Feature Engineering in Machine Learning

Ayush Singh[1]

Antern Department of Artificial Intelligence
ayush@antern.co

Abstract. This document contains contents on data preparation in machine learning and we also cover several components of data preparation like feature engineering, feature selection, dimensionality reduction, etc. We provide readers with several traditional and modern techniques to handle complicated data tasks.

Key words: data preparation, feature engineering, feature selection, data cleansing, data transformation, dimensionality reduction

1 Introduction

The process of producing, changing, or choosing features (sometimes referred to as variables or attributes) from raw data in order to enhance the performance of machine learning algorithms is known as feature engineering. It entails the extraction of pertinent data and the development of fresh features that can aid algorithms in better comprehending the data and producing more precise predictions.

2 Examples:

- **Age from date of birth:-**If you have a dataset including an individual's date of birth, you can compute the age of that individual by subtracting their current date from their date of birth. Particularly in jobs like forecasting health risks, insurance premiums, or consumer segmentation, this characteristic may have greater relevance for a machine learning system than the raw date of birth.
- **Text length:-**The length of the text may prove to be a helpful characteristic when solving a text categorization challenge. For instance, you can add a new feature called "text length" to categorise movie reviews as favourable or unfavourable based on how many words or characters are contained in each review. This characteristic can aid the computer in comprehending the connection between a review's length and sentiment.
- **Average purchase amount:-** Assume you have a dataset of client transactions, each of which includes the date, item, and amount spent. Based on their spending patterns, you should be able to estimate client attrition. By figuring out the typical purchase amount for each customer, you may add a

new feature named "average purchase amount". With the help of this new functionality, a machine learning system may be able to better identify client spending habits and forecast customer attrition.

3 How Feature Engineering helps?

Take into account a dataset of fruit samples with the attributes weight and colour (red or green). Predicting whether a particular fruit is an apple or a watermelon is the objective. The dataset appears as follows:

Fruit	Color	Weight
Apple	Red	150g
Apple	Red	170g
Apple	Green	160g
Watermelon	Green	4,000g
Watermelon	Green	4,500g
Watermelon	Red	4,200g

Apples and watermelons can both be red or green, therefore if you were to employ a machine learning method with these two features directly, it might have trouble telling them apart.

Here, feature engineering may be useful. By dividing each fruit's weight by the total weight of all the fruits in the dataset, you may produce a new feature called "weight ratio." The new dataset would seem as follows:

Fruit	Color	Weight	Weight_Ratio
Apple	Red	150g	0.0326
Apple	Red	170g	0.0369
Apple	Green	160g	0.0348
Watermelon	Green	4,000g	0.8696
Watermelon	Green	4,500g	0.9783
Watermelon	Red	4,200g	0.9130

The machine learning system can now quickly distinguish between apples and watermelons based on their weight ratios thanks to this new functionality. Compared to watermelons, apples have far smaller weight ratios, which facilitates accurate classification by the algorithm.