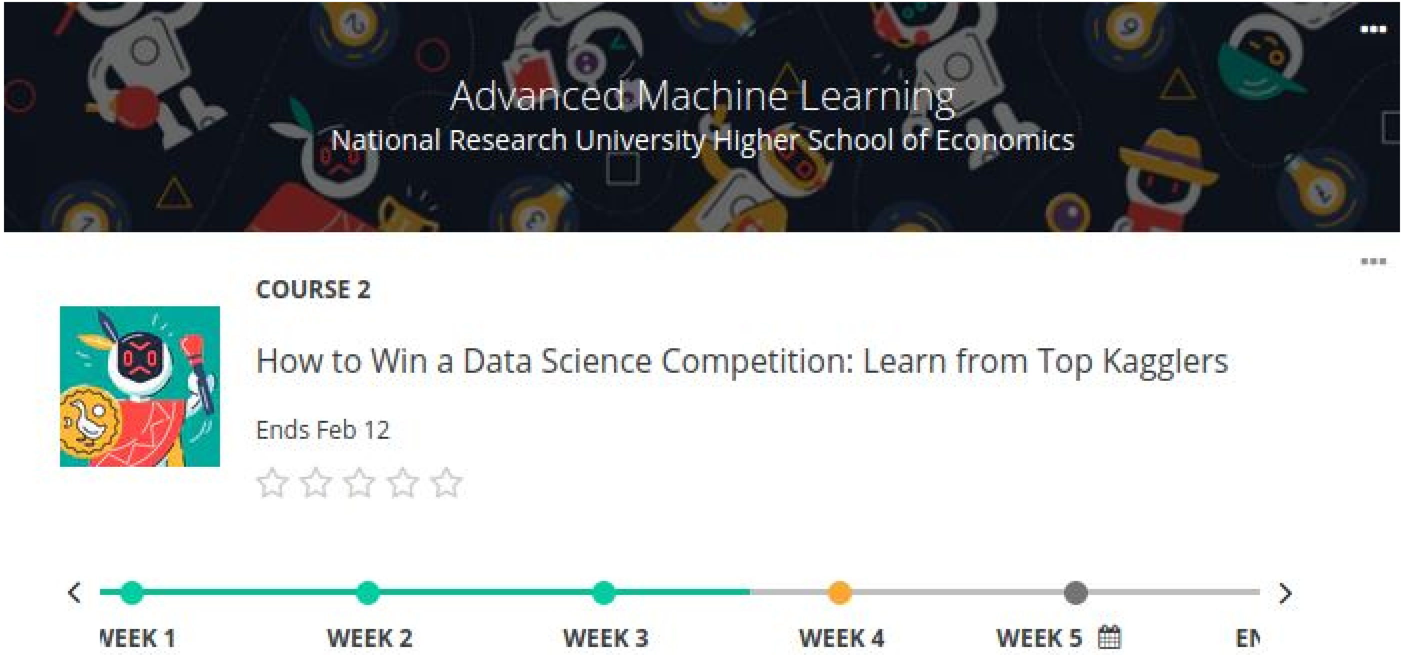
Feature Preprocessing and Generation with Respect to Models Part 2

Cinyoung Hur
Seoul AI

2018.01.06

I’m taking this course



Advanced Machine Learning

National Research University Higher School of Economics

COURSE 2

How to Win a Data Science Competition: Learn from Top Kagglers

Ends Feb 12

☆☆☆☆☆

disclaimer: This material is based on coursera course https://www.coursera.org/learn/competitive-data-science

Preprocessing issues

Feature types

Date and time

1. Periodicity
    a. Day number in week, month, season, year, second, minute, hour

Date and time

1. Periodicity
    a. Day number in week, month, season, year, second, minute, hour
2. Time since
    a. Row-independent moment
    E.g. since 00:00:00 UTC, 1 January 1970
    b. Row-dependent important moment
    E.g. Number of days left until next holidays
    Time passed after last holiday

