Data Science
Study Group

Feature engineering

OUARDINI OUSSAMA
@oussamaouardini
Data & AI Engineering Student

Feature engineering is an informal topic, but one that is absolutely known and agreed to be key to success in applied machine learning

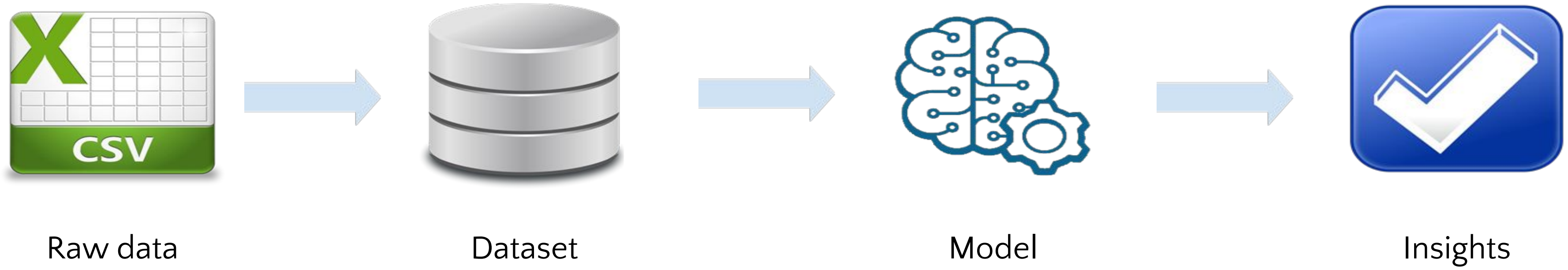
– Jason Brownlee

“

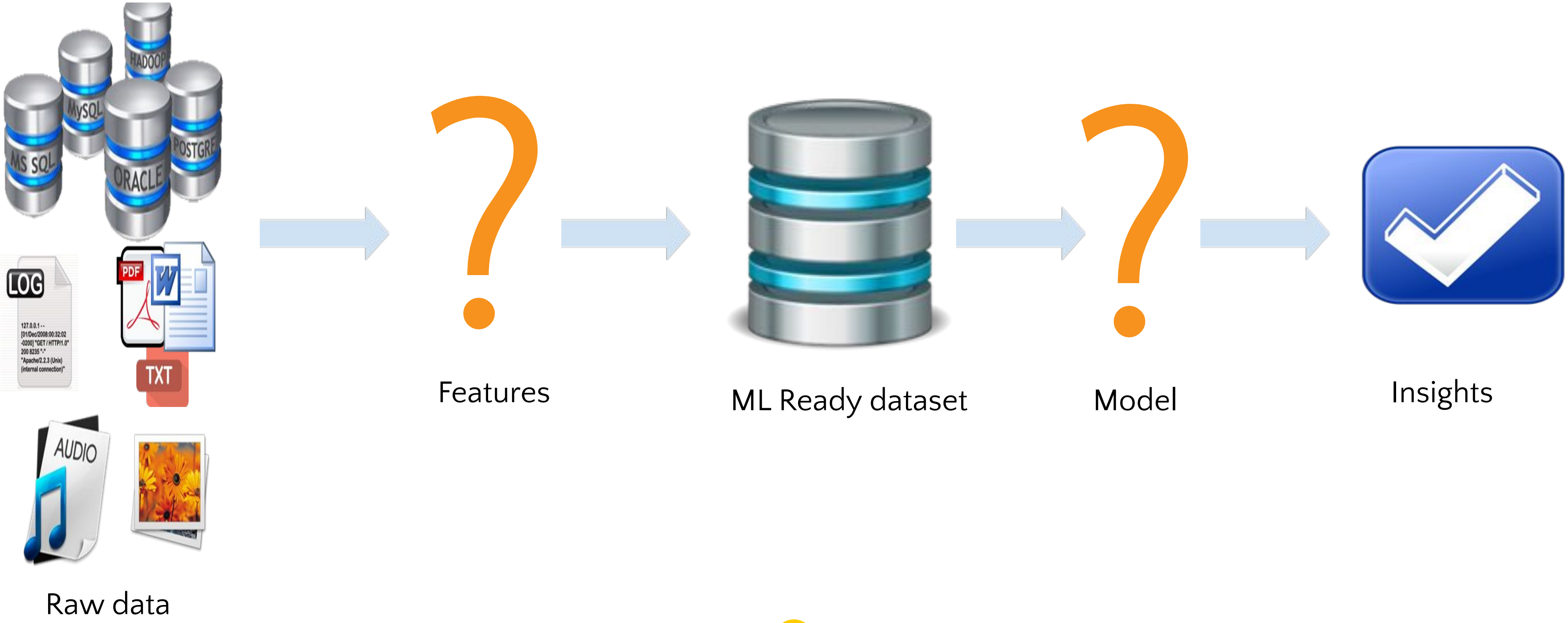
Coming up with features is difficult,
time-consuming,
requires expert knowledge.
'Applied machine learning' is basically
feature engineering

