Predictive Maintenance to Reduce Machine Downtime

Wei Yuan
Purdue University, Daniels School of Business
yuan389@purdue.edu
Instructor: Prof. Matthew A. Lanham

ABSTRACT

This project analyzes the factors that affect Machine failure rates such as temperature, rotational speed, torque, tool wear, etc. This is very important for all manufacturing companies that need to produce machines as it helps the company to identify and proactively contact customers at risk of churn and try to repair the relationship in advance to reduce the risk of reduced revenue.

Machine maintenance is one of the major expenses in terms of cost and downtime due to machine breakdown affecting the entire manufacturing process. Building a highly accurate predictive model can help a company identify problem areas in advance and maximize cost reduction. The failure rate is predicted by using random forest, logistic regression, KNN, and decision trees to get a model with high accuracy.

BUSINESS PROBLEM

- Industries relying on machinery face significant challenges due to unexpected machine failures, which lead to costly downtimes. based on DISPEL's survey, hardware (45%) and software (39%) failures/malfunctions are the causes of downtime. As a result, the key to solving the majority of downtime lies in enabling technologies that prevent malfunctions, and facilitate reacting to malfunctions faster.

- According to James Chan, facilities using predictive maintenance can reduce mean time to repair (MTTR) by as much as 60 percent, highlighting the significant financial and operational benefits of this approach. In addition to maintenance costs, the accuracy and timely intervention of predictive modeling can significantly reduce maintenance costs.

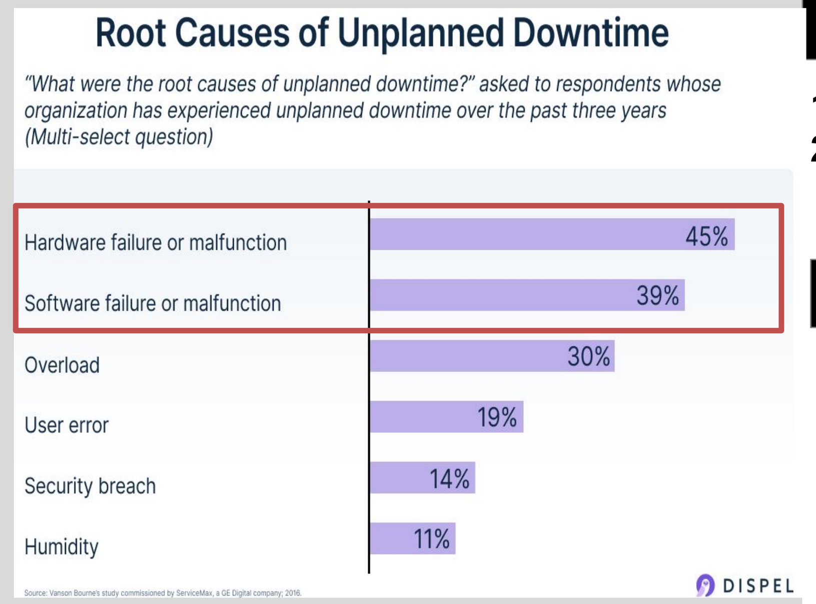
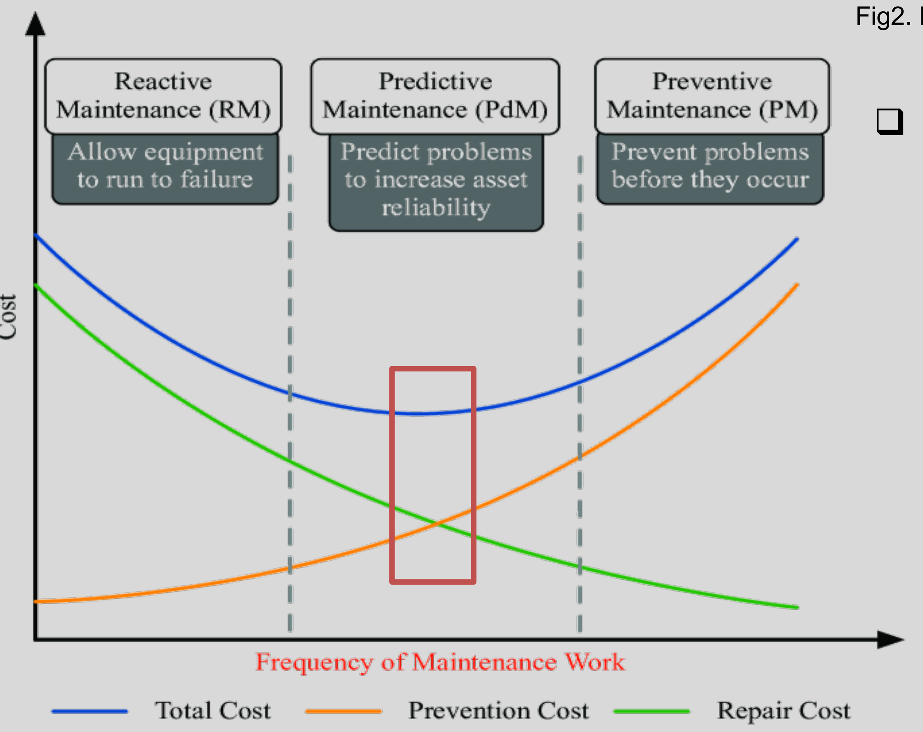


