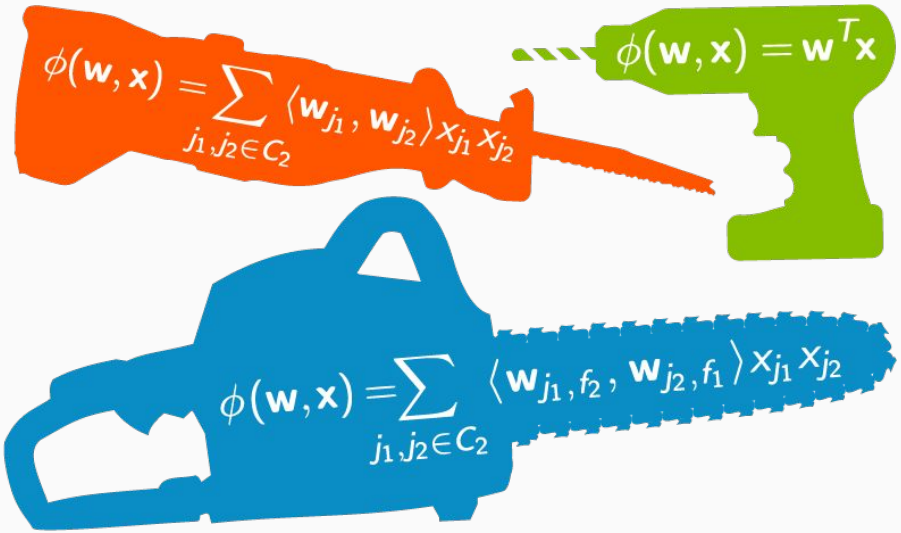
Feature Engineering

Creating new input features

What is Feature Engineering?

Process of transforming data from its raw form to something more useful to the predictive model.

In general, you can think of data cleaning as a process of subtraction and feature engineering as the process of addition.



What makes a good feature?

Infuse Domain Knowledge

Domain knowledge is knowledge of a specific, specialized discipline or field, in contrast to general knowledge.

You can often engineer informative features by tapping into your (or others') expertise about the domain.

Interaction Features

Interaction features are combinations of two or more features.

They can be products, sums, or differences between two features.

Example: Given how much a person made this week and hours worked, you could calculate their hourly wage.

Combine Sparse Classes

Sparse classes (in categorical features) are those that have very few total observations. They can be problematic for certain machine learning algorithms, causing models to be overfit.

- There's no formal rule of how many each class needs.
- As a rule of thumb, we recommend combining classes until each one has at least ~50 observations. As with any "rule" of thumb, use this as a guideline (not actually as a rule).

Add Dummy Variables

Most machine learning algorithms cannot directly handle categorical features. Specifically, they cannot handle text values.

Therefore, we need to create dummy variables for our categorical features.

Dummy variables are a set of binary (0 or 1) variables that each represent a single class from a categorical feature.

Original class	Dummy variable	E.g.
Wood	exterior_walls_Wood	= 0
Brick	exterior_walls_Brick	**= 1**
Other	exterior_walls_Other	= 0
Siding (Alum/Vinyl)	exterior_walls_Siding (Alum/Vinyl)	= 0
Missing	exterior_walls_Missing	= 0
Metal	exterior_walls_Metal	= 0
Brick veneer	exterior_walls_Brick veneer	= 0
Combination	exterior_walls_Combination	= 0

Encoding Categorical Features
Find and Replace

There are two columns of data where the values are words used to represent numbers. Specifically the number of cylinders in the engine and number of doors on the car.

	make	fuel_type	aspiration	num_doors	body_style	drive_wheels	engine_location	engine_type	num_cylinders
0	alfa-romero	gas	std	two	convertible	rwd	front	dohc	four
1	alfa-romero	gas	std	two	convertible	rwd	front	dohc	four
2	alfa-romero	gas	std	two	hatchback	rwd	front	ohcv	six
3	audi	gas	std	four	sedan	fwd	front	ohc	four
4	audi	gas	std	four	sedan	4wd	front	ohc	five

Encoding Categorical Features
Find and Replace

	make	fuel_type	aspiration	num_doors	body_style	drive_wheels	engine_location	engine_type	num_cylinders
0	alfa-romero	gas	std	2	convertible	rwd	front	dohc	4
1	alfa-romero	gas	std	2	convertible	rwd	front	dohc	4
2	alfa-romero	gas	std	2	hatchback	rwd	front	ohcv	6
3	audi	gas	std	4	sedan	fwd	front	ohc	4
4	audi	gas	std	4	sedan	4wd	front	ohc	5

Encoding Categorical Features
Label Encoding

Label encoding is simply converting each value in a column to a number. For example, the body_style column contains 5 different values.

	make	fuel_type	aspiration	num_doors	body_style	drive_wheels	engine_location	engine_type	num_cylinders
0	alfa-romero	gas	std	two	convertible	rwd	front	dohc	four
1	alfa-romero	gas	std	two	convertible	rwd	front	dohc	four
2	alfa-romero	gas	std	two	hatchback	rwd	front	ohcv	six
3	audi	gas	std	four	sedan	fwd	front	ohc	four
4	audi	gas	std	four	sedan	4wd	front	ohc	five