Date and time

1. Periodicity
    a. Day number in week, month, season, year, second, minute, hour
2. Time since
    a. Row-independent moment
    E.g. since 00:00:00 UTC, 1 January 1970
    b. Row-dependent important moment
    E.g. Number of days left until next holidays
    Time passed after last holiday
3. Difference between dates
    a. Datetime_feature_1 - datetime_feature_2

Periodicity & Time since

Date	sales
01.01.14	1213
02.01.14	938
03.01.14	2448
04.01.14	1744
05.01.14	1732
06.01.14	1022

Periodicity & Time since

Date	weekday	daynumber_since_year_2014	is_holiday	days_till_holidays	sales
01.01.14	5	0	True	0	1213
02.01.14	6	1	False	3	938
03.01.14	0	2	False	2	2448
04.01.14	1	3	False	1	1744
05.01.14	2	4	True	0	1732
06.01.14	3	5	False	9	1022

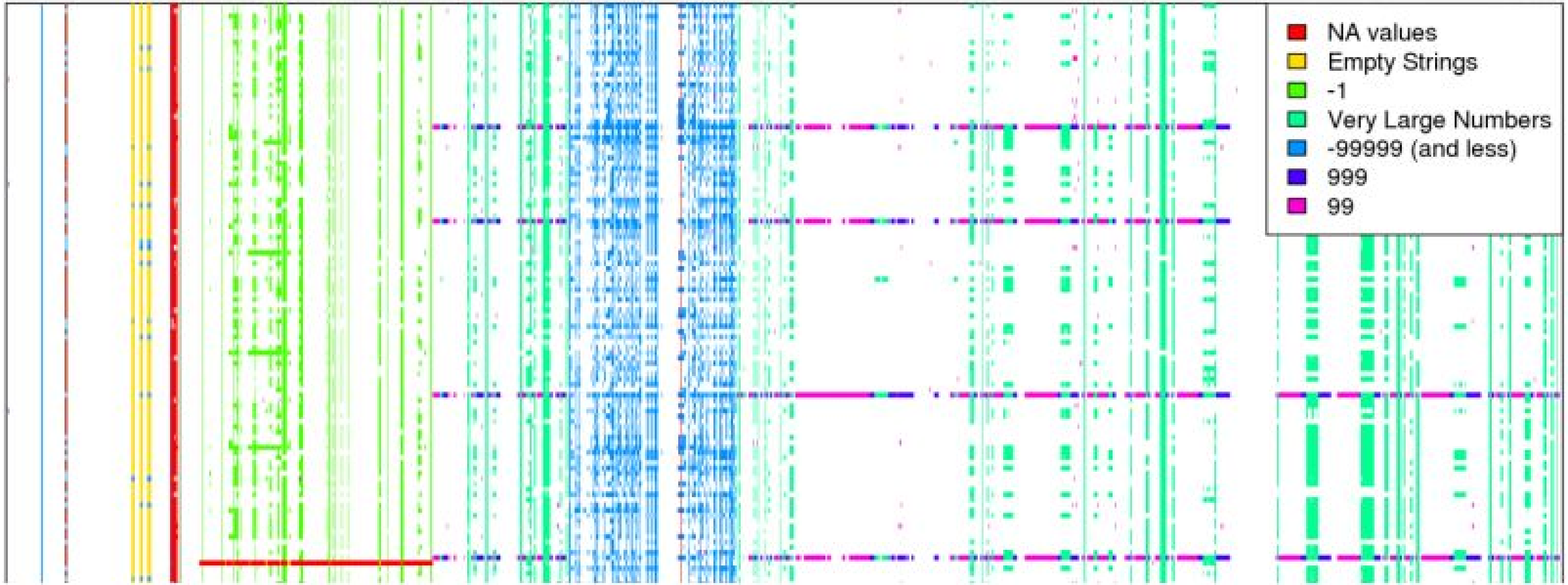
Difference between dates

user_id	registration_date	last_purchase_date	last_call_date	date_diff	churn
14	10.02.2016	21.04.2016	26.04.2016	5	0
15	10.02.2016	03.06.2016	01.06.2016	-2	1
16	11.01.2017	11.01.2017	12.01.2017	1	1
20	06.11.2016	06.11.2016	08.02.2017	94	0

