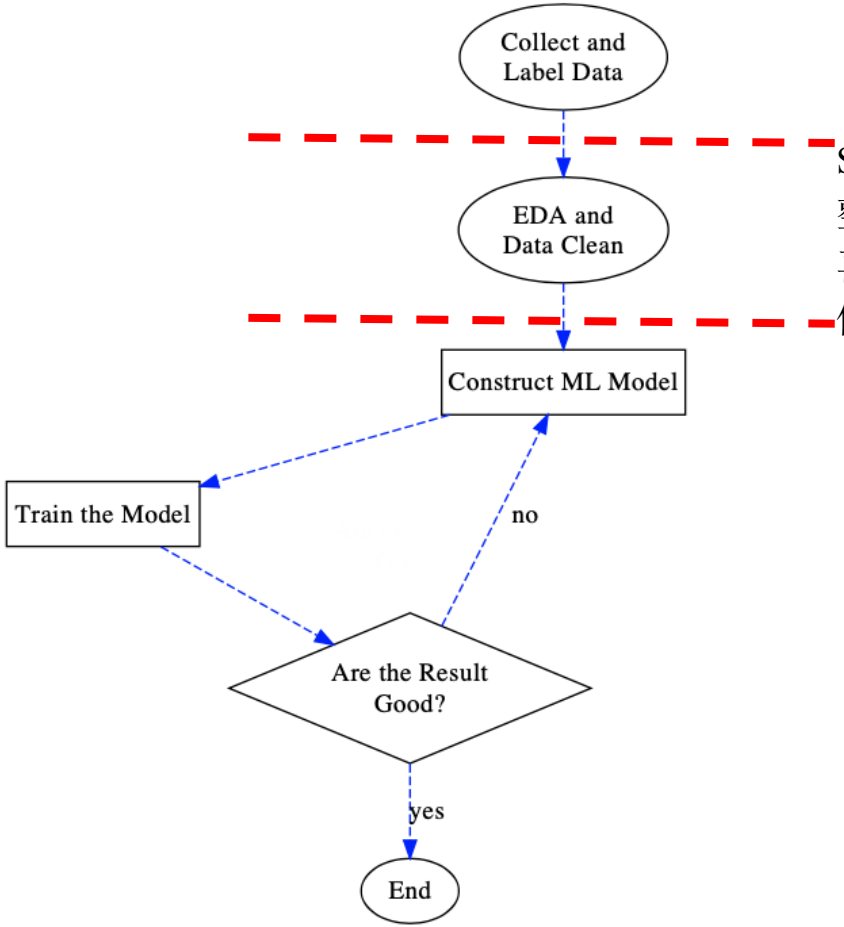
ML and MLOPs

黄世豪

Contents

講解 ML Workflow

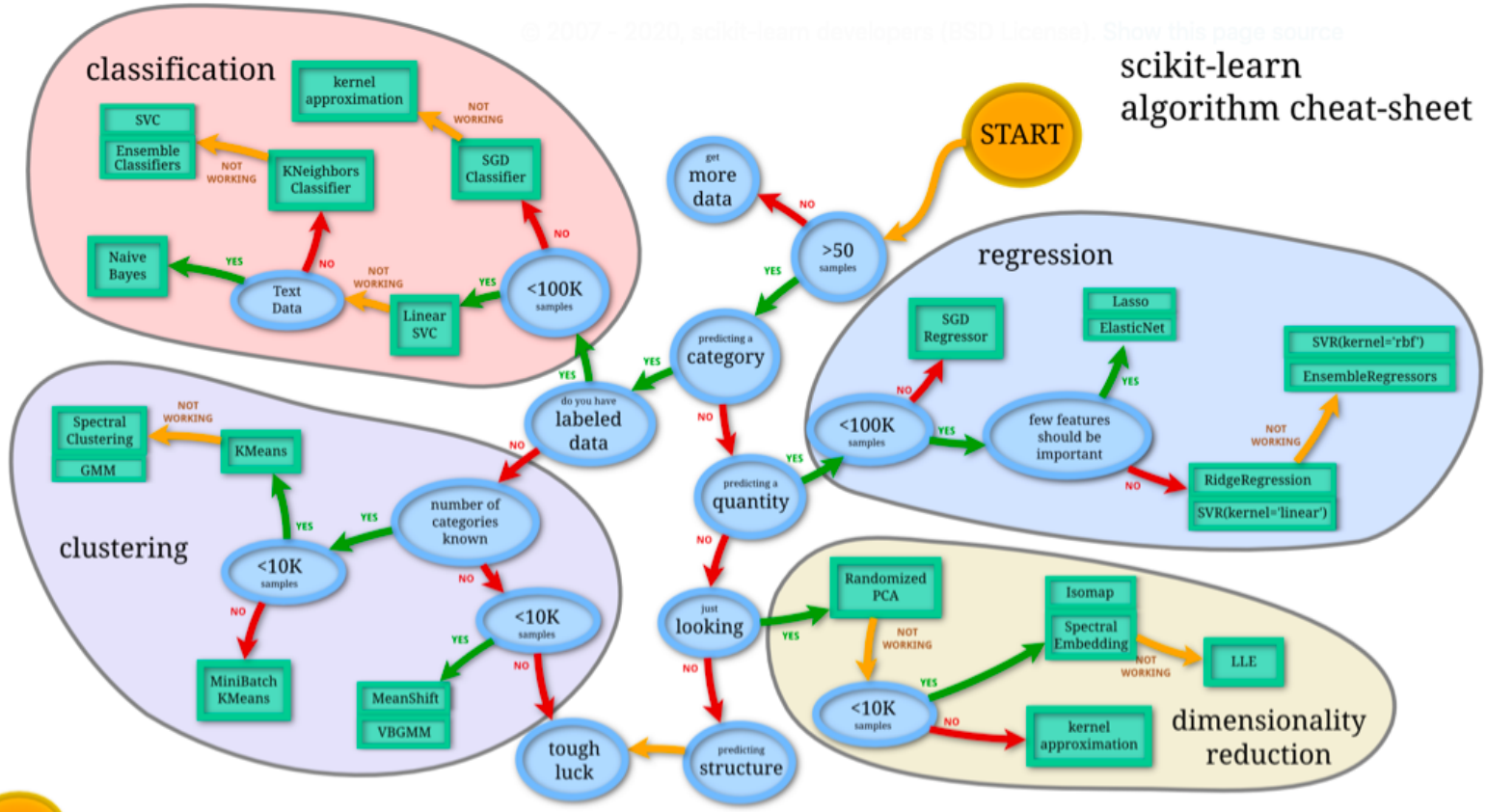
Machine Learning Workflow



Step1: 收集資料，和為這些資料做標籤

Step2: Exploratory Data Analysis and Data Clean
整理圖表或畫數據圖以了解整份數據的特性，再根據這些統計結果來調整數據，可能會是整個流程中最吃重和繁複的部分。

Step3: 根據資料類型和任務取向選取模型，這部分大部分時間是花在調整模型參數上，使結果更好（loss 降低，acc上升）



https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Demo -> The Importance of EDA, Different ML Models

Getting Started Prediction Competition

House Prices - Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

Kaggle · 8,668 teams · Ongoing

Overview Data Code Discussion Leaderboard Rules Team My Submissions **Submit Predictions**

Overview

- Description
- Evaluation
- Tutorials
- Frequently Asked Questions

Start here if...

You have some experience with R or Python and machine learning basics. This is a perfect competition for data science students who have completed an online course in machine learning and are looking to expand their skill set before trying a featured competition.

Competition Description

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

Total 8668 teams

Data

One row, one data.

Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1
1	60	RL	65	8450	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm
2	20	RL	80	9600	Pave	NA	Reg	Lvl	AllPub	FR2	Gtl	Veenker	Feedr
3	60	RL	68	11250	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm
4	70	RL	60	9550	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	Crawfor	Norm
5	60	RL	84	14260	Pave	NA	IR1	Lvl	AllPub	FR2	Gtl	NoRidge	Norm

Total 1460 rows

Id1 data's label

EnclosedPorch	3SsnPorch	ScreenPorch	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice
0	0	0	0	NA	NA	NA	0	2	2008	WD	Normal	208500
0	0	0	0	NA	NA	NA	0	5	2007	WD	Normal	181500
0	0	0	0	NA	NA	NA	0	9	2008	WD	Normal	223500
272	0	0	0	NA	NA	NA	0	2	2006	WD	Abnorml	140000
0	0	0	0	NA	NA	NA	0	12	2008	WD	Normal	250000

Evaluation

house_linear_noEDA

Id	SalePrice
1461	27722.298597542300
1462	8516.89604525684
1463	188412.59316105300
1464	68885.33072748160
