Project Report: Equipment Failure Prediction Using Machine Learning

Kartik

Enrollment No. 22117065

Objective

The objective of this project is to develop machine learning models for predicting equipment failures based on sensor data. By accurately predicting failures in advance, the project aims to minimize downtime and maintenance costs in industrial settings.

Overview of the Project

In industries where equipment failure can lead to significant losses in productivity and revenue, predicting failures before they occur is crucial. This project leverages historical sensor data from industrial machinery to build robust machine learning models. These models are designed to predict equipment failures proactively, enabling timely maintenance and reducing operational disruptions.

Approach and Steps

1. Data Collection and Preprocessing

- **Data Collection**: Collected sensor data from various industrial equipment, including timestamps and multiple sensor readings.
- **Data Cleaning**: Cleaned the data to handle missing values, outliers, and ensure consistency across sensor readings.
- **Feature Engineering**: Engineered features such as statistical moments, time-based features, and rolling averages to capture equipment behaviour patterns.

2. Exploratory Data Analysis (EDA)

- **Target Distribution**: Analysed the distribution of the target variable ('failure') to understand class balance.
- **Feature Analysis**: Visualized sensor data distributions using line plots, PDFs, and box plots to identify trends and outliers.
- **Correlation Matrix**: Examined correlations between sensor readings to detect multicollinearity and understand feature relationships.

3. Model Selection and Evaluation

- **Baseline Models**: Implemented baseline models (e.g., Decision Trees, Random Forests) to establish initial performance benchmarks.
- **Handling Class Imbalance**: Addressed class imbalance using techniques like SMOTE (Synthetic Minority Over-sampling Technique) and SMOTE TOMEK.
- **Advanced Models**: Utilized ensemble methods (Random Forest, LightGBM) and gradient boosting algorithms (XGBoost) known for their accuracy in complex datasets.

4. Model Training and Hyperparameter Tuning

- **Cross-Validation**: Employed k-fold cross-validation to optimize model hyperparameters and ensure robust performance evaluation.
- **Performance Metrics**: Evaluated models based on metrics such as F1-score, precision, recall, and accuracy to measure predictive performance effectively.

5. Model Deployment and Monitoring

- **Deployment Strategy**: Discussed strategies for deploying models into operational systems, considering scalability and real-time prediction capabilities.
- **Monitoring**: Outlined methods for monitoring model performance post-deployment, including drift detection and periodic model retraining.

Defining Factors

1. Data Quality and Availability

The quality and availability of sensor data significantly influence model accuracy and reliability. Clean, consistent data without missing values or outliers is crucial for robust predictions.

2. Feature Engineering

Effective feature engineering, including the selection of relevant sensor data and creation of meaningful features, enhances the model's ability to capture equipment failure patterns accurately.

3. Sampling Techniques

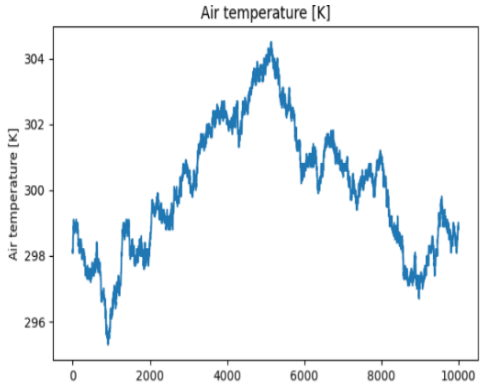
Implementing appropriate sampling techniques to handle class imbalance ensures models can predict both majority and minority class instances effectively, improving overall prediction performance.

