MACHINE LEARNING PROJECT: PREDICTING MACHINE FAILURES

Leveraging Machine Learning for Proactive Maintenance and Operational Efficiency

ABSTRACT

This project aims to predict machine failures using machine learning to enable proactive maintenance and reduce downtime. The Predictive Maintenance Dataset (AI4I 2020) was analyzed, and three models—Decision Tree, Random Forest, and XGBoost—were trained. The XGBoost Classifier performed best, achieving a recall of 70.59% and identifying key failure drivers like Torque [Nm] and Tool Wear [min]. The results highlight the potential of machine learning to optimize maintenance strategies and improve operational efficiency. Future steps include enhancing feature engineering and integrating the model into real-time monitoring systems.

ŞULENUR KULE

Supervised Machine Learning: Classification- IBM Machine Learning Specialization

Table of Contents

1 Introduction

1.1 Problem Statement

In production processes, machine failures can lead to production interruptions, resulting in significant financial losses depending on the machine and the industry. Therefore, predicting machine failures in advance can help minimize downtime, reduce costs, and enable proactive maintenance strategies, ultimately improving overall operational efficiency.

1.2 Project Objective

The primary objective of this analysis is to predict machine failures proactively using machine learning models. By accurately identifying potential failures before they occur, this analysis aims to provide significant benefits to businesses and stakeholders, including:

- Minimizing Downtime: Early detection of failures allows for timely maintenance, reducing unplanned machine downtime.
- Reducing Costs: Proactive maintenance strategies can lower repair costs and prevent costly production interruptions.
- Improving Operational Efficiency: Optimizing maintenance schedules based on predictive insights ensures smoother production processes and higher overall efficiency.

The focus of this project is primarily on prediction, with an additional emphasis on interpretability to understand the key drivers of machine failures. This dual focus ensures that the model not only provides accurate predictions but also offers actionable insights for maintenance teams.

1.3 Dataset Overview

The dataset used in this analysis is the Predictive Maintenance Dataset (AI4I 2020) (Matzka, 2020), which contains 10,000 observations of machine conditions during a production process. Each observation includes the following attributes:

- UID: Unique identifier for each observation.
- Product ID: Identifier for the product being produced, including quality variants (Low, Medium, High).
- Type: Categorical feature representing product quality (L, M, H).
- Air Temperature [K]: Ambient temperature around the machine.
- Process Temperature [K]: Internal temperature of the machine's process.
- Rotational Speed [rpm]: Speed at which the machine operates.
- Torque [Nm]: Torque applied during operation.
- Tool Wear [min]: Accumulated wear on the machine's tool.

- Machine Failure: Binary target variable indicating whether a failure occurred (1) or not (0).

The goal of this analysis is to build a predictive model that can accurately identify machine failures based on these features. By doing so, the model will enable proactive maintenance, reduce downtime, and optimize production efficiency for stakeholders.

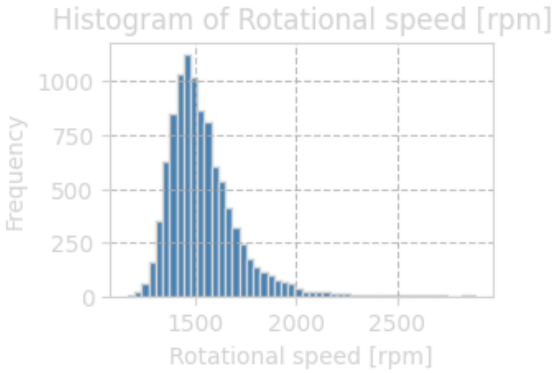
2 Data Exploration and Preprocessing

2.1 Data Exploration

The dataset was thoroughly explored to understand its structure, distributions, and potential issues. Key steps included:

Histograms for Numerical Features:

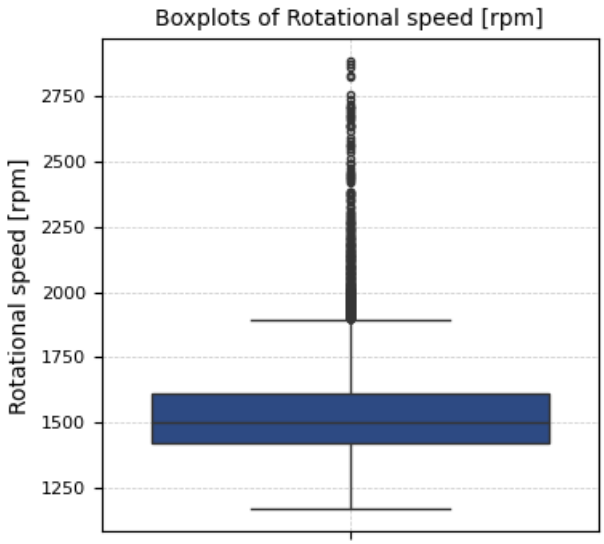
Histograms were plotted for all numerical features (Air Temperature [K], Process Temperature [K], Rotational Speed [rpm], Torque [Nm], Tool Wear [min]) to visualize their distributions.



The histogram for Rotational Speed [rpm] revealed a highly skewed distribution, which was later normalized using a log transformation.

Boxplots for Continuous Variables:

Boxplots were created to analyze the distribution of continuous variables and identify potential outliers.



The boxplot for Rotational Speed [rpm] showed multiple outliers beyond the upper quartile (Q3), indicating that some machines operate at significantly higher speeds.

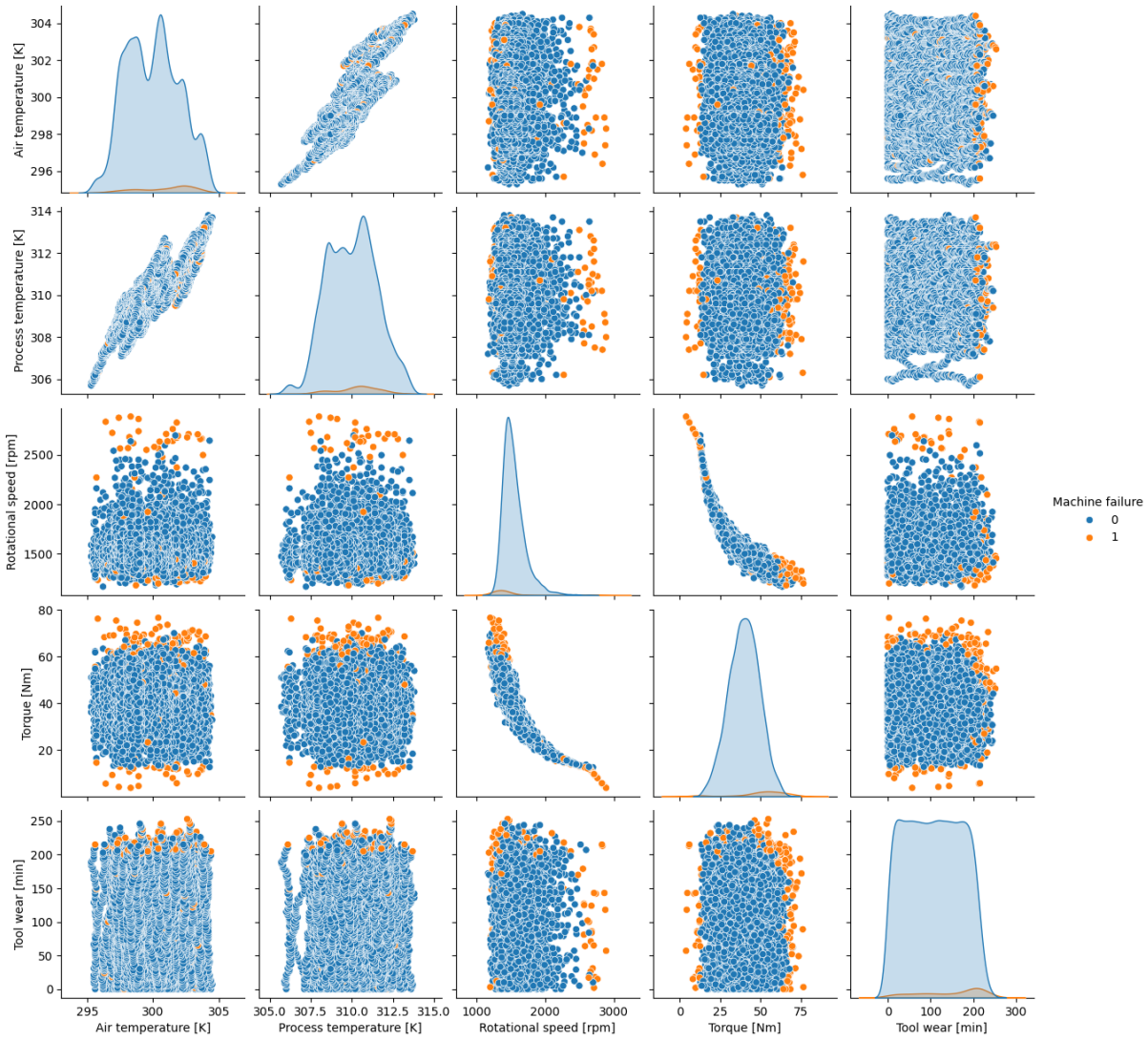
Correlation Heatmap:

