**Feature Engineering**

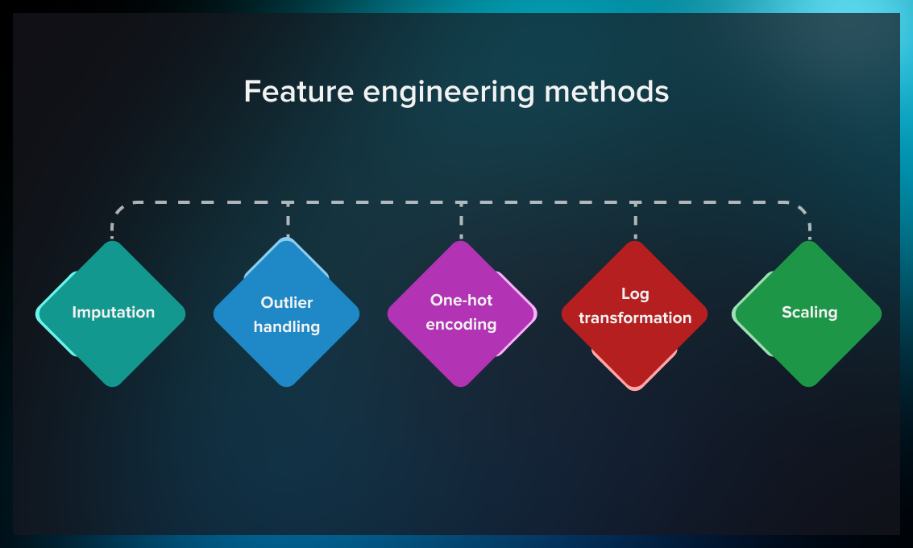
**Feature engineering is a very important process in the field of data analysis and machine learning. Data Scientists spends majority of their time in data cleaning because in the real world, the data looks pretty messy and need to clean the data before feeding to machine learning models. Even, if you are not using machine learning and doing simple data analysis still you need to clean the data. Feature engineering is something that helps in data cleaning.**

**Outlier detection, one hot encoding, handling missing values are few basic examples of feature engineering that helps in data cleaning, data preparation or data manipulation.**

**Feature engineering is a process of extracting useful features from raw data using mathematics, statistics and domain knowledge.**

**We have seen advancements especially in the area of neural networks where we can auto detect the meaningful features.**

## **Feature engineering methods**



### **Imputation**

Imputation is the process of managing missing values, which is one of the most common problems when it comes to preparing data for machine learning. By missing values, we mean places where information is missing in some cells of a respective row.

There may be different causes for missing values, including human error, data flow interruptions, cross-datasets errors, etc. Since data completeness impacts how well machine learning models perform, imputation is quite important.

Here are some ways how you can solve the issue of missing values:

* If a row is less than 20-30% complete, it’s recommended to dismiss such a record.
* A standard approach to assigning values to the missing cells is to calculate a mode, mean, or median for a column and replace the missing values with it.
* In other cases, there are possibilities to reconstruct the value based on other entries. For example, we can find out the name of a country if we have the name of a city and an administrative unit. Conversely, we can often determine the country/city by a postal code.

You can find more sophisticated approaches to imputation [in this post](https://bookdown.org/max/FES/imputation-methods.html).

### **Outlier handling**

Outlier handling is another way to increase the accuracy of data representation. Outliers are data points that are significantly different from other observations.

This graph shows how outliers can influence the ML model. By dismissing the outliers, we can achieve more accurate results.



It can be done by removing or replacing outliers. Check out [this post](https://towardsdatascience.com/5-ways-to-detect-outliers-that-every-data-scientist-should-know-python-code-70a54335a623) for an overview of the five most popular approaches to handling outliers.

### **One-hot encoding**

Categorical values (often referred to as nominal) such as gender, seasons, pets, brand names, or age groups often require transformation, depending on the ML algorithm used. For example, decision trees can work with categorical data. However, many others need the introduction of additional artificial categories with a binary representation.

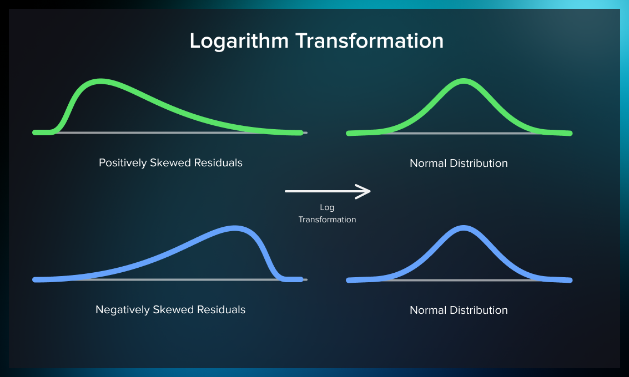
Binary representation means you assign a value of 1 if the feature is valid and 0 if it is not.

|  |  |  |
| --- | --- | --- |
| **ID** | **Male** | **Female** |
| User 1 | 1 | 0 |
| User 2 | 0 | 1 |

One-hot encoding is a technique of preprocessing categorical features for machine learning models. For each category, it designs a new binary feature, often called a “**dummy variable**.”

