Elevator Vibration Predictive Maintenance Documentation

Project Overview

The main goal of this project is to predict abnormal vibrations for an elevator system by using sensor data, before too much damage is done and repair costs are higher. By accurately predicting these abnormalities, we can anticipate maintenance needs and prevent potential system failures, thereby improving the efficiency and safety of the elevator.

Data Acquisition and Cleaning

The data used in this project is public data from five sensors installed in an elevator system in Brazil. The primary target variable is 'vibration'. The dataset initially contained NaN values in the 'vibration' column, which were replaced with zero to clean the data. Targeting vibration allows for a simplification of the project as only one variable needs to be considered. Vibration gives a good insight into the health of the governor, the hoist ropes, and other safety features. Possibly even the sump pump. All these components of course depend on the type of elevator.

Exploratory Data Analysis

The dataset was split into training and test sets with an 80:20 ratio. This allowed for sufficient data for training the models, while also ensuring a substantial amount of unseen data for testing and validation.

Model Selection and Training

Three different machine learning models were used in this project: Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor. These models were chosen for their versatility and effectiveness in dealing with regression tasks. Each model was trained on the training data and used to make predictions on the test data. The three models were then able to be compared to get a final average of peaks of interest.

Model Evaluation and Optimization

The performance of the models was evaluated using the Mean Absolute Error (MAE). The lower the MAE, the better the model's performance. Initial runs used a fixed number of estimators for the Random Forest and Gradient Boosting models. Further optimization was achieved by increasing the number of estimators in the Random Forest model to 1000. The models' predictions were plotted against actual values to visually assess their performance.

Optimal Estimator Conclusion

The Random Forest model was initially run with 100 estimators. However, to achieve a balance between runtime and accuracy, the number of estimators was later increased to 1000. This increased the model's accuracy without significantly impacting the runtime, making it the optimal choice for this project.

Model Advantages and Limitations

Each model used in this project has its advantages and limitations.

Linear Regression is straightforward and computationally efficient, but it may not capture complex relationships in the data.

The Random Forest model is highly flexible and capable of modeling complex patterns, but it can overfit if not properly tuned and may be slower to train than simpler models.

Gradient Boosting can provide high accuracy and handle a mixture of feature types, but it can also be prone to overfitting and requires careful tuning.

It is also important to point out that the quality of the data was not the best as I had to infer what the variables meant, as well as use an arbitrary time scale as an index as there was no measure. This of course would give greater insight into the efficiency and effectiveness of the model in predicting failures and suggesting maintenance.

In this project, the Random Forest model provided the best balance of accuracy and computational efficiency.

Project Limitations

Some limitations of this project include the assumption that the NaN values in the vibration data can be replaced with zero, which may not accurately reflect real-world conditions. The models used also assume that the relationships between the features and the target variable remain constant over time. In the real world, these relationships may change, requiring the model to be updated or retrained- or realistically more than a single variable being the criteria. Finally, the models were evaluated based on a single metric, the Mean Absolute Error. Other metrics, including the Mean Squared Error or R2 score, might provide different perspectives on model performance.