Weekly Progress Report

Name: Agamya David

Domain: Data Science and Machine Learning

Date of submission: 5 August 2025

Week Ending: 03

I. Tasks Completed & Milestones Achieved

This week, the focus was on the end-to-end development of predictive models for all four turbofan engine datasets. The process began with a thorough exploratory data analysis and concluded with a comparative analysis between a baseline Random Forest model and a more advanced LSTM network.

1. **Exploratory Data Analysis & Preprocessing:**
    - For each dataset (FD001-FD004), an initial analysis was conducted by plotting sensor trends over time. This visual inspection provided an initial hypothesis about which sensors were indicative of engine wear.
    - To confirm these findings, the standard deviation for each sensor was calculated. Based on the plots and the distribution of the standard deviation values, a suitable threshold was determined for each dataset individually to programmatically identify "useless" sensors with little to no variance.
    - Before training, all feature data was normalized using Min-Max Scaling to ensure that all inputs were on a consistent [0, 1] scale, a crucial step for model performance.
2. **Baseline Model (Random Forest) Evaluation:**
    - A Random Forest Regressor was trained on all four datasets to establish a performance benchmark.
    - The model performed well on datasets with varied operating conditions, achieving a strong RMSE of **31.37** on FD002.
    - However, its performance degraded significantly on the FD003 dataset, which features multiple fault modes, resulting in a high RMSE of **50.42**. This highlighted the model's limitations and set a clear target for improvement.

3. **Advanced Model (LSTM) Implementation & Tuning:**
    - An LSTM (Long Short-Term Memory) model was implemented to better capture the time-series nature of the data. Initial results were significantly worse than the Random Forest baseline, with the model failing to learn effectively.
    - After a detailed iterative tuning process—which included scaling the target variable (RUL), clipping the RUL values to focus on the critical failure period, and adjusting the network architecture—the model's performance improved.
    - The tuned LSTM showed a remarkable improvement on the complex FD003 dataset, reducing the RMSE from the Random Forest's **50.42** down to **14.26**. This demonstrated the LSTM's superior ability to learn from data with different types of engine failure.

II. Challenges and Hurdles

1. **Initial LSTM Performance:**
    - **Challenge:** The initial versions of the LSTM model performed significantly worse than the Random Forest baseline, often predicting the same average value for all engines or producing very high error rates.
    - **Solution:** This was resolved through an iterative tuning process. Key solutions included:
        - Scaling the target variable (RUL) in addition to the input features.
        - Clipping the maximum RUL value to help the model focus on the more critical degradation phase.
        - Adjusting the model architecture (e.g., adding Dropout layers, removing the final activation function) and increasing the number of training epochs.
2. **Data Reshaping Complexity:**
    - **Challenge:** Correctly reshaping the data into sequences for the LSTM, especially for the test set where sequences could be of varying lengths, was a technical hurdle that initially caused errors.
    - **Solution:** Implemented a robust data preparation pipeline that handles engines with short histories by padding the sequences with zeros, ensuring a consistent input shape for the model.

III. Lessons Learned

1. **Model Selection is Context-Dependent:** The key lesson from this week is that the "best" model depends on the complexity of the data. The Random Forest was effective for simpler cases, but the LSTM's ability to understand sequences was necessary to achieve high performance on the dataset with multiple fault modes.
2. **The Importance of Iterative Tuning:** This week reinforced that building a deep learning model is not a one-shot process. The initial poor performance of the LSTM was not a failure of the model itself, but an indication that it required careful tuning of its architecture, preprocessing steps (like target scaling), and training parameters to be effective.
3. **Data-Driven Decision Making:** I learned the importance of using a simple baseline model to set a performance benchmark. This allowed me to make an informed, data-driven decision that the initial LSTM was not an improvement and required further work, guiding the tuning process effectively.