Sandia ML Team 2 Final Report

Turbofan Degradation Analysis

I. Introduction

i. Background Information

Renowned for their efficiency and reliability, turbofan engines are the primary propulsion system for both commercial and military aircraft. A typical turbofan engine configuration includes several key components: a fan and low-pressure compressor (LPC) module, a high-pressure compressor (HPC), a combustor, a high-pressure turbine (HPT), a low-pressure turbine (LPT), and an exit nozzle. These components work together to generate thrust and power the aircraft.

In commercial aircraft, the thrust primarily comes from the fan airflow that bypasses the engine core, known as bypass flow. In contrast, military fighter jets generate most of their thrust from the core flow. The fan, LPC, and LPT usually rotate on a low spool shaft. In some advanced configurations, a gearbox separates the low spool from the fan, allowing each to rotate at optimal speeds independently. The HPC and HPT rotate on a separate, counter-rotating shaft called the high spool.

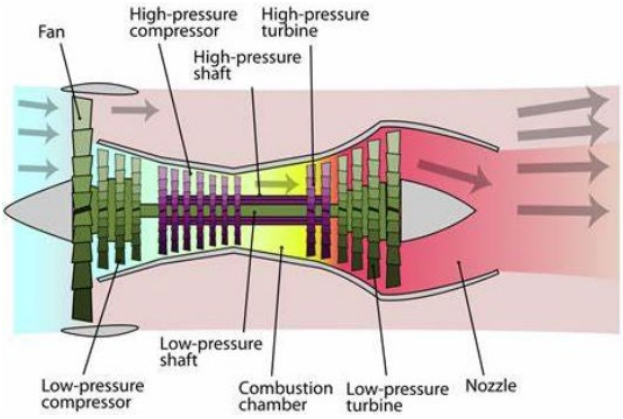

Figure 1: Turbofan Engine Configuration

ii. Problem Statement/Objective

In extreme environments, turbofan engine components gradually degrade over time. Accurately predicting when an engine will no longer meet its original design specifications is crucial for effective maintenance planning and operational efficiency. This analysis uses simulated engine sensor data to predict the remaining useful life (RUL) of a fleet of engines with various fault modes. This study seeks to answer two main research questions:

- ***Can the remaining useful life be accurately predicted using information from the sensor channels? (Regression Problem)***
- ***Given the sensor information, can we predict whether an engine is faulty? (Classification Problem)***

iii. Description of Engine Features

Understanding the various sensor measurements is crucial for predicting engine degradation. Below are the descriptions of key engine features measured by different sensors:

Feature	Description	Units
T2	Measures the temperature of the air entering the fan	°R
T24	Measures the temperature of the air exiting the Low-Pressure Compressor (LPC)	°R
T30	Measures the temperature of the air exiting the High-Pressure Compressor (HPC)	°R
T50	Measures the temperature of the air exiting the Low-Pressure Turbine (LPT)	°R
P2	Measures the pressure of the air entering the fan	psia
P15	Measures the pressure within the bypass duct	psia
P30	Measures the pressure of the air exiting the High-Pressure Compressor (HPC)	psia
Nf	Measures the rotational speed of the fan	rpm
Nc	Measures the rotational speed of the core	rpm
epr	Ratio of the pressure at the outlet of the engine to the pressure at the inlet (P50/P2)	-
Ps30	Measures the static pressure of the air exiting the High-Pressure Compressor (HPC)	psia
phi	Ratio of the fuel flow to the static pressure at the HPC outlet	pps/psi
NRf	Measures the fan speed corrected for standard conditions	rpm
NRc	Measures the core speed corrected for standard conditions	rpm
BPR	Ratio of the mass of air bypassing the engine core to the mass of air passing through the engine core	-
farB	Ratio of the fuel mass flow rate to the air mass flow rate in the burner	-
htBleed	Measures the enthalpy of the bleed air taken from the engine	-
Nf_dmd	The desired or commanded fan speed	rpm
PCNfR_dmd	The desired or commanded fan speed corrected for standard conditions	rpm
W31	Measures the mass flow rate of the coolant bleed air from the High-Pressure Turbine (HPT)	lbm/s
W32	Measures the mass flow rate of the coolant bleed air from the Low-Pressure Turbine (LPT)	lbm/s

Table 1: Sensor Descriptions

II. Data Exploration and Feature Selection for FD001

i. Feature Selection Process

A comprehensive feature selection process was conducted to develop a robust predictive model for engine health management. This process involved multiple analyses to ensure that only the most relevant and impactful features were included in the model. The initial step involved examining the feature distribution to identify features with normal variability, which is crucial for effective predictive modeling.

Next, a correlation heatmap was generated to reveal the relationships between sensor measurements and operational settings, highlighting clusters of highly correlated features. Following this, the correlation of individual features with RUL was assessed to pinpoint which features are directly associated with engine degradation.

Finally, the VIF (Variance Inflation Factor) test was employed to detect multicollinearity among features, ensuring that highly collinear features did not adversely affect the model's performance. This rigorous selection process identified and eliminated irrelevant and redundant features, streamlining the model, and enhancing its predictive power.

ii. Feature Distribution

Almost all features that exhibited some degree of correlation with RUL showed a normal distribution pattern. These features included settings and sensor measurements such as T24, T30, T50, P30, Nf, Nc, Ps30, phi, NRf, NRc, BPR, htBleed, W31, and W32. On the other hand, features like Setting_3, T2, P2, epr, farB, Nf_dmd, and PCNfR_dmd showed no correlation with RUL, as they all had constant values. This lack of variability suggested that these features were uninformative for predicting RUL. Hence, they were excluded from the analysis.

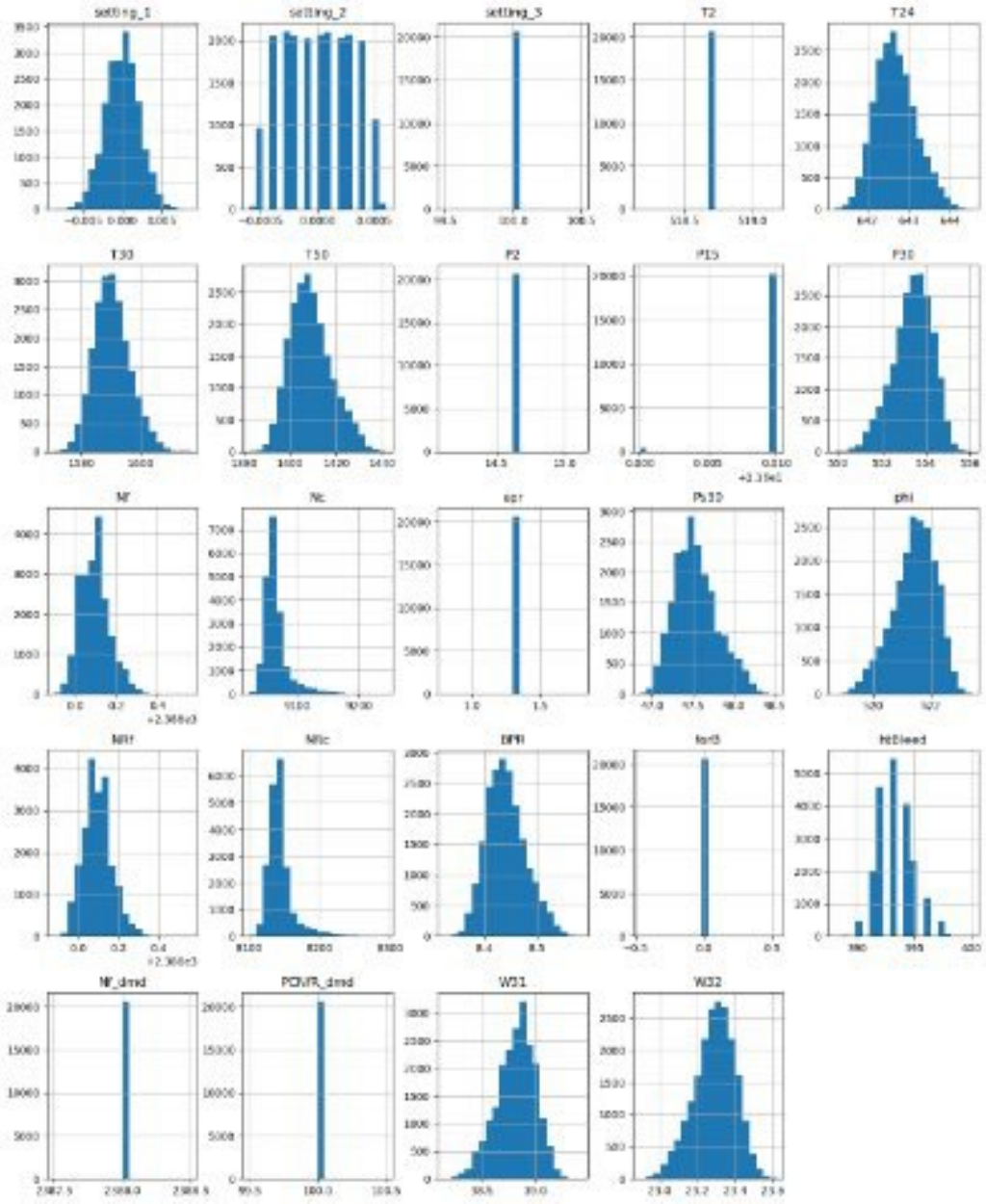

Figure 2: FD001 Sensor Data Distributions

iii. Correlation Heatmap and Analysis

The correlation heatmap revealed strong correlations between certain sensors and operational settings. Sensors like T2, T24, T30, T50, P2, P15, and P30 showed significant correlations. However, operational settings (Setting_1, Setting_2, Setting_3) had minimal correlation with these sensor measurements. This minimal correlation suggested that operational settings had a distinct influence on engine performance, which was crucial for isolating specific fault modes during analysis.

