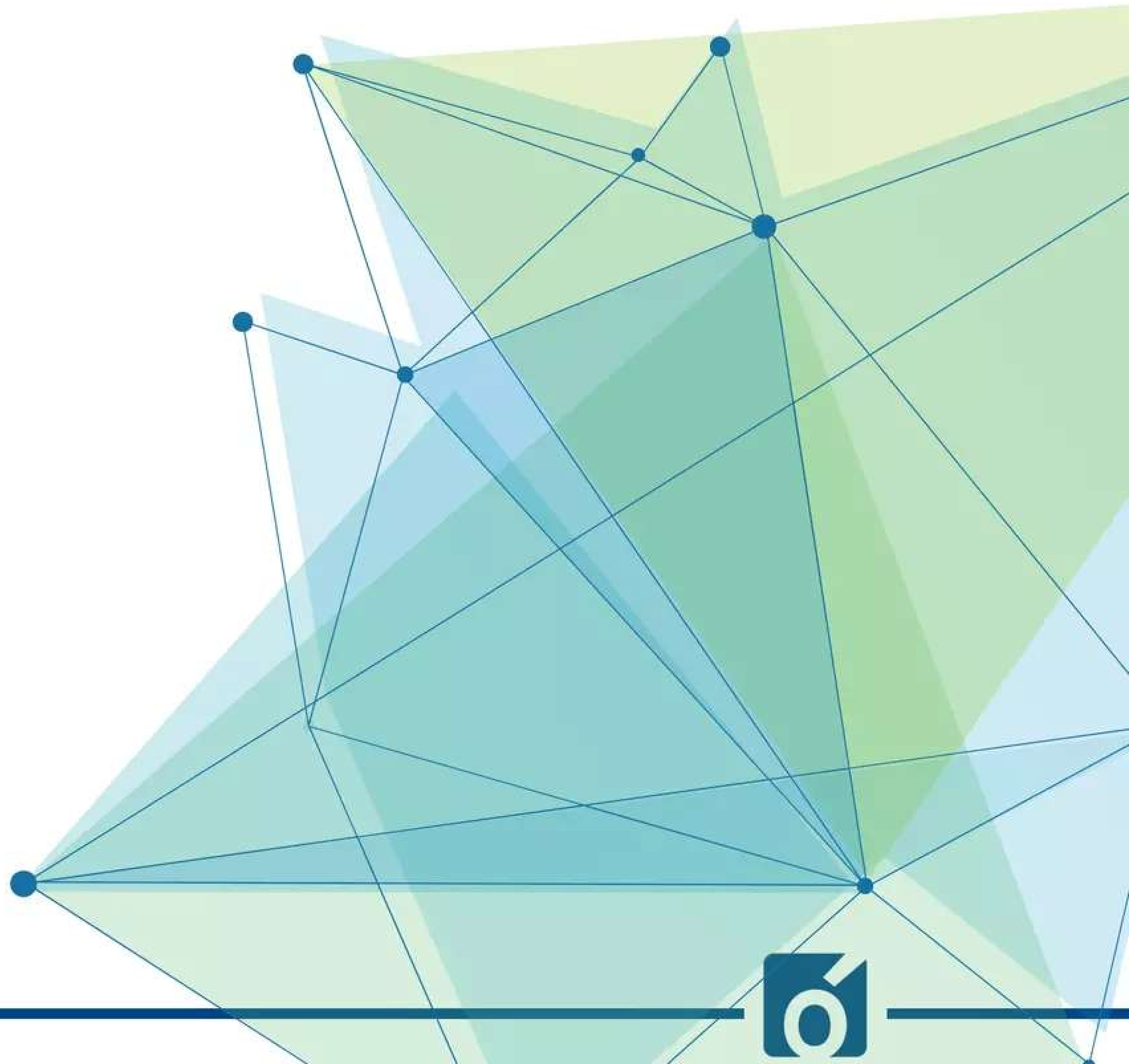
knóldus

Feature Engineering in Machine Learning

Tanishka Garg
Mayura Zadane
08th April 2022

Agenda

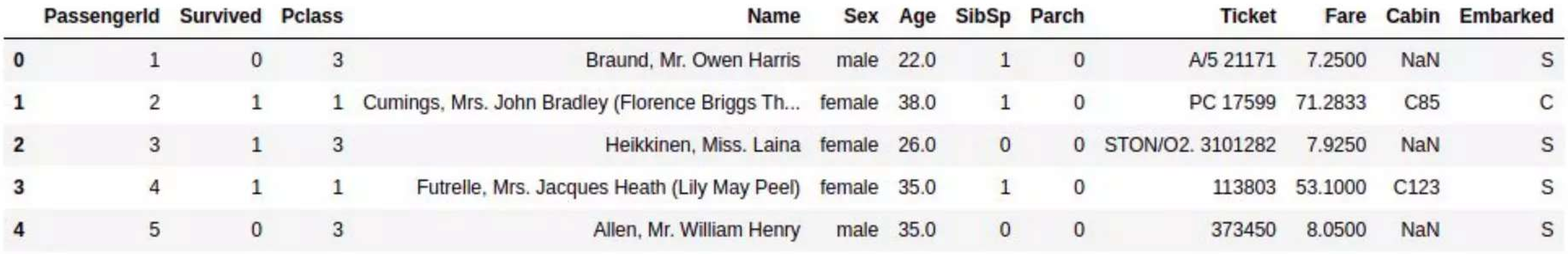
1. What is a Feature? What is feature Engineering?
2. What is its importance and why it is used?
3. Main processes of Feature Engineering
4. Feature Engineering Techniques
    - Imputation
    - Handling Outliers
    - Transformations
    - Encoding
    - Scaling (Normalization & Standardization)
    - Binning



What is a Feature? What is Feature Engineering?

What is a Feature?

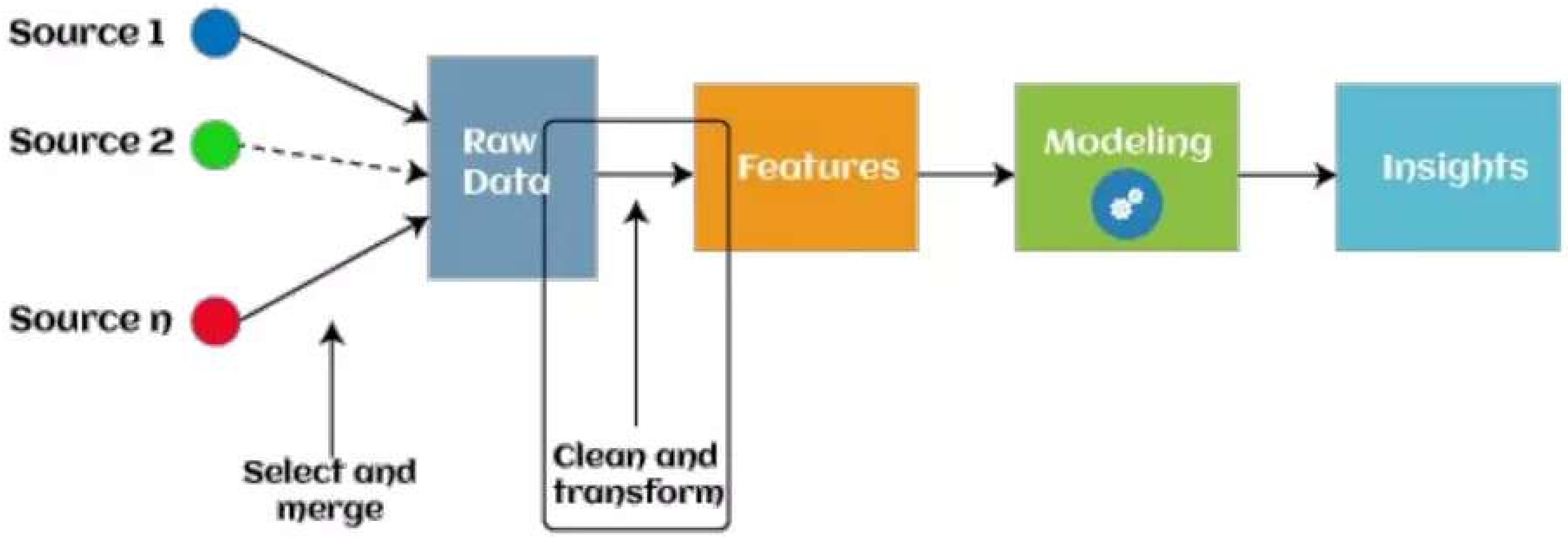
- Generally, all machine learning algorithms take input data to generate the output. The input data remains in a tabular form consisting of rows (instances or observations) and columns (variable or attributes), and these attributes are often known as **features**. For example, an image is an instance in computer vision, but a line in the image could be the feature. Similarly, in NLP, a document can be an observation, and the word count could be the feature. So, we can say **a feature is an attribute that impacts a problem or is useful for the problem**.
- The features you use influence more than everything else then the result. No algorithm alone, to my knowledge, can supplement the information gain given by correct feature engineering.



