Machine Learning Challenge: Day 7

Welcome to the seventh day of our 30 days Machine learning Challenge.

Data Pre-processing Techniques for Machine Learning: Standardization, Scaling, Encoding, and Feature Engineering

1. Pre-processing Data: Data pre-processing is an essential step in machine learning, it is the process of cleaning, transforming and preparing the data for a model to learn from it.
2. Standardize: Standardization is a technique to transform the data so that it has a mean of zero and a standard deviation of one. It is used to bring all the variables to the same scale so that one variable does not dominate the others.
3. Scale to Range: Scaling to a range is a technique to transform the data to a specific range. It is used to normalize the data so that all the values are within a specific range.
4. Dummy Variables: Dummy variables are used to handle categorical variables in the data. It is used to convert categorical variables into numerical variables that can be used in the model.
5. Label Encoder: Label Encoding is a technique to convert categorical variables into numerical variables. It assigns a unique number to each category.
6. Frequency Encoding: Frequency Encoding is a technique to handle categorical variables in the data by replacing the categorical variables with their frequencies.
7. Pulling Categories from Strings: This technique is used to extract categorical variables from strings. It is used to convert free-text variables into categorical variables.
8. Other Categorical Encoding: There are several other categorical encoding techniques used such as one-hot encoding, ordinal encoding, and more.
9. Date Feature Engineering: Date Feature Engineering is used to extract features from date variables such as day of the week, month, year, and more.
10. Add col _na Feature: Adding a column with the number of missing values is a technique to capture the missing values in the data.
11. Manual Feature Engineering: Manual Feature Engineering is the process of creating new features from the existing data by applying domain knowledge.