Furthermore, significant negative correlations with RUL were observed in sensors such as T24, T30, T50, P30, Nf, and Nc. This indicated that these sensors were critical for predicting the RUL of the engine. Clusters of high correlations were particularly noticeable among temperature sensors (T-series) and pressure sensors (P-series). EPR (Engine Pressure Ratio) and Ps30 are also strongly correlated with multiple sensors, especially those in the P-series. The limited correlation between operational settings and most sensors ensured clarity in analyzing sensor data for specific fault modes. The heatmap identified critical sensors essential for predictive engine health and RUL modeling.

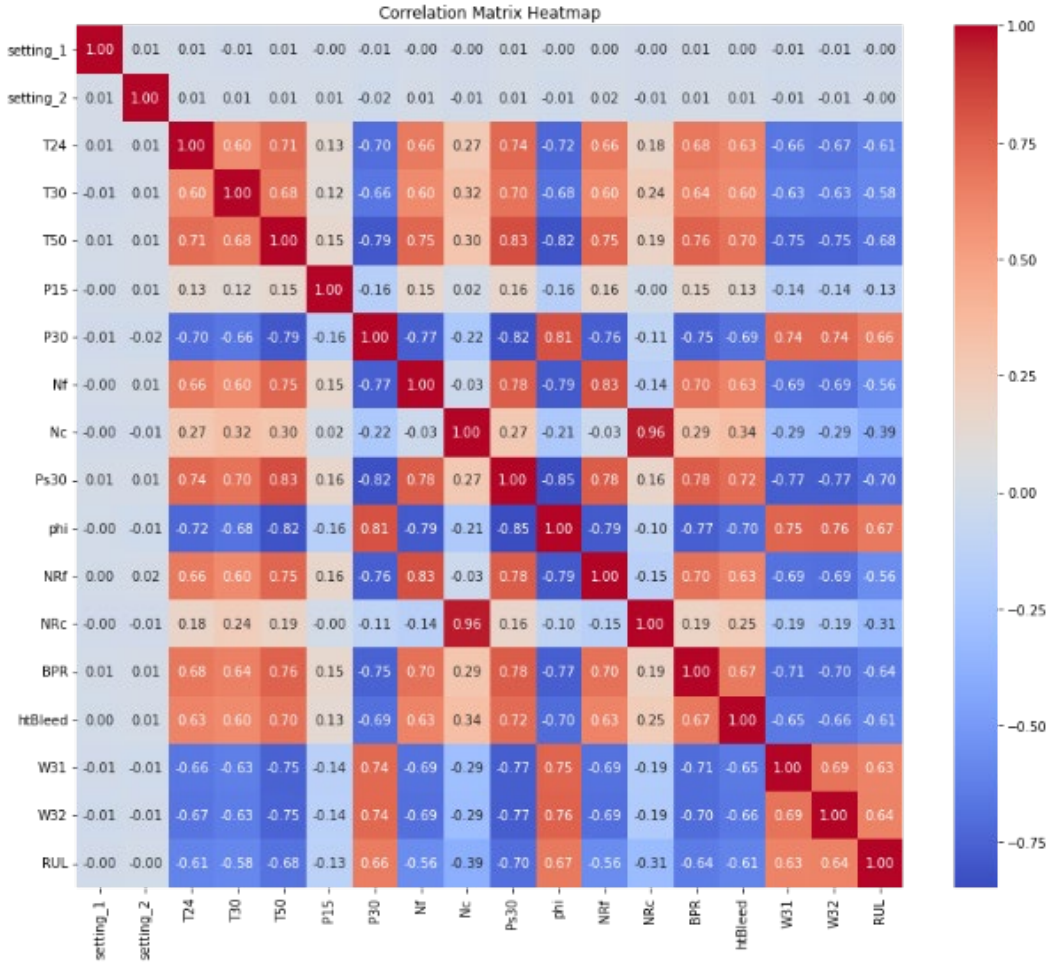

Figure 3: FD001 Correlation Matrix Heatmap

iv. Feature Correlation with RUL

The analysis of feature correlations with RUL revealed varying degrees of correlation across different sensor types.

Operational Settings:

Setting_1 and Setting_2 exhibited exceptionally low negative correlations with RUL, with correlation coefficients of -0.003198 and -0.001948, respectively. This minimal correlation suggested that these operational settings did not significantly influence RUL predictions.

Temperature Sensors:

Temperature sensors showed stronger negative correlations with RUL. T24 had a strong negative correlation (-0.606484), T30 had a moderate negative correlation (-0.584520), and T50 showed a strong negative correlation (-0.678948). These strong negative correlations indicated that as the temperature readings increased, the RUL decreased, making these sensors critical for predicting engine degradation.

Pressure Sensors:

Among the pressure sensors, P15 showed a weak negative correlation with RUL (-0.128348), while P30 displayed a strong positive correlation (0.657223). Ps30, like the temperature sensors, had a strong negative correlation (-0.696228). The strong correlations of P30 and Ps30 highlighted their importance in RUL prediction models.

Speed Sensors:

Speed sensors also displayed moderate to weak negative correlations with RUL. Nf had a moderate negative correlation (-0.563968), Nc a moderate negative correlation (-0.390102), NRf a moderate negative correlation (-0.562569), and NRc a weak negative correlation (-0.306769). These correlations suggested that higher speeds were associated with a shorter RUL.

Ratio and Flow Sensors:

BPR had a strong negative correlation with RUL (-0.642667) for ratio and flow sensors, whereas phi showed a strong positive correlation (0.671983). These sensors were important for understanding the fuel efficiency and airflow dynamics affecting the engine's RUL.

Bleed and Demand Sensors:

The htBleed sensor exhibited a strong negative correlation with RUL (-0.606154), indicating its relevance in predicting engine health.

Other Sensors:

Other sensors, such as W31 and W32, showed moderate positive correlations with RUL, with values of 0.629428 and 0.635662, respectively. These correlations suggested that these sensors could also contribute valuable information for RUL predictions.

Certain predictors had constant values in the dataset, making it impossible to assess their correlation with RUL. These included Setting_3, T2, P2, epr, farB, Nf_dmd, and PCNfR_dmd. Due to their lack of variability, these features did not provide meaningful information for predicting engine degradation and were thus excluded from further analysis.