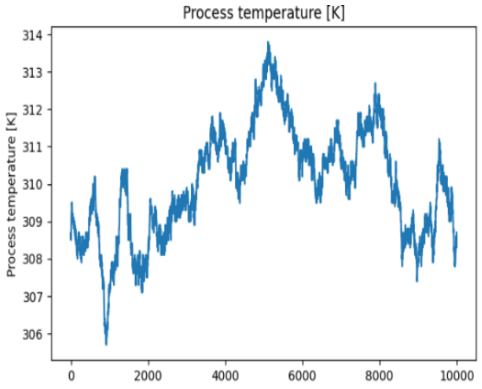
Models Used

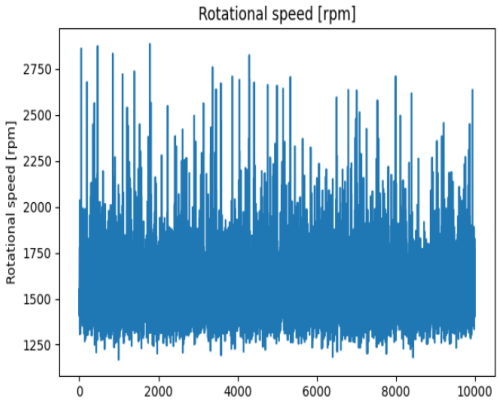
- **Random Forest**: Ensemble learning method known for its robustness and ability to handle complex relationships in data.
- **LightGBM**: Gradient boosting framework that excels in processing large datasets and achieving high accuracy.
- **XGBoost**: Another gradient boosting library known for its efficiency and speed.
- **Decision Tree**: Simple yet powerful tree-based model used for classification tasks.

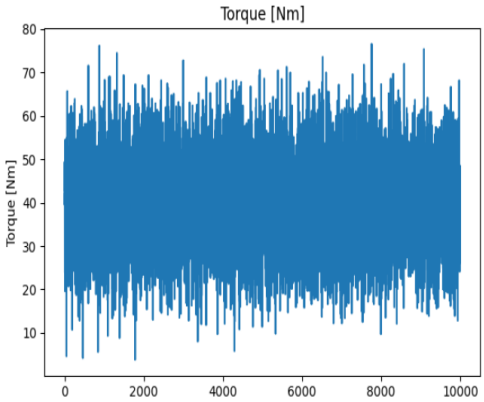
Exploratory Data Analysis (EDA) Plots

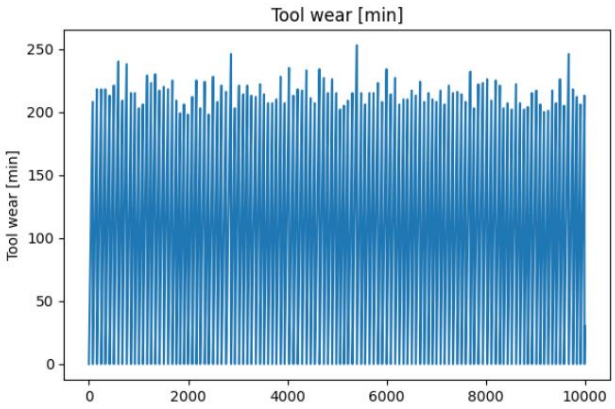
Line Plots:



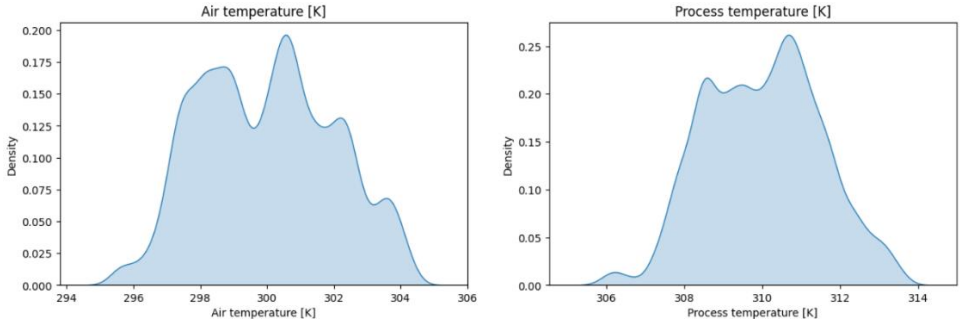


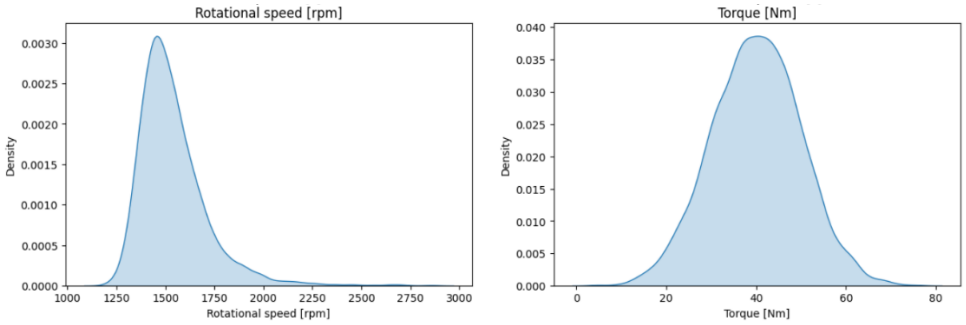


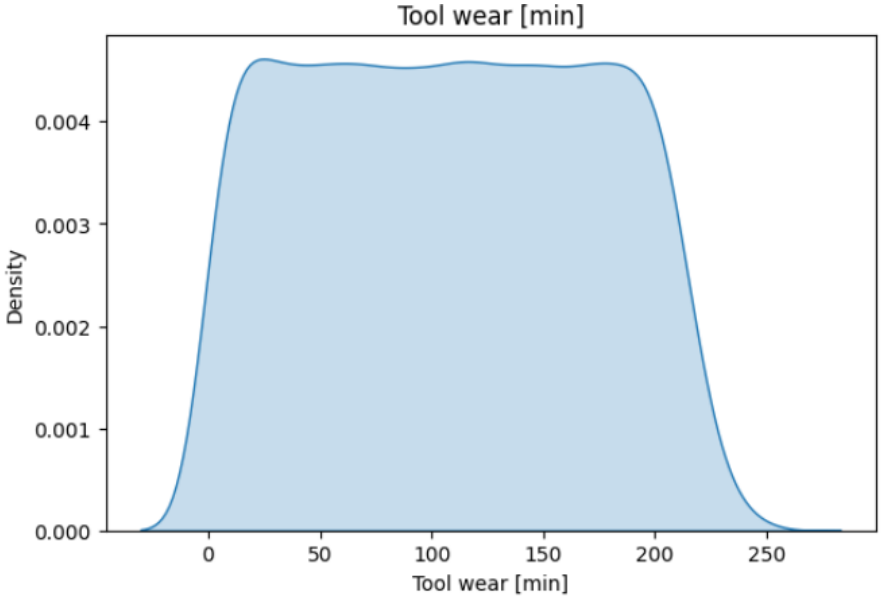



Tool wear [min]
Tool wear [min]
250
200
150
100
50
0
0
2000
4000
6000
8000
10000

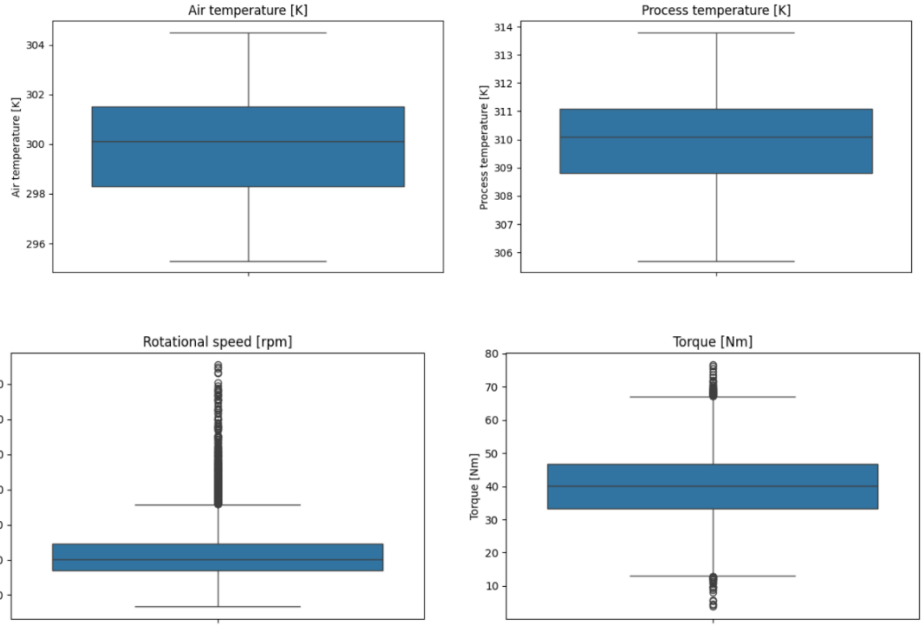