	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

What is Feature Engineering?

- Feature engineering is the pre-processing step of machine learning, which extracts features from raw data.
- It helps to represent an underlying problem to predictive models in a better way, which as a result, improve the accuracy of the model for unseen data.
- The predictive model contains predictor variables and an outcome variable, and while the feature engineering process selects the most useful predictor variables for the model.



What is its importance and why it is used?

What is its importance and why it is used?

Feature engineering in machine learning improves the model's performance. Below are some points that explain the need for feature engineering:

- **Better features mean flexibility.**

  In machine learning, we always try to choose the optimal model to get good results. However, sometimes after choosing the wrong model, still, we can get better predictions, and this is because of better features. The flexibility in features will enable you to select the less complex models. Because less complex models are faster to run, easier to understand and maintain, which is always desirable.

- **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. After feature engineering, it is not necessary to do hard for picking the right model with the most optimized parameters. If we have good features, we can better represent the complete data and use it to best characterize the given problem.

- **Better features mean better results.**

  As already discussed, in machine learning, as data we will provide will get the same output. So, to obtain better results, we must need to use better features.

Main processes of Feature Engineering

Main processes of Feature Engineering

The steps of feature engineering may vary as per different data scientists and ML engineers. However, there are some common steps that are involved in most machine learning algorithms, and these steps are as follows:

- **Data Preparation:** The first step is data preparation. In this step, 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. The data preparation may contain cleaning of data, delivery, data augmentation, fusion, ingestion, or loading.
- **Exploratory Analysis:** Exploratory analysis or Exploratory data analysis (EDA) is an important step of features engineering, which is mainly used by data scientists. This step involves 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, to find the most appropriate statistical technique for data analysis, and to select the best features for the data.
- **Benchmark**: Benchmarking is a process of setting a standard baseline for accuracy to compare all the variables from this baseline. The benchmarking process is used to improve the predictability of the model and reduce the error rate.

Feature Engineering Techniques

Feature Engineering Techniques

1. 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.**

For example, removing the missing values from the complete row or complete column by a huge percentage of missing values. But at the same time, to maintain the data size & to prevent loss of information, it is required to impute the missing data, which can be done as:

- For numerical data imputation, a default value can be imputed in a column, and missing values can be filled with means or medians of the columns.
- For categorical data imputation, missing values can be interchanged with the maximum occurred value in a column.

2. Handling Outliers

Outliers are the deviated values or data points that are observed too away from other data points in such a way that they badly affect the performance of the model. Outliers can be handled with this feature engineering technique. This technique first identifies the outliers and then remove them out.

Standard deviation can be used to identify the outliers. For example, each value within a space has a definite to an average distance, but if a value is greater distant than a certain value, it can be considered as an outlier.

Feature Engineering Techniques

3. 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. It also reduces the effects of outliers on the data, as because of the normalization of magnitude differences, a model becomes much robust.

4. Encoding

One hot encoding is the popular encoding technique in machine learning. It is a technique that converts the categorical data in a form so that they can be easily understood by machine learning algorithms and hence can make a good prediction. It enables group the of categorical data without losing any information.

5. Scaling

In most cases, the numerical features of the dataset do not have a certain range and they differ from each other. In real life, it is nonsense to expect age and income columns to have the same range. But from the machine learning point of view, how these two columns can be compared?

Scaling solves this problem. The continuous features become identical in terms of the range, after a scaling process.

Any Questions??

Reference :

1. https://www.repath.in/gallery/feature_engineering_for_machine_learning.pdf
2. https://www.analyticssteps.com/blogs/feature-engineering-method-machine- learning

Thank You!