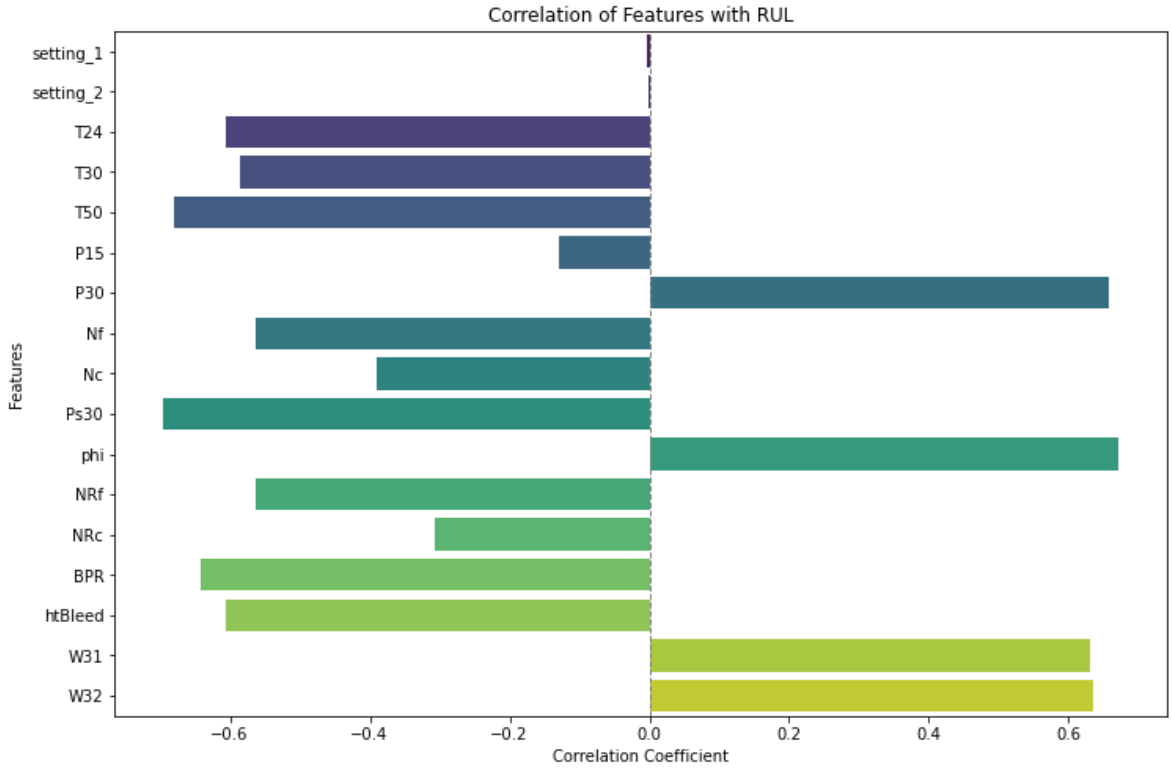

Figure 4: FD001 Sensor Correlation with Remaining Useful Life

v. VIF Analysis

The VIF (Variance Inflation Factor) analysis revealed high collinearity between certain sensor measurements. For example, high collinearity was observed between Nc (Physical core speed) and NRc (Corrected core speed), with VIF values of 17.844748 and 17.255307, respectively. Both sensors measured core speed, but NRc was adjusted for standard conditions, resulting in inherently correlated values.

The high collinearity observed in this case was understandable due to the physical and operational relationships between the parameters measured. Sensors that measure similar or related physical quantities naturally exhibit collinearity. Given the nature of these correlations, no feature reduction due to collinearity was necessary. This collinearity was expected and did not adversely impact the predictive modeling process, allowing the model to effectively utilize the available sensor data.

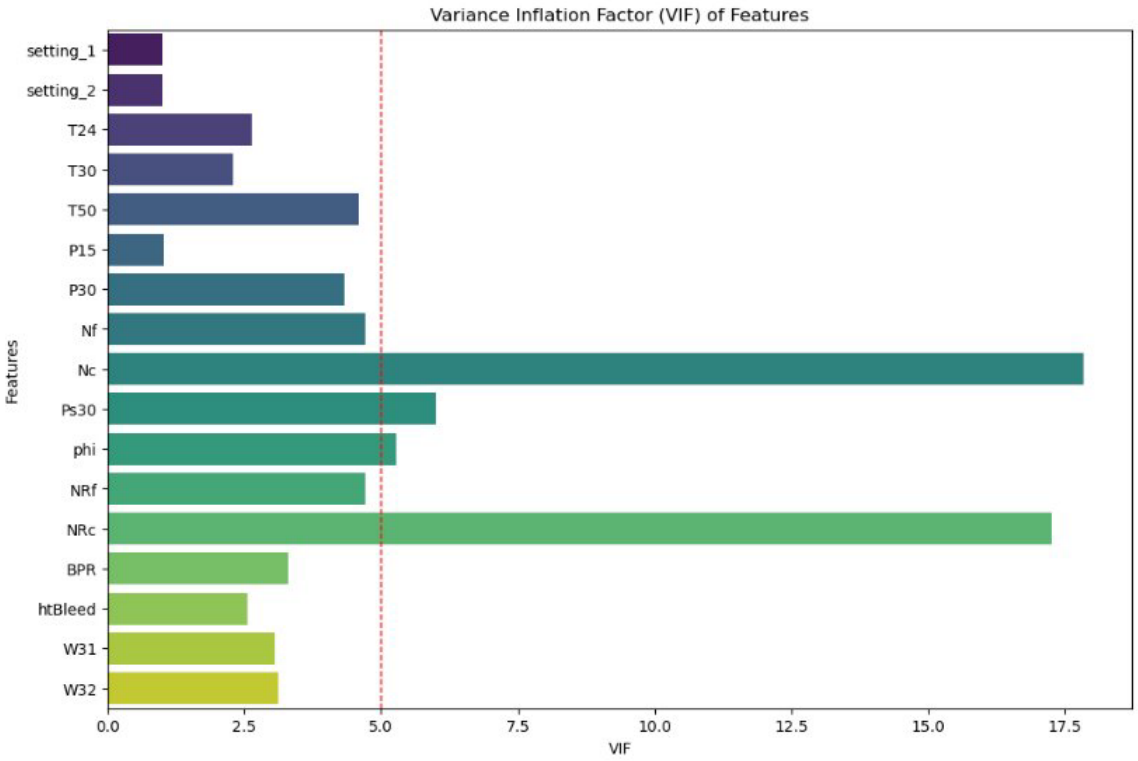

Figure 5: FD001 Sensor VIF

vi. Feature Elimination

The feature elimination process identified specific features that did not contribute meaningful information for predicting RUL and were removed to enhance model performance.

Features to be Eliminated:

Certain features were found to have no correlation with RUL and thus did not provide valuable data for predictive modeling. These features included Setting_3, T2 (Total Temperature at Fan Inlet), P2 (Pressure at Fan Inlet), epr (Engine Pressure Ratio), farB (Burner Fuel-Air Ratio), Nf_dmd (Demanded Fan Speed), and PCNfR_dmd (Demanded Corrected Fan Speed).

Benefits of Feature Elimination:

- Features unrelated to RUL did not offer meaningful insights for predicting engine degradation.
- These features remained constant and showed no variation with engine health, rendering them irrelevant to our predictive model.
- Removing irrelevant features led to improved model accuracy and efficiency.

- This process allowed the model to concentrate on the most relevant and impactful features, enhancing its predictive capabilities.

Conclusion:

Purposeful feature selection is essential for developing an effective predictive model. The team streamlined the model by eliminating features unrelated to RUL and focused on the most predictive variables. This approach aimed to enhance the overall performance and reliability of the predictive model for engine health management, ensuring accurate and actionable insights for maintenance and operational decisions.

III. Predicting RUL of an Engine (FD001) (Regression)

i. Performance Evaluation in Engine Degradation Prediction

A comprehensive performance evaluation process was implemented to ensure the robustness and reliability of the predictive models. This process thoroughly assessed whether the models met task-specific requirements and effectively predicted engine degradation, instilling confidence in the model's reliability. All models were trained in the training set, and their performance was evaluated in the test set. As the Data Description chapter mentioned, NASA provided both test and training sets.

Overview: The performance evaluation was based on a formula created by Saxena, Goebel, Simon, and Eklund [2], which evaluated if the model's prognosis met the necessary criteria. In Prognostics and Health Management (PHM), early predictions are preferred to avoid failures. However, early predictions can sometimes lead to economic burdens. This dual consideration was incorporated into the scoring function to evaluate each model.

Scoring Algorithm: For engine degradation scenarios, NASA provided the scoring system, along with the specific scoring equation, alongside the data. This system preferred early predictions over late predictions and employed an asymmetric penalty, penalizing late predictions more heavily than early ones. This approach was crucial for ensuring timely maintenance and avoiding unexpected engine failures. In this NASA-provided scoring system, a lower score indicates better predictive performance, underlining the significance of accurate and timely predictions.

Scoring Function: The scoring function was formulated as follows:

$$s = \begin{cases} \sum_{i=1}^{n} e^{-\left(\frac{d}{10}\right)} - 1 & d < 0 \\ \sum_{i=1}^{n} e^{\left(\frac{d}{13}\right)} - 1 & d \geq 0 \end{cases}$$

- s: Computed score
- n: Number of Units Under Test (UUTs)
- d: Error (Estimated RUL - True RUL)

Performance Metrics: The following metrics were used to evaluate the models:

- **Mean Absolute Error (MAE):** Measured the average absolute error between predicted and actual values.
- **R-Squared (R^2):** Indicated the proportion of variance explained by the model.
- **Mean Squared Error (MSE):** Measured the average squared differences between predicted and actual values.
- **Exponential Penalty Growth:** The penalty for prediction errors increased exponentially with the magnitude of the error, emphasizing the importance of accurate predictions.

Using these metrics and the scoring function, the performance of each predictive model was meticulously and comprehensively evaluated. This rigorous process instills confidence in selecting the most reliable and accurate engine health management model, enhancing the models' predictive capabilities, and providing valuable insights for maintenance and operational planning.

Initial Hypotheses:

The team hypothesized that the model would achieve the best results for the FD001 dataset due to its simplicity and lower complexity. FD001 involved only one failure mode and operating condition, allowing the model to adjust more effectively and accurately predict the engines' RUL.

In contrast, FD002, with its six operating conditions, introduced significantly more variability into the data. Different atmospheric temperatures and other environmental factors affected the engine performance and complicated the predictions. Since the operating condition was not a parameter we could account for, we had to treat all data equally, which might have reduced the model's accuracy for FD002.