PDF Plots:

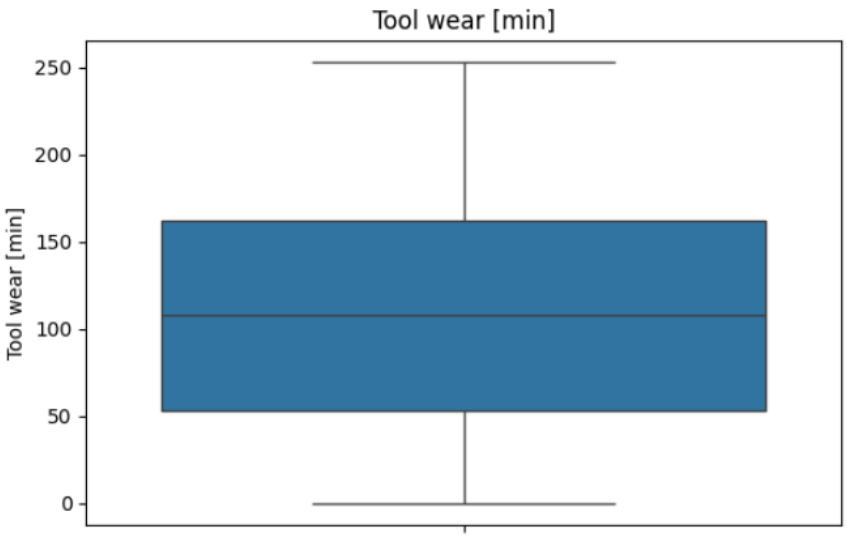

Air temperature [K]
Density
Air temperature [K]
Process temperature [K]
Density
Process temperature [K]


Rotational speed [rpm]
Density
Rotational speed [rpm]
Torque [Nm]
Density
Torque [Nm]