Encoding Categorical Features
Label Encoding

convertible -> 0
hardtop -> 1
hatchback -> 2
sedan -> 3
wagon -> 4

	make	fuel_type	aspiration	num_doors	body_style	body_style_cat
0	alfa-romero	gas	std	2	convertible	0
1	alfa-romero	gas	std	2	convertible	0
2	alfa-romero	gas	std	2	hatchback	2
3	audi	gas	std	4	sedan	3
4	audi	gas	std	4	sedan	3

Encoding Categorical Features
One Hot Encoding

Label encoding has the advantage that it is straightforward but it has the disadvantage that the numeric values can be “misinterpreted” by the algorithms. For example, the value of 0 is obviously less than the value of 4 but does that really correspond to the data set in real life?

A common alternative approach is called one hot encoding. The basic strategy is to convert each category value into a new column and assigns a 1 or 0 (True/False) value to the column.

Encoding Categorical Features
One Hot Encoding

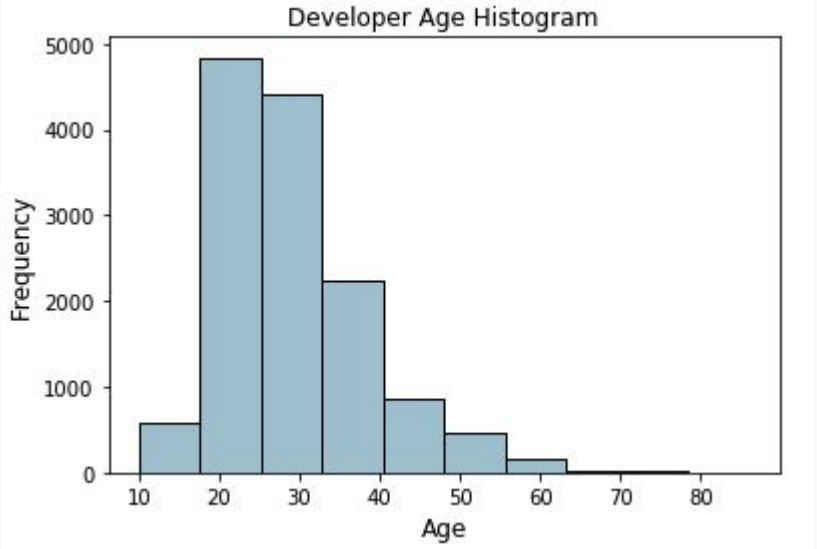
	make	fuel_type	aspiration	num_doors	body_style	drive_wheels	engine_location
0	alfa-romero	gas	std	2	convertible	rwd	front
1	alfa-romero	gas	std	2	convertible	rwd	front
2	alfa-romero	gas	std	2	hatchback	rwd	front
3	audi	gas	std	4	sedan	fwd	front
4	audi	gas	std	4	sedan	4wd	front

Encoding Categorical Features
One Hot Encoding

	make	fuel_type	aspiration	num_doors	body_style	drive_wheels	engine_location	drive_4wd	drive_fwd	drive_rwd
0	alfa-romero	gas	std	2	convertible	rwd	front	0	0	1
1	alfa-romero	gas	std	2	convertible	rwd	front	0	0	1
2	alfa-romero	gas	std	2	hatchback	rwd	front	0	0	1
3	audi	gas	std	4	sedan	fwd	front	0	1	0
4	audi	gas	std	4	sedan	4wd	front	1	0	0

Encoding Numerical Features
Binning

Each bin has a pre-fixed range of values which should be assigned to that bin on the basis of some domain knowledge, rules or constraints.



External Data

A Feature could also be created through the use of a foreign dataset; By finding some relationship between a column(s) in your dataset and the column(s) of a foreign dataset.

External Data

Local Data

median_age	married	college_grad	property_tax	insurance	median_school	num_schools	tx_year	city
33.0	65.0	84.0	234.0	81.0	9.0	3.0	2013	San Diego
39.0	73.0	69.0	169.0	51.0	3.0	3.0	2006	Los Angeles
28.0	15.0	86.0	216.0	74.0	8.0	3.0	2012	San Francisco
36.0	25.0	91.0	265.0	92.0	9.0	3.0	2005	Sacramento
37.0	20.0	75.0	88.0	30.0	9.0	3.0	2002	La Jolla

External Data

	city	population
0	San Diego	1.4 million
1	Los Angeles	3.9 million
2	San Francisco	800 thousand
3	Sacramento	495 thousand
4	La Jolla	46 thousand

Removed Unused Features

Unused features are those that don’t make sense to pass into our machine learning algorithms. Examples include:

- ID columns
- Features that wouldn't be available at the time of prediction
- Other text descriptions

Redundant features would typically be those that have been replaced by other features that you’ve added during feature engineering.