FEATURE ENGINEERING

MIP-ML-08

VINAYAK GUPTA

MENTORNESS

INTRODUCTION

Feature engineering is the pre-processing step of machine learning, which is used to transform raw data into features that can be used for creating a predictive model using Machine learning or statistical Modelling. Feature engineering in machine learning aims to improve the performance of models.

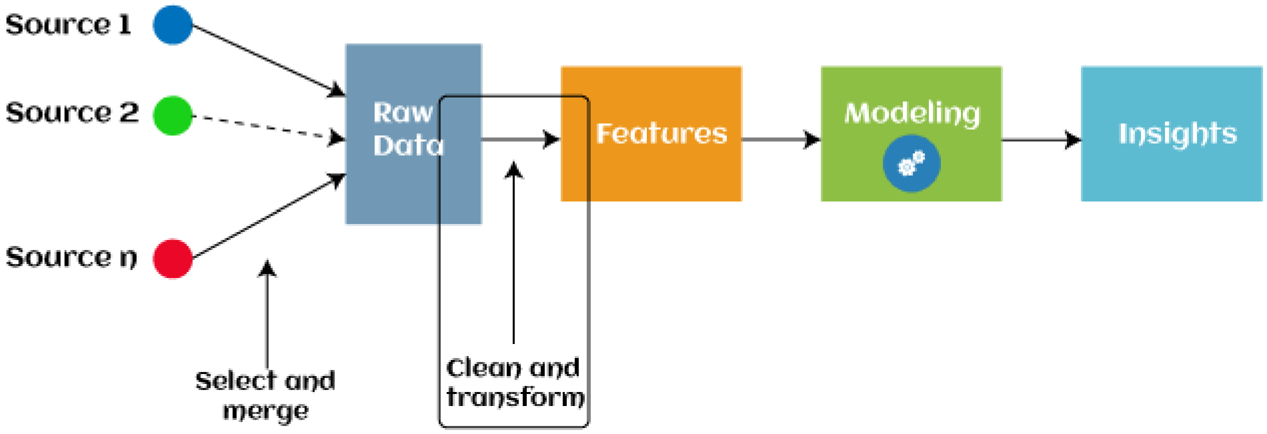
What is a feature?

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**.

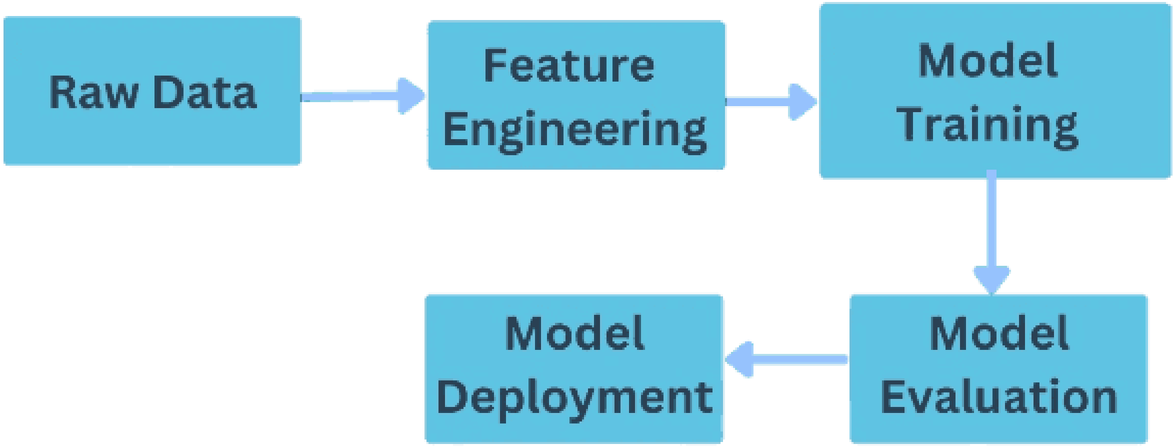
What is Feature Engineering?

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.

Feature engineering acts as an intermediary between the raw data and the model.



Let us demonstrate using a diagram so that we can understand how feature engineering works and where it falls in the process of building models.



Explanations of steps are:

- **Raw Data:** This is the original dataset with various features and the matching target variable.

- **Feature Engineering:** This is the process in which important features from the raw data are selected, created, transformed, and dimensionally reduced.

- **Feature Transformation:** After features are selected, they are manipulated and transformed. This is a step where new features can be created, dimensions are reduced, and categorical values are handled.

- **Model Training:** In this stage, the data with engineered features is fit into the model as input. The model then learns patterns and relationships between the engineered features and the target variable.

- **Model Evaluation:** This is a crucial step in Machine Learning where the model generalization and prediction accuracy are evaluated. The evaluation helps identify if the model has overfitting or underfitting issues.