1465	1039168.37910609
1466	241324.4159601520
1467	984029.3493541860
1468	623388.3976040320
1469	23295.94945903130
1470	336489.658600901
1471	379733.3930456250
1472	151970.75618570700
1473	364435.52467853800
1474	31698.520327679200
1475	133746.86146303700

對值取log後做
RMSE

$$\text{RMSD} = \sqrt{\frac{\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}{T}}.$$

sample_submission

Id	SalePrice
1461	169277.0524984
1462	187758.393988768
1463	183583.683569555
1464	179317.47751083
1465	150730.079976501
1466	177150.989247307
1467	172070.659229164
1468	175110.956519547
1469	162011.698831665
1470	160726.247831419
1471	157933.279456005
1472	145291.245020389
1473	159672.017631819
1474	164167.518301885
1475	150891.638244053

No EDA-> Replace missing value with NONE

```python
def na_check(df_data):
  data_na = (df_data.isnull().sum() / len(df_data)) * 100
  data_na = data_na.drop(data_na[data_na == 0].index).sort_values(ascending = False)
  missing_data = pd.DataFrame({'Missing Ratio' :data_na})
  display(missing_data.head(10))
na_check(df)
```

	Missing Ratio
PoolQC	99.657417
MiscFeature	96.402878
Alley	93.216855
Fence	80.438506
FireplaceQu	48.646797
LotFrontage	16.649538
GarageFinish	5.447071
GarageYrBlt	5.447071
GarageQual	5.447071
GarageCond	5.447071

```python
# 缺值補 'None'
none_cols = ['PoolQC', 'MiscFeature', 'Alley', 'Fence', 'FireplaceQu', 'FireplaceQu', 'FireplaceQu', 'FireplaceQu
             'GarageType', 'GarageFinish', 'GarageQual', 'GarageCond', 'BsmtQual', 'BsmtCond', 'BsmtExposure',
              'BsmtFinType1', 'BsmtFinType2', 'MasVnrType', 'Functional', 'MSSubClass','GarageYrBlt', 'GarageArea'
              'BsmtFullBath', 'BsmtHalfBath', 'MasVnrArea','MSZoning', 'Electrical', 'KitchenQual', 'Exterior1st',
for col in none_cols:
    df[col] = df[col].fillna("None")

# Utilities 參考資訊很少, 所以直接捨棄
df = df.drop(['Utilities'], axis=1)
```

[30]
```python
na_check(df)
```

Missing Ratio

No EDA Results

Algorithm: Linear Regression

名次 RMSE

8517	Shih Hao	1.43337	1	1m

Your First Entry

Welcome to the leaderboard!

