Feature Engineering

What is Feature Engineering?

Feature engineering is the process of using domain knowledge to extract features from raw data via data mining techniques. These features can be used to improve the performance of machine learning algorithms.

OneHot Encoding

- One-of-K encoding on an array of length K.
- Sparse format is memory-friendly

Onehot encoding

```
Sample: ["BR"]

country      country=NL country=BR country=US
-------      ---------- ---------- ----------
NL       => [          0,         1,         0]
BR
US
```

Label encoding

- Give every categorical variable a unique numeric ID
- Does not increase dimensionality

Label encoding

```
Sample: ["Queenstown"]

city                 city
-----------          ----
Cherbourg               1
Queenstown    =>        2
Southampton             3
```

Consolidation encoding

- **Map different categorical variables to the same variable**
- Spelling errors, slightly different job descriptions, full names vs. abbreviations
- Real data is messy, free text especially so

Expansion encoding

company_desc		desc1	company_desc2
Shell		Shell	Gas station
shel		Shell	Gas station
SHELL		Shell	Gas station
Shell Gasoline		Shell	Gas station
BP	=>	BP	Gas station
British Petr.		BP	Gas station
B&P		BP	Gas station
BP Gas Station		BP	Gas station
bp		BP	Gas station
Procter&Gamble		P&G	Manufacturer

Rounding

age		age1	age2
23.6671		23	2
23.8891		23	2
22.1261	=>	22	2
19.5506		19	1
18.2114		18	1

n=165	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	