FD003, while having only one operating condition at sea level like FD001, introduced another failure mode. This added complexity might have impacted the model's performance. We anticipated that FD003 might perform better than FD002 due to the reduced variability from a single operating condition but might not achieve the same level of accuracy as FD001 due to the added failure mode.

Finally, FD004, with multiple operating conditions and fault modes, was expected to present the greatest challenge to the model. Its high variability and complexity would probably result in the lowest predictive performance among the datasets.

ii. Initial Models Used

The team explored various approaches for our initial predictive modeling to determine which would best predict engines' RUL. The models assessed include a Neural Network, K-Nearest Neighbors (KNN), Ridge Regression, XGBoost, Lasso Regression, and linear regression.

The Neural Network, inspired by the human brain's neural networks, excels at capturing complex non-linear relationships within the data. KNN, a simple, instance-based learning model, predicts values based on the average of the k-nearest neighbors, making it a practical and applicable choice for capturing local patterns. Ridge Regression, which includes a regularization term to penalize significant coefficients, is particularly useful for managing multicollinearity and preventing overfitting. XGBoost, an ensemble learning method, combines multiple models to enhance predictive accuracy and effectively manage bias and variance. Lasso Regression, like Ridge Regression but with L1 regularization, can shrink some coefficients to zero, performing variable selection and regularization to improve model interpretability. Lastly, linear regression, a basic predictive modeling technique assuming a linear relationship between variables, is a baseline model that can be compared against more complex methods.

Each model brings unique strengths, from managing simple linear relationships to capturing intricate non-linear interactions, providing a comprehensive foundation for our predictive modeling efforts.

The models were trained on the training dataset's relevant sensor readings for each cycle and engine. A key limitation noticed early on was that RUL was only provided for the final row of each test, so the team manually calculated the RUL for each test row to enable learning. An additional column was added to track the RUL for each engine, acting as the response variable. Since each engine in the training datasets is run to failure, the RUL is simply the difference between the cycle at which failure occurs and the current cycle. Only the last row of data for each engine in the test dataset was used to predict the RUL.

Table 2 shows the results for the initial models.

Initial Model Results

Model	Adjusted R^2	MAE	MSE
Lasso	0.4037	25.6551	1029.8136
Ridge	0.4083	25.5543	1021.8608
Neural Network	**0.5246**	**21.2**	**820.9232**
KNN (k=35)	0.4871	21.5029	885.6284
XGBoost	0.4053	22.9047	1026.8844
Linear Regression	0.4055	25.5918	1026.6318

Table 2: Performance of Initial Regression Models

The Neural Network demonstrated the best overall performance, achieving the highest R^2 and the lowest error metrics, indicating superior predictive accuracy and generalization ability. The KNN model performed well, with a substantial R^2 value and low error metrics, making it a viable option for RUL prediction.

Ridge Regression showed decent performance, which is particularly useful for handling multicollinearity in the dataset. XGBoost provided a balance between complexity and performance, showing competitive error metrics. Lasso Regression performed similarly to Ridge Regression, with slightly higher error metrics.

Lastly, linear regression served as a baseline model, with higher error metrics than the more advanced models. Overall, the Neural Network stood out as the most effective model, highlighting its capability to predict engine degradation accurately.

Scoring and Performance of Best Initial Model

The performance of the Neural Network model was evaluated using several key metrics. It achieved a Mean Absolute Error (MAE) of 21.2001, indicating the average absolute difference between the predicted and actual values. The Mean Squared Error (MSE) was 820.9232, reflecting the average squared differences and highlighting the model's ability to manage more significant errors effectively. Additionally, the model attained an R^2 score of 0.5246, demonstrating its ability to explain over 52% of the variance in the data. However, according to the weighted scoring system, the Neural Network scored 1960, indicating significant room for improvement, as a "good" score is typically below 1000.

Following the preliminary analysis, the team decided to use a Neural Network for all additional tests, as it demonstrated the best potential for performance in the preliminary analysis. The chart below highlights the general accuracy of the Neural Network model and shows how the model can be improved.

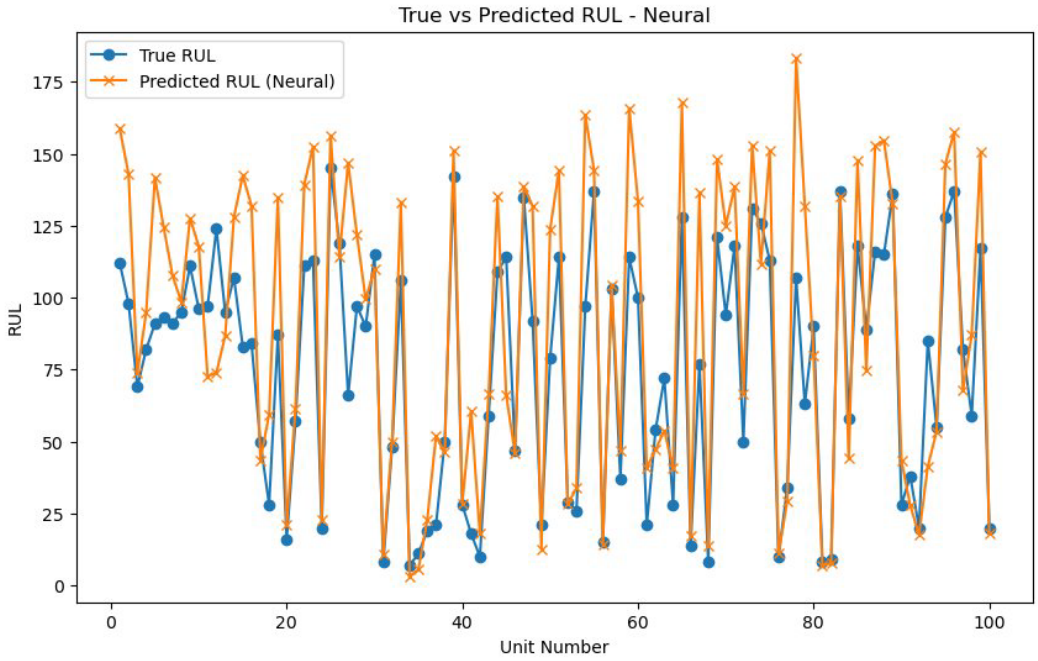

Figure 6: FD001 Neural Network Remaining Useful Life Predictions

iii. Improving the Regression Models: Adding Indicators

To enhance the predictive accuracy of the models, the team sought to augment the existing data by applying indicators that track strength and variability of the sensors. By incorporating technical indicators commonly used in stock trading, the team aimed to provide additional insights and dimensions to the dataset. The indicators applied to the data included the Commodity Channel Index (CCI), Relative Strength Index (RSI), Bollinger Bands Percentage (BB%), and Stochastic Oscillators.

Commodity Channel Index (CCI): The CCI is calculated by taking the difference between the current value of a data point and its moving average, divided by the mean absolute deviation of the data points. The formula is:

$$CCI = \frac{X - MA}{0.015 * MAD}$$

X is the current data value, MA is the moving average of the data values, and MAD is the mean absolute deviation. The constant 0.015 ensures that most CCI values fall within the range of -100 to +100.

Relative Strength Index (RSI): The RSI is calculated by taking the average of the gains and losses of the data points over a specified period and then creating an index from these averages. The formula is:

$$RSI = 100 - \frac{100}{1 - \frac{AG}{AL}}$$

Where the average gain (AG) is the sum of all gains over the past n periods divided by n, and the average loss (AL) is the sum of all losses over the past n periods divided by n.

Bollinger Bands Percentage (BB%): BB%, derived from Bollinger Bands, are calculated by taking a moving average and adding/subtracting a multiple of the standard deviation. BB% is then calculated as:

$$BB\% = \frac{X - LB}{UB - LB}$$

Where X is the current data value, and the upper and lower bands (UB, LB) are defined as the moving average plus/minus k times the standard deviation, respectively.

Stochastic Oscillators: Stochastic Oscillators are calculated by comparing the current data value to the range of values over a certain period. The formula is:

$$K\% = 100 * \frac{X - L}{H - L}$$

Where X is the current data value, L is the lowest value over the past n periods, H is the highest value over the past n periods, and D% is the simple moving average of K%.

The team enriched the feature set by adopting these trading indicators to the engine data, enhancing the predictive models. This innovative approach allowed for a more comprehensive analysis, improving the prediction of engines' RUL.

IV. Updated Results

As discussed in the previous section, the FD001 dataset was augmented with technical indicators for relevant sensors. Adding these indicators led to significant improvements in the model's accuracy. The R^2 increased from 0.5246 to 0.786, the MAE decreased from 21.2 to 14.25, the

MSE decreased from 820.92 to 364.04, and the score decreased from 1960 to 583. These improvements highlight the effectiveness of incorporating technical indicators into the predictive model.