Algorithm: Random Forest Regression

7410	Shih Hao	0.23895	2	~10s

Your Best Entry

Your submission scored 0.23895, which is not an improvement of your best score. Keep trying!

Algorithm: Gradient Boosting Regression

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
house_gdbt_noEDA.csv	a minute ago	1 seconds	0 seconds	0.19878

Complete

Basic EDA

```python
# 部分欄位缺值補 'None'
none_cols = ['PoolQC', 'MiscFeature', 'Alley', 'Fence', 'FireplaceQu', 'FireplaceQu', 'FireplaceQu', 'FireplaceQu',
             'GarageType', 'GarageFinish', 'GarageQual', 'GarageCond', 'BsmtQual', 'BsmtCond', 'BsmtExposure',
              'BsmtFinType1', 'BsmtFinType2', 'MasVnrType', 'Functional', 'MSSubClass']
for col in none_cols:
    df[col] = df[col].fillna("None")

# 部分欄位缺值填補 0
zero_cols = ['GarageYrBlt', 'GarageArea', 'GarageCars', 'BsmtFinSF1', 'BsmtFinSF2', 'BsmtUnfSF','TotalBsmtSF',
              'BsmtFullBath', 'BsmtHalfBath', 'MasVnrArea']
for col in zero_cols:
    df[col] = df[col].fillna(0)

```

```python
# 部分欄位缺值補眾數
mode_cols = ['MSZoning', 'Electrical', 'KitchenQual', 'Exterior1st', 'Exterior2nd', 'SaleType']
for col in mode_cols:
    df[col] = df[col].fillna(df[col].mode()[0])

# 'LotFrontage' 有空缺時, 以同一區 (Neighborhood) 的 LotFrontage 中位數填補 (可以視為填補一種群聚編碼 )
df["LotFrontage"] = df.groupby("Neighborhood")["LotFrontage"].transform(lambda x: x.fillna(x.median()))

# Utilities 參考資訊很少, 所以直接捨棄
df = df.drop(['Utilities'], axis=1)
```

Basic EDA

```python
# 做標籤編碼
cols = ('FireplaceQu', 'BsmtQual', 'BsmtCond', 'GarageQual', 'GarageCond',
        'ExterQual', 'ExterCond','HeatingQC', 'PoolQC', 'KitchenQual', 'BsmtFinType1',
        'BsmtFinType2', 'Functional', 'Fence', 'BsmtExposure', 'GarageFinish', 'LandSlope',
        'LotShape', 'PavedDrive', 'Street', 'Alley', 'CentralAir', 'MSSubClass', 'OverallCond',
        'YrSold', 'MoSold')
for c in cols:
    lbl = LabelEncoder()
    lbl.fit(list(df[c].values))
    df[c] = lbl.transform(list(df[c].values))

# 由地下室面積 + 1樓面積 + 2樓面積, 計算總坪數特徵
df['TotalSF'] = df['TotalBsmtSF'] + df['1stFlrSF'] + df['2ndFlrSF']
```

增加特徵值會增加模型預測能力

No EDA Results

Algorithm: Linear Regression

4751	Shih Hao	0.14580	5	1m

Your Best Entry ↑

Your submission scored 0.14580, which is not an improvement of your best score. Keep trying!

Algorithm: Random Forest Regression

6045	Shih Hao	0.16383	4	~10s

Your Best Entry ↑

Your submission scored 0.16383, which is not an improvement of your best score. Keep trying!