- Andrew Ng

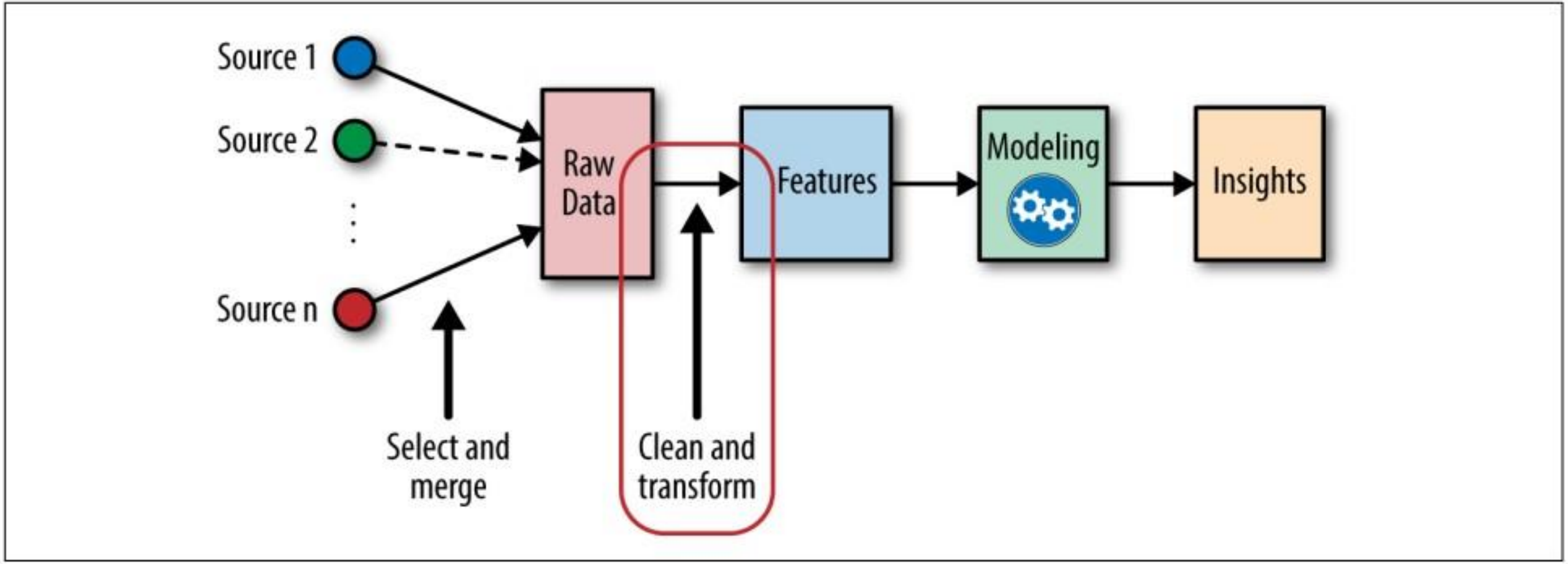
The Dream...



The Reality...



HADOOP
MySQL
MS SQL
ORACLE
POSTGRE
LOG
PDF
TXT
AUDIO
Raw data
Features
ML Ready dataset
Model
Insights


Source 1
Source 2
Source n
Raw Data
Features
Modeling
Insights
Select and merge
Clean and transform

How do we get the most out of our data for predictive modeling?

This is the problem that the process and practice of feature engineering solves.

