**Design Choices Overview**

**Data Pre-processing:**

* **Handling Missing Values:**
  + Missing values were identified and rows containing them were removed due to their minimal impact on the dataset (<5%).
* **Outlier Detection and Removal:**
  + Outliers in numerical features were identified and eliminated using the Interquartile Range (IQR) method.
* **Feature Standardization:**
  + Numerical features were standardized to have a mean of zero and a variance of one using StandardScaler to ensure uniform scaling across features.

**Feature Engineering:**

* **New Feature Creation:**
  + New features were created, including the sum and average of multiple features to capture aggregated information.
  + Additional features, such as the square of feature x1 and the hour from the timestamp, were derived to uncover non-linear relationships and temporal patterns.

**Model Selection:**

* **Choice of Algorithm:**
  + Logistic regression was chosen as the baseline model for binary classification because of its simplicity, interpretability, and effectiveness in handling linear relationships.

**Model Training:**

* **Data Division:**
  + The dataset was split into training and testing subsets in an 80-20 ratio, allowing the model to be trained on a substantial amount of data while providing a robust evaluation.
* **Feature Standardization:**
  + Features were standardized prior to training the logistic regression model to ensure convergence and enhance model performance.

**Model Performance Evaluation:**

* **Metric Selection:**
  + **Accuracy:** Used as the primary metric due to its straightforward nature and effectiveness in assessing overall model performance.
  + **Precision, Recall, F1 Score:** These metrics were used to evaluate the model's capability to identify anomalies correctly and minimize false positives.
* **Evaluation Results:**
  + **Training Performance:** The model exhibited high accuracy, precision, recall, and F1 score on the training set, demonstrating its effectiveness in learning data patterns.
  + **Test Performance:** The model was evaluated on unseen test data, achieving similar performance metrics, which confirmed its generalizability.
  + **ROC AUC Score:** The ROC AUC score was calculated to measure the model's ability to differentiate between classes, resulting in a satisfactory score.

**Conclusion:**

* The logistic regression model performed well in predicting machine breakdowns with the given features.
* It met the project's success criterion by achieving over 75% accuracy on both the training and test datasets.
* The evaluation metrics provided a thorough understanding of the model’s performance, affirming its practicality for real-world applications.