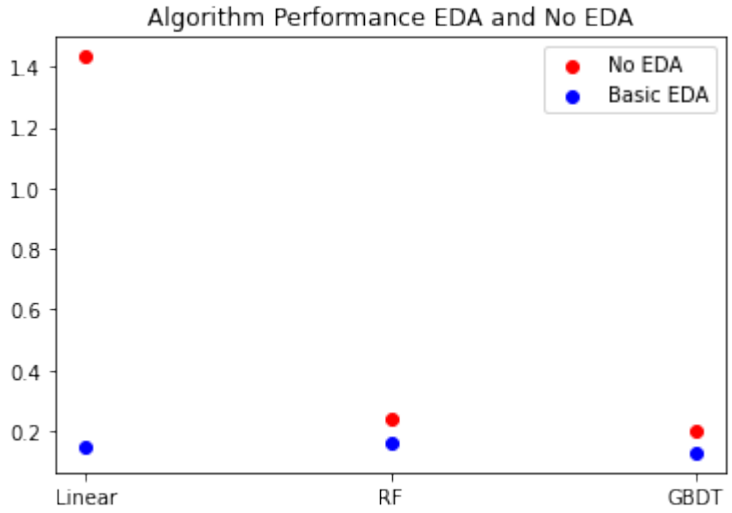
Algorithm: Gradient Boosting Regression

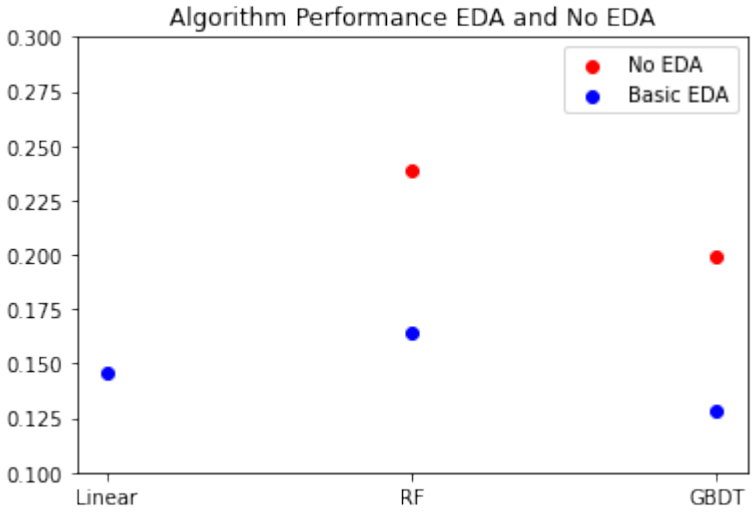
2196	Shih Hao	0.12834	7	~10s

Your Best Entry ↑

Your submission scored 0.12834, which is not an improvement of your best score. Keep trying!

Performance





論文（Data Driven Chiller Plant Energy Optimization with Domain Knowledge) 重點整理

論文重點

Most of these works, however, do not consider the varying factors, such as ageing equipments and indoor activities.

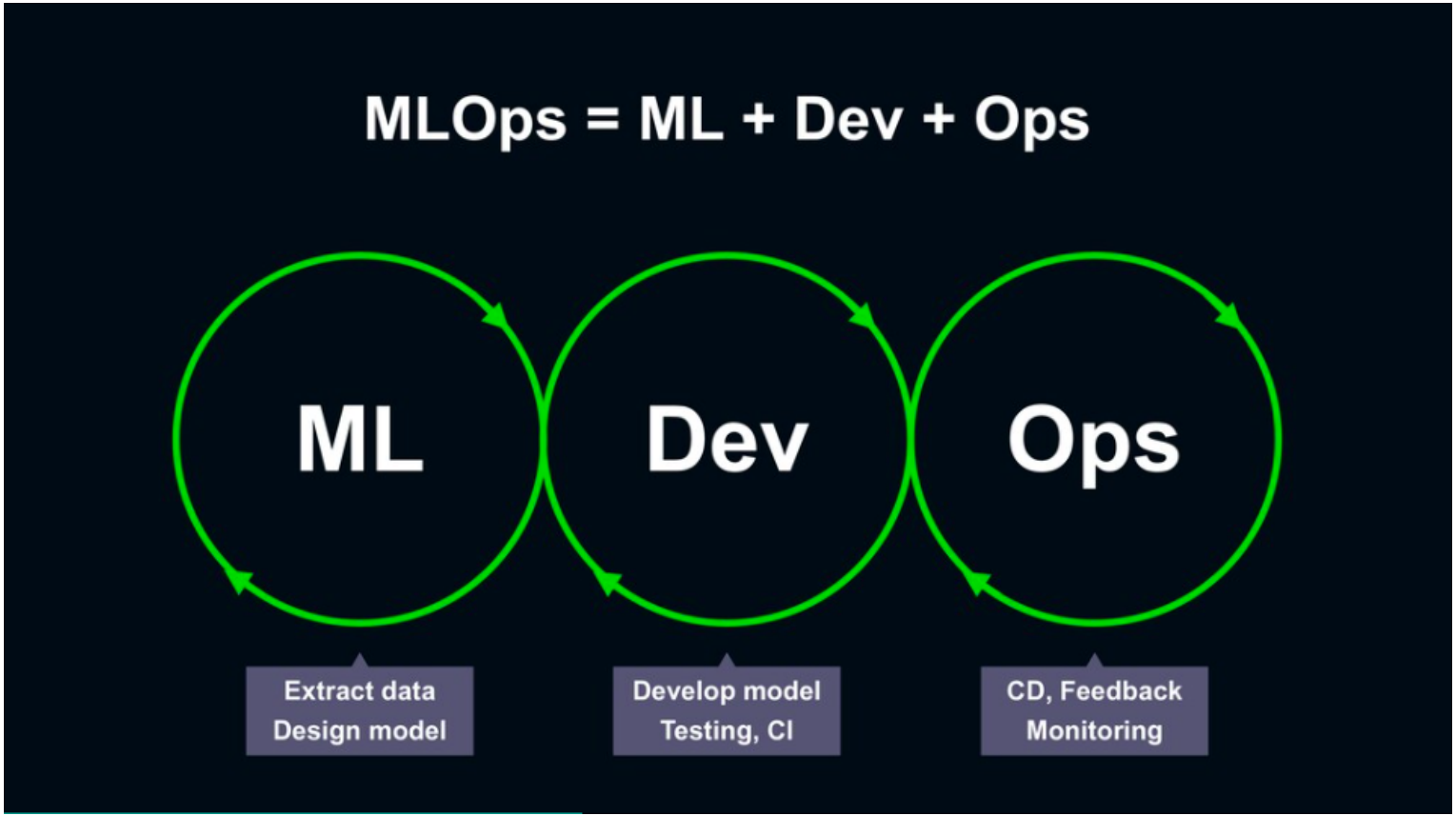
環境會變，資料不能是固定的，模型也應該要更動才能應付？

In our work, we solve these problems with active data enrichment and select models based on domain knowledge over the equipments.