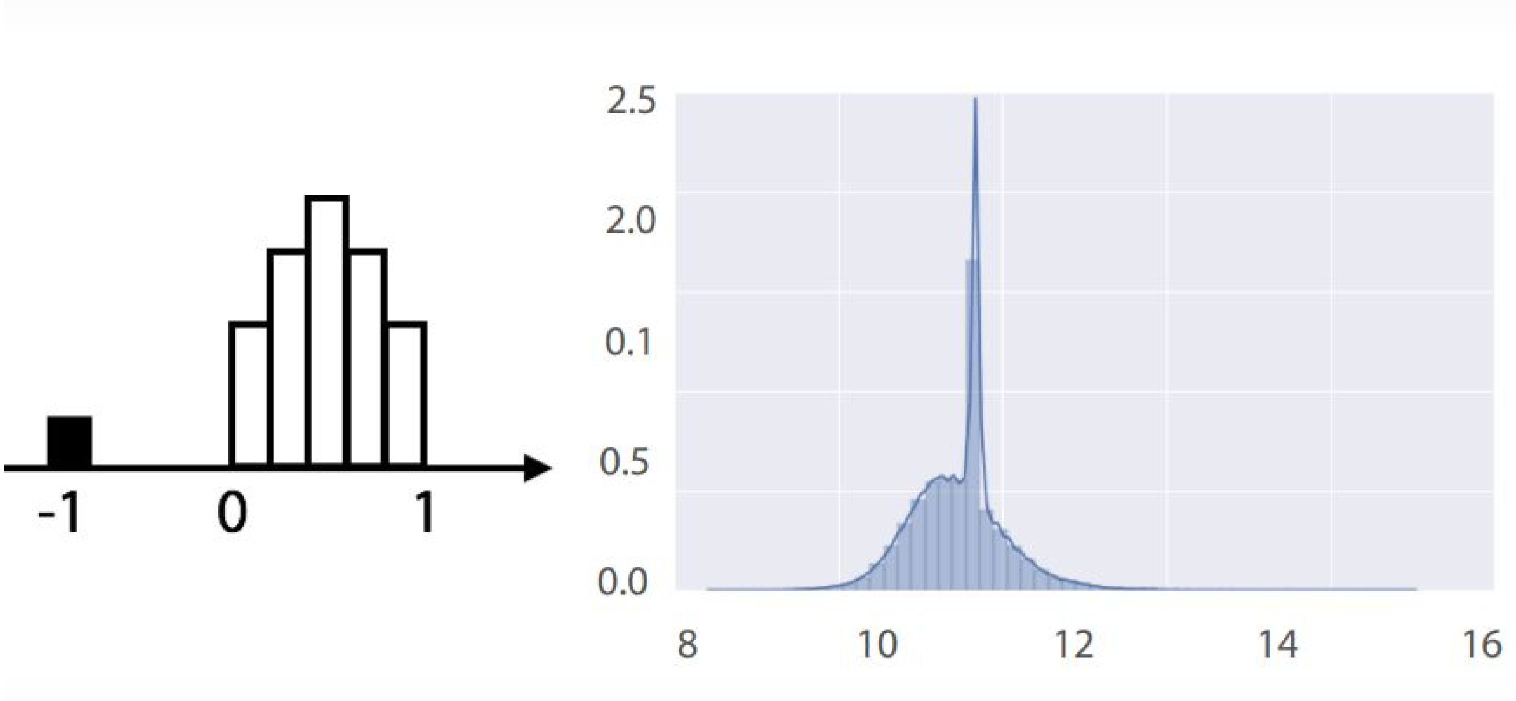
Preprocessing issues

Missing values

Missing data, numeric



Hidden NaNs

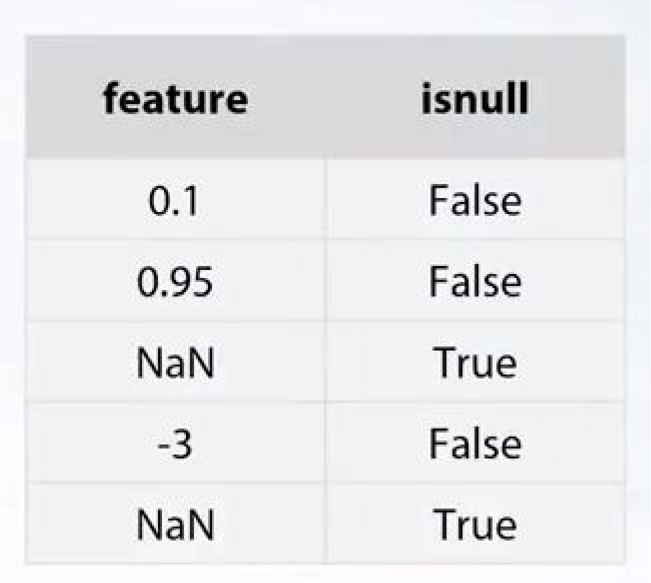


Fillna approaches

1. -999, -1, etc
2. Mean, median
3. Reconstruct value

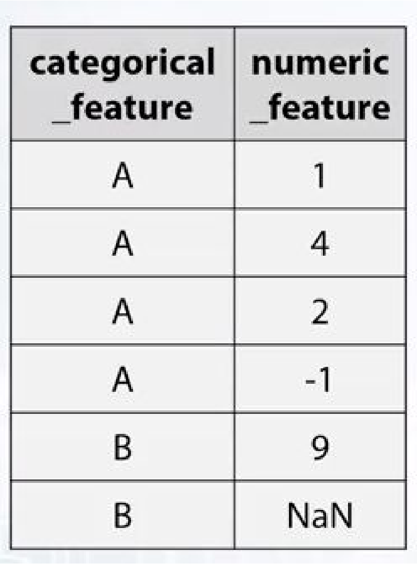