Fig2. Root Causes of Unplanned Downtime



- The goal of predictive maintenance is to develop a model that accurately predicts these failures. With accurate predictions, companies can optimize their maintenance schedules to minimize downtime and reduce overall maintenance costs. Therefore, building an effective and accurate predictive maintenance modeling application and optimizing it during testing is critical for the manufacturing industry.

Fig1. Comparison of RM, PM and PdM on the cost and frequency

Source:
A Survey of Predictive Maintenance: Systems, Purposes and Approaches. (n.d.-a). https://arxiv.org/html/1912.07383v2
Chan, J. (2024, April 9). *7 Benefits of Predictive Maintenance | Limble CMMS*. Limble CMMS. https://limblecmms.com/blog/benefits-of-predictive-maintenance/
How digital transformation reduces unplanned downtime in the energy sector. (n.d.). Dispel. https://dispel.com/blog/how-digital-transformation-reduces-unplanned-downtime-in-the-energy-sector

ANALYTICS PROBLEM FRAMING

Based on the business problem of unexpected machine failures leading to costly downtimes, predictive models will be developed using the provided maintenance dataset. These models will explore the relationship between input variables (air temperature, process temperature, rotational speed, torque, tool wear, etc.) and the target variable (machine failure). Key factors impacting machine failure will be identified, and different predictive models will be compared to determine the most accurate one.

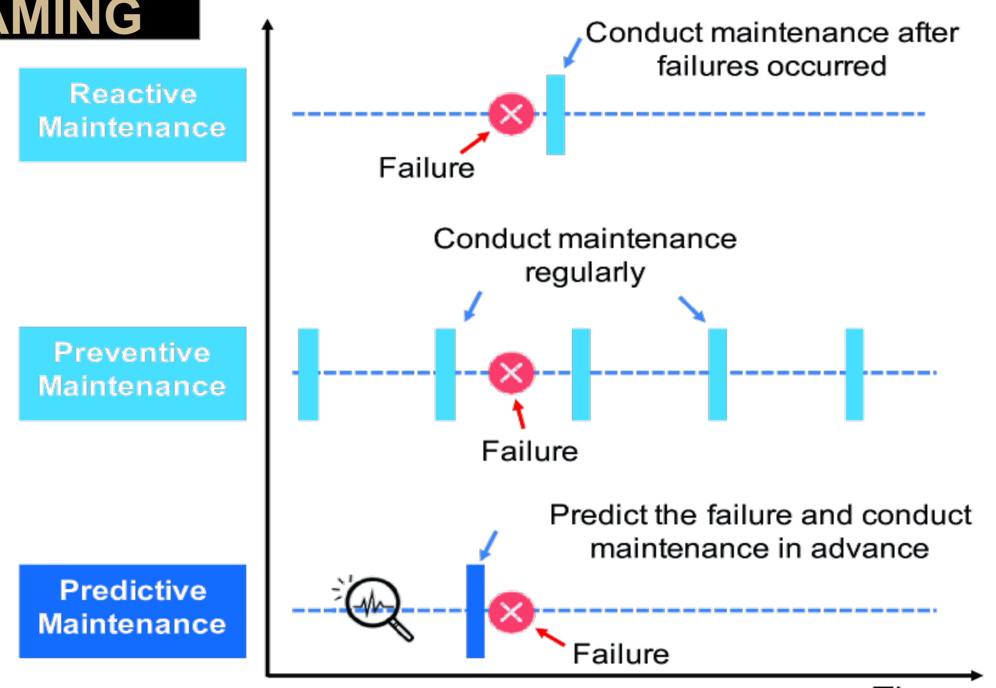


Fig3. Maintenance plans of RM, PM and PdM.