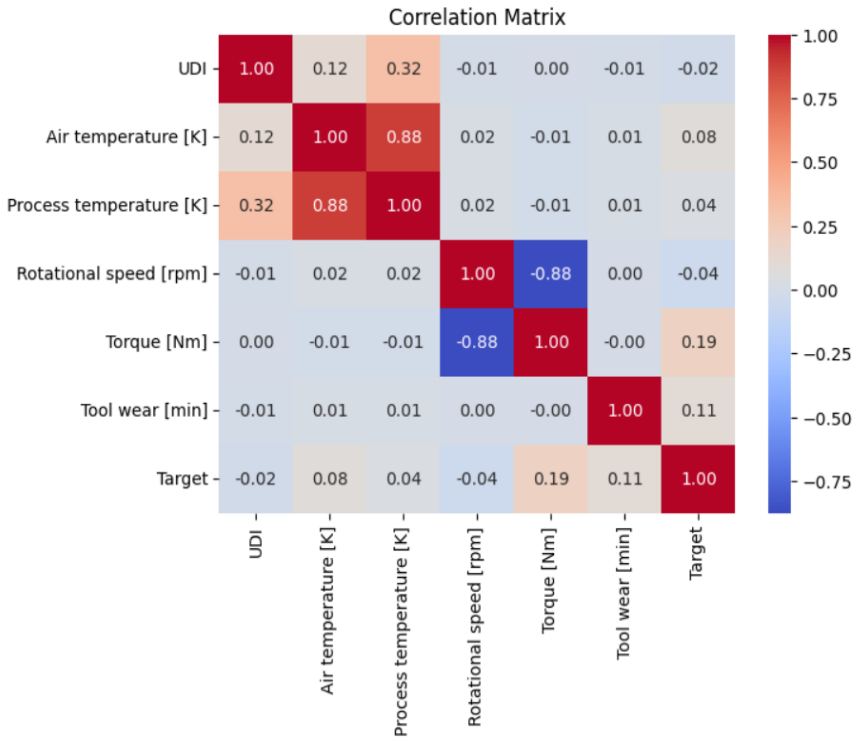

Tool wear [min]
Density
0.004
0.003
0.002
0.001
0.000
0
50
100
150
200
250
Tool wear [min]

Box Plots:


Air temperature [K]
Air temperature [K]
304
302
300
298
296
Process temperature [K]
Process temperature [K]
314
313
312
311
310
309
308
307
306
Rotational speed [rpm]
Rotational speed [rpm]
2750
2500
2250
2000
1750
1500
1250
Torque [Nm]
Torque [Nm]
80
70
60
50
40
30
20
10


Tool wear [min]
Tool wear [min]
250
200
150
100
50
0

Correlation Matrix:



Results and analysis :

1.Random Forest with No Sampling

```
Confusion Matrix:
 [[1928    4]
 [  19   49]]
F1 Score: 0.987809425194787
Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      0.99      1932
           1       0.92      0.72      0.81        68

    accuracy                           0.99      2000
   macro avg       0.96      0.86      0.90      2000
weighted avg       0.99      0.99      0.99      2000
```

2. Random Forest with RandomOverSampler

```
Confusion Matrix:
 [[1925    7]
 [  22   46]]
F1 Score: 0.984629275245601
Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      0.99      1932
           1       0.87      0.68      0.76        68

    accuracy                           0.99      2000
   macro avg       0.93      0.84      0.88      2000
weighted avg       0.98      0.99      0.98      2000
```

3. Random Forest with SMOTE

```
Confusion Matrix:
 [[1890   42]
 [  12   56]]
F1 Score: 0.9753341252333276
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.98      0.99      1932
           1       0.57      0.82      0.67        68

    accuracy                           0.97      2000
   macro avg       0.78      0.90      0.83      2000
weighted avg       0.98      0.97      0.98      2000
```

4. Random Forest with SMOTE TOMEK

```
Confusion Matrix:
 [[1889   43]
 [  13   55]]
F1 Score: 0.9744205743160435
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.98      0.99      1932
           1       0.56      0.81      0.66        68

    accuracy                           0.97      2000
   macro avg       0.78      0.89      0.82      2000
weighted avg       0.98      0.97      0.97      2000
```

5. LightGBM with No Sampling

```
Confusion Matrix:
 [[1931    1]
 [  16   52]]
F1 Score: 0.9909895751439731
Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      1.00      1932
           1       0.98      0.76      0.86        68

    accuracy                           0.99      2000
   macro avg       0.99      0.88      0.93      2000
weighted avg       0.99      0.99      0.99      2000
```

6. LightGBM with RandomOverSampler

```
Confusion Matrix:
 [[1919   13]
 [  15   53]]
F1 Score: 0.9858991437020794
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      1932
           1       0.80      0.78      0.79        68

    accuracy                           0.99      2000
   macro avg       0.90      0.89      0.89      2000
weighted avg       0.99      0.99      0.99      2000
```

