

Feature Engineering

Intro

For Machine Learning related tasks, it isn't possible to apply a machine learning model directly to the data. Feature engineering is the process of applying domain knowledge, analyzing relationships between features, and performing transformations on features. The goal of feature engineering is to improve the data to increase the performance of the model.

Feature engineering is an iterative, time-consuming process where data scientists continuously experiment to see what works and what doesn't. Feature engineering is the stage where data scientists spend most of their time. One of the significant advantages AutoAI provides is automated feature engineering, which removes the need for time-consuming manual feature engineering.

Types of Feature Engineering

Transformation

1. Log Transform



- The log transform is used to handle skewed data to make it more normal. In order for a log transform to work, the data values must all be positive. Common techniques to resolve negative values is to add 1 or subtract the smallest value and then add 1.

x	$\log(x + 1)$	$\log(x - \min + 1)$
-5	NaN	0
-4	NaN	0.3010299957
3	0.60	0.95





2. One-Hot Encoding

- Categorical data must be turned into numbers so that the machine learning models can interpret it. One-Hot Encoding is a technique to do this by creating a column for *each* unique value and for each row placing a 1 in the corresponding column for the categorical variable.

Before One-Hot Encoding

 Row	 City
<u>1</u>	Dallas
<u>2</u>	Austin
<u>3</u>	Houston

After One Hot Encoding




 Row	 Dallas	 Austin	 Houston
1	1	0	0
2	0	1	0
3	0	0	1

The disadvantage with One-Hot Encoding is that it leads to an explosion in the number of columns, and the data is very sparse.

3. Binning

1. Binning is the process of grouping the data together into... bins. Binning is done on both numerical and categorical data. Binning regularizes the model and prevents overfitting at the cost of performance, also known as the *bias-variance tradeoff*.

Numerical Binning

 Row	 Age	 Age - Binned
1	15	Teenager
2	13	Teenager
3	22	Twenties
4	28	Twenties
5	45	Middle Aged
6	55	Middle Aged

Categorical Binning

Company Name	Company - Binned
IBM	Tech
Facebook	Tech
Honda	Auto
Mercedes	Auto
United Healthcare	Healthcare
Humana	Healthcare

Creating New Features

- Using Domain Knowledge
 - It is possible to augment your data using domain knowledge with additional features that are combinations of the provided features. For example, if your dataset is about predicting the probability of heart disease, and it includes weight and height, then using height and weight, a BMI column can be created. Note, the new feature must be a non-linear transformation for it to affect the performance. Linear combinations of features do not provide added value.
- Looking at Feature Relationships
 - By plotting features against one another, unexpected relationships can surface. New features can be created by taking advantage of these relationships. In the example below, the two features have a parabolic relationship. A new feature of form $B_A = -A^2$ is added to the data.

