Datasheet

Motivation

- The dataset is a synthetic dataset that reflects real predictive maintenance encountered in industries that employ the use of machines and equipment. Due to trade secrets issues, maintenance datasets are generally difficult to obtain. For this reason, the synthetic data was created for training purposes.

- The AI4I 2020 Predictive Maintenance dataset is from the UC Irvine repository and was donated on 29 August 2020.

Composition

- The dataset consists of 10,000 instances. Each instance contains readings, which are usually collected by sensors fitted on the machine parts. These readings are information that can signal the health status of the machines. The instances contain (1) product ID (L, M, H for Low, Medium and Low) for the quality of the product being machined, (2) air temperature [K], (3) process temperature [K], (4) rotational speed [rpm], calculated from a power of 2860W overlaid with normally distributed noise, (5) Torque [Nm]; normally distributed around 40Nm, (6) Tool wear [min]; the quality variants H/M/L add 5/3/2 minutes of tool wear to the processing tool, (7) Machine failure. The dataset consists of one-hot encoded failure modes corresponding to tool wear failure (TWF), heat dissipation failure (HDF), power failure (PWF), overstrain failure (OSF), and random failure (RNF).

- There are 10,000 instances.

- The dataset is a synthetic dataset of predictive maintenance records. The dataset exhibits records of a multivariate time series. This dataset is not a subset of a larger dataset.

- Each instance consists of numerical data for air temperature, process temperature, rotational speed, torque, and tool wear. The remaining entries of machine failure, tool wear failure, heat dissipation failure, power failure, overstrain failure, and random failures are all binary.

- Yes. There exist labels in the dataset. They represent when the machine fails or does not fail in a binary sense.

- There are no missing data. Data is complete.

- There are relationships between the instances of the output - machine failures and the failure modes descriptors. At least one of the failure modes must be true for machine failure to occur; hence, the label is set to 1. However, it is not transparent to the machine learning method which of the failure modes has caused the process of the machine to fail.

- There is no recommended data split of the dataset.

- The dataset is self-contained. It does not rely on any external resources.

- The dataset is comprised of synthetic data. Although it is not real, it reflects the data collected from predictive maintenance records.

- The dataset is synthetic and does not contain any confidential information. The dataset is not a subpopulation.

- AI4I 2020 Predictive Maintenance Data is a predictive maintenance dataset that does not relate to humans.

- The data relates to the status of machines only. There are no personal identifiers in the data relating to race, ethnicity, criminal history, and financial or health data.

Collection process

- The data is synthetic data, which reflects real predictive maintenance records. The aim is to provide publicly available

predictive maintenance data for any purpose - like training or practice.
- The data is not a subsample.
- The data represents synthetic preventive maintenance data for the year 2020.
- It is unknown if an ethical review process was conducted on the dataset.
- AI4I 2020 Predictive Maintenance Data is only synthetic predictive maintenance data for many purposes.
- AI4I 2020 Predictive Maintenance Data is only synthetic predictive maintenance data for many purposes.
- AI4I 2020 Predictive Maintenance Data is only synthetic predictive maintenance data for many purposes.
- It's unknown if an analysis of the potential impact of the data has been conducted.

Preprocessing/cleaning/labeling

- AI4I 2020 Predictive Maintenance Data is synthetic data that reflects predictive maintenance in the industry. The air temperature and process temperature features have normalised to a standard deviation of 2K and around 300K, and a standard deviation of 1K added to the air temperature plus 10K. The machine failure consists of five independent failure modes: tool wear failure, heat dissipation failure, power failure, overstrain failure, and random failure. All the failure modes are one-hot encoded.

- There is no publicly available information on the raw data.

Uses

- The AI4I 2020 Predictive Maintenance Data is a publicly available multivariate, time series data used to build, validate and test classification models. The data is also frequently used to benchmark classification algorithms.

- It's unknown to the author of this datasheet if there are links to any papers that have used this dataset, however, see below a link to the dataset: AI4I 2020 Predictive Maintenance Dataset - UCI Machine Learning Repository. The dataset has over 17,000 views already.

- Apart from benchmarking classification algorithms, the dataset could also be used to determine the highly correlated features with machine failures.

-AI4I 2020 Predictive Maintenance Data is a synthetic dataset. Although it reflects the expectations of a real predictive maintenance dataset, it does not reveal any organisation's trade secrets.

- Predictive maintenance datasets are proprietary. Therefore, such datasets relate to particular machines and organisations. This dataset, therefore, does not apply to every machine or organisation.

