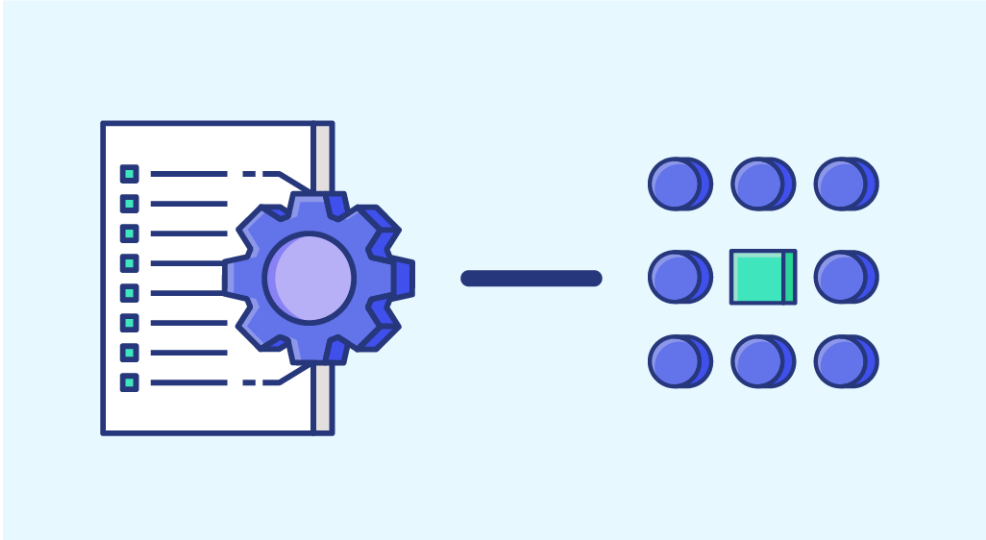
Feature Engineering: Importance for Machine Learning

October 5, 2023(https://datascientest.com/en/2023/10/05) - Reading Time: 6 minutes

Machine Learning (https://datascientest.com/en/category/machine-learning)



Feature Engineering involves extracting features from raw data to solve specific domain-specific problems using machine learning. Discover everything you need to know: definition, algorithms, use cases, training courses…

Artificial intelligence (https://datascientest.com/en/artificial-intelligence-definition) is increasingly used in all fields. However, to fully unleash its potential, a predictive analysis model (https://datascientest.com/en/knn-what-is-the-knn-algorithm) requires leveraging the available data.

To achieve this, it's essential to choose the right algorithm (https://datascientest.com/en/algorithm-what-is-it) and train machine learning models. (https://datascientest.com/en/unlock-your-future-dive-into-machine-learning-engineer-training) In reality, the most crucial aspect is utilizing **"Feature Engineering."**

Indeed, the features of the data have a direct impact on predictive models and their results. The more carefully prepared and chosen the features are, the more accurate the results will be. They should describe the inherent structure within the data.

In general, results depend on the chosen model, available data, and prepared features. **Problem framing and the metrics** used to estimate accuracy also play a **significant role.**

Even if a model isn't optimal, it can still yield good results. The key is to use good features, which allows for the use of less complex, faster-to-run models that are simpler to understand and maintain.

Likewise, good **feature engineering** can yield good results even if the chosen parameters aren't optimal. So, there's no need to endlessly search for the best model and the most optimized parameters, as long as you have the right features.

These features allow you to get closer to the underlying problem and represent the data accurately. So, what is Feature Engineering?

What is Feature Engineering ?

Feature Engineering is a process that involves transforming raw data into features that more precisely represent the underlying problem for a predictive model

Simply put, it's about applying domain knowledge to extract analytical representations from raw data and preparing them for machine learning.

This is the **first step in developing a predictive machine learning model.** It helps increase the model's accuracy on new, unseen data.

It's important to remember that **machine learning algorithms** learn a solution to a problem from sample data. Thus, Feature Engineering determines the best representation of the **sample data for learning** the solution to the problem.

This is highly significant because the success of an artificial intelligence or machine learning project (https://datascientest.com/en/artificial-intelligence-definition) often depends on the data representation. The algorithms must be able to understand the inputs.