主動數據充實說明了資料要能隨時補充，而且模型也要能夠接受隨時調整參數

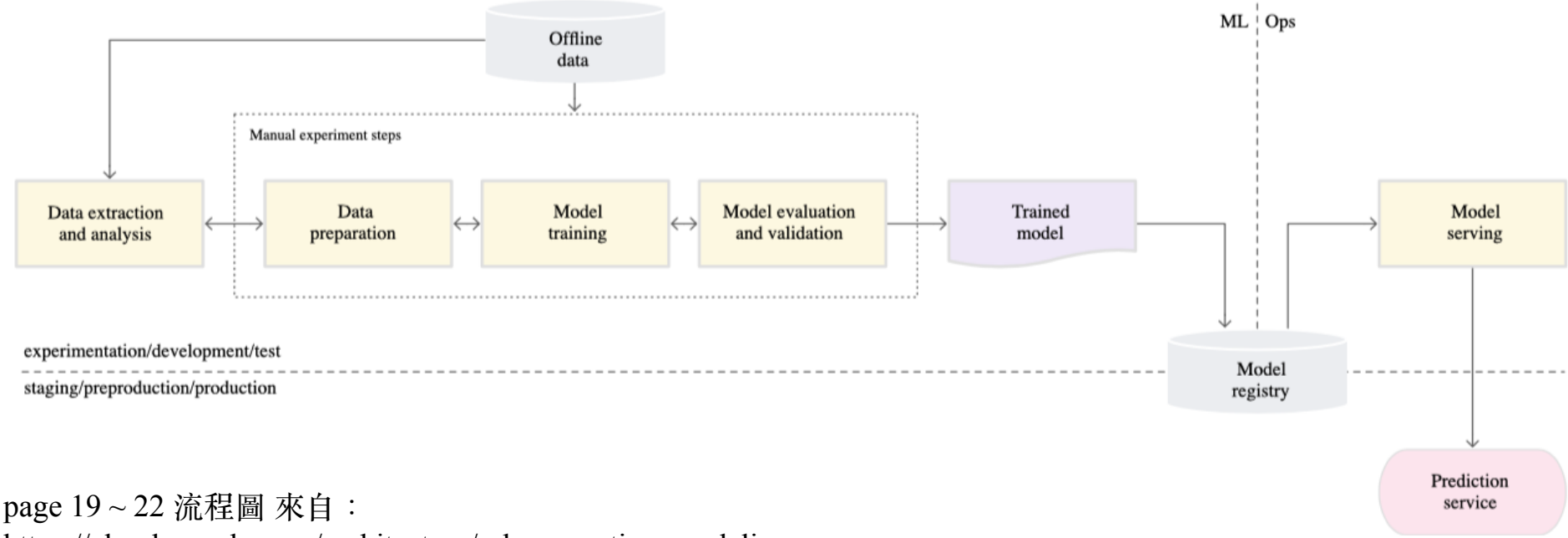
說明 MLOps

What is MLOps?



