Day001:

談談機器學習中的Evaluation Metrics

參考網站:

<https://blog.csdn.net/aws3217150/article/details/50479457>

了解常用的評價指標

機器學習中的回歸(regression)與分類(classification)問題

參考網站:

<https://blog.csdn.net/wspba/article/details/61927105>

了解回歸與分類可以應用在哪些方面

Day004:

What is Exploratory Data Analysis?

<https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15>

透過EDA(探索式資料分析)來初步了解的相關資訊並檢查資料是否有誤(ex:異常數值)

BookData(數據科學學習筆記):

<https://bookdata.readthedocs.io/en/latest/base/01_pandas.html>



Day005:

pandas\_exercises:

<https://github.com/guipsamora/pandas_exercises>

Day006:

label-encoder-vs-one-hot-encoder:

<https://medium.com/@contactsunny/label-encoder-vs-one-hot-encoder-in-machine-learning-3fc273365621>

Day007:

k- fold cross validation:

<https://zhuanlan.zhihu.com/p/24825503>

Day008:

Descriptive Statistics For pandas Dataframe:

<https://chrisalbon.com/python/data_wrangling/pandas_dataframe_descriptive_stats/>

Pandas的統計相關內容，裡面有max、min、describe...等EDA的相關操作

Pandas中的繪圖函數:

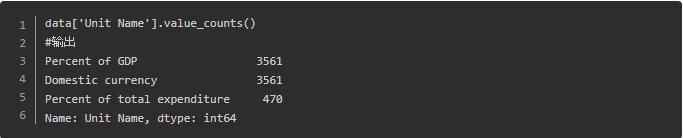
<https://amaozhao.gitbooks.io/pandas-notebook/content/pandas%E4%B8%AD%E7%9A%84%E7%BB%98%E5%9B%BE%E5%87%BD%E6%95%B0.html>

裡面有許多Python相關的教學，有空可以多看看。(中文)

Day009:

data['Unit Name'].value\_counts()

以Series形式返回指定列的不同取值的頻率



df.sort\_index() 按索引排序, 默認為生序

df.cumsum() 用於計算累加量

Day010:

調整離群值

df.clip(min, max)

低於min的數值全部調整為min, 高於max的數值全部調整為max

捨棄離群值

df = (df[‘1stfloor’] > 500) & (df[‘1stfloor’] < 2300)

只選擇大於500、小於2300區間內的資料

Day011:

計算分位數

np.percentile(a, 90) – 求a數列第90%分位的數值

計算眾數

from collections import defaultdict

範例:

mode\_dict = defaultdict(lambda:0) – 可以放自己想放的方法(例:list)

for value in app\_train[~app\_train['AMT\_GOODS\_PRICE'].isnull()]['AMT\_GOODS\_PRICE']:

mode\_dict[value] += 1

value\_most = sorted(mode\_dict.items(), key=lambda kv: kv[1], reverse=True)

print(value\_most[0]) – 第一個為最多的元組

參考文獻:

<https://ithelp.ithome.com.tw/articles/10193094>

透過作業解答學到pandas中value\_counts()的用法

可以用來查看有哪些不同的值, 並計算每個值有多少個重複值

參考文獻:

<https://www.jianshu.com/p/f773b4b82c66>

Day012:

延伸閱讀

數據缺失值處理:

<https://juejin.im/post/5b5c4e6c6fb9a04f90791e0c>

數據的標準化和歸一化:

<https://blog.csdn.net/pipisorry/article/details/52247379>