



Case study: Missing Value Imputation

IKMLab



Contents

- 1. What is missing values?How to identify missing values?**
- 2. Quick classification of missing data**
- 3. Different Methods to Handle Missing Values**
 - a. Deletion Methods**
 - b. Imputation Methods**

What is missing values ?

How to identify missing values?

What is missing values ?

- **Missing values**, is where some of the observations in a dataset are blank , NaN , -999,or any other placeholder.
- In the example below, the first and third rows contain missing data. First and third rows have missing values for Cabin, and represent for NaN.

	0	1	2		3	4	5	6	7		8	9	10	11
0	PassengerId	Survived	Pclass		Name	Sex	Age	SibSp	Parch		Ticket	Fare	Cabin	Embarked
1	1	0	3		Braund, Mr. Owen Harris	male	22	1	0		A/5 21171	7.25	NaN	S
2	2	1	1		Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38	1	0		PC 17599	71.2833	C85	C
3	3	1	3		Heikkinen, Miss. Laina	female	26	0	0		STON/O2. 3101282	7.925	NaN	S



How to identify missing values?

We can check for missing values in a dataset using **pandas** function as:

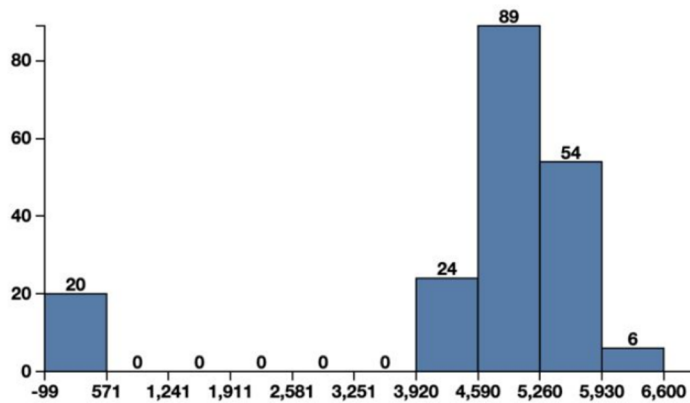
```
df.isnull().sum()
```

```
PassengerId      0
Survived          0
Pclass           0
Name             0
Sex              0
Age             177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin           687
Embarked         2
dtype: int64
```

How to identify missing values?

Also, Histograms are a great tool to find the placeholder character, if any.

Histogram



- In this example, we see that most values fall between 3900 and 6600. The value -99 looks rather displaced and, in this case, could be a placeholder for missing values.

Quick classification of missing data



Three different types of missing values

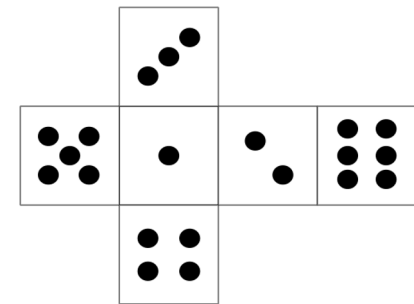
1. **MCAR:** Missing Completely At Random.
2. **MAR:** Missing At Random.
3. **MNAR:** Missing Not At Random.

1. random

MCAR: Missing Completely At Random.

- The probability of an instance being missing does not depend on known values or the missing value itself.
- *Example:* In a questionnaire, the respondent decide whether or not answer the question by throwing a dice.

gender	age	job	weight
male	18	student	100
female	no answer	model	39
no answer	4	baby	20



2. 某些欄位會跟隨別的欄位有缺失值ex: 30 歲女性不填體重

MAR: Missing At Random.

- The probability of an instance being missing may depend on known values but not on the missing value itself.
- *Example:* In a questionnaire, we find women wouldn't answer the weight when their ages are under 30.

gender	age	job	weight
female	18	student	no answer
female	65	president	62
female	24	software engineer	no answer

3. 沒有random的特性，ex: 設計有問題，這個屬性很容易產生缺失值

MNAR: Missing Not At Random

- the probability of an instance being missing could depend on the value of the variable itself.
- Example:* In a questionnaire, we cannot figure out why the jobs feature has NA about 60%.

gender	age	job	weight
male	18	student	100
female	41	no answer	39
male	4	no answer	20
female	75	retire	60
male	25	no answer	80



Note

- Imputing **NMAR** missing values is more complicated, since additional factors to just statistical distributions and statistical parameters have to be taken into account.
- Only the knowledge of the data collection process and the business experience can tell whether the missing values we have found are of type MAR, MCAR, or NMAR.