The objective is to predict machine failures accurately and optimize maintenance schedules, thereby minimizing downtime and reducing maintenance costs. Assumptions include the representativeness and quality of the dataset, and success will be measured using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. This approach ensures timely interventions and significant operational improvements.

RESEARCH QUESTIONS

1. **Key Predictors**: Which features are the most significant predictors of machine failures?
2. **Predictive Model and Accuracy**: Which model can better fit our needs and how accurately can we predict machine failures using the given maintenance dataset features?

DATA

- **Insight**: This synthetic dataset is from Kaggle which is modeled after an existing milling machine and consists of 10,000 data points stored as rows with 14 features in columns
  - 9 features of the milling machine: Product ID, Type, Air Temperature [K], Process Temperature [K], Rotational Speed [rpm], Torque [Nm], Tool Wear [min] with UDI
  - 5 independent Machine failure modes: tool wear failure (TWF), heat dissipation failure (HDF), power failure (PWF), overstrain failure (OSF), and random failures (RNF). These are binary variables, and if it is 1, it indicates machine failure because of non-compliance.
  - By analyzing the failure modes, we know that the total machine failure is 339, for each one TWF: 46 HDF: 115 PWF: 95 OSF: 98 RNF: 19. The highest one is 'Heat dissipation failure (HDF)', based on we could a conjecture that tools rotational speed, air temperature or torque, maybe the key factors.

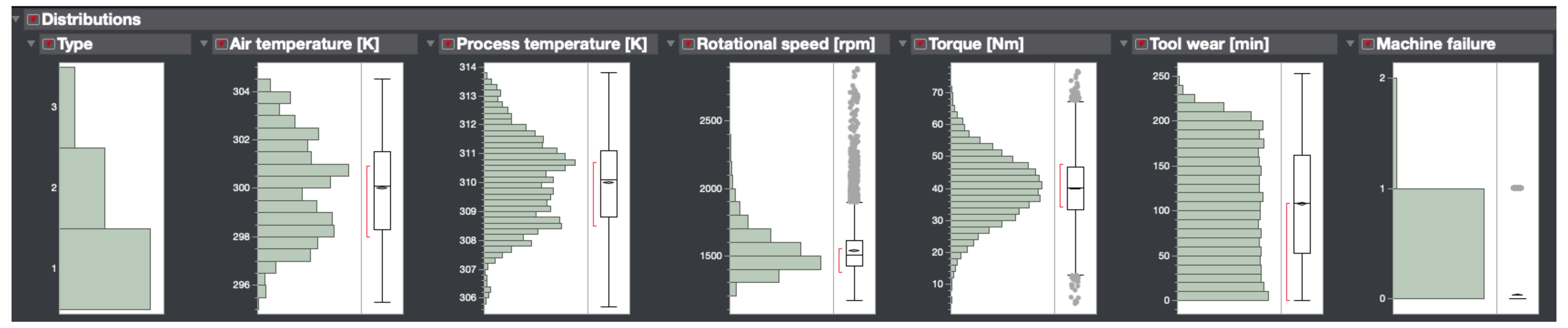


Fig4. Variables Distribution By JMP Pro

- **Data Preprocessing**
  - Missing Value & Outliers: Checked all numerical values with no missing values and outliers.
  - Label encoding: I changed the categorical variable 'Type' (e.g., 'low', 'medium', 'high'), and re-coding it to numerical values (e.g., 1, 2, 3) can help the model understand the order and make better predictions.

PURDUE UNIVERSITY®

Mitchell E. Daniels, Jr. School of Business

METHODOLOGY

1. Data Insight

Data Collection — Dataset From Kaggle and is part of the publication of Matzka (2020) → **Raw Data** 10,000 data points and 14 feature columns → **EDA**
- Visualization Distribution
- Understanding data meaning

2. Data Preparation and Partitioning

- Identify Missing values & Outliers
- Standardization
- Transformation
- Variable Selection

→ Data Split: Training Set 60% | Validation Set 20% | Testing Set 20%

3. Predict Model

Random Forest | Logistic Regression | Decision Tree | KNN

jmp STATISTICAL DISCOVERY

Model Evaluation and Selection
- Accuracy
- Precision
- Recall
- F1
- ROC_AUC
- Confusion Matrix