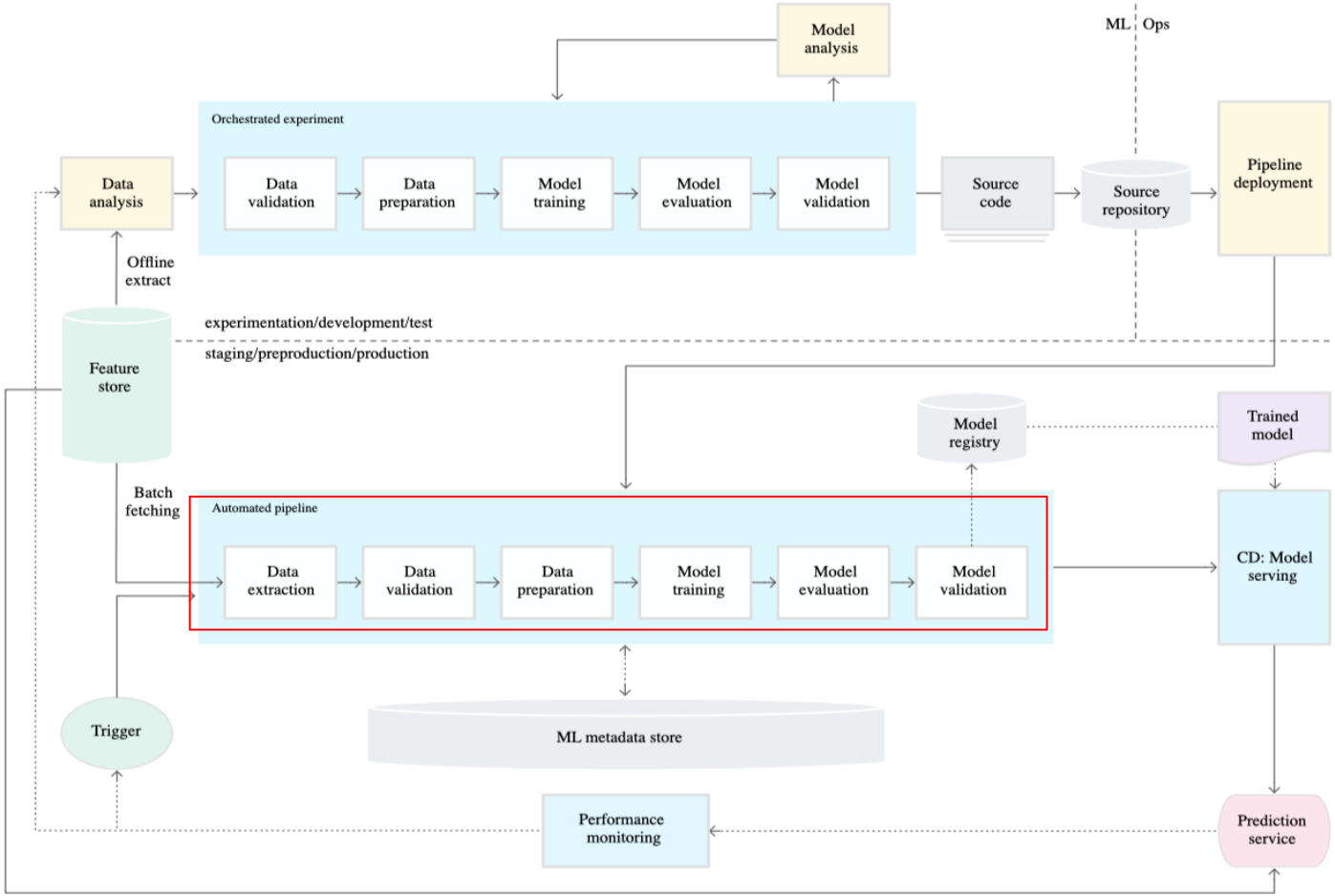
https://www.ithome.com.tw/news/141774

level 0

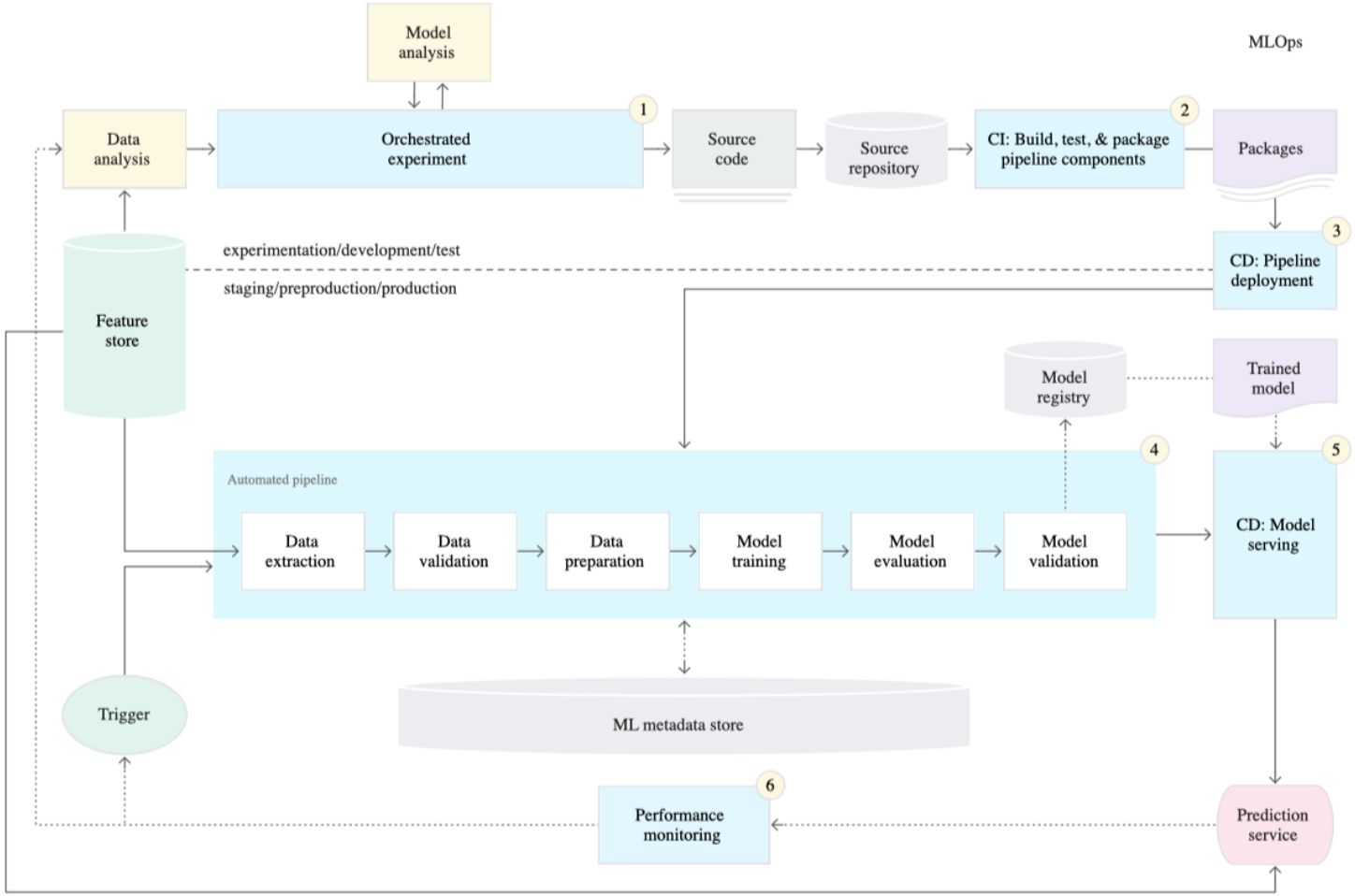


page 19 ~ 22 流程圖 來自：
https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

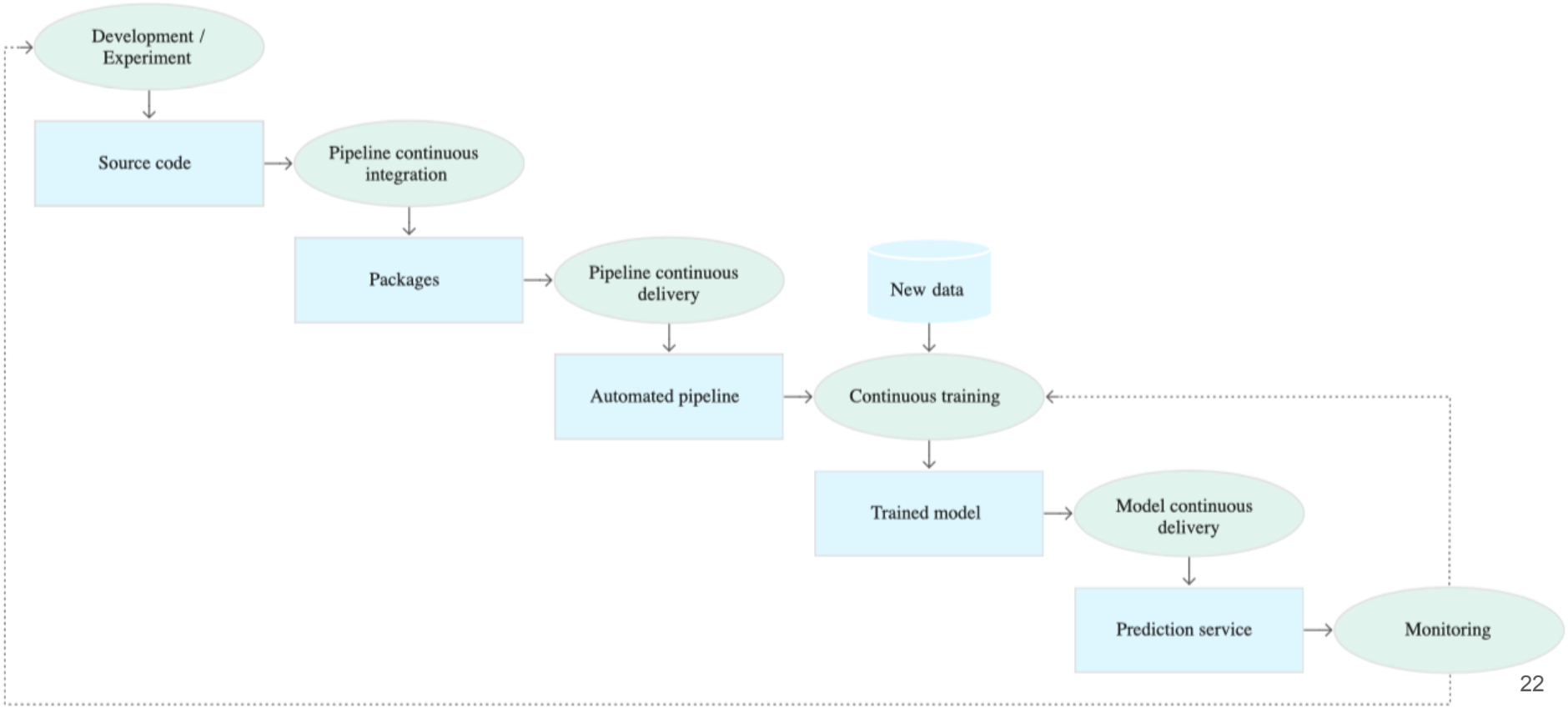
level 1

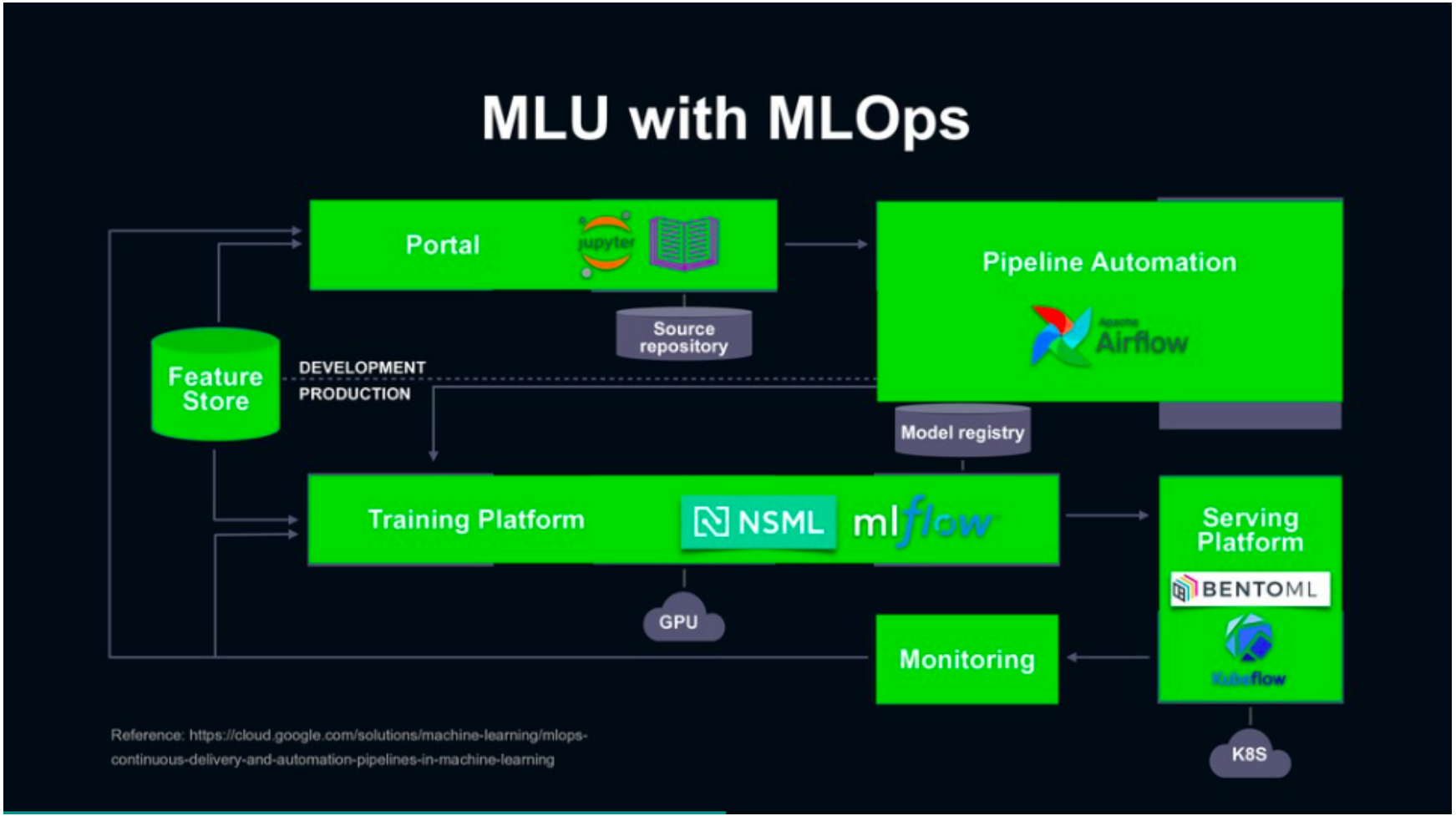

Model analysis
ML
Ops
Orchestrated experiment
Data analysis
Data validation
Data preparation
Model training
Model evaluation
Model validation
Source code
Source repository
Pipeline deployment
Offline extract
experimentation/development/test
staging/preproduction/production
Feature store
Model registry
Trained model
Batch fetching
Automated pipeline
Data extraction
Data validation
Data preparation
Model training
Model evaluation
Model validation
CD: Model serving
Trigger
ML metadata store
Performance monitoring
Prediction service

level 2


Model analysis
MLOps
Data analysis
Orchestrated experiment
1
Source code
Source repository
CI: Build, test, & package pipeline components
