Predictive Maintenance Analytics: Enhancing Industrial Machinery Reliability

Samriddha Adhikary

Chennai Mathematical Institute

30th April 2024

Introduction

- **Motivation:** Reliability is critical for new generation industrial machinery. Unplanned downtime due to machinery failures can lead to substantial costs and productivity losses. Proactive strategies are needed to predict and prevent such failures.
- **Problem Statement:** This project addresses the challenge of predicting machinery failures, using data science techniques. The goal is to develop predictive maintenance solutions that minimize downtime and maintenance costs.
- **Objective:** The primary objective is to design predictive maintenance analytics that can accurately forecast machinery failures. By integrating sensor data and machine learning algorithms, the aim is to empower industries with proactive machinery management capabilities.

Dataset Overview

- We utilize AI4I Predictive Maintenance Dataset from UCI Repository.
- It is a synthetic dataset reflecting real-world predictive maintenance scenarios.
- 10,000 data points with 6 features in columns.
- Features include: UID, productID, type, air temperature [K], process temperature [K], rotational speed [rpm], torque [Nm], tool wear [min]. There are two target variables, 'Target' and 'Failure Type' .
- Two main tasks: Predict machine failure and determine fault type.
- Evaluate performance and interpretability of results obtained for both tasks.

Dataset Overview

Description of Features

- **UID:** unique identifier ranging from 1 to 10,000
- **productID:** consisting of a letter L, M, or H for low (50% of all products), medium (30%), and high (20%) as product quality variants and a variant-specific serial number
- **air temperature [K]:** generated using a random walk process later normalized to a standard deviation of 2 K around 300 K
- **process temperature [K]:** generated using a random walk process normalized to a standard deviation of 1 K, added to the air temperature plus 10 K
- **rotational speed [rpm]:** calculated from power of 2860 W, overlaid with a normally distributed noise
- **torque [Nm]:** torque values are normally distributed around 40 Nm with an $\sigma = 10$ Nm and no negative values
- **tool wear [min]:** The quality variants H/M/L add 5/3/2 minutes of tool wear to the used tool in the process.

Dataset Overview

Machine Failure Modes

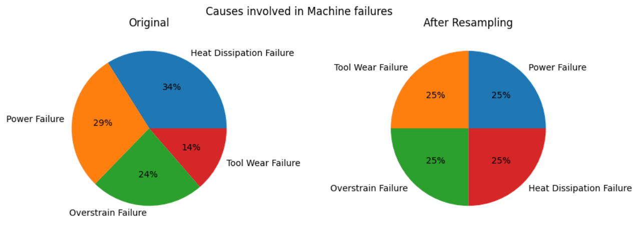
- **Tool Wear Failure (TWF):** The tool will be replaced if it fails at a randomly selected tool wear time between 200 - 240 mins.
- **Heat Dissipation Failure (HDF):** Heat dissipation causes a process failure if the difference between air and process temperature is below 8.6 K and the tool's rotational speed is below 1380 rpm.
- **Power Failure (PWF):** The product of torque and rotational speed (in rad/s) equals the power required for the process. If this power is below 3500 W or above 9000 W, the process fails.
- **Overstrain Failure (OSF):** If the product of tool wear and torque exceeds 11,000 minNm for the L product variant (12,000 M, 13,000 H), the process fails due to overstrain.
- **Random Failures (RNF):** Each process has a 0.1% chance to fail regardless of its process parameters. If at least one of the above failure modes is true, the process fails, and the 'machine failure' label is set to 1.

Data Preprocessing

- **ID Columns:** Machine failures are unlikely to depend on identifiers. The Product ID, composed of a letter followed by numbers, mirrors machine type and can be dropped due to redundancy with the machine type feature.
- **Feature Encoding:** A label encoding is applied to the categorical columns, since Type is an ordinal feature and Failure Type must be represented in one column. The mapping follows this scheme: Type: L=0, M=1, H=2 Failure Type: Working=0, PWF=1, OSF=2, HDF=3, TWF=4

Data Preprocessing

- **Target anomalies:** It is observed that, with respect to dataset's description when the failure is random (RNF), 'Machine Failure' is not set to 1. With only **18** RNF occurrences, and given its unpredictable random nature, we remove these rows, constituting a negligible **0.18%** loss of data.
- **Resampling with SMOTE:** Machine failures constitute a mere **3.4%** of the dataset, indicating a significant imbalance. Additionally, a pie plot depicting the causes of each failure highlights further imbalance in the dataset.After resampling, we get a **20.77%** increment of observations and **20.01%** cases of machine failures.



Model Comparison and Evaluation

Metric of Evaluation: F2 score.

Model	Parameters	F2 Score (Validation) - Binary	F2 Score (Validation) - Multiclass
KNN	KNeighborsClassifier() {'n_neighbors': [1, 3, 5, 8, 10]} Best Parameters: {'n_neighbors': 1}	0.9459	0.9415
SVC	SVC() {'C': [1, 10, 100], 'gamma': [0.1, 1], 'kernel': ['rbf'], 'probability': [True], 'random_state': [0]} Best Parameters: {'C': 1, 'gamma': 0.1, 'kernel': 'rbf', 'probability': True, 'random_state': 0}	0.9142	0.9019
RFC	RandomForestClassifier() {'n_estimators': [100, 300, 500, 700], 'max_depth': [5, 7, 10], 'random_state': [0]} Best Parameters: {'max_depth': 5, 'n_estimators': 100, 'random_state': 0}	0.9397	0.9340
XGB	XGBClassifier() {'n_estimators': [100, 300, 500], 'max_depth': [5, 7, 10], 'learning_rate': [0.01, 0.1], 'objective': ['multi:softprob']} Best Parameters: {'learning_rate': 0.01, 'max_depth': 5, 'n_estimators': 100, 'objective': 'binary:logistic'(for binary classification task),'multi:softprob'(for multiclass classification task)}	0.9628	0.9557

Observations

- Across both the binary and multiclass classification tasks, the XGBoost model consistently demonstrated superior performance, achieving the highest F2 scores among the evaluated models.
- It is noteworthy, however, that the training duration for the XGBoost model was comparatively longer than that of the other models assessed.

References

- Link to the dataset: Machine Predictive Maintenance Classification.
- Link to GitHub repository: Applied Machine Learning Final Project.

Thank You!