Features Engineering

20CP401T

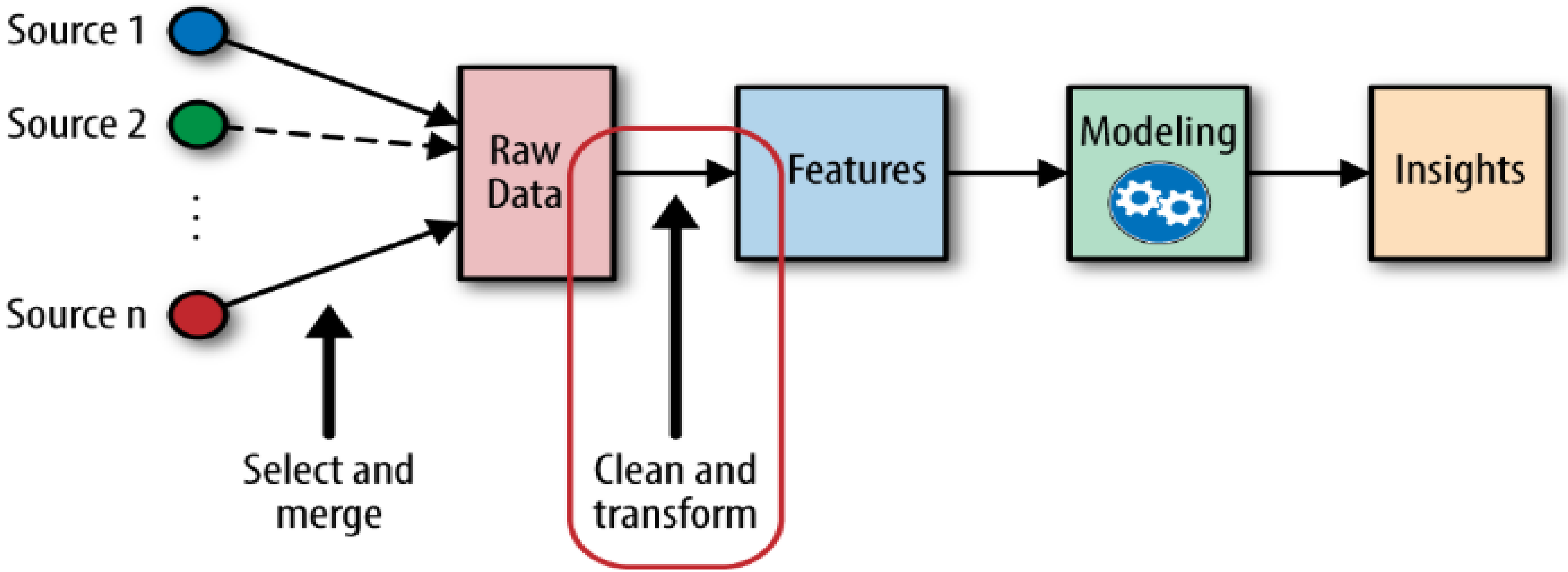
Himanshu K. Gajera
Department of Computer Science & Engineering
Pandit Deendayal Energy University, Gandhinagar

What is Feature Engineering?

What is Feature Engineering?

What is Feature Engineering?

- Feature Engineering is the process of extracting and organizing the important features from raw data in such a way that it fits the purpose of the machine learning model.



Source: https://www.omnisci.com/technical-glossary/feature-engineering

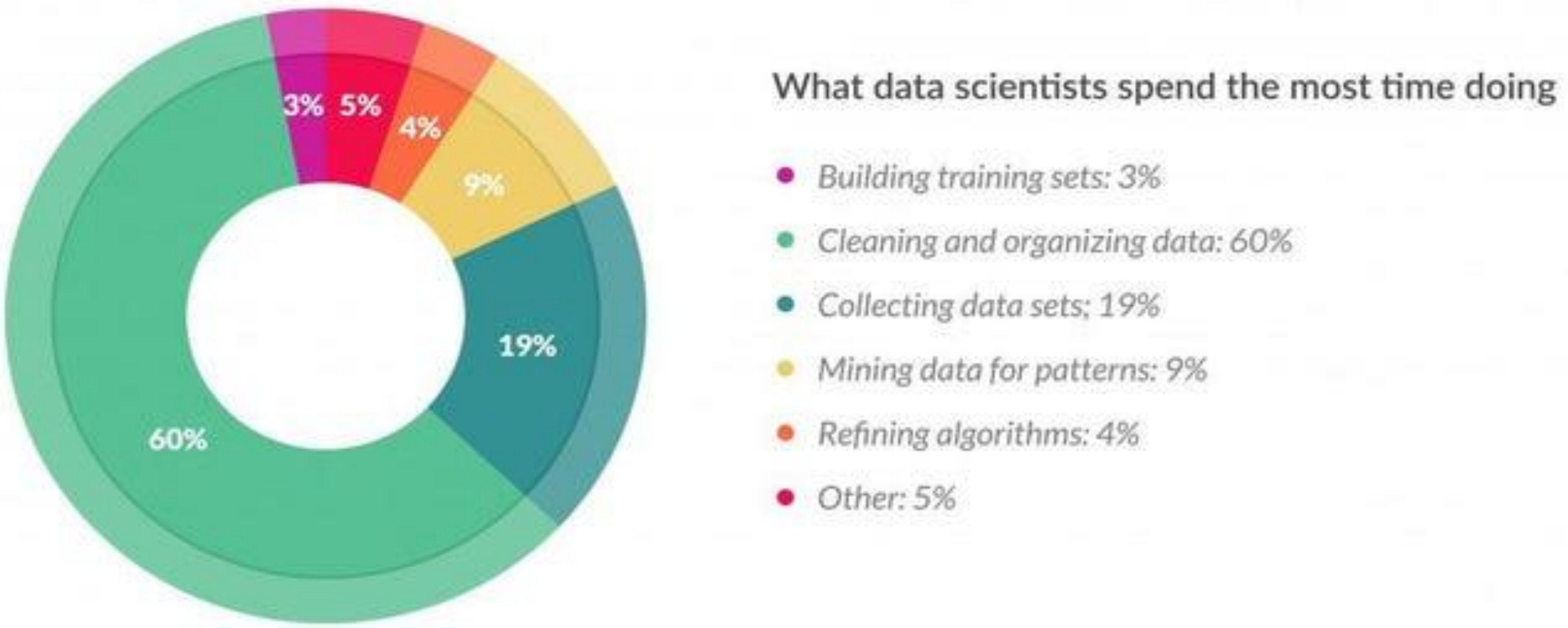
What is Feature Engineering?