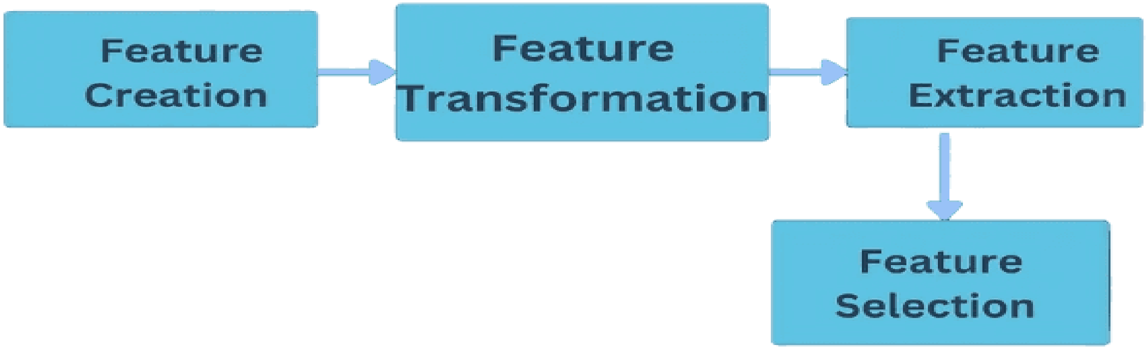
- **Model Deployment:** This is the final step where the model is realized to perform real-world tasks.

Since 2016, automated feature engineering is also used in different machine learning software that helps in automatically extracting features from raw data. Feature engineering in ML contains mainly four processes: **Feature Creation, Transformations, Feature Extraction, and Feature Selection.**

Feature Engineering Process

The main goal of feature engineering is to provide the model with important features of data that will help the model learn and make accurate predictions.

Feature engineering is an iterative process involving experimentation, model evaluation, and refinement to find the best feature set.



Feature Creation

Feature creation is the process of developing new features from existing ones in order to capture complicated relationships and patterns in data.

Let’s discuss the ways in which features are created.

- **Interaction features:** You can create new features by capturing the combined influence of two or more features.
- **Arithmetic operations:** New features can be created by doing simple calculations using arithmetic operations. For example, you can calculate the sum of values from existing features. That way, you’ll have created a new feature with the sum of the original values.
- **Polynomial features:** You can create features by raising values in existing features to power. This technique is beneficial for linear algorithms because it captures nonlinear correlations between characteristics and target variables.
- **Aggregate features:** You can create new features by using functions to summarize or aggregate data from numerous entries per entity. For example, you can calculate the average of values across different entities.
- **Feature scaling and normalizing:** New features can be created by modifying the distribution of features for specific algorithms. Techniques like z-score normalization or min-max scaling are used to perform feature scaling and normalization.
- **Feature binning:** You can create new features by separating continuous features into discrete bins or intervals.

Feature Transformation

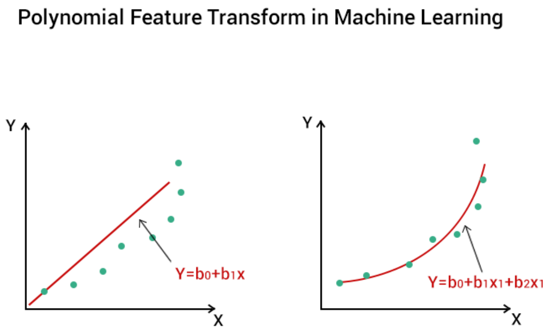
Feature transformation is a step-in feature engineering that involves utilizing mathematical approaches to change the values of features in order to improve the performance of machine learning models.

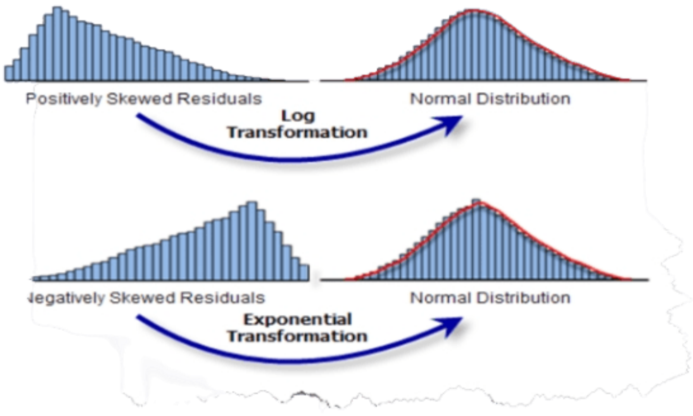
It serves numerous functions, including enhancing algorithm convergence during training, matching data distributions with the assumptions of particular algorithms, and permitting a better fit between the model and the data.

Let’s discuss methods you can use to transform features.

- **Normalization:** Features can be transformed by normalizing them. This can be done by decreasing the range of feature values to a preset range, typically between 0 and 1.
- **Standardization:** Features can be standardized by converting feature values to have a mean of 0 and a standard deviation, ensuring that all features contribute equally.
- **Logarithmic scaling:** Features can be transformed by taking the logarithm of feature values, allowing it to handle skewed data distributions and reduce the impact of big data sets.

In feature transformation, other mathematical operations such as square root, exponentiation, and division can also be applied.

