Description of Design Choices for the Model

DATA PRE-PROCESSING:

MISSING VALUES HANDLING:

- **Identification of Missing Values:** There was a scan of missing values throughout the dataset.
- **Decision:** Instances with missing values, less than 5% of the total amount of data, were dropped in order to ensure that only instances free of missing values remained and were representative.

OUTLIER REMOVAL:

- **Method:** The Interquartile Range (IQR) method was employed for the detection of outliers in the case of numerical features.
- **Decision:** Outliers were removed from the dataset so that they didn't affect the model in its understanding of the underlying patterns and relationships in the data.

STANDARDIZATION:

- **Approach:** StandardScaler was used for the standardization of numerical features to a mean of zero and a variance of one.
- **Purpose**: Standardization ensured that features were on a comparable scale, and this is important for algorithms sensitive to feature magnitude, such as logistic regression.

FEATURE ENGINEERING:

CREATION OF NEW FEATURES:

- **Aggregated Information:** New features were engineered to calculate the sum and mean of many existing features, thereby capturing aggregated information that could enhance the predictive power of the model.
- **Non-linear Relationships:** Additional features, including the square of a feature ('x1') and the hour component of timestamp data, were derived to catch non-linear relationships and temporal patterns that may exist within the dataset.

MODEL SELECTION:

ALGORITHM CHOICE:

- **Selection:** Logistic regression has been selected as the primary model for binary classification.

- **Reasoning:** Simple, interpretable, and effective in handling linear relationships, which was suited to the nature of the problem and features available.

MODEL TRAINING:

DATA SPLITTING:

- **Ratio:** The dataset was divided into a training and testing set in an 80-20 ratio.

- **Purpose**: This ensured the model was trained on a sufficiently large portion of the data, with a portion reserved for evaluating the performance of the model, thus reducing the risk of overfitting.

STANDARDIZATION:

Preprocessing: Numerical features were standardized with the StandardScaler before model training to mitigate the impact of varying scales and ease convergence during the model training process.

PERFORMANCE EVALUATION OF THE MODEL:

METRIC CHOICE:

- **Main Metric:** Accuracy was chosen as the main metric because it is straightforward and effective to use while measuring the general correctness of the predictions.

- **Other Metrics:** Additionally, precision, recall, and F1-score were evaluated to reveal a more detailed performance of the model in its ability to classify instances correctly and reduce false positives.

MODEL EVALUATION:

- **Training Set Performance:** The model has been evaluated on the training set, and high accuracy, precision, recall, and F1 score were achieved, showing that the model actually learned well the underlying trends in the data.

- **Test Set Performance:** The model was tested on unseen test data, and similar performance metrics were achieved, showing that it could generalize well to new instances.

- **ROC AUC Score:** The ROC AUC score was computed as a measure of how well the model could separate the two classes, and the results point to satisfactory discriminative power.

CONCLUSION:

- **Model Performance:** In this project, a logistic regression model achieved strong performance in predicting machine breakdowns with a set of provided features. A high accuracy was achieved on both the training and test datasets.

- **Success Metric:** The success metric for this project, set at 75%, has been met through an accuracy that exceeds it, and the validity of the approach used is thereby confirmed.

- **Insights:** From the thorough performance evaluation of the model using a suite of various metrics, very valuable insights were drawn, confirming that it was suitable for real-world applications.