Different Methods to Handle Missing Values

1. 拿掉，不要亂補可能造成更多問題。
評估要丟掉的Data分佈跟完整屬性的分布像不像，項就丟掉



Deletion Methods

- In this method, cases which have missing values for one or more features are deleted. If the cases having missing values are **small** in number, it is better to drop them.
- Though this is an easy approach, it might lead to a significant **decrease** in the sample **size**. Also, the data may not always be missing completely at random. This may lead to **biased estimation of parameters**.



Deletion Methods

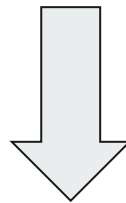
There are three common deletion approaches: **listwise deletion**, **pairwise deletion**, and **dropping features**.

- **Listwise Deletion:** Delete all rows where one or more values are missing.
- **Pairwise Deletion:** Delete only the rows that have missing values in the columns used for the analysis. It is only recommended to use this method if the missing data are **MCAR**.
- **Dropping Features:** Drop entire columns with more missing values than a given threshold, e.g. 60%.

只要有欄位是沒有的就把那一筆直接丟掉

Listwise Deletion Example

F1	F2	F3	F4
12	11	535	1
		777	0
15		987	



F1	F2	F3	F4
12	11	535	1

看哪些屬性對我最後是沒用的ex: 只是填個人資料，內容無關

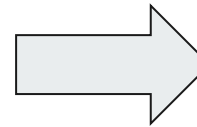
Pairwise Deletion

- A table with missing values, where only F1, F2, and F3 are used in the analysis.

F1	F2	F3	F4
12	11	535	1
		777	0
15		987	
17	3	689	

Used

Not Used

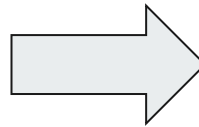


F1	F2	F3	F4
12	11	535	1
17	3	689	

Dropping Features Example

- Drop entire columns with more missing values than a given threshold, for 60%.

F1	F2	F3	F4
12	11	535	1
		777	0
15		987	



F1	F3	F4
12	535	1
	777	0
15	987	

Imputation Methods

自己去猜這裡沒有填可能是什麼值，猜錯會產生更多的noise



Imputation Methods

- The idea behind the imputation approach is to replace missing values with other sensible values.
- As you always lose information with the deletion approach when dropping either samples (rows) or entire features (columns), **imputation is often the preferred approach.**
- The many imputation techniques can be divided into two subgroups: **univariate imputation** or **multivariate imputation.**



Imputation Methods

- **Univariate imputation:** a single / one imputation value for each of the missing observations is generated.

You can try by yourself : [sklearn.impute.SimpleImputer](#)

- **Multivariate imputation:** many imputed values for each of the missing observations are generated. This means many complete datasets with different imputed values are created. The analysis (e.g. training a linear regression to predict a target column) is performed on each of these datasets and the results are polled.

You can try by yourself : [sklearn.impute.IterativeImputer](#)

Univariate Imputation

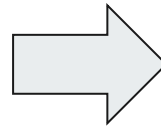
—

有缺就補某一個值：補眾數...

Constant Value Imputation

- Constant value imputation is a general method that works for all data types and consists of substituting the missing value with a fixed value.

F1	F2	F3	F4
12	11	535	1
		777	0
15		987	



F1	F2	F3	F4
12	11	535	1
2	2	777	0
15	2	987	2

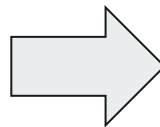
補眾數、平均值、中位數

要先看一下後面的值正不正常，不然可能把outlyar變成一般的值

Mean/Median Value Imputation

- common imputation methods for numerical features are mean, or median imputation. In this case, the method substitutes the missing value with the mean, or the median value calculated for that feature on the whole dataset.

F1	F2	F3	F4
12	11	535	1
		777	0
15		987	

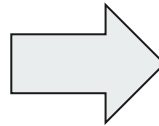


F1	F2	F3	F4
12	11	535	1
13.5	11	777	0
15	11	987	0.5

Most Frequent Value Imputation

- Another common method that works for both numerical and nominal features uses the most frequent value in the column to replace the missing values.