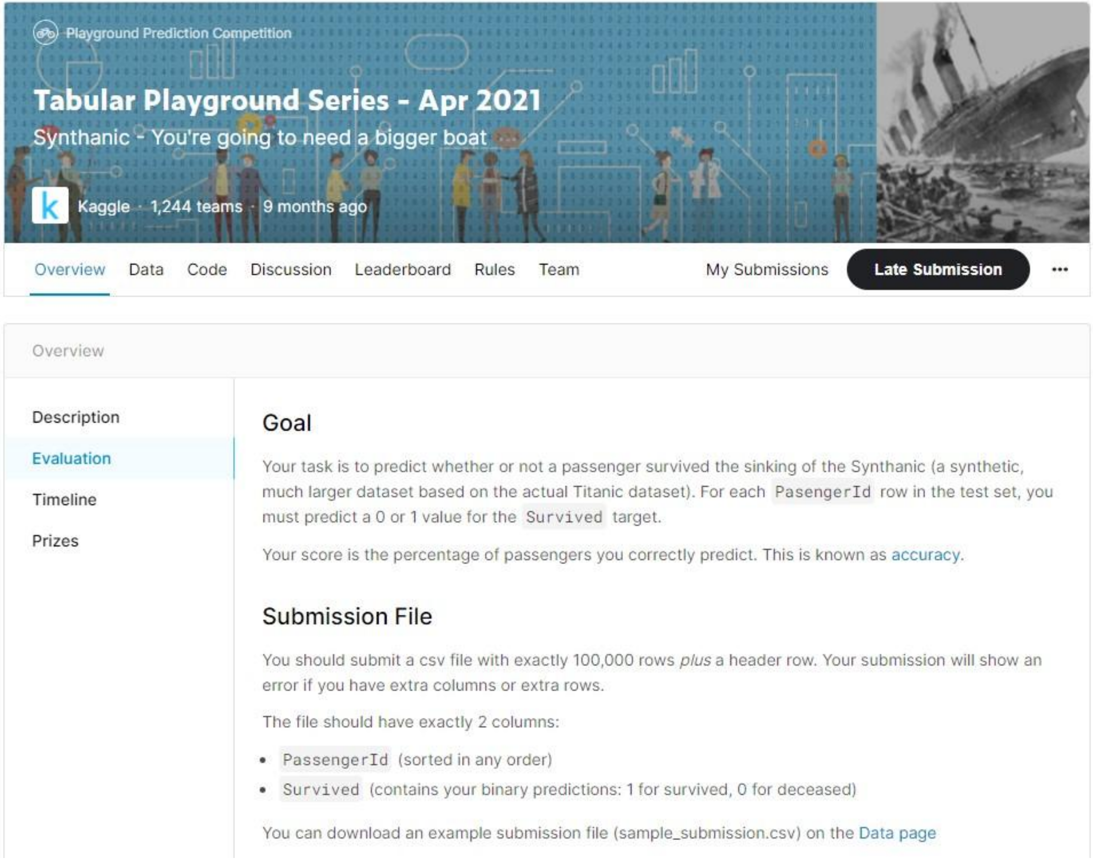
Actually the success of all Machine Learning algorithms depends on how you present the data. - Mohammad Pezeshk

Hmm, But How ???

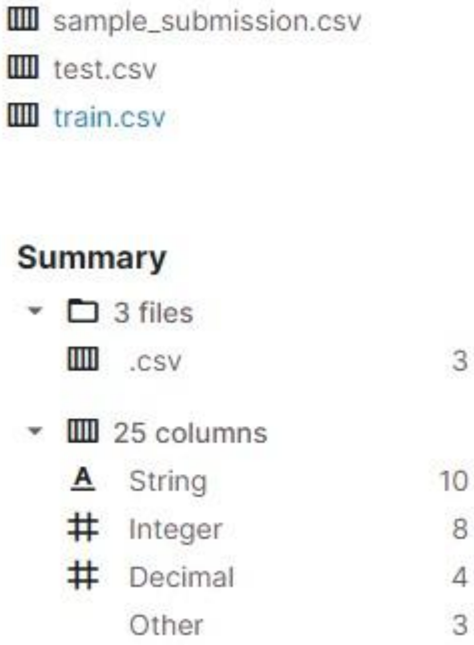
Let’s see some Feature Engineering techniques for your Data Science toolbox...

