資料前處理：

1. 缺失值填補
2. 離群值處理
3. 標準化

定義目標：

1. 回歸？分類？
2. 預測目標是什麼？
3. 用什麼資料預測？
4. Training set/ Validation set/ Test set

評估標準：

1. 回歸問題：
   1. RMSE, Root Mean Square Error
   2. Mean Absolute Error
   3. R-Square
2. 分類問題：
   1. Accuracy
   2. F1-score
   3. AUC, Area Under Curve

根據設定目標建立機器學習模型：

數據分析流程：

1. 收集數據

---

1. 數據清理
2. 特徵萃取
3. 資料視覺化>>>了解資料, 發現outliers 或異常數值, 分析個變數間的關練性
4. 建立模型
5. 驗證模型

---

1. 決策應用

Supervised Learning

1. 前處理：
   1. 資料讀取
   2. 格式調整
      1. 資料類型
         1. Discrete variable: 整數單位
         2. Continuous variable
         3. 字串/類別轉換：
            1. Label encoding: 有序 (sklearn.preprocessing.LabelEncoder) Text

               Description automatically generated
            2. One hot encoding: 無序 (pd.get\_dummies)
      2. 特徵類型
         1. 數值型特徵
         2. 類別型特徵：一種類別對應一種分數
         3. 時間型特徵：週期性
   3. 填補缺值
      1. Median: np.median(value\_array)
      2. Quantiles: np.quantile(value\_arrar,q = ..)
      3. Mode: scipy.stats.mode(value\_array) or dictionary method
      4. Mean: np.mean(value\_array)
      5. 連續行數值標準化：
         1. Z 轉換
         2. 空間壓縮 (Ｙ＝0-1, -1~1）
         3. Regression model: 有差
         4. Tree-based model: 沒有太大關係
   4. 去離群值
      1. 透過scatter, histogram 檢查
      2. 透過describe() 檢查
      3. Emprical Cumulative Density Plot

A picture containing text

Description automatically generatedChart, line chart

Description automatically generated(009)

* + 1. 可捨棄或是調整 (需分析離群值之原因)
       1. 對數轉換 ln / log
       2. 縮尾 (將離群值換為 較中間的數)
       3. 截尾 （刪除）
       4. 插值 （賦予較合理之數）（隨機,均值,回歸等）＊＊＊
    2. 減少標準化/最小最大化（特徵縮放）之問題
  1. 特徵縮放
     1. Good:
        1. 對權重敏感或損失函數平滑程度有幫助者
        2. 特徵間的量級差異甚大
     2. Bad:
        1. 有些指標不適合在標準化空間進行
        2. 量的單位在某些特徵上有意義

1. Exploratory Data Analysis：  
   plt.style.use(‘default/ggplot/seaborn’)  
   統計值視覺化
   1. 相關係數
      1. Pd\_df.corr()[“Target”]
   2. 核密度函數KDE (Kernel Density Estimation)
      1. 線下面積和為1
      2. 對稱：K(-u) = K(u)
      3. 常用的Kernel function
         1. Gaussian (normal dist)
         2. Cosine
      4. 不同組間的分布差異
   3. 離散化
      1. 減少Outlier 對分析與模型的影響
      2. 等寬pd.cut(df,num\_group)
      3. 等頻pd.qcut(df,num\_group)
      4. 劇類
   4. 繪圖排版:subplot
      1. The index for subplot starts with 1
      2. 圖形必須直覺
      3. EDA讓人看懂資料
   5. 常用圖形
      1. Heatmap
         1. Sns.heatmap()
      2. Gridplot
         1. Sns.pairplot()
   6. 模型體驗
      1. 使用最簡單的模型當作baseline
2. 特徵工程:
   1. 從事實到對應分數的轉換
      1. 用於資料彙整之後,擬合模型之前
      2. 類別編碼
      3. 特徵縮放
   2. 數值型特徵：若離群值比例太高，或平均值不具代表性，執行去偏態
      1. 目標在於讓數值更接近常態分佈，讓平均值更具代表性
      2. 對數去偏：np.log1p()
      3. 方根去偏：減去min 後開根號 （sqrt）
      4. 分佈去偏(Boxcox):   
         from scipy import stats  
         stats.boxcox(df,lambda= λ)
         1. λ介於0~0.5之間
         2. df內之值必須>0
   3. 類別型特徵：
      1. 標籤編碼(Label Encoding)：單一col輸出
         1. 適合非深度學習模型
         2. 有序資料時可使用
         3. Pd.get\_dummies()
         4. Sklearn.preprocessing.OneHotEncoder()
      2. 獨熱編碼(One Hot Encoding)：多col輸出
         1. 適合深度學習模型
         2. 特徵重要性高，且可能值較少時使用
      3. 若特徵值與目標值有顯著相關：
         1. 均值編碼：容易overfitting，可參考出現頻率
         2. 平滑化：
         3. 計數編碼 (Counting)：類別筆數與目標值有高度相關，筆數可作為特徵
         4. 特徵雜湊 (Feature Hash)：鄉議類別的數量非常龐大時
   4. 時間型特徵：
      1. 時間特徵分解
      2. 週期循環特徵
      3. 時段特徵

EDA – 統計量化的方式

1. 集中趨勢：
   1. 平均值 (mean)
   2. 中位數 (median)
   3. 眾數 (mode)
2. 資料分散程度：
   1. Min
   2. Max
   3. Range
   4. Quartiles
   5. Variance
   6. Standard Deviation
3. pd.hist(bin=n)  
   Chart

   Description automatically generated
4. pd. value\_counts().plot(kind="bar")

Chart, waterfall chart

Description automatically generated