F1	F2	F3	F4
12	11	535	1
		777	0
15		987	
12	5	580	1



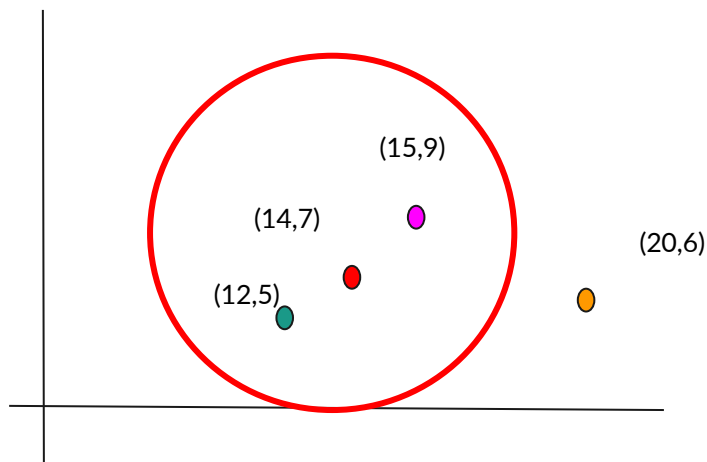
F1	F2	F3	F4
12	11	535	1
12	5	777	0
15	5	987	1
12	5	580	1

精神：把資料用的糊一點

K Nearest Neighbors Imputation

- This method uses k-nearest neighbour algorithms to estimate and replace missing data. The k-neighbours are chosen using some distance measure and their average is used as an imputation estimate.

F1	F2	F3
12	11	5
14	10.5	7
15	10	9
20	12	6



- For example $K = 2$, find the 2 nearest neighbors and get the average of their F2 features.

Multivariate Imputation

—



Multiple Imputation by Chained Equations (MICE)

- Multiple Imputation by Chained Equations (MICE) is a robust, informative method for dealing with missing values in datasets.
- MICE operates under the assumption that the missing data are **Missing At Random (MAR)** or **Missing Completely At Random (MCAR)** .
- The procedure is an extension of the single imputation procedure by “Missing Value Prediction” (eg. KNN or linear regression): this is step 1. However, there are two additional steps in the MICE procedure.



MICE procedure

step 1: A simple **imputation**, such as imputing the mean, is performed for every missing value in the dataset. These mean imputations can be thought of as “place holders.”

step 2: The “place holder” mean imputations for one variable (“var”) are set back to missing.

step 3: “var” is the dependent variable in a regression model and all the other variables are independent variables in the regression model. The missing values for “var” are then replaced with predictions (e.g. Bayesian linear regression) from the regression model.



MICE procedure

Step 4: Moving on to the next variable with the next fewest missing values, steps 2–4 are then repeated for each variable that has missing data.


Step 5: Steps 2 through 4 are repeated for a number of cycles, with the imputations being updated at each cycle. In the end of the cycles, the distribution of the parameters governing the imputations (e.g., the coefficients in the regression models) should have converged in the sense of becoming stable.

把其中一筆先拿掉，根據其他回歸的結果再填回這一筆缺失的值

MICE procedure

age	salary	gender
33	NA	female
18	12000	NA
NA	13542	male

mean
imputation




age	salary	gender
33	12771	female
18	12000	female
25.5	13542	male

age back to NA



age	salary	gender
33	12771	female
18	12000	female
NA	13542	male

Bayesian Linear
Regression
 $\text{age} \sim \text{salary}, \text{gender}$



age	salary	gender
33	12771	female
18	12000	female
NA	13542	male

MICE procedure

Bayesian
Linear Reg.
Predict age



age	salary	gender
33	12771	female
18	12000	female
35.3	13542	male

salary back
to NA



age	salary	gender
33	NA	female
18	12000	female
35.3	13542	male

Bayesian Linear
Regression
salary ~ age,gender



age	salary	gender
33	NA	female
18	12000	female
35.3	13542	male

Bayesian
Linear Reg.
Predict salary



age	salary	gender
33	13103	female
18	12000	female
35.3	13542	male

MICE procedure

...the same for gender



age	salary	gender
33	13103	female
18	12000	male
35.3	13542	male

repeat until the model is stable



- The observed data and the final set of imputed values would then constitute one “complete” data set.



Conclusion

- Use listwise deletion (“deletion”) carefully, especially on small datasets. When removing data, you are removing information. Not all datasets have redundant information to spare!
- When using fixed value imputation, you need to know what that fixed value means in the data domain and in the business problem. Here, you are injecting arbitrary information into the data, which can bias the predictions of the final model.
- If you want to impute missing values without prior knowledge it is hard to say which imputation method works best, as it is heavily dependent on the data itself.