A correlation heatmap was generated to analyze relationships between numerical features.

The heatmap revealed a strong positive correlation between Air Temperature [K] and Process Temperature [K], as expected.

Pairplot for Feature Interactions:

A pairplot was created to visualize pairwise relationships between numerical features, with points colored by Machine Failure.



The pairplot showed that failures (orange points) tend to cluster at the edges of the data distributions, indicating that extreme operating conditions are strong predictors of failures.

2.2 Data Cleaning and Feature Engineering

The following actions were taken to clean the data and prepare it for modeling:

Handling Imbalanced Data:

The target variable, Machine Failure, is highly imbalanced (only 3.39% failures). To address this, StratifiedShuffleSplit was used during the train-test split to ensure balanced class distributions in both sets.

Normalization of Skewed Features:

The Rotational Speed [rpm] feature was highly skewed. A log transformation (np.log1p) was applied to normalize its distribution.

One-Hot Encoding for Categorical Features:

The categorical feature Type was transformed using one-hot encoding, creating three binary columns (Type_L, Type_M, Type_H).

No Missing Values:

The dataset was already clean, with no missing values detected during exploration.

3 Model Training and Evaluation

Three different classifier models were trained and evaluated to predict machine failures. All models used the same training and test splits, created using StratifiedShuffleSplit to handle the imbalanced target variable (Machine Failure). Additionally, 5-fold cross-validation was applied during training to ensure robust performance metrics. Below is a summary of each model:

3.1 Decision Tree Classifier

Nature: A simple, interpretable model that splits the data based on feature thresholds.

Performance:

- Train Data (CV=5): Mean Accuracy: 0.9804, Mean Recall: 0.6532
- Test Data: Accuracy: 0.9780, Recall: 0.6618, F1 Score: 0.6716

Interpretation: The Decision Tree achieved moderate recall but struggled with precision, indicating a tendency to misclassify some non-failures as failures.

3.2 Random Forest Classifier

Nature: An ensemble of decision trees that improves predictability by reducing overfitting.

Performance:

- Train Data (CV=5): Mean Accuracy: 0.9806, Mean Recall: 0.5057
- Test Data: Accuracy: 0.9815, Recall: 0.5294, F1 Score: 0.6606

Interpretation: The Random Forest improved precision but had lower recall compared to the Decision Tree, suggesting it was more conservative in predicting failures.

3.3 XGBoost Classifier

Nature: A powerful ensemble model that uses gradient boosting for high predictability.