Distribution

- The dataset is publicly available.
- The dataset is publicly available on the UC Irvine Machine Learning Repository - Home - UCI Machine Learning Repository.
- There might also be other places where the dataset is available.
- The dataset is currently available. According to the UC Irvine repository, the dataset has been open since 29 August 2020.
- The dataset is licensed under Creative Commons Attribution 4.0 International.

Maintenance

- AI4I 2020 Predictive Maintenance Data is a synthetic dataset. As far as the author of this datasheet knows, the dataset is unmaintained.

Model Card

Model Description, input, output

- The inputs to the model include Type, Air temperature, process temperature, rotational speed, torque, tool wear, tool wear failure, heat dissipation failure, power failure, overstrain failure, and random failure. The binary output (Machine failures) is predicted using two models: an LSTM neural network and a decision tree model.

- A total of 2 models were built and compared with each other on specific metrics.

- An LSTM neural network was constructed and fitted on the original, oversampled, and undersampled data, while the AUC score was used to measure the model's performance. The LSTM was built on six layers - the first layer has 100 units, the second layer serves as the dropout layer, the third and fourth layer has 50 units, the fifth is the flattening layer, and the sixth is the dense layer.

- A decision tree model was built and trained on the original, oversampled, and undersampled data. The metrics used were Accuracy, Recall, Precision and F1 scores. However, more emphasis was placed on the F1 score.

- Oversampling and undersampling were used to balance the dataset to ensure the models were not biased to any classes.

Performance

- Both the LSTM neural network and decision tree models were trained using the training dataset; the hyperparameters were analysed and chosen using the validation dataset, and the test set was used to determine the performance of each of the models.

LSTM Neural Network Performance Comparison:

	AUC Perf. Original Train Data	AUC Perf. Original Val. Data	AUC Perf. Original Test Data	AUC Perf. Undersampled Train Data	AUC Perf. Undersampled Val. Data	AUC Perf. Undersampled Test Data	AUC Perf. Oversampled Train Data	AUC Perf. Oversampled Val. Data	AUC Perf. Oversampled Test Data
0	0.991556	0.995438	0.985674	0.994155	0.996198	0.982765	0.990002	0.989957	0.979261

- The LSTM neural network trained on the undersampled trained dataset performed the best compared to the Bayesian-optimisation-tuned LSTM neural network.

Bayesian Optimisation LSTM Comparison

	AUC Perf. Bayes Train Data	AUC Perf. Bayes Val. Data	AUC Perf. Bayes Test Data
0	0.879346	0.880962	0.874792

- The decision tree performed excellently on the original train set as expected, without the oversampling and undersampling because of the imbalanced nature of the dataset.

- The decision tree trained on the oversampled data performed well as it had the most reasonable F1 score, which is one of the suitable metrics for this classification case.

- The decision tree model does not perform well after random and grid searches.

Decision Tree Performance Comparison:

	DTree Perf. Original Train Data	DTree Perf. Original Val. Data	DTree Perf. Original Test Data	DTree Perf. Undersampled Train Data	DTree Perf. Undersampled Val. Data	DTree Perf. Undersampled Test Data	DTree Perf. Oversampled Train Data	DTree Perf. Oversampled Val. Data	DTree Perf. Oversampled Test Data
Accuracy	1.0	0.998500	0.996000	0.943167	0.938500	0.937500	1.0	0.967000	0.96650
Recall	1.0	0.983333	0.967213	1.000000	0.933333	0.950820	1.0	0.983333	0.95082
Precision	1.0	0.967213	0.907692	0.389982	0.320000	0.322222	1.0	0.475806	0.47541
F1	1.0	0.975207	0.936508	0.561133	0.476596	0.481328	1.0	0.641304	0.63388

Random-Search-Tuned Decision Tree Performance Comparison:

	Tuned DTree Perf. Original Val. Data	Tuned DTree Perf. Original Test Data
Accuracy	0.829935	0.796000
Recall	0.873165	0.909091
Precision	0.803679	0.116959
F1	0.836982	0.207254

Grid-Search-Tuned Decision Tree Performance Comparison:

	Tuned DTree Perf. Original Train Data	Tuned DTree Perf. Original Val. Data	Tuned DTree Perf. Original Test Data
Accuracy	0.838000	0.829000	0.828500
Recall	0.954128	0.966667	0.934426
Precision	0.177778	0.145729	0.143939
F1	0.299712	0.253275	0.249453

Limitations

- Unlike the decision trees, the LSTM neural network is not explainable and is sensitive to different random weight initialisations.

- The LSTM neural network and decision tree model are computationally expensive. They require large memory requirements.

Trade-offs

- The decision tree model trained on the oversampled train dataset has a high recall but low precision. Therefore, the model has a high false positive, which affects the F1 score. However, the decision tree trained on the oversampled data still has the highest F1 score.