7. LightGBM with SMOTE

```
Confusion Matrix:
 [[1895   37]
 [  11   57]]
F1 Score: 0.9778446335861656
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.98      0.99      1932
           1       0.61      0.84      0.70        68

    accuracy                           0.98      2000
   macro avg       0.80      0.91      0.85      2000
weighted avg       0.98      0.98      0.98      2000
```

8. LightGBM with SMOTE TOMEK

```
Confusion Matrix:
 [[1893   39]
 [  12   56]]
F1 Score: 0.9765222446600824
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.98      0.99      1932
           1       0.59      0.82      0.69        68

    accuracy                           0.97      2000
   macro avg       0.79      0.90      0.84      2000
weighted avg       0.98      0.97      0.98      2000
```

9. XGBoost with No Sampling

```
Confusion Matrix:
 [[1926    6]
 [  21   47]]
F1 Score: 0.9856893252286629
Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      0.99      1932
           1       0.89      0.69      0.78        68

    accuracy                           0.99      2000
   macro avg       0.94      0.84      0.88      2000
weighted avg       0.99      0.99      0.99      2000
```

10. XGBoost with RandomOverSampler

```
Confusion Matrix:
 [[1918   14]
 [  15   53]]
F1 Score: 0.9854481721048345
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      1932
           1       0.79      0.78      0.79        68

    accuracy                           0.99      2000
   macro avg       0.89      0.89      0.89      2000
weighted avg       0.99      0.99      0.99      2000
```

11. Decision Tree with No Sampling

```
Confusion Matrix:
 [[1913   19]
 [  17   51]]
F1 Score: 0.9821257739850945
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      1932
           1       0.73      0.75      0.74        68

    accuracy                           0.98      2000
   macro avg       0.86      0.87      0.86      2000
weighted avg       0.98      0.98      0.98      2000
```

12. Decision Tree with RandomOverSampler

```
Confusion Matrix:
 [[1915   17]
 [  23   45]]
F1 Score: 0.9795539654144305
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      1932
           1       0.73      0.66      0.69        68

    accuracy                           0.98      2000
   macro avg       0.86      0.83      0.84      2000
weighted avg       0.98      0.98      0.98      2000
```

13. Decision Tree with SMOTE

```
Confusion Matrix:
 [[1870   62]
 [  14   54]]
F1 Score: 0.9667175280284387
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.97      0.98      1932
           1       0.47      0.79      0.59        68

    accuracy                           0.96      2000
   macro avg       0.73      0.88      0.78      2000
weighted avg       0.97      0.96      0.97      2000
```

14. Decision Tree with SMOTE TOMEK

```
Confusion Matrix:
 [[1858   74]
 [  15   53]]
F1 Score: 0.9618870447117491
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.96      0.98      1932
           1       0.42      0.78      0.54        68

    accuracy                           0.96      2000
   macro avg       0.70      0.87      0.76      2000
weighted avg       0.97      0.96      0.96      2000
```

- **Based on the provided performance metrics, the LightGBM model with no sampling demonstrates the highest effectiveness, achieving an F1 score of 0.9909 and high precision and recall for both classes. Random Forest models with no sampling and RandomOverSampler also perform well, though slightly less so in minority class detection. Decision Tree models, while generally effective, show a notable drop in minority class performance, particularly with SMOTE and SMOTE TOMEK, indicating challenges in balancing precision and recall for the minority class.**

Conclusion

The equipment failure prediction models show that LightGBM without sampling delivers the best overall performance, achieving a high F1 score and balanced precision and recall. Random Forest models also perform robustly, with no significant advantage gained from sampling techniques. Users gain a reliable method for predicting equipment failures, enhancing maintenance schedules and minimizing downtime through data-driven insights.

Drawback: The models show reduced performance in detecting minority class failures, leading to potential missed failure predictions. Additionally, some sampling techniques introduce imbalances that impact overall model accuracy.

Recommendation: Future work should explore more advanced imbalance handling techniques, such as ensemble methods or anomaly detection algorithms. Enhancing feature engineering and incorporating real-time data could further improve prediction accuracy and reliability.