Data Preprocessing for Machine Learning Modules

Rawan El Moghrabi

Faculty Supervisor: Dr. Miriam Capretz

Introduction

This project revolves around the data preprocessing methodology implemented for a data-driven predictive maintenance solution. Developing the predictive maintenance solution involved acquiring historical electrical data from assets and creating a health index for each asset. Before development began the data needed several processing steps to improve the solution's accuracy and efficiency [1].

Objectives

The objectives of this research project include:

1. Identify the required data processing steps to prepare data for the machine learning module
2. Sanitize data by removing null values and outliers
3. Transform data to account for real-world variabilities such as Daylight Savings Time
4. Encode cyclical features into the data set
5. Normalize data to ensure each feature has an equal contribution

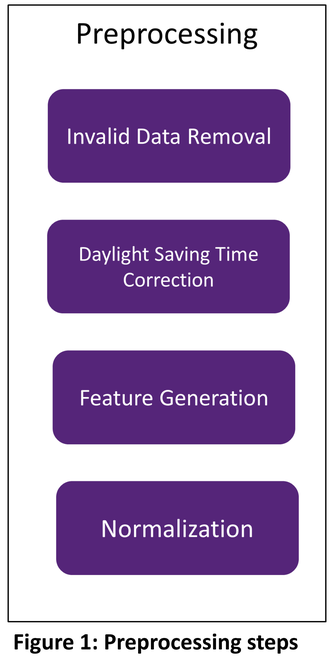

Figure 1: Preprocessing steps

Research Methodology

The following steps were followed to correctly prepare the data to be used in the machine learning module. All steps were implemented using custom Python scripts that utilized various libraries such as the PANDAS and NumPy libraries.

A. Data Sanitization

Data Sanitization performs noise treatment, reduce redundancies, and identify outliers [2].

A. Daylight Savings Correction

To support the predictive maintenance solution, the data collected had accurately represent local time for each asset. This meant that a transformation had to be applied to account for the variation due to daylight savings

A. Feature Generation

Feature Generation is used to create specific features from the raw data available. To do that, a sine-cosine coordinate system was developed to preserve the cyclical aspects of date and time.

A. Normalization

Data normalization is the process in which the available data set is scaled or transformed to fit a specific range giving each feature an equal contribution [3].

Results

Table 1 shows an example of the date after data sanitization occurred. All null values and outliers were removed.

Table 1: First five entries after Data Sanitization and Daylight Savings Correction

Timecode	Year	Month	Day	Hour	Minute	Weekday	Volts_avg
1.56E+09	2019	7	17	8	0	3	0
1.56E+09	2019	7	17	8	15	3	0
1.56E+09	2019	7	17	8	30	3	0
1.56E+09	2019	7	17	8	45	3	0
1.56E+09	2019	7	17	9	0	3	0

Table 2 shows an example of the sine and cosine values that were developed during feature generation. They represent some time and date features.

Table 2: Example of the sine and cosine coordinates for select date and time features

month_sin	month_cos	day_sin	day_cos	hour_sin	hour_cos
-0.23932	-0.97094	-0.40674	-0.91355	0.866025	-0.5
-0.23932	-0.97094	-0.40674	-0.91355	0.866025	-0.5
-0.23932	-0.97094	-0.40674	-0.91355	0.866025	-0.5
-0.23932	-0.97094	-0.40674	-0.91355	0.866025	-0.5
-0.23932	-0.97094	-0.40674	-0.91355	0.707107	-0.70711

Conclusions

Using modern Python scripts and software engineering principles, the data set was improved by eliminating outliers, applying real world patterns, and encoding cyclical attributes to the data features. Data preprocessing represents one of the significant steps of machine learning projects as it directly impacts the success of learning modules.

References

[1] J. Nalić and A. Švraka, "Importance of data preprocessing in credit scoring models based on data mining approaches," 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2018, pp. 1046-1051, DOI: 10.23919/MIPRO.2018.8400191.

[2] Y. Huang, M. Milani, and F. Chiang, "PACAS: Privacy-aware, data cleaning-as-a-service," in 2018 IEEE International Conference on Big Data (Big Data). IEEE, dec 2018.

[3] D. Singh, B. Singh, 'Investigating the impact of data normalization on classification performance', *Applied Soft Computing*, τ. 97, σ. 105524, 2020.