Feature generation: Missing values

“Isnull” feature



feature	isnull
0.1	False
0.95	False
NaN	True
-3	False
NaN	True

Missing values: fillna

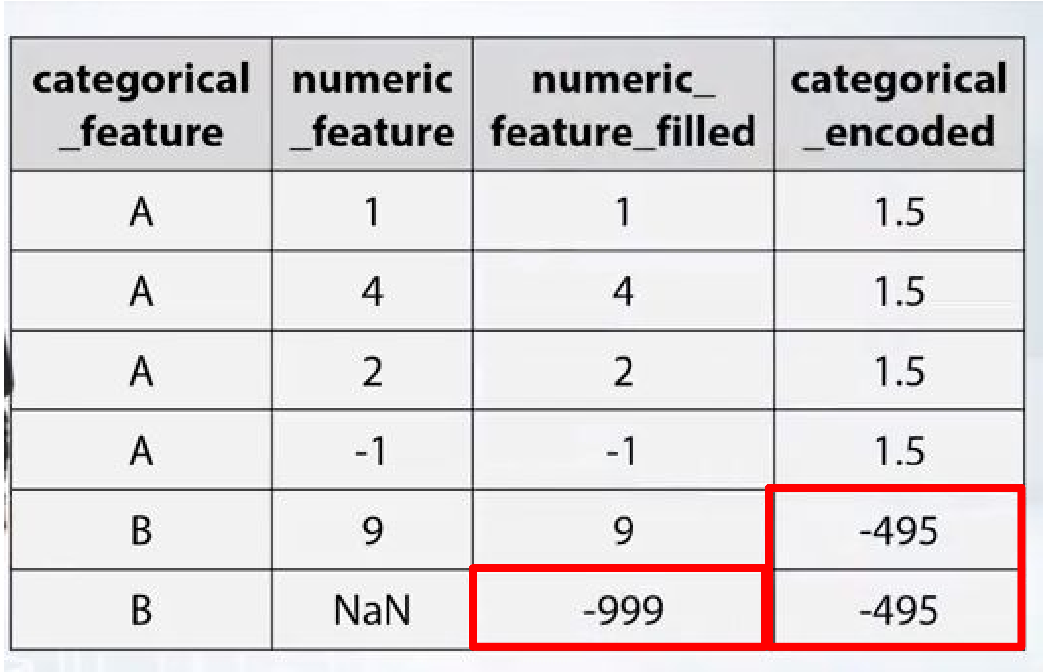


categorical _feature	numeric _feature
A	1
A	4
A	2
A	-1
B	9
B	NaN

Missing values: fillna