Feature Engineering relies on a set of well-defined procedures and methods. The procedures to use vary depending on the data, and it's through experience and practice that one learns which ones to use in a given context.

What is a "feature" ?

Data is presented online in tables, with their attributes and variables presented in columns. An attribute can be a feature.

However, in the context of a problem, a feature is a useful or relevant attribute with respect to that problem. It's an important part of an observation aimed at understanding the structure of the modeled problem.

For example, in a computer vision problem, an image is an observation, while a feature could be a line within that image. In natural language processing, the observation could be a document, while a feature could be a sentence or a word from that document. In speech recognition, a complete utterance could be an observation, while an individual word could be a feature.

Learn Feature Engineering
(https://datascientest.com/en/machine-learning-engineer-course)

The different approaches to Feature Engineering

Feature Engineering is not a one-size-fits-all process. There are multiple approaches, and the one to adopt depends on the specific subproblem you are trying to solve.

"Feature Importance" involves objectively estimating the utility of a feature. This can be useful for feature selection. Each feature is assigned a score, and they can be ranked based on these scores. Features with the highest scores can be chosen to be included in the dataset.

This importance score can also be used to extract or construct new features that are similar but different from those already considered useful.

In general, a feature can be considered important if it is highly correlated with the dependent variable, which is what you are trying to predict.

Correlation coefficients are commonly used to measure feature importance.

Some more complex **predictive modeling** algorithms perform this selection internally alongside model construction. This is the case with **algorithms like MARS or Random Forests.**

Feature Extraction involves automatically constructing new features from raw data. This is very useful when observations in their raw form are too voluminous to be directly modeled by **predictive algorithms.**

Examples include textual, audio, and image data. It also applies to tabular data with millions of attributes.

The goal of **Feature Extraction** is to automatically reduce the dimensionality of these types of observations into a smaller set that can be modeled. Methods like **Principal Component Analysis** or unsupervised clustering can be used for tabular data (https://datascientest.com/en/k-means-clustering-in-machine-learning-a-deep-dive), while edge detection can be used for images.

Feature Selection is another method that involves removing unnecessary or redundant attributes from the data in the context of the problem being solved. This approach automatically selects the most useful subset for solving the problem.

Algorithms can use methods like correlation or other feature importance methods to rank and select features. A more advanced technique is to create and evaluate models automatically until the most appropriate one for prediction is found.

Feature Construction involves manually creating new features from raw data. This requires structuring sample data and exposing it to predictive modeling algorithms based on the problem being solved.

For **tabular data,** this might involve aggregating and combining features to create new ones or decomposing them. This task requires a lot of time and thought but can make a significant difference in the performance of a machine learning model.

Feature Learning involves automatically identifying and using features from raw data. The goal is to avoid the need for manual feature construction or extraction.

Modern deep learning methods (https://datascientest.com/en/all-about-deep-learning) can achieve this. Autoencoders and Restricted Boltzmann Machines are examples. These techniques can automatically learn abstract feature representations in an unsupervised or semi-supervised manner.

These compressed feature representations can then be used for speech recognition, image classification, or object recognition. Unfortunately, this approach works as a “black box” and doesn’t provide insight into how the representations were learned. Feature Engineering cannot be entirely automated.

The Feature Engineering process

Feature Engineering is part of the Machine Learning process. After defining a problem, the next step is to select and prepare the data. Data is collected, aggregated, cleaned, and formatted to be usable.

Feature Engineering occurs during the data transformation step, where data is converted from its raw state to a format suitable for modeling. Before this step, the data is in a format that doesn’t **allow for manipulation.**

The rest of the Machine Learning process (https://datascientest.com/en/gan-machine-learning-putting-fictitious-faces-into-practice) involves modeling the data by creating models, evaluating them, and configuring them.

The final step is presenting the results. Whenever new insights are identified in the data, this process must be repeated in the same order.

The **Feature Engineering** process is not independent. It’s an iterative process closely tied to data selection and model evaluation.