Apple Candidate Model → Feedback ⇄ Refine & Optimize

MODEL BUILDING AND EVALUATION – STATISTCIAL PERFORMANCE

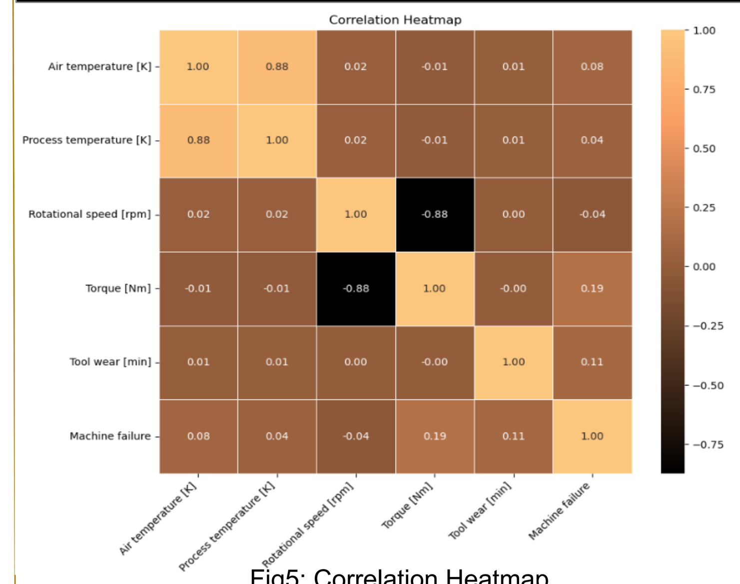


Fig5: Correlation Heatmap

- As we find one the data insight, we found Rotational Speed, Torque, and temperature have a high correlation. This is where we can conduct further analysis to confirm our first research question about key predictors.

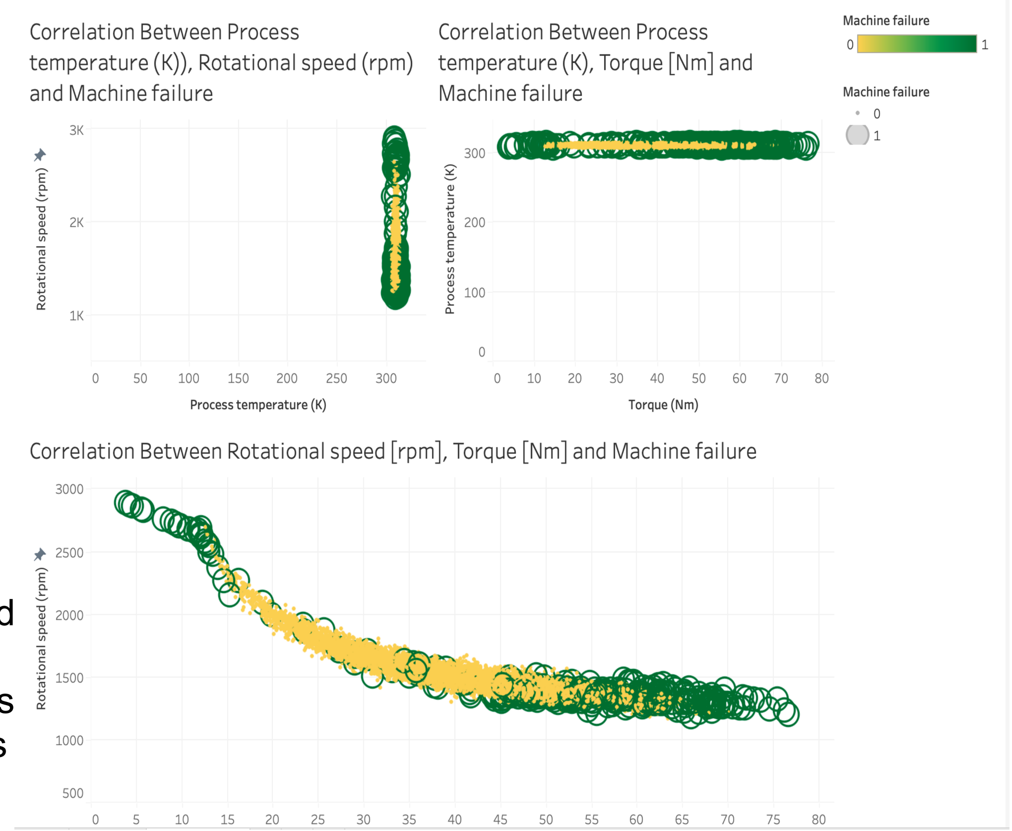


Fig6:Scatter Plot of variable by Tableau

Predict Model Evaluation Summary for Test Set

	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.8190	0.1284	0.8525	0.2232
Decision Tree	0.9765	0.6296	0.5574	0.5913
Random Forest	0.9845	0.8125	0.6393	0.7156
K-Nearest Neighbors	0.9460	0.3206	0.6885	0.4375

$$F_1 = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

- **Random Forest** performs well in accuracy, precision, recall, and F1 score, especially its F1 score (0.7156) and recall (0.6393), indicating that it performs well in both accurate identification and comprehensive fault detection.