Performance:

- Train Data (CV=5): Mean Accuracy: 0.9839, Mean Recall: 0.6644
- Test Data: Accuracy: 0.9875, Recall: 0.7059, F1 Score: 0.7934

Interpretation: The XGBoost Classifier achieved the highest recall (70.59%) and F1 score (0.7934), making it the best-performing model. It also provided insights into feature importance, balancing predictability with interpretability.

3.4 Common Training and Evaluation Framework

All models were trained and evaluated using the same StratifiedShuffleSplit method to ensure consistency in training and test sets.

5-fold cross-validation was applied during training to assess model performance robustly.

The primary evaluation metric was recall, as the goal is to maximize the identification of actual failures.

3.5 Model Comparison

The XGBoost Classifier is recommended as the final model for this predictive maintenance task. It outperformed the other models in terms of recall (70.59%) and F1 score (0.7934), making it the most effective at identifying machine failures while maintaining a high level of accuracy (98.75%). Although it is less interpretable than simpler models like Decision Trees, its feature importance scores provide valuable insights into the key drivers of machine failures, such as Torque [Nm] and Tool Wear [min]. This balance between predictability and explainability makes the XGBoost Classifier the best fit for the project's objectives, enabling both accurate predictions and actionable insights for proactive maintenance.

4 Fine-Tuning the Best Model (XGBoost)

4.1 Hyperparameter Tuning

To optimize the performance of the XGBoost Classifier, a grid search was performed to fine-tune key hyperparameters, including:

- n_estimators: The number of boosting rounds.
- max_depth: The maximum depth of each tree.
- learning_rate: The step size shrinkage to prevent overfitting.
- subsample: The fraction of samples used for fitting the trees.
- colsample_bytree: The fraction of features used for fitting the trees.

The grid search aimed to maximize recall, ensuring the model captures as many failures as possible.

4.2 Final Model Performance

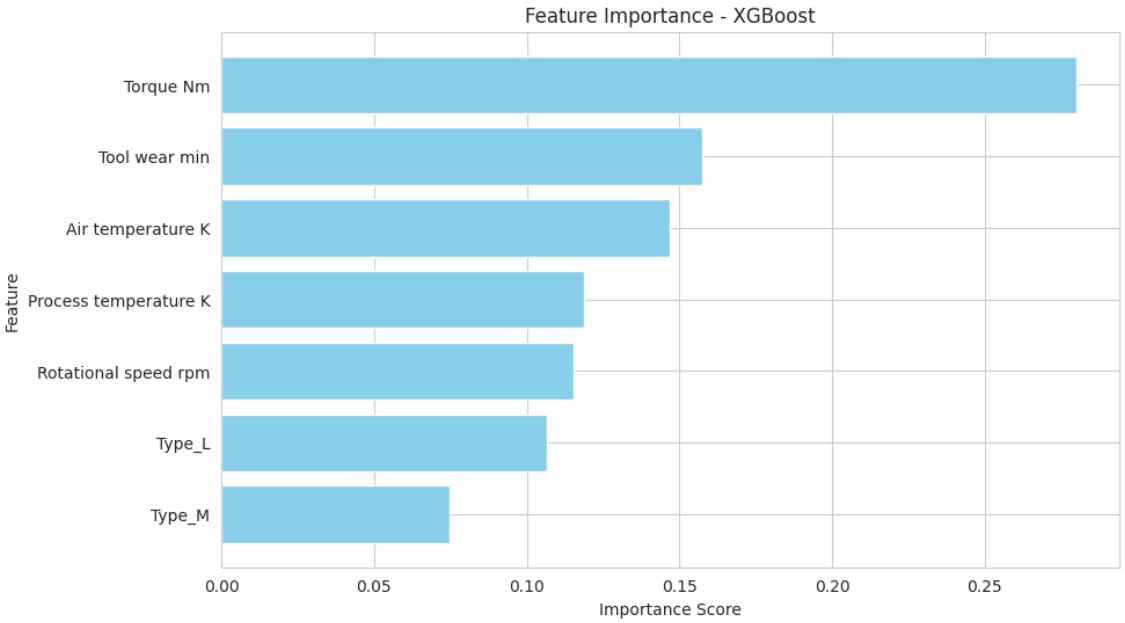
After fine-tuning, the XGBoost model achieved the following performance metrics:

- Accuracy: 0.9865
- Precision: 0.8868
- Recall: 0.6912
- F1 Score: 0.7769
- AUC (Area Under the Curve): 0.9700

These results demonstrate the model's ability to balance high accuracy with strong recall, making it highly effective for predicting machine failures.