Depending on the problem at hand, different **Feature Engineering** methods are used. After selecting the appropriate features, the model’s accuracy is assessed by testing it on new data using the chosen features.

It's crucial to define the problem properly so that different models, configurations, and model sets can be tried. The testing method should accurately measure performance.

What's the point of Feature Engineering?

Feature Engineering can be used for various purposes. It can, for example, involve decomposing categorical attributes, breaking down date-time information, or scaling numeric quantities.

Here are some concrete use cases to better understand it. In the KDD Cup 2010 Machine Learning competition, participants had to model how students learn. A dataset of student performance (https://datascientest.com/en/what-is-a-dataset-how-do-i-work-with-it) on algebra problems was provided, and it had to be used to predict future performance. The winners of the competition were a group of students from National Taiwan University, who simplified the problem's structure through **Feature Engineering** by creating millions of binary features.

This structure allowed the team to use very simple but highly performing linear methods to create the best **predictive model**. Non-linear elements like temporality were reduced to binary indicators. This demonstrates the possibilities offered by binary indicators.

Another example is the Heritage Health Prize, a three-million-dollar prize awarded to the team capable of predicting which patients would be admitted to the hospital in the following year. Many participants in this competition used Feature Engineering techniques.

Why automate Feature Engineering?

Feature Engineering is an iterative process that requires a lot of time, resources, and technical expertise. A Data Science team (https://datascientest.com/en/test-annika-data-analyst) also needs to collaborate with domain experts to provide them with machine learning models tailored to their needs.

The automation of this process has the potential to disrupt the field of Data Science. It simplifies access to machine learning, eliminates the need for manual SQL query creation, and accelerates Data Science projects even without domain knowledge.

With automation, millions of hypotheses can be explored in a matter of hours.

Thanks to **AutoML products**, automation of Feature Engineering is now possible. With AutoML 2.0, the entire cycle from raw data to machine learning model development can be reduced to a few days instead of several months. This allows Data Science teams to deliver numerous machine learning models.

How do I learn Feature Engineering?

Feature Engineering is at the core of Data Science and Machine Learning. By choosing DataScientest's training programs, you can learn to master this discipline along with all the techniques and tools of data science.

Indeed, Machine Learning is an essential part of our Data Scientist, Data Analyst, or ML Engineer programs. You will also learn Python programming, database manipulation techniques, Deep Learning, and Data Visualization.

Our training programs are designed by professionals and directly address the needs of businesses. Learners receive a diploma certified by the University of Sorbonne, and 93% of them find employment immediately.

Each of our courses takes an innovative approach to Blended Learning, combining distance learning with in-person instruction. These programs can be taken as Continuing Education or in an intensive BootCamp mode.

Our courses can be financed through the Personal Training Account (CPF) or through Pôle Emploi via AIF or the **Bildungsgutschein** is you are in Germany. Don't wait any longer and discover our Data Science training programs now!

DataScientest Courses
(https://datascientest.com/en/data-scientist-course)

Related articles

(https://datascientest.com/en/cloudera-course-what-you-should-know)

Cloudera Course: What you should know (https://datascientest.com/en/cloudera-course-what-you-should-know)

Melanie - December 26, 2023

(https://datascientest.com/en/records-manager-a-key-data-governance-job)

Records Manager: A key Data Governance job (https://datascientest.com/en/records-manager-a-key-data-governance-job)

Melanie - December 23, 2023

(https://datascientest.com/en/amazon-sns-the-messaging-service-connected-to-aws)

Amazon SNS: The messaging service connected to AWS (https://datascientest.com/en/amazon-sns-the-messaging-service-connected-to-aws)

Melanie - December 22, 2023

(https://datascientest.com/en/aws-fargate-the-cloud-solution-for-running-containers-2)

AWS Fargate: The Cloud solution for running containers (https://datascientest.com/en/aws-fargate-the-cloud-solution-for-running-containers-2)

Melanie - December 22, 2023

Want to follow the news of Datascientest?

Subscribe to our newsletter ! (/en/blog-en#formnewsletter)

Sign up (/en/blog-en#formnewsletter)