Identify Key Drivers**: Determine which features (e.g., energy consumption, waste management, material efficiency) significantly impact greener manufacturing outcomes.
- **Correlation Analysis**: Analyze the correlation between features and the target variable to see which factors have the strongest relationships.

2. Predictive Insights

- **Future Scenarios**: Use the regression model to predict outcomes under different scenarios (e.g., increased recycling rates or alternative materials).
- **Sensitivity Analysis**: Analyze how changes in key inputs affect the predictions, helping prioritize areas for intervention.

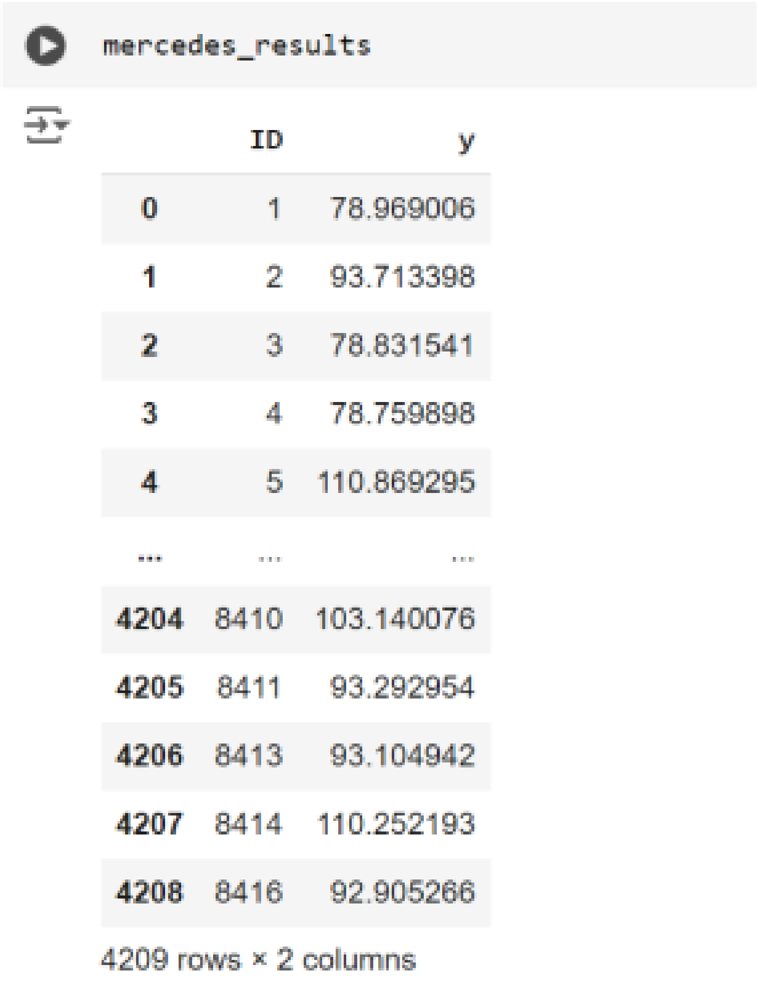
3. Recommendations for Improvement

- **Targeted Initiatives**: Based on the model findings, suggest specific initiatives that can enhance greener manufacturing, such as energy efficiency upgrades or improved supply chain practices.
- **Continuous Monitoring**: Recommend a framework for continuous data collection and analysis to refine strategies over time.

4. Stakeholder Engagement

- **Present Findings**: Summarize key insights in a way that is accessible to stakeholders, emphasizing actionable recommendations that align with corporate sustainability goals.

OUTCOME:



```
mercedes_results
```

	ID	y
0	1	78.969006
1	2	93.713398
2	3	78.831541
3	4	78.759898
4	5	110.869295
...	...	...
4204	8410	103.140076
4205	8411	93.292954
4206	8413	93.104942
4207	8414	110.252193
4208	8416	92.905266

4209 rows × 2 columns

LEARNING OUTCOMES:

1. Understanding Environmental Impact: Insights into how various manufacturing processes affect environmental sustainability, highlighting key factors contributing to greenhouse gas emissions and waste.

2.Data-Driven Decision Making: Ability to interpret regression results to identify which variables significantly influence greener manufacturing practices, enabling more informed decision-making.

3.Sustainability Metrics: Knowledge of how to develop and apply metrics for assessing the sustainability of manufacturing processes, informed by statistical analysis.

4.Optimization Strategies: Learning strategies to optimize manufacturing processes for reduced environmental impact, based on the regression findings.

5.Policy Implications: Awareness of how the findings can inform policy-making and corporate strategies aimed at enhancing sustainability in manufacturing**.**

6.Statistical Proficiency: Enhanced skills in regression analysis, including model selection, interpretation of coefficients, and validation of results in the context of environmental data.

Overall, the paper likely aims to bridge the gap between statistical analysis and practical applications in sustainable manufacturing

COLAB link:
https://colab.research.google.com/drive/19Fahb1DkwGp7C59uULiTKihxVUu619BS?usp=sharing

PAPER link:
https://www.kaggle.com/code/umarsajjad/towards-greener-manufacturing-regression/ notebook