Case study

Titanic survival prediction – Kaggle competition



Dataset



Titanic survival prediction – Kaggle competition

978	swatanabe1234		0.78542	3	9mo
979	Joseph Hoon		0.78537	3	9mo
980	Phuchit Chadasin		0.78537	4	9mo
981	HarieshJay		0.78533	1	9mo
982	Betesss		0.78533	1	9mo
983	Bruno Leite		0.78529	14	9mo
984	Thoughtful Monkey		0.78529	2	9mo
985	ike		0.78525	3	9mo
986	Ahmed Kachkach	</> Titanic - live stream.	0.78521	3	9mo
987	Oussama Ouardini		0.78521	1	9mo

Your First Entry ↑

Welcome to the leaderboard!

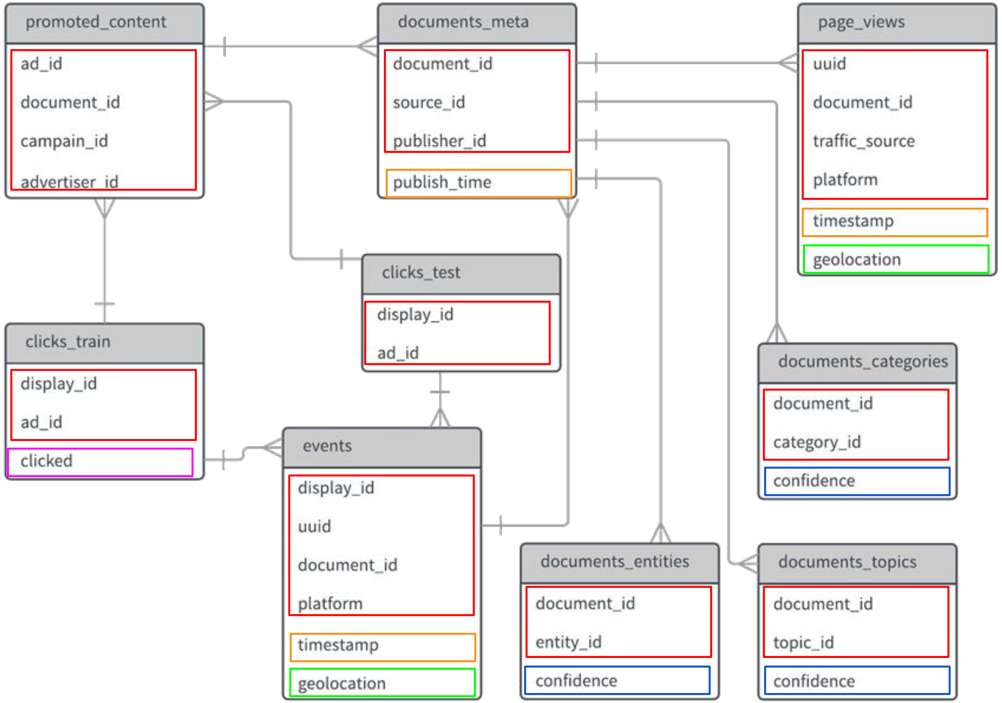
988	Thomas Ocallaghan	0.78509	3	9mo
989	Atharv Chaudhari	0.78509	12	9mo
990	[Deleted] dd42c536-fbbf-4a8...	0.78505	2	9mo
991	AdarGalili	0.78505	1	9mo
992	Marília Prata	0.78505	1	9mo
993	Grayson Felt	0.78505	2	9mo
994	tawa7029	0.78505	5	9mo
995	Nerwosolek	0.78505	5	9mo

I got 987th position
From about
1244 competitors :(

Mostly due to
Feature Engineering
techniques

First of all ... a closer look at your data

What does the data model look like?