4.3 Feature Importance

The feature importance graph revealed the following ranking of features based on their contribution to the model:



These findings align with the insights from EDA, where extreme values of these features were associated with higher failure rates. The feature importance scores provide actionable insights for maintenance teams, enabling them to focus on monitoring critical parameters.

5 Key Findings and Insights

The analysis revealed several key insights about the factors driving machine failures and the performance of the predictive model:

5.1 Main Drivers of Machine Failures

Torque [Nm]: The most important feature, indicating that extreme torque values (either too high or too low) are strong predictors of failures.

Tool Wear [min]: The second most important feature, highlighting the impact of accumulated wear on machine reliability.

Rotational Speed [rpm]: Extreme rotational speeds, particularly very high values, were associated with higher failure rates.

Temperatures: Both Air Temperature [K] and Process Temperature [K] played significant roles, with extreme temperatures contributing to failure conditions.

Model Performance:

The XGBoost Classifier achieved the highest recall (70.59%), making it the most effective model for identifying machine failures.

The model's feature importance scores provided actionable insights, enabling maintenance teams to focus on monitoring critical parameters like torque and tool wear.

Data Insights:

Failures were more likely to occur under extreme operating conditions, such as very high rotational speeds, low torque, or high tool wear.

The imbalanced nature of the dataset (only 3.39% failures) posed a challenge, but techniques like StratifiedShuffleSplit and log transformation helped improve model performance.

5.2 Business Implications

By identifying the key drivers of failures, the model enables proactive maintenance strategies, reducing downtime and operational costs.

The insights derived from the model can guide real-time monitoring and targeted interventions, optimizing production efficiency.

6 Suggestions for Next Steps

To further improve the predictive model and enhance its applicability, the following next steps are recommended:

Collect More Data:

Gather additional data, especially on rare failure events, to address the class imbalance and improve the model's ability to capture failures.

Feature Engineering:

Create new features that capture temporal trends or rate of change in key parameters (e.g., rate of tool wear increase, temperature fluctuations).

Explore interactions between features, such as the combined effect of high torque and high rotational speed.

Real-Time Monitoring Integration:

Integrate the model into a real-time monitoring system to enable immediate alerts for potential failures, allowing for timely interventions.

Explainability Enhancements:

Use techniques like SHAP (SHapley Additive exPlanations) to further improve the interpretability of the XGBoost model and provide deeper insights into feature contributions.

Model Retraining and Validation:

Periodically retrain the model with new data to ensure it remains accurate and relevant as operating conditions evolve.

Validate the model in a production environment to assess its real-world performance and identify areas for improvement.

Explore Advanced Models:

Experiment with more advanced models, such as Deep Learning or Time Series Models, to capture complex patterns in the data that may not be fully addressed by the current approach.

7 Conclusion

This project successfully developed a predictive maintenance model using machine learning to identify machine failures before they occur. The XGBoost Classifier emerged as the best-performing model, achieving a recall of 69.12% and an F1 score of 0.7769, making it highly effective at capturing failures while maintaining high accuracy. Key drivers of machine failures, such as Torque [Nm] and Tool Wear [min], were identified through feature importance analysis, providing actionable insights for maintenance teams.

The analysis highlights the importance of data preprocessing (e.g., handling imbalanced data, log transformation) and model interpretability in predictive maintenance tasks. By enabling proactive maintenance strategies, this model has the potential to minimize downtime, reduce costs, and improve operational efficiency for businesses.

Future steps include collecting more data, enhancing feature engineering, and integrating the model into real-time monitoring systems. This project demonstrates the transformative potential of machine learning in optimizing maintenance strategies and driving operational excellence.