CS 7313.001 Project Proposal

Aakash Dhakal, Mirna Elizondo

Introduction

In our project, we will be working on binary classification of machine failures. It is a basic classification problem but requires better accuracy for the competition. We have planned to use boosting models along with Transfer learning to improve its accuracy.

1 Problem Description

1.1 Data

The data provided for the competition Reade and Chow (2023) includes a training and testing dataset generated from a deep learning model trained on Machine Failure Predictions. Feature distributions are close to, but not exactly the same, as the original. A summary of the unique values for machine failure attributes presented in Table 1
Kaggle Competition Link:
https://www.kaggle.com/competitions/playground-series-s3e17/discussion

1.2 Task

This is a binary classification problem. The goal of this competition is to predict whether a machine will fail or not, based on a set of features such as air temperature, process temperature, rotational speed, torque, tool wear, and product ID. Submissions are evaluated on area under the ROC curve between the predicted probability and the observed target. The provided training dataset has 136429 rows x 13 columns and the testing dataset 90954 rows x 12 columns.

1.3 Methods

Support Vector Machine (SVM) is a versatile machine learning algorithm used for classification and regression. It works by finding a hyperplane that best separates data points into distinct classes, making it effective for both linear and non-linear problems. **Random Forest Classifier** are an ensemble learning method that combines multiple decision trees to make predictions. They are known for their robustness and ability to handle complex datasets. **Gradient Boosting Classifier** is another ensemble learning method that combines the predictions of weak learners iteratively to improve accuracy. It is often used in tasks where high predictive accuracy is essential. **LightGBM** is a gradient boosting framework designed for efficiency and scalability. It employs tree-based learning algorithms, making it well-suited for handling large datasets and achieving high-performance results. **Transfer Learning** is one of techniques used for transferring the pre-built results and models for the better accuracy of classification.

1.4 Evaluation Measures

Accuracy measures the percentage of predictions that are correct and is calculated as the ratio of the number of correct predictions to the total number of predictions. **Precision** represents the percentage of positive predictions that are actually positive. It is computed as the ratio of true positives to the sum of true positives and false positives. **Recall** quantifies the percentage of actual positive examples that are correctly predicted. It is determined by dividing the number of true positives by the sum of true positives and false negatives. **F1 Score** is a metric that combines precision and recall into a single value. It is calculated as the harmonic mean of precision and recall, providing a balanced measure of a model's performance. Additionally, the **Area Under the Receiver Operating Characteristic Curve (AUC-ROC)** is a common metric used to assess the performance of binary classification models. It measures the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity).

2 Expected Outcome

In this advanced machine learning project, the goal is to develop a robust predictive model for machine failures, which can have far-reaching benefits. By accurately predicting machine failures before they occur, businesses can enhance reliability, reduce costs, and improve maintenance strategies. This data-driven approach empowers decision-makers to optimize operations and gain a competitive advantage. Furthermore, it contributes to sustainability efforts by minimizing resource consumption and waste generation. In summary, the successful implementation of this model has the potential to revolutionize machinery management, driving efficiency, cost savings, and environmental responsibility.

3 Additional Considerations

One of the challenges of this problem is the class imbalance. The dataset contains significantly more negative examples (machines that did not fail) than positive examples (machines that failed). This can make it difficult for machine learning models to learn to identify the positive examples. Another challenge is the lack of information about the machine failures. The dataset does not provide any information about the cause of the machine failures. This makes it difficult to develop a machine learning model that can accurately predict machine failures. Despite these challenges, I believe that it is possible to develop a machine learning model that can accurately predict machine failures. I am excited to work on this project and to see what I can achieve.

References

Walter Reade and Ashley Chow. 2023. Binary Classification of Machine Failures. `https://www.kaggle.com/competitions/playground-series-s3e17`

Feature	Description	Train	Test
Product ID	Unique failure ID	9976	9909
Type	Type of failure	3	3
Air Temperature	Temperature of the surrounding air where a machine or process is operating	95	92
Process Temperature	Temperature of a substance or material within a manufacturing or industrial process	81	84
Rotational Speed	Speed at which a component revolves	952	946
Torque	Measure of the rotational force applied to a component	611	595
Tool wear	Gradual deterioration of cutting tools or abrasive materials used in machining or manufacturing processes	246	246
Machine Failure	Binary variable indicating whether or not the machine failed	2	2
Time to Failure (TWF)	Amount of time it takes for a machine or system to fail after being put into operation	2	2
Health Diagnostic Feature (HDF)	Features or data points indicative of the health or condition of a machine	2	2
Predictive Warning Flag (PWF)	System or indicator that issues warnings or alerts when it predicts a potential machine failure	2	2
Operational Status Flag (OSF)	Used to indicate the current operational status of a machine (e.g., 0 for operational, 1 for non-operational)	2	2
Remaining Useful Life (RNF)	Estimated or predicted remaining lifespan of a machine or component before it is expected to fail	2	2

Table 1: Unique Values for Machine Failure Attributes for Train and Test Sets