Project Proposal (COMP 7150/8150)

Group Members

Neehanth Reddy Maramreddy (U00924209)
Sandeep Rao Gandra (U00911593)

Problem statement

Predict machine failures using sensor data and various machine parameters. The target variable is Machine failure (0 = No, 1 = Yes), which is a binary classification problem. The goal is to build a model to predict machine failure and potentially reduce unplanned downtimes.

Objective:

Build a machine learning model that accurately predicts whether a machine will fail, using input features such as temperature, rotational speed, torque, tool wear, and several other machine parameters.

Dataset Information:

Training data (train.csv): 136,429 entries and 14 columns, containing machine parameters and failure status. Test data (test.csv): 90,954 entries and 13 columns, contains the same features (without the target variable) for prediction.
(Dataset source: Walter Reade and Ashley Chow. Binary Classification of Machine Failures. https://kaggle.com/competitions/playground-series-s3e17, 2023. Kaggle.)

Pre-processing challenges:

- Handling categorical variables (ProductID, Type) appropriately.
- Checking for class imbalance in machine failures.
- Identifying and dealing with outliers in numerical features.
- Scaling features (Rotational speed [rpm], Torque [Nm], etc)

Questions we will answer using the dataset

- What are the key factors contributing to machine failures?
- How do different machine types compare in terms of failure rates?
- How well do different classification models perform in predicting failures?
- What is the best model for classifying the machine failures?
- What are the important features for classification?

Methods we plan to use to answer the questions

- Exploratory Data Analysis (EDA)
- Feature selection and Engineering (If needed)
- Machine learning models (Logistic Regression, Tree-Based models, other)
- Model evaluation (Evaluation metrics, Hyperparameter tuning, Future importance analysis)

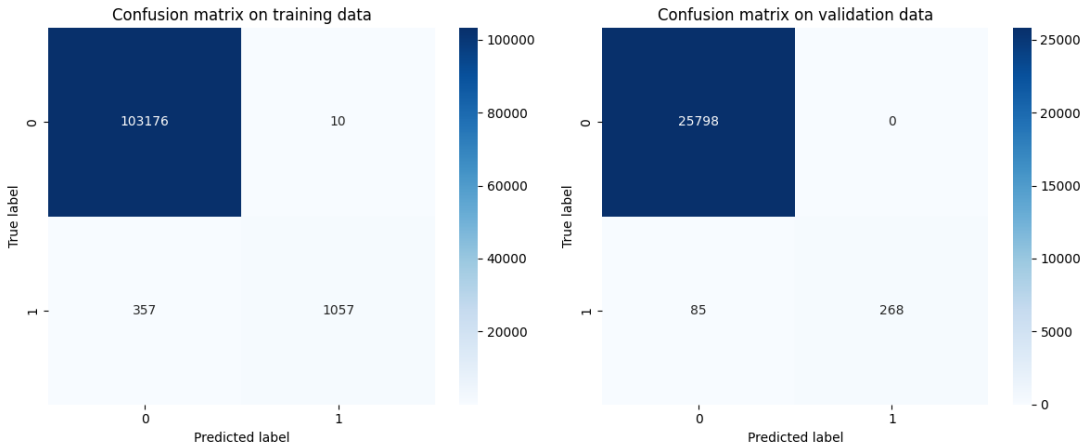
Baseline Model

We built a baseline performance with Logistic Regression. This model will serve as a basis for comparison with more sophisticated models.

The baseline model is trained on the preprocessed training data, and its performance will be evaluated using standard classification metrics.

Initial Results

	Training Data		Validation Data	
Model	**Accuracy**	**ROC-AUC score**	**Accuracy**	**ROC-AUC score**
Logistic Regression	0.996491	0.941341	0.996750	0.937451



The model demonstrates high accuracy on both training and validation sets. The data exhibits significant class imbalance, which is leading model to perform well on majority class but poorly on minority class. This results in potential overfitting, which will be investigated by using SMOTE or any other methods in future.

Deliverables

Project Report
Code (EDA, Preprocessing, Baseline model)