Despite these improvements, there are still a few engines where the model significantly overpredicts RUL, indicating room for further model tuning and feature engineering. Most enhancements were attributed to adding the RSI indicator, which improved the R^2 to 0.73-0.75 and reduced the score to 1000-1200. This suggests that other engineered features could offer an even better complexity-accuracy tradeoff, warranting further investigation.

The figure below shows the predicted and actual RUL for the updated model. Like the baseline model, the model demonstrates exceptionally high accuracy for engines with little remaining life. However, the added indicators significantly enhance the model's ability to accurately predict RUL for engines with moderate to high remaining life. This enhancement is evident in the overall improvement of all model evaluation metrics.

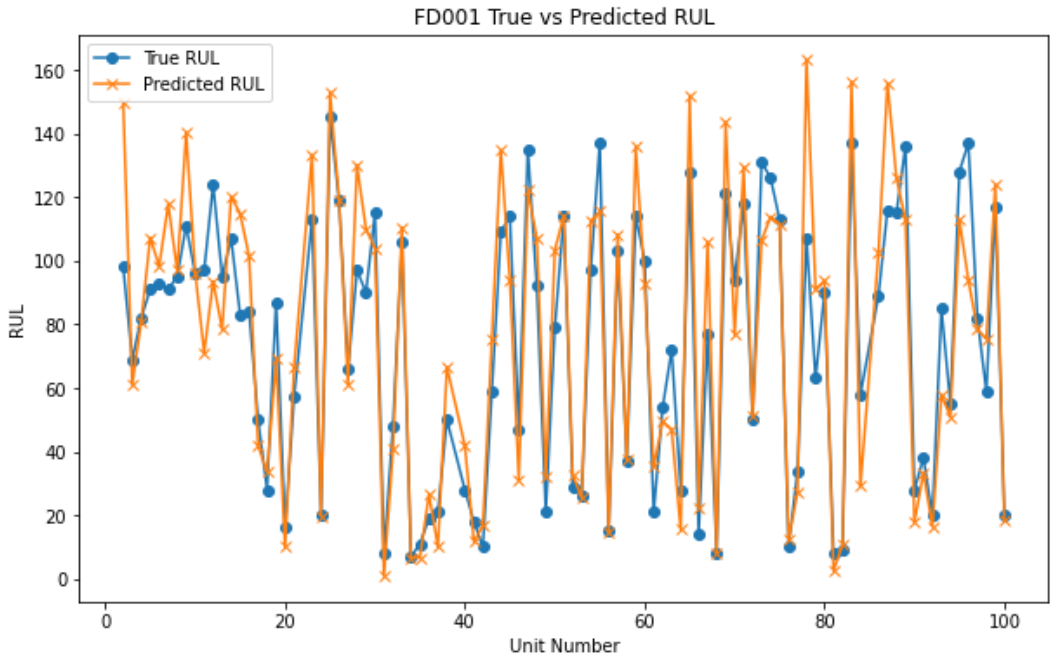

Figure 7: FD001 RUL Predictions with Indicators

IV. Predicting RUL of an Engine (FD002, FD003, FD004)

The same model and approach were subsequently applied to the remaining datasets and evaluated as previously described. Slightly different feature selection methods were employed for each dataset. Compared to FD001, all three datasets exhibited higher levels of multicollinearity, indicating that the initial feature set chosen for FD001 was not as effective for the other datasets. Initially, features were selected based on the Variance Inflation Factor (VIF). However, experimentation identified a few key features with predictive power across all datasets: Nf, Nc, NRf, NRc, T24, T30, P30, W32, phi, BPR, and htBleed.

As expected, FD002 and FD004 exhibit significantly more variability and noise compared to FD001, primarily due to having six operational settings instead of one. FD003 is more like FD001 but includes the added complexity of an additional fault mode.

i. FD002 Output:

As mentioned above, FD002 has significant multicollinearity, making the removal of variables purely based on VIF ineffective within the framework of the neural network (NN) model. Additionally, unlike FD001, none of the variables in FD002 exhibit a significant correlation with remaining life. From Figure 8, the highest correlation factor was -.07, and almost all other factors hover around 0.0. Therefore, an iterative approach was used to add variables for this dataset, rather than eliminating variables based on VIF and including variables based on their correlation with the response. The variables Nf, Nc, NRf, NRc, T24, T30, P30, W32, phi, BPR, and htBleed resulted in the best performance.

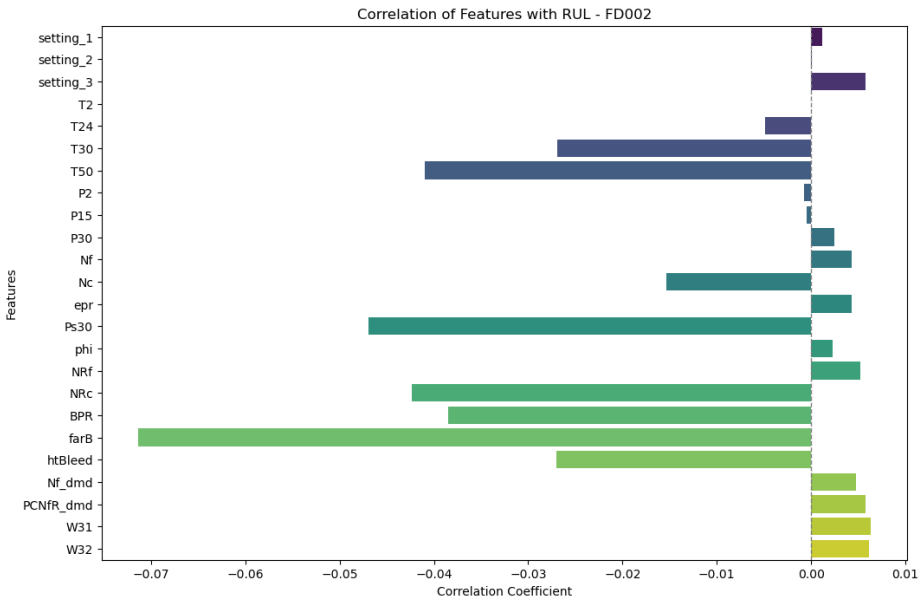

Figure 8: FD002 Sensor Correlation with Remaining Useful Life

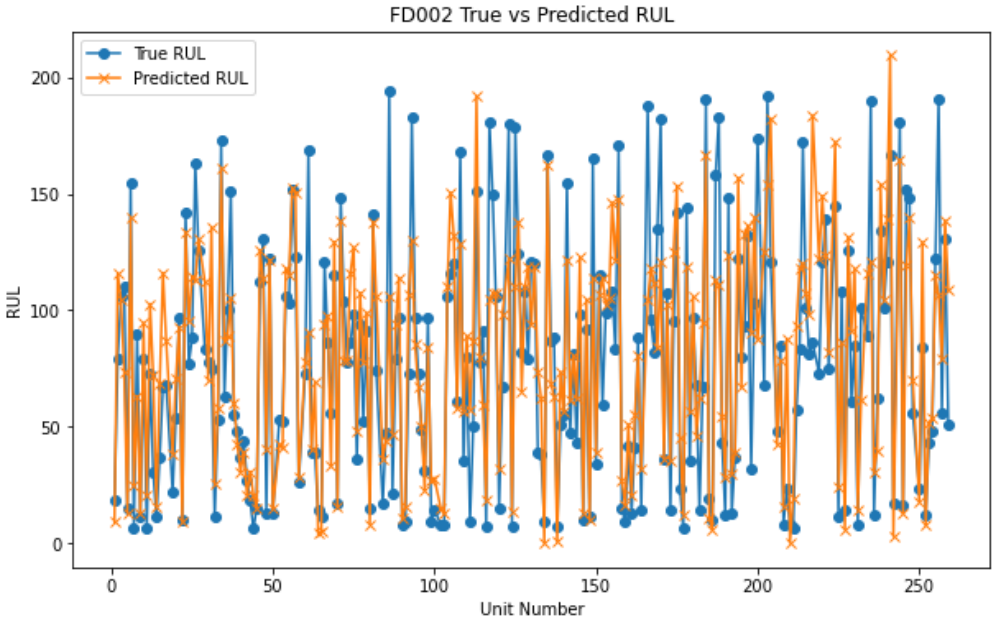

Figure 9: FD002 True vs. Predicted RUL

The model achieved an R^2 of 0.689, a Mean Absolute Error (MAE) of 22.55, a Mean Squared Error (MSE) of 880, and a score of 25,616. Like FD001, the model tends to fit engines with little remaining life more accurately than those with higher life. There is a relatively small subset of engines where the discrepancy between predicted and true Remaining Useful Life (RUL) is significant; however, in this case, the predicted life is much lower than the actual life. These significant discrepancies are primarily responsible for the substantial increase in the score despite a comparable R^2.

ii. FD003 Output:

FD003, which is like FD001 but with an additional failure mode, follows the same general trends as FD001. Additionally, FD003 has strong, but mostly negative, correlations between sensors and RUL like those of FD001, which can be seen in Figure 10 below. This led to the selection of the same features for FD003 as were selected for FD001: setting_1, T24, T30, T50, P30, Nf, Nc, Ps30, phi, NRf, NRc, BPR, htBleed, W31, and W32

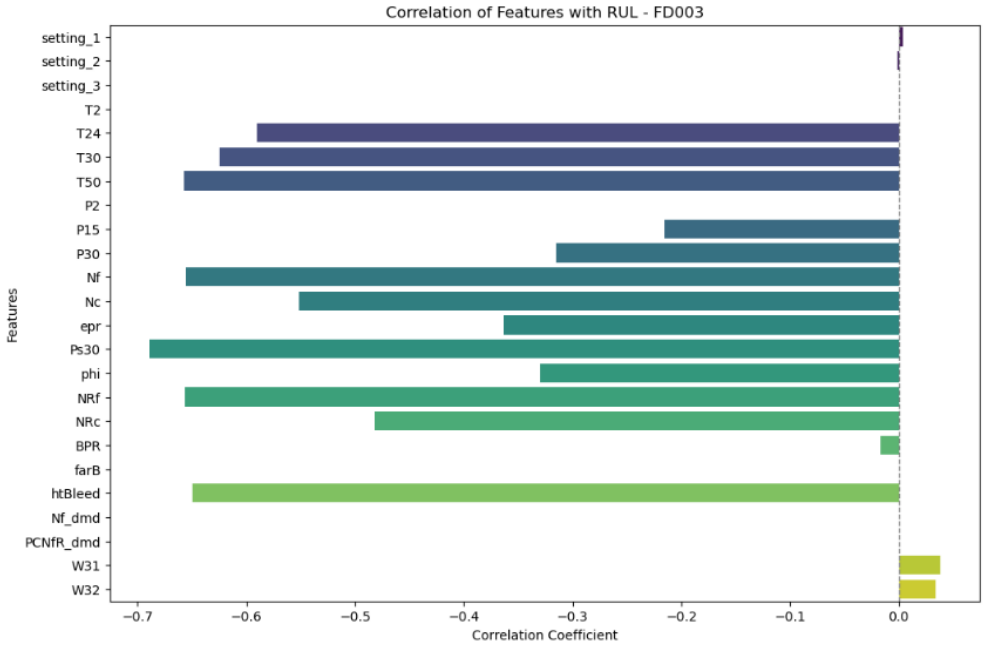

Figure 10: FD003 Sensor Correlation with Remaining Useful Life

.

The team observed the same general trends in predictions for FD001 and FD003, as shown in Figure 11. The model accurately predicts lower RUL engines but tends to overshoot for higher RUL.

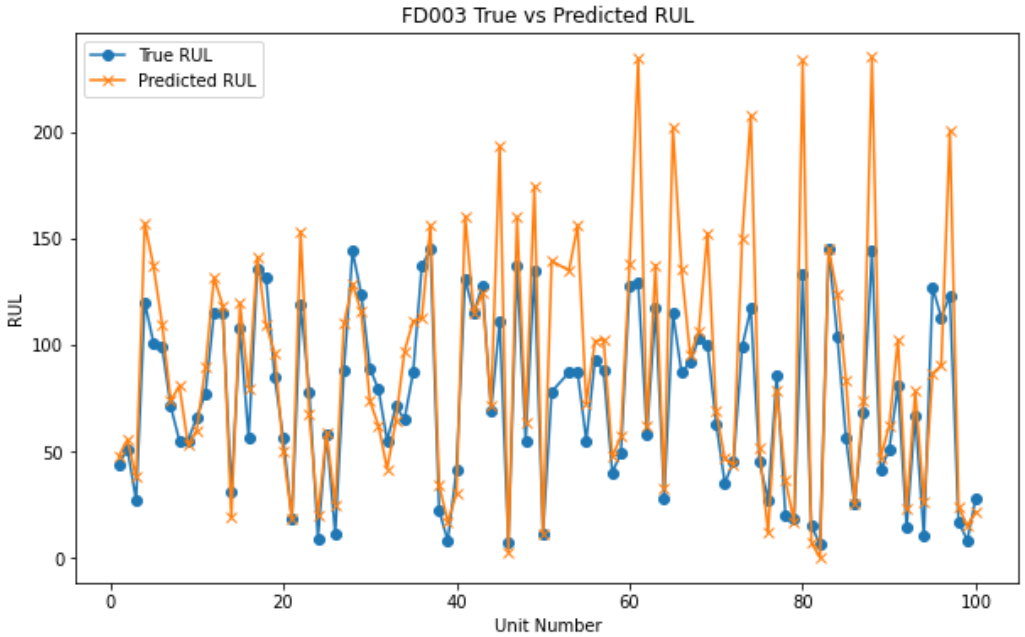

Figure 11: FD003 True vs. Predicted RUL

With the additional failure mode introduced by FD003, the model performance drops in every metric. R^2 decreased from 0.786 to 0.56, MAE increased from 14.25 to 18.18, MSE increased from 364 to 752, and most notably, the score increased from 583 to 4677, driven by the more significant errors on the higher RUL engines.

iii. FD004 Output:

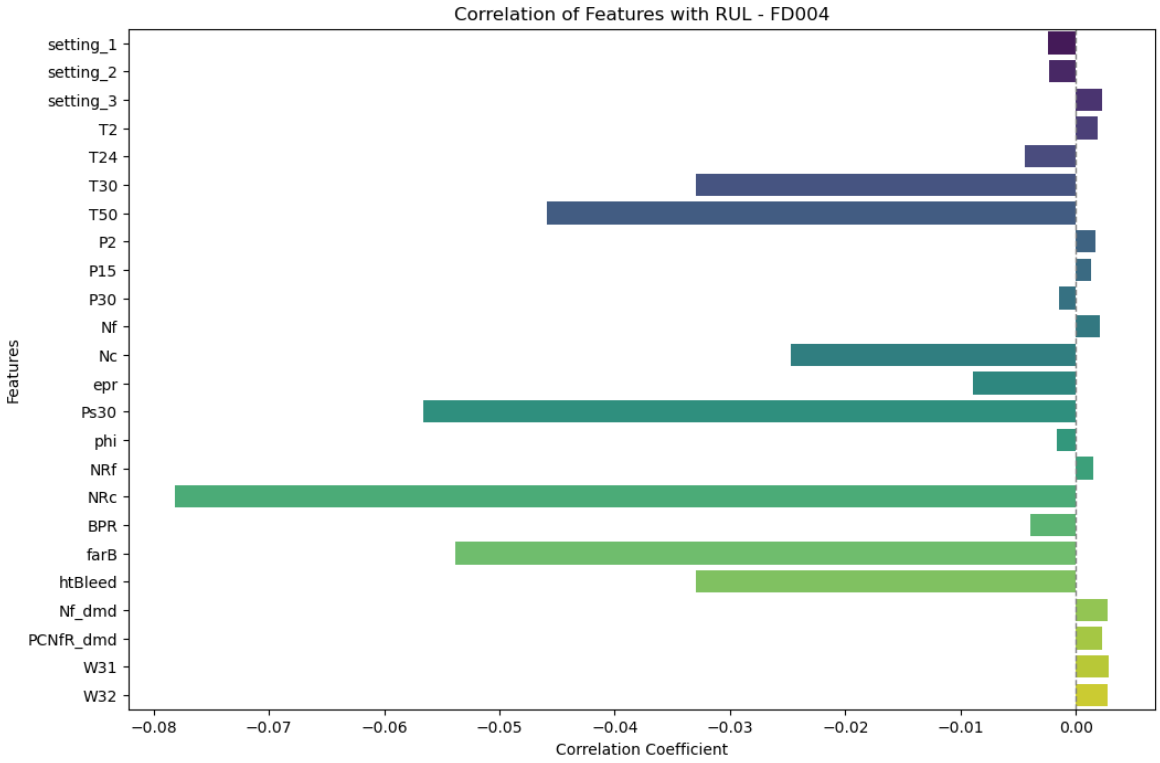

Figure 12: FD004 Sensor Correlation with Remaining Useful Life

FD004 is the most complex of the datasets, representing a mix of two fault modes and six operational settings. This complexity is reflected in the sensor data's correlation to RUL, the presence of noise, and the sensors' high multicollinearity. Due to these similarities, the same features used for FD002 were also applied to FD004.