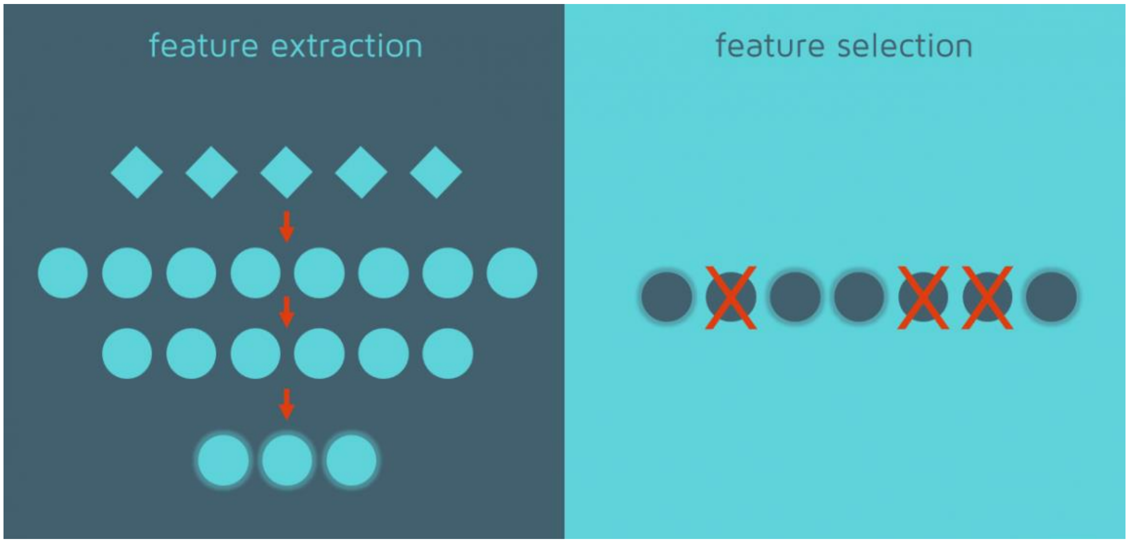




Feature Extraction

Feature extraction is an important step in feature engineering since it aims to minimize the dimensionality of data by translating it into a lower-dimensional representation while maintaining useful information.

This technique is useful for high-dimensional data or compact representations that retain crucial data features. Feature extraction techniques reduce complexity and improve visualization. Reducing dimensions may result in interpretability loss. What determines if the feature can be reduced is the nature of the data, the problem, and the trade-offs.



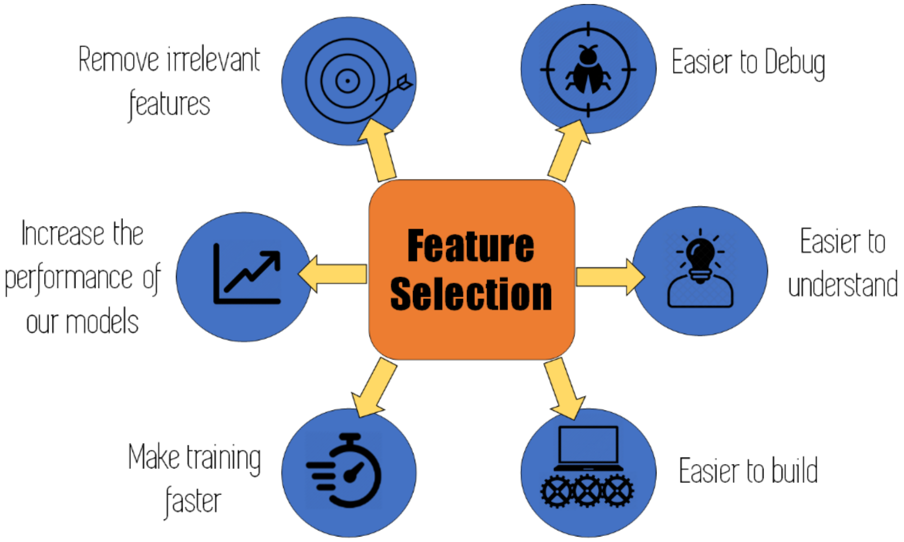
Feature Selection

Feature selection is a crucial step in machine learning, involving the identification and selection of relevant and meaningful features from raw data to enhance the predictive power and accuracy of the model.

The goal of feature selection is to increase the model's performance by minimizing input complexity while maintaining the most significant and informative elements, reducing overfitting, speeding up training, and increasing interpretability.

There are three types of feature selection methods:

- **Filter methods:** Examine the importance of features independently of any given model, employing statistical measures to rank or score features. Wrapper approaches include
- **Wrapper methods:** Involve training and evaluating the model multiple times with different feature subsets
- **Embedded methods:** Incorporate feature selection into the model-building process.



Features in a dataset are not equally valuable. Selecting unnecessary or redundant features might result in overfitting, increased computational cost, and decreased model interpretability.

Conclusion

In this article, we delve into feature engineering in machine learning. We discussed the concepts, processes, and techniques of feature engineering. The art of feature engineering emerges as a crucial force affecting the landscape of data analysis and model creation in the area of machine learning.