Categorical

Temporal

Spacial

Numerical

Target

Data Cleaning : process of detecting and correcting corrupt or inaccurate records

Name	Date	Duration (s)	Genre	Plays
Highway star	1984-05-24	-	Rock	139
Blues alive	1990/03/01	281	Blues	239
Lonely planet	2002-11-19	5:32s	Techno	42
Dance, dance	02/23/1983	312	Disco	N/A
The wall	1943-01-20	218	Reagge	83
Offside down	1965-02-19	4 minutes	Techno	895
The alchemist	2001-11-21	418	Bluesss	178
Bring me down	18-10-98	328	Classic	21
The scarecrow	1994-10-12	269	Rock	734

Original Data

Name	**Date**	**Duration (s)**	**Genre**	**Plays**
Highway star	1984-05-24		Rock	139
Blues alive	1990-03-01	281	Blues	239
Lonely planet	2002-11-19	332	Techno	42
Dance, dance	1983-02-23	312	Disco	
The wall	1943-01-20	218	Reagge	83
Offside down	1965-02-19	240	Techno	895
The alchemist	2001-11-21	418	Blues	178
Bring me down	1998-10-18	328	Classic	21
The scarecrow	1994-10-12	269	Rock	734

Cleaned Data

Feature Engineering

1

Numerical Features

Let’s start with the first set of slides

Imputation for missing values

- Datasets contain missing values, often encoded as blanks, NaNs or other placeholders
- Ignoring or deleting rows and/or columns with missing values is possible, but at the price of losing data which might be valuable (Not recommended if data is too small)
- Better strategy is to infer them from the known part of data
- Strategies
  - **Mean**: Basic approach
  - **Median**: More robust to outliers
  - **Mode**: Most frequent value
  - **Using a model** (Predicting missing values of Data by Linear Regression Model): Can expose algorithmic bias

Binarization

- Transform discrete or continuous numeric features in binary features

Example: Number of user views of the same document

document_id	uuid	views_count
25792	6d82e412aa0f0d	8
25792	571016386ffee7	6
25792	6a91157d820e37	6
25792	ad45fc764587b0	6
25792	a743b03f2b8ddc	3

document_id	uuid	viewed
25792	6d82e412aa0f0d	1
25792	571016386ffee7	1
25792	6a91157d820e37	1
25792	ad45fc764587b0	1
25792	8d87becfb35857	1
25792	abcdefg1234567	0

```
[1]  from sklearn import preprocessing
     ✓ 0.1s

     X = [
         [1., 8, 2.],
         [2., 0, 0.],
         [0, 1, -1]
     ]
[2]  ✓ 0.6s

     binarizer = preprocessing.Binarizer(threshold=1.0)
[3]  ✓ 0.4s

     binarizer.transform(X)
[4]  ✓ 0.6s
...  array([[0., 1., 1.],
            [1., 0., 0.],
            [0., 0., 0.]])
```

Log transformation

- Compresses the range of large numbers and expand the range of small numbers.
  Eg. The larger x is, the slower log(x) increments

user_id	views_count
a	1000
b	500
c	300
d	200
e	150
f	100
g	70
h	50
i	30
j	20
k	10
l	5
m	1

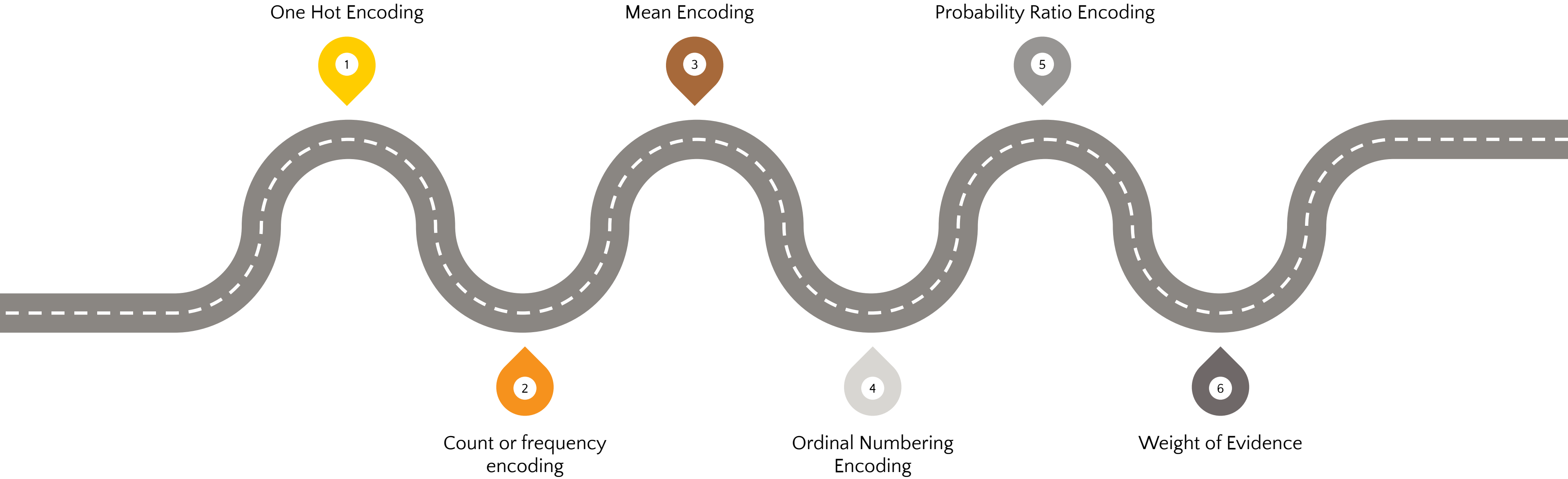
log(1+views_count)
6.91
6.22
5.71
5.30
5.02
4.62
4.26
3.93
3.43
3.04
2.40
1.79
0.69