Figure 13: FD004 Sensor Data Distributions

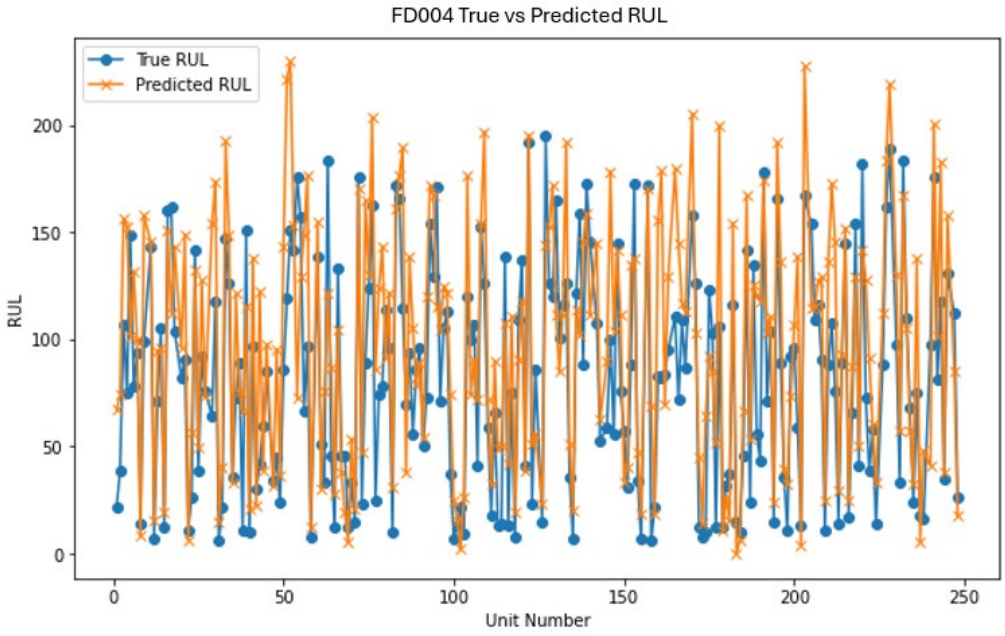

Figure 14: FD004 True vs. Predicted RUL

The model unsurprisingly performed the worst on this dataset. The R^2 was 0.461, the MAE was 30.9, the MSE was 1496, and the score was 36,608. These performance metrics indicate significant challenges in accurately predicting the Remaining Useful Life (RUL) for FD004.

Data Trends and Challenges

Lack of Normal Distribution

One of the primary challenges with FD004 is the lack of a normal data distribution. As seen in the feature distribution chart, none of the features display the normal distribution in the FD001 dataset. The sensor readings exhibit high variability and noise, complicating the modeling process. Unlike datasets with a normal distribution, where data points are symmetrically distributed around the mean, FD004's data is highly skewed and irregular. This irregularity makes it difficult for the model to learn and generalize patterns effectively.

High Variability and Noise

The presence of six different operational settings introduces significant variability in the dataset. Each operational setting affects the engine differently, leading to a wide range of sensor readings and degradation patterns. This variability is further exacerbated by the two fault modes present in FD004, adding layers of complexity to the data. The noise in the sensor data, caused by external factors and inherent measurement errors, further hampers the model's ability to make accurate predictions.

Multicollinearity among the sensors is another significant issue in FD004. High multicollinearity means that several sensor readings are highly correlated, making it challenging to isolate the individual impact of each sensor on the RUL prediction. This interdependence of sensor readings can lead to redundant information, confusing the model and reducing its predictive power.

The complexity of FD004, with its mix of fault modes and operational settings, presents significant challenges for predictive modeling. The data's high variability, noise, and multicollinearity make it almost impossible to predict RUL accurately using standard modeling techniques. However, by understanding these challenges, the team was able to develop strategies to improve the model's performance and reliability. Addressing these issues is crucial for developing robust predictive maintenance models capable of handling the complexities of real-world turbofan engine data.

Conclusion:

Incorporating technical indicators into the predictive model for the FD001 dataset led to substantial improvements in model performance. The R^2 increased from 0.5246 to 0.786, the MAE decreased from 21.2 to 14.25, the MSE decreased from 820.92 to 364.04, and the score decreased from 1960 to 583. These improvements underscore the effectiveness of using technical indicators to enhance the model's predictive accuracy. Most of these enhancements were primarily attributed to adding the RSI indicator, which significantly improved the R^2 and reduced the score.

When applying the same model and approach to the other datasets (FD002, FD003, and FD004), distinct differences in performance were observed. FD001 and FD003, which operated under a single operating condition, exhibited less data variability and better model performance. FD003, despite having an additional failure mode, followed similar trends to FD001 and showed strong correlations between sensors and remaining life, resulting in a decent R^2 of 0.56.

In contrast, FD002 and FD004, which operated under six different conditions, presented significantly more variable and noisy data. This increased complexity led to challenges in tracking and predicting RUL accurately. For FD002, the model achieved an R^2 of 0.689, but the overall score was much higher due to the variability and noise. FD004, the most complex dataset, performed the worst with an R^2 of 0.461 and the highest score of 36,608, indicating the need for further feature engineering and signal filtering to improve model reliability.

The technical indicators significantly improved model performance for datasets with single operating conditions (FD001 and FD003). However, for datasets with multiple operating conditions (FD002 and FD004), the increased data variability posed challenges that required further refinement of the model and feature selection techniques. This analysis highlights the importance of considering operating conditions and data variability in predictive modeling for more accurate and reliable results.

V. Determining if an Engine is Faulty (Classification)

Through classification techniques, the team aimed to predict whether a turbofan engine is degraded (faulty) based on sensor measurements. Our primary focus was on the first simulation, FD001, which showed a stronger correlation between remaining proper cycles and sensor measurements. To facilitate this, we created a binary categorical variable: "Faulty" if it is the last cycle for a specific engine and "Not Faulty" otherwise.

The FD001 simulation has 20,631-time series snapshots for 100 different engines. Of these, only 100 snapshots correspond to the last cycle, indicating only 100 faulty engines. Consequently, our binary variable "Not Faulty" appeared 20,531 times, while "Faulty" appeared only 100 times, resulting in a highly imbalanced distribution. To address this imbalance, we used "balanced accuracy" to evaluate our models' performance, accounting for the class disparity and providing a more accurate assessment of the model's ability to predict faulty engines.

$$Balanced\ Accuracy = \frac{\frac{TP}{P} + \frac{TN}{N}}{2} = \frac{\frac{Detected\ faulty}{Faulty} + \frac{Detected\ not\ faulty}{Not\ faulty}}{2}$$

Our study prioritized minimizing Type II error (1-Recall), understanding that incorrectly classifying a faulty engine as not faulty (false negative) is significantly more costly and dangerous than a false positive. During maintenance, an undetected faulty turbofan engine could lead to severe in-flight issues, endangering flight safety.

To evaluate and compare different classification models, we employed 10-fold cross-validation. All models used the same features as those applied in the regression models for FD001. Features with zero standard deviation were excluded, and the predictor "P15" was omitted as it did not enhance the predictive power of the models. Below are the summarized results of the models, rounded to three decimal places:

Model	Accuracy	Balanced Accuracy	Type 2 Error
Logistic Regression	**0.995**	0.54	0.92
SVM	**0.995**	0.51	0.98
LDA	0.991	0.622	0.75
QDA	0.99	0.697	0.6
Naïve Bayes	0.923	**0.963**	**0**
Random Forest	**0.995**	0.519	0.95
Gradient Boosting	0.994	0.568	0.863

Table 3: Performance of Classification Models

The table demonstrates that models with higher balanced accuracy exhibit lower Type II error and overall accuracy. This indicates a tradeoff between recall and precision. Models with high

Type II errors have low Type I errors, and since most observations are not faulty (N), these models achieve high overall accuracy by minimizing Type I errors.

Our priority is to maximize the balanced accuracy of the classification model. Consequently, we performed threshold adjustments on Quadratic Discriminant Analysis (QDA) and Naïve Bayes models to minimize Type II error while potentially increasing overall accuracy.

Using an 80-20% training-test split, we observed the behavior of various metrics as we tuned the threshold for classifying observations. QDA and Naïve Bayes classify based on posterior probability, with a default threshold of 0.5. By adjusting the threshold across different ranges, we monitored changes in accuracy, balanced accuracy, and Type II error. This approach allowed us to optimize the models' performance by detecting faulty engines while maintaining reasonable overall accuracy.

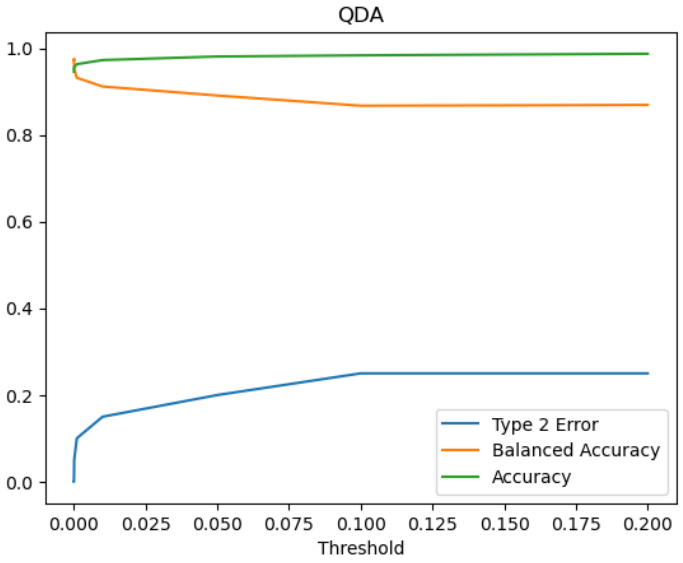

Figure 15: QDA's Performance for Different Thresholds

Figure 15 shows that when the threshold is set to 0.00001, the Type II error is zero, and balanced accuracy is maximized at 0.9755. The corresponding overall accuracy is 0.9513.