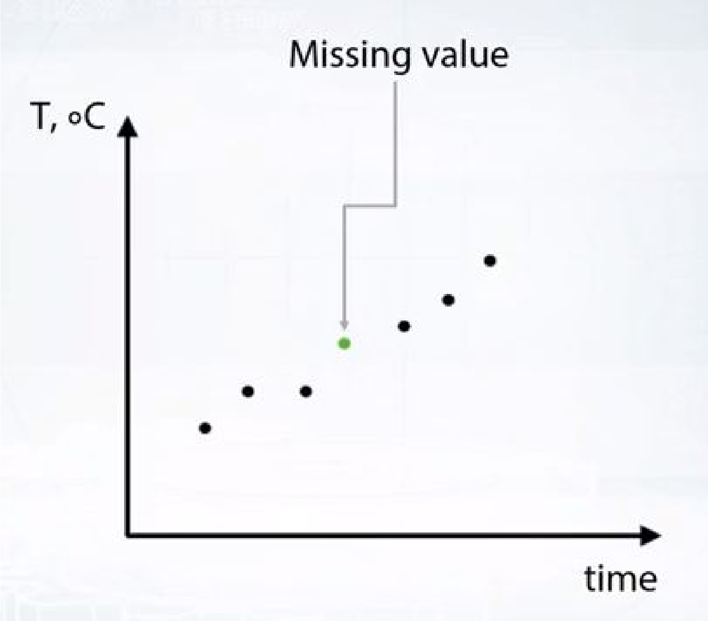
For categorical features: mean/median

Missing values: fillna

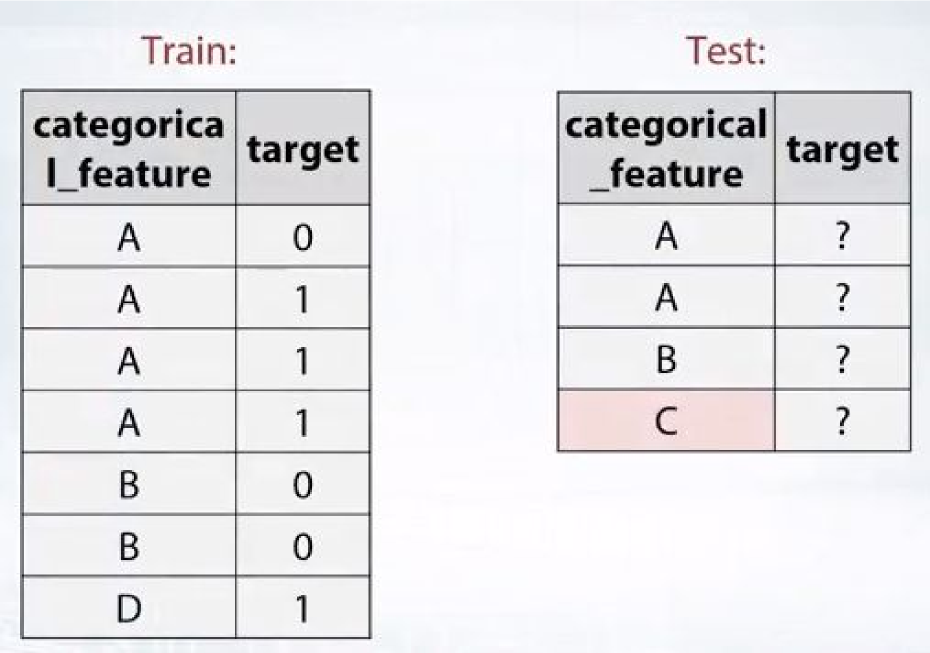


categorical _feature	numeric _feature	numeric_ feature_filled	categorical _encoded
A	1	1	1.5
A	4	4	1.5
A	2	2	1.5
A	-1	-1	1.5
B	9	9	-495
B	NaN	-999	-495

