Advanced Data Preprocessing for Machine Learning

DigitalSkola
Uncover The World Of Digital Skills With Us

Trainer

Hi there! I am Agil Haykal, just call me Agil. I am a curious guy who end up involved in data technology.

I have experienced Data Science as a trainer, consultant, and developer. I have taught +300 Data Scientist, Engineer, and Business Intelligence in total.

I handled several industries from manufacturing, banking, telecommunication, government, and Insurance. Please feel free to contact me to discuss anything about data technology.

Linkedin: Agil Haykal
Instagram/twitter: haykalatas

DigitalSkola
Uncover The World Of Digital Skills With Us

Quote of The Day

Garbage in Garbage out.

Golden in Golden out.

- Unknown

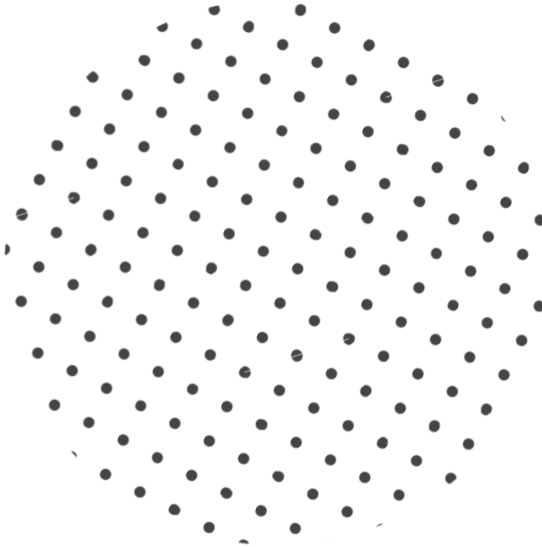
DigitalSkola
Uncover The World Of Digital Skills With Us

Table of Content

What will We Learn Today?

1. Feature Engineering
2. Advanced Label Encoder
3. Handling Text Data
4. Imbalaced Data

Feature Engineering



After we do exploratory data analysis, sometimes we ask "what next?" Feature engineering is one of follow up method to answer EDA. So what is this?

Feature engineering is a method to create new feature from other features. Why this is matter? Data is limited and might have unprovided potential information. Here is example of Feature engineering:

we can calculate several features that is ready from dataset
(price - discount) / quantity

Advanced Label Encoder

When we have a feature that own multiple categories, normal label encoder might abuse the memory of the engine. Let's say city feature has 100 categories, should we encode it to 100 columns? it will be a burden for the engine and not scalable. This method tackle the problem.

This advanced label encoder does not transform a feature into multiple features, instead we create harmony between a categorical column with other numerical ones. This method also seldom mentioned in online courses or articles. Let's try it in google colab!

Handling Text Data

Have you ever heard of sentiment analysis in twitter? It basically get people's opinion toward specific topic, like politics, brand, or even individual. This kind of subject is called Natural Language Processing (NLP). NLP is Data Science's sub subject that is specialized in language in general.

Then how to fit text data into NLP machine learning while it can only accept numerical data? This topic will help us to manipulate text data into numerical.

Disclaimer: Bahasa Indonesia is one of hardest language to start with, so we use English instead :D

DigitalSkola
Uncover The World Of Digital Skills With Us

Text Cleansing

When we cleanse text data, symbol like . , : ") (are not utilized and must be cleaned. That is why this step is important to start NLP process.

Tokenization

Tokenization is an NLP method to change a text into a list of words. This will help machine to separate each content.

Stop Words

Conjunction (kata sambung), repetitive words (I, you, is, am, are, ...) need to be removed from dataset. These words are called stop words.

Imbalanced Data

Why we need to do this? Imagine we work in an ecommerce and want to predict which one is the fraud account. And then, the customer service team reported that there are several accounts tagged as fraud. But the fraud account only 0.0001% from total users. If we use machine learning to predict this small number of account, it will prefer to be lazy since lacking the data. Then how to solve it?

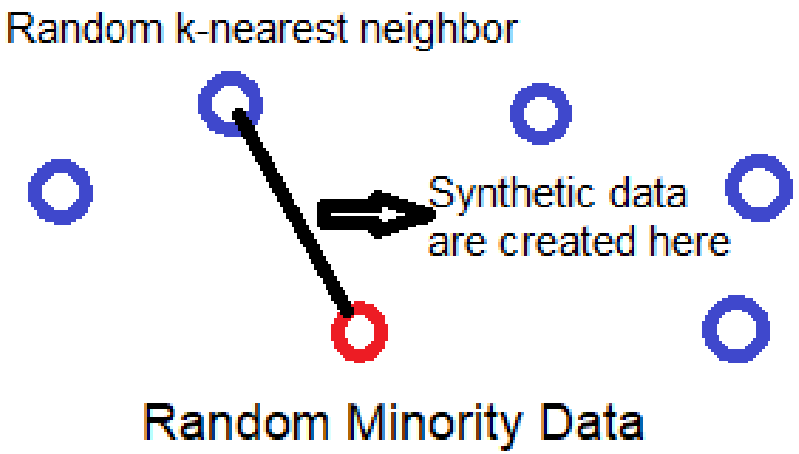
We can use undersampling and oversampling simultaneously. Undersampling is considerably easy because we choose data randomly. However, oversampling need extra effort to generate dataset. That is why algorithm like SMOTE comes for the help.

NOTE!

We oversample or undersample data for training data only! please keep in mind. Because we want to teach machine with anything we have. We should not create synthetic data to be predicted.

SMOTE

How smote works? SMOTE or (Synthetic Minority Oversampling Technique) basically creates fake/synthetic data that has similar characteristics with real data. It is used only to balance the proportion of minority data.



SMOTE works by utilizing a k-nearest neighbor algorithm to create synthetic data. SMOTE first start by choosing random data from the minority class, then k-nearest neighbors from the data are set. Synthetic data would then made between the random data and the randomly selected k-nearest neighbor. Still confused with the algorithm works? no problem, we will touch this again in ML class :)

When to use oversampling?

There are three arbitrary threshold that can be used:

1. Mild: Proportion of minority is 20-40%
2. Moderate: Proportion of minority is 1-20%
3. Extreme: Proportion of minority is < 1 %

Downside of Oversampling

Do not oversample data more than >50% of minority size because the synthetic data will dominate the information and lead to bias interpretation. Unless we have extremely little size of minority class, we can oversample up to 5x its size (it can be tested many times).

DigitalSkola

Uncover The World Of Digital Skills With Us

Thank YOU