Figure 16 indicates that the Type II error is consistently zero as we adjust the threshold between 0.9 and 1 for Naïve Bayes. Maximum balanced accuracy and overall accuracy are reached when the threshold is 0.99999. At this point, the balanced accuracy is 0.9761, and the overall accuracy is 0.9525.

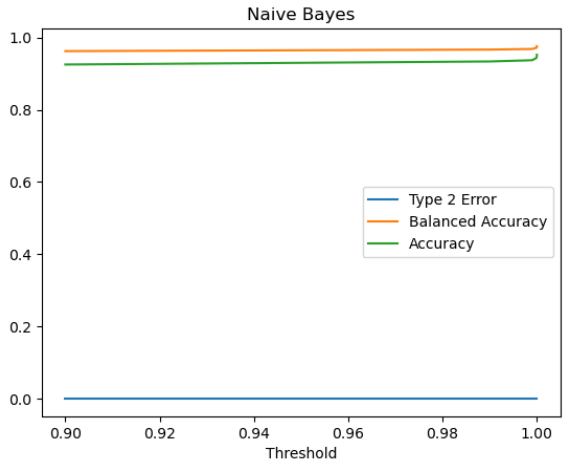

Figure 16: Naive Bayes' Performance for Different Thresholds

Note that the overall accuracy values obtained are lower than some of those seen in Table 3, such as Logistic Regression and Random Forest, which had an accuracy of 0.995. We prioritized minimizing False Negatives (Type II error) over overall accuracy because the cost of a late prediction significantly outweighs the cost of an early prediction.

Another way to observe the effects of changing thresholds is by examining the behavior of False Positives versus True Positives:

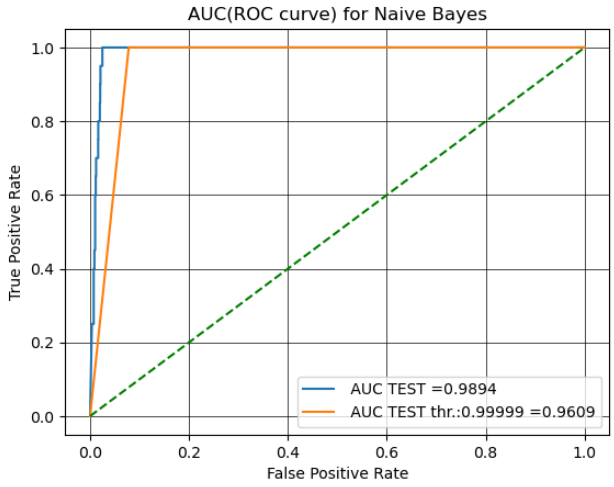

Figure 17: ROC curve for Different Thresholds

We observe that the area under the curve decreases as we change the regular threshold from 0.5 to 0.99999.

In conclusion, our selected model to classify whether an engine is faulty or not is Naïve Bayes, with a posterior probability threshold of 0.99999. We can expect this model to successfully detect when an engine is about to become faulty with zero to exceptionally low error.

VI. Team Efforts

The team efforts and responsibilities were evenly distributed, and each member significantly contributed to the project's success. Each team member contributed to the data analysis and feature selection process utilizing techniques learned throughout the analytics program. However, the team also engaged in individual project work. Cory developed the indicator logic, which helped the team extrapolate more significant insights from the data. Baris created the classification modeling systems that allowed the team to determine engine error likelihood, and Michael developed the various models used for the regression RUL analysis and prediction. Attached below is a Gantt chart to help visualize the timeline of the project and the distribution of responsibilities.

		Schedule							
Tasks/Deliverables	**Lead(s)**	**5/31/2024**	**6/7/2024**	**6/14/2024**	**6/21/2024**	**6/28/2024**	**7/5/2024**	**7/12/2024**	**7/19/2024**
Proposal and Presentation	Team								
Literature Survey/Research	Team								
Data Cleaning	Baris								
Data Exploration	Michael/Cory								
Midterm Progress Report	Team								
Regression Model Creation	Michael/Cory								
Classification Model Creation	Baris								
Model Validation	Team								
Final Report	Team								

VII. Conclusion

The primary objective of this project was to predict the Remaining Useful Life (RUL) of turbofan engines and determine if an engine is faulty based on sensor measurements. The team successfully addressed these objectives through rigorous data analysis, feature selection, and model development.

Key Achievements:

1. **Accurate RUL Prediction for FD001:**
    - By incorporating technical indicators, the team significantly improved the model's predictive accuracy for the FD001 dataset. The R^2 increased from 0.5246 to 0.786, the MAE decreased from 21.2 to 14.25, the MSE decreased from 820.92 to 364.04, and the score decreased from 1960 to 583. These results demonstrate the model's enhanced ability to predict RUL accurately for engines operating under a single condition and fault mode.
2. **Adaptation to More Complex Datasets:**
    - The same model and approach were applied to FD002, FD003, and FD004, each with increasing complexity. Despite the challenges posed by higher levels of multicollinearity and multiple operating conditions, the team identified key features that maintained predictive power across all datasets. For FD002, the model achieved an R^2 of 0.689, an MAE of 22.55, and an MSE of 880. For FD003, the R^2 was 0.56, with an MAE of 18.18 and an MSE of 752. FD004, the most complex dataset, presented the greatest challenge, with an R^2 of 0.461, an MAE of 30.9, and an MSE of 1496.
3. **Effective Fault Classification:**
    - For the classification problem, the team prioritized minimizing Type II errors. By adjusting the thresholds for Quadratic Discriminant Analysis (QDA) and Naïve Bayes models, they achieved a balanced accuracy of 0.9761 and an overall accuracy of 0.9525 with Naïve Bayes, successfully detecting faulty engines with zero to deficient error.
4. **Handling Data Variability:**
    - The team effectively managed the variability in FD002 and FD004, acknowledging the increased complexity due to multiple operating conditions. They highlighted the need for further feature engineering and signal filtering to improve the performance and reliability of these datasets.

Overall Impact:

The project demonstrated the importance of considering operating conditions and data variability in predictive modeling. The team significantly improved model performance for datasets with single operating conditions (FD001 and FD003) while also addressing the complexities of datasets with multiple conditions (FD002 and FD004) utilizing time series indicators. The

analysis underscores the effectiveness of technical indicators in enhancing model accuracy and provides a foundation for further refinement and application in real-world scenarios.

In conclusion, the team successfully met its objectives by accurately predicting RUL and detecting faulty engines across different datasets. The insights gained from this analysis pave the way for more robust predictive maintenance models, contributing to safer and more efficient aircraft operations.

VIII. References

1. NASA, Prognostics Center of Excellence Data Repository, http://ti.arc.nasa.gov/projects/data_prognostics

2. Saxena, K. Goebel, D. Simon, and N. Eklund, "*Damage Propagation Modeling for Aircraft Engine Run-to-Failure Simulation",* in the Proceedings of the 1st International Conference on Prognostics and Health Management (PHM08), Denver CO, Oct 2008.

3. Heimes, F.O., "*Recurrent neural networks for remaining useful life estimation",* in the Proceedings of the 1st International Conference on Prognostics and Health Management (PHM08), Denver CO, Oct 2008.

4. T. Wang, J. Yu, D. Siegel, J. Lee, "*A similarity-based prognostics approach for Remaining Useful Life estimation of engineered systems*", in the Proceedings of the 1st International Conference on Prognostics and Health Management (PHM08), Denver CO, Oct 2008.

5. L. Peel, "*Recurrent neural networks for remaining useful life estimation"*, in the Proceedings of the 1st International Conference on Prognostics and Health Management (PHM08), Denver CO, Oct 2008.

6. B. A. Roth, D. L. Doel, and J. J. Cissel, "Probabilistic Matching of Turbofan Engine Performance Models to Test Data*",* in ASME Turbo Expo 2005: Land Sea & Air, Reno-Tahoe, NV,2005.

7. N. Eklund, "Using Synthetic Data to Train an Accurate Real-World Fault Detection System*"*, in IMACS Multiconference on Computational Engineering in Systems Applications, pp. 483-488, 2006.

8. H. Madsen, G. Kariniotakis, H. A. Nielsen, T. S. Nielsen, and P. Pinson, "*A Protocol for Standardizing the Performance Evaluation of Short-Term Wind Power Prediction Models",* Technical University of Denmark, IMM, Lyngby, Denmark, Deliverable ENK5-CT-2002-00665, 2004.