- The recall of logistic regression is very high (0.8525), meaning that it detects most of the failures, but the precision (0.1284) is very low and may generate many false positives, could because of the imbalance in our categories, the total number of MACHINE failures is 339, but the overall sample size is 10, 000

- Based on the comparison of ROC curves and AUC values, the Random Forest model (AUC=0.96) performed the best in predicting machine failures and is recommended to be used as the primary model.

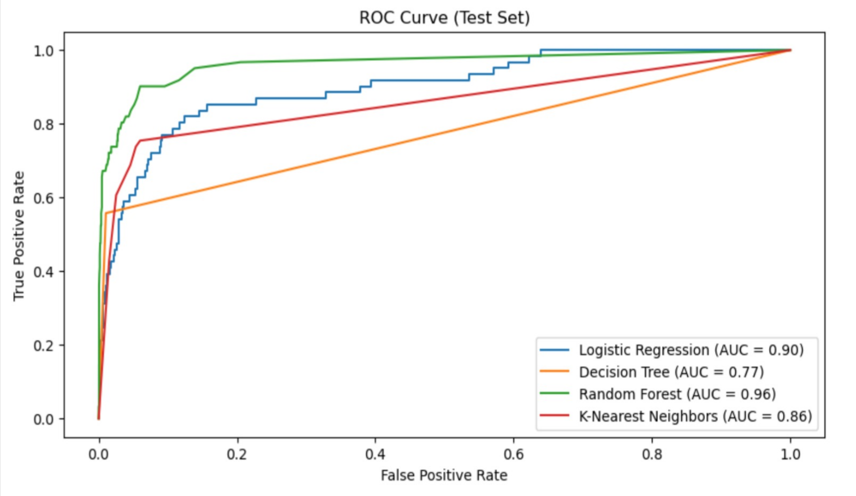


Fig7. ROC Curve for Test Set

MODEL EVALUATION – BUSINESS IMPLICATIONS

- Although the original Bagged Trees Ensemble has a higher accuracy (0.86) for faulty classes than Random Forest (0.81), Random Forest has a higher accuracy (0.99) for non-faulty classes, suggesting that Random Forest performs better in reducing false positives (false positives).

TABLE I. CONFUSION MATRIX OF THE BAGGED TREES ENSEMBLE CLASSIFIER USING 5-FOLD CROSS VALIDATION.

		true class: failure	true class: operation
predicted class	failure	294 (86.7 %)	45 (13.3 %)
	operation	121 (1.3 %)	9,540 (98.7 %)

Random Forest 5-Fold Cross Validation Accuracy:	[0.9755 0.974 0.5645 0.973 0.9815]
Bagged Trees 5-Fold Cross Validation Accuracy:	[0.962 0.9735 0.5845 0.9645 0.979]
Random Forest Mean Accuracy:	0.8937
Bagged Trees Mean Accuracy:	0.8927
Random Forest Accuracy Standard Deviation:	0.164626426
Bagged Trees Accuracy Standard Deviation:	0.154221464

Table2. Cross validation result

- Evaluating the performance of Random Forest and Bagged Trees Ensemble by further 5-fold cross validation, we can see that although the standard deviation of Random Forest is slightly higher than that of Bagged Trees, its average accuracy is slightly higher and it performs better in most of the cases, showing strong robustness

CONCLUSIONS

Predicting machine failures is critical to reducing downtime, lowering maintenance costs and improving overall operational efficiency. Predictive maintenance can greatly improve machine reliability and prevent unexpected breakdowns.

- Rotational speed [rpm], and Torque [Nm] have a strong correlation, and machine failure occurs when either is too fast. They are critical factors for machine failure
- Random Forest usually performs better in handling unbalanced datasets due to its randomness in feature selection and higher robustness.

Limitations:

1. The dataset is severely imbalanced having only 339 data points labeled as machine failure. It is not match with mass production environments real case. Therefore， in some further anal y tic we need more data or made some assumptions to fit it.
2. False rate: the recall of Random Forest is low on faulty classes, which can be further improved by using SMOTE for oversampling or combining with other sampling methods to improve the recall on a few classes (faulty classes)