Feature engineering in ML contains mainly four processes:

- Feature Creation
  - Feature creation is finding the most useful variables to be used in a predictive model.
- Transformations
  - Adjusting the predictor variable to improve the accuracy and performance of the model.
- Feature Extraction
  - Automated feature engineering process that generates new variables by extracting them from the raw data.
- Feature Selection
  - Few variables in the dataset are useful for building the model, and the rest features are either redundant or irrelevant.

Why is Feature Engineering so important?

- Do you know what takes the maximum amount of time and effort in a Machine Learning workflow?



Source: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

Why is Feature Engineering so important?

Data Engineering is an extremely important part of a Machine Learning Pipeline, but why is it needed in the first place?

- How we collect the data.
- Open data sources such as the internet, surveys, or reviews.
- This data is known as **raw data**.
- It may contain missing values, unstructured data, incorrect inputs, and outliers.
- If we directly use this raw, un-processed data to train our models, we will land up with a model having a very poor efficiency.

Benefits of Feature Engineering

- Higher efficiency of the model
- Easier Algorithms that fit the data
- Easier for Algorithms to detect patterns in the data
- Greater Flexibility of the features
- It helps in avoiding the curse of dimensionality.
- It helps in the simplification of the model so that the researchers can easily interpret it.
- It reduces the training time.
- It reduces overfitting hence enhancing the generalization.

Need for Feature Engineering in Machine Learning

- **Better features mean flexibility**
  - In machine learning, we always try to choose the optimal model to get good results.
  - Sometimes after choosing the wrong model, still, we can get better predictions, and this is because of better features.

- **Better features mean simpler models**
  - If we input the well-engineered features to our model, then even after selecting the wrong parameters (Not much optimal), we can have good outcomes.

- **Better features mean better results**

Steps in Feature Engineering

- **Data Preparation:**
  - Raw data acquired from different resources are prepared to make it in a suitable format so that it can be used in the ML model.
  - Contain cleaning of data, delivery, data augmentation, fusion, ingestion, or loading.
- **Exploratory Analysis:**
  - Analysis, investing data set, and summarization of the main characteristics of data.
  - Different data visualization techniques are used to better understand the manipulation of data sources
- **Benchmark:**
  - Benchmarking is a process of setting a standard baseline for accuracy to compare all the variables from this baseline

Feature Engineering Techniques

- **Imputation:**
  - Feature engineering deals with inappropriate data, missing values, human interruption, general errors, insufficient data sources, etc.
  - Missing values within the dataset highly affect the performance of the algorithm, and to deal with them "Imputation" technique is used.
  - Imputation is responsible for handling irregularities within the dataset.

- **Handling Outliers:**
  - Outliers are the deviated values or data points that are observed too away from other data points
  - Standard deviation can be used to identify the outliers. **Z-score** can also be used to detect outliers.

Feature Engineering Techniques

- **Log transform:**
  - Logarithm transformation or log transform is one of the commonly used mathematical techniques in machine learning.
  - Log transform helps in handling the skewed data, and it makes the distribution more approximate to normal after transformation.

- **Binning:**
  - Overfitting is one of the main issues that degrade the performance of the model and which occurs due to a greater number of parameters and noisy data.
  - One of the popular techniques of feature engineering, "**binning**", can be used to normalize the noisy data.
  - This process involves segmenting different features into bins.

Feature Engineering Techniques

- **Feature Split:**
  - Feature split is the process of splitting features intimately into two or more parts and performing to make new features.
  - This technique helps the algorithms to better understand and learn the patterns in the dataset.

- **One hot encoding:**
  - It is a technique that converts the categorical data in a form so that they can be easily understood by machine learning algorithms.
  - It enables group the of categorical data without losing any information.

Handling Missing Data

- Deleting Rows with missing values
- Impute missing values for continuous variable
- Impute missing values for categorical variable
- Other Imputation Methods
- Using Algorithms that support missing values
- Prediction of missing values
- Imputation using Deep Learning Library — Datawig

Feature Scaling

- Standardization

Standardization:

$$z = \frac{x-\mu}{\sigma}$$

with mean:

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i)$$

and standard deviation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

- Normalization

$$x_{\mathrm{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Conclusion

- Feature engineering helps in increasing the accuracy and performance of the model, there are also other methods that can increase prediction accuracy.

- There are many more available techniques of feature engineering, but mentioned the most commonly used techniques.