Missing values: fillna with reconstruct value



Missing value: value only exists in test set



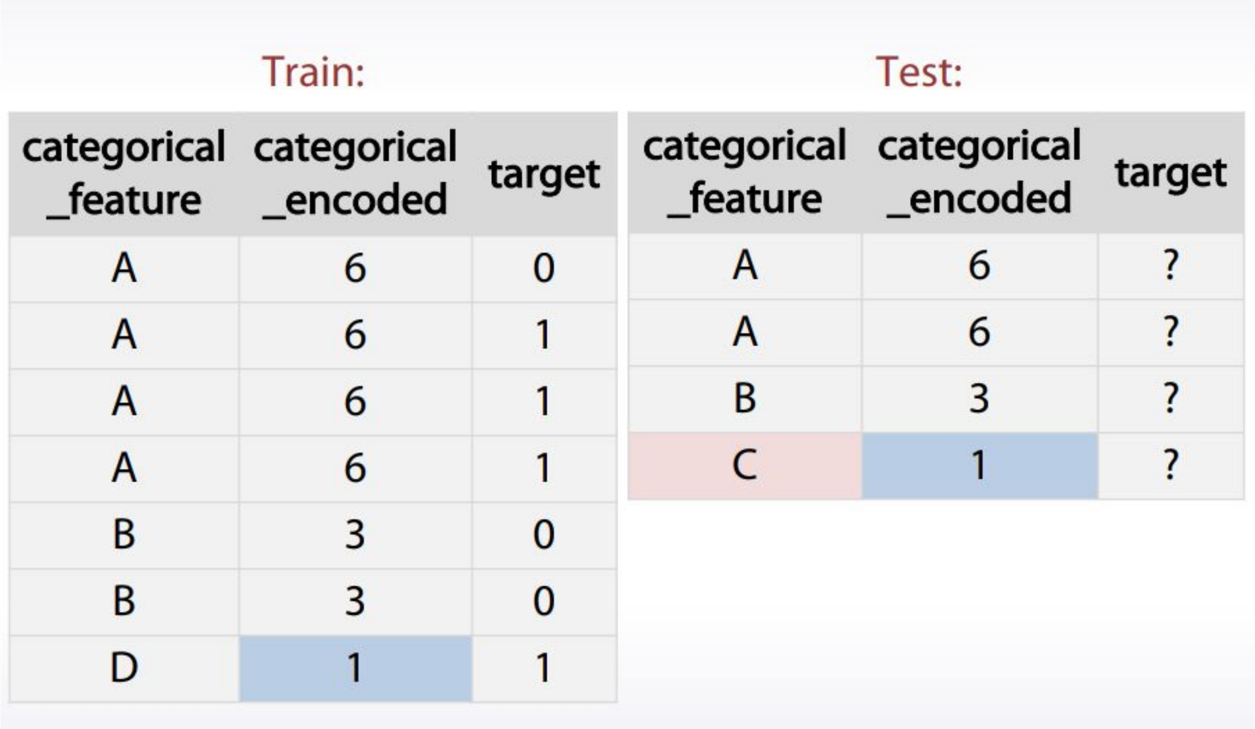
Train:

categorical_feature	target
A	0
A	1
A	1
A	1
B	0
B	0
D	1

Test:

categorical_feature	target
A	?
A	?
B	?
C	?

Missing value: value only exists in test set



Train:

categorical _feature	categorical _encoded	target
A	6	0
A	6	1
A	6	1
A	6	1
B	3	0
B	3	0
D	1	1

Test:

categorical _feature	categorical _encoded	target
A	6	?
A	6	?
B	3	?
C	1	?

References

- Feature preprocessing
    - Preprocessing in Sklearn
    - Andrew NG about gradient descent and feature scaling
    - Feature Scaling and the effect of standardization for machine learning algorithms
- Feature generation
    - Discover Feature Engineering, How to Engineer Features and How to Get Good at It
    - Discussion of feature engineering on Quora