### **Log transformation**

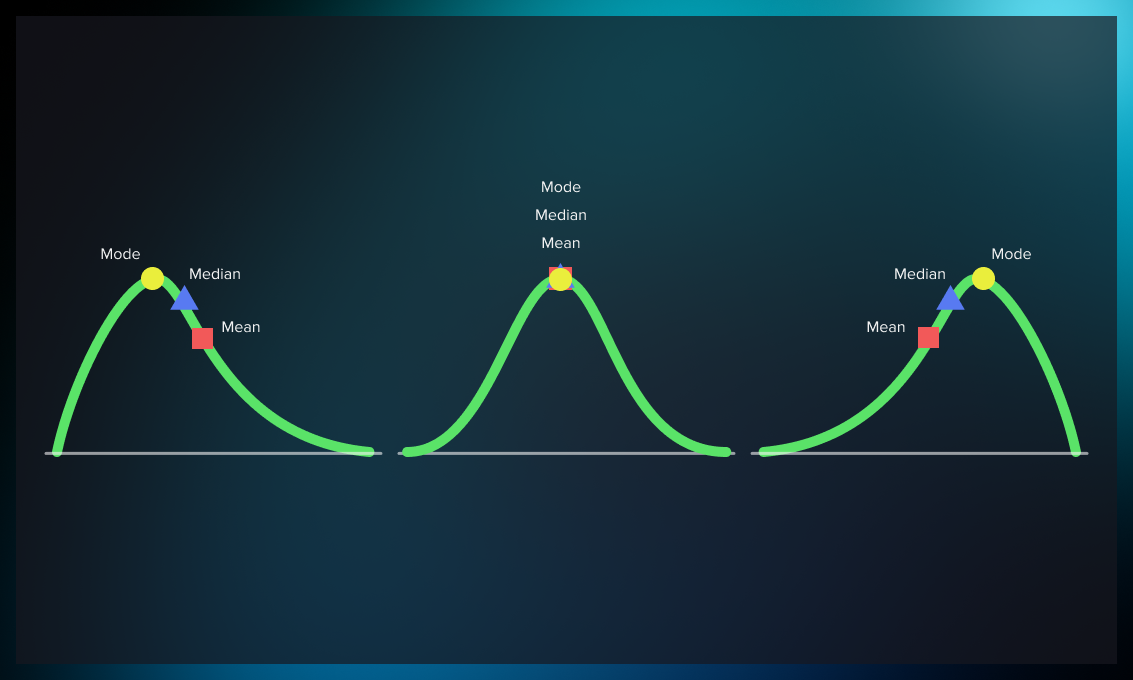
This method can approximate a skewed distribution to a normal one. Logarithm transformation (or log transformation) replaces each variable x with a log(x).



**The benefits of log transform:**

* Data magnitude within a range often varies. For example, magnitude between ages 10 and 20 is not the same as that between ages 60 and 70. Differences in this type of data are normalized by log transformation.
* Normalizing magnitude differences and increasing the robustness of the model also reduces the negative effect of outliers.

If we compare graphs on top and bottom, we’ll see that mode, mean, and median are slightly different for normal and skewed distribution. Normal distribution has undeniable advantages, but note that in some cases it can affect the model’s robustness and accuracy of results.



### **Scaling**

Scaling is a data calibration technique that facilitates the comparison of different types of data. It is useful for measurements to correct the way the model handles small and large numbers.

For example, despite its small value, the floor number in a building is as important as the square footage.

As another example, it is easier to perform a comparative analysis of the planets if we normalize values using their proportions against each other instead of actual diameters.



The most popular scaling techniques include:

* min-max scaling
* absolute maximum
* standardization
* normalization

## **Tools for feature engineering**

Below, you will find an overview of some of the best libraries and frameworks you can use for automating feature engineering.

[Featuretools](https://www.featuretools.com/) is one of the most widely used libraries for feature engineering automation. It supports a wide range of operations such as selecting features and constructing new ones with relational databases, etc. In addition, it offers simple conversions utilizing max, sum, mode, and other terms. But one of its most important functionalities is the possibility to build features using deep feature synthesis (DFS).

Some other useful tools for feature engineering include:

* the [NumPy](https://numpy.org/) library with numeric and matrix operations;
* [Pandas](https://pandas.pydata.org/) where you can find the Data Frame, one of the most important elements of data science in Python;
* [Scikit-learn](https://scikit-learn.org/) framework, a general ML package with multiple models and feature transformers;
* [Matplotlib](https://matplotlib.org/) and [Seaborn](https://seaborn.pydata.org/) that will help you with plotting and visualization.