Feature Engineering

2

Categorical Features

Categorical Encoding



3

Feature Engineering

Temporal Features

Time binning

- Apply binning on time data to make it categorial and more general.
  Binning a time in hours or periods of day, like below.

Hour range	Bin ID	Bin Description
[5, 8]	1	Early Morning
[8, 11]	2	Morning
[11, 14]	3	Midday
[14, 19]	4	Afternoon
[19, 22]	5	Evening
[22, 00] and [00, 5]	6	night

Feature Engineering

4

Textual Features

Natural Language Processing

Cleaning
- Lowercasing
- Convert accented characters
- Removing non-alphanumeric
- Repairing

Tokenizing
- Encode punctuation marks
- Tokenize
- N-Grams
- Skip-grams
- Char-grams
- Affixes

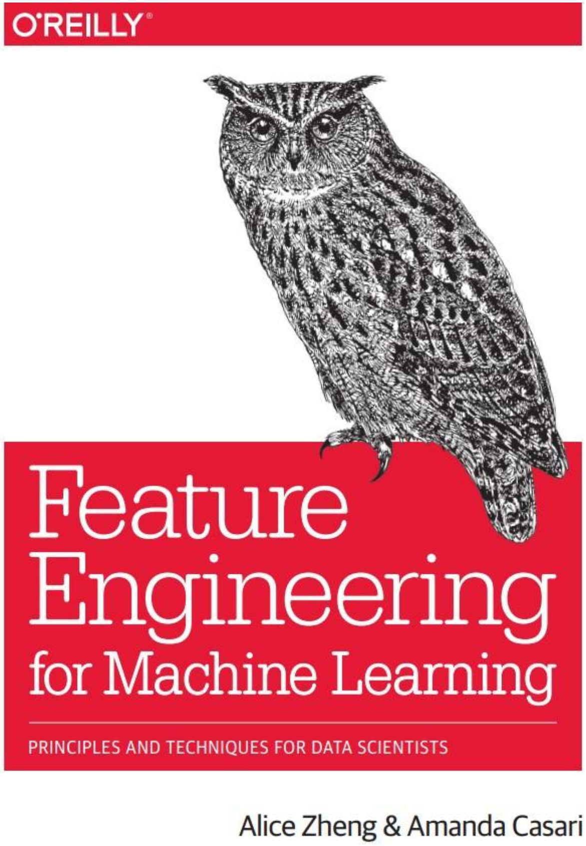
Removing
- Stopwords
- Rare words
- Common words

Roots
- Spelling correction
- Chop
- Stem
- Lemmatize

Enrich
- Entity Insertion / Extraction
- Parse Trees
- Reading Level

5 Practical Session

Références



Discover Feature Engineering, How to Engineer Features and How to Get Good at It

Scikit-learn

Feature Engineering for Machine Learning

Thanks!

Any questions ?

You can find me at

- @oussamaouardini
- ouss.ouardini@gmail.com