2
Packages
3
CD: Pipeline deployment
experimentation/development/test
staging/preproduction/production
Feature store
Model registry
Trained model
4
5
Automated pipeline
Data extraction
Data validation
Data preparation
Model training
Model evaluation
Model validation
CD: Model serving
Trigger
ML metadata store
6
Performance monitoring
Prediction service

ML CI (continuous integration)/ CD (continuous delivery) automation pipeline





https://www.ithome.com.tw/news/141774

Brief MLOps Demo

https://youtu.be/9BgIDqAzfuA

Wine Data

wine_quality

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
7.4	0.7	0.0	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
7.8	0.88	0.0	2.6	0.098	25.0	67.0	0.9968	3.2	0.68	9.8	5
7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.997	3.26	0.65	9.8	5
11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.998	3.16	0.58	9.8	6
7.4	0.7	0.0	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
7.4	0.66	0.0	1.8	0.075	13.0	40.0	0.9978	3.51	0.56	9.4	5
7.9	0.6	0.06	1.6	0.069	15.0	59.0	0.9964	3.3	0.46	9.4	5
7.3	0.65	0.0	1.2	0.065	15.0	21.0	0.9946	3.39	0.47	10.0	7
7.8	0.58	0.02	2.0	0.073	9.0	18.0	0.9968	3.36	0.57	9.5	7
7.5	0.5	0.36	6.1	0.071	17.0	102.0	0.9978	3.35	0.8	10.5	5

One row one data, one data 11 features. Total 1600 rows

```
################################
########## MODELLING ############
################################

# Fit a model on the train section
regr = RandomForestRegressor(max_depth=2, random_state=seed)
regr.fit(X_train, y_train)
```

Model: Random Forest Regressor, depth: 2

Github Actions: Script

MLOps CI Demo

CI (Continuous integration):

Continuous integration is an idea from dev ops that's all about connecting changes to your code to fast feedback, to testing how that's affected your ultimate project.

experiment-has... wine_kaggle / .github / workflows / cml.yaml

hao134 Update cml.yaml

1 contributor

25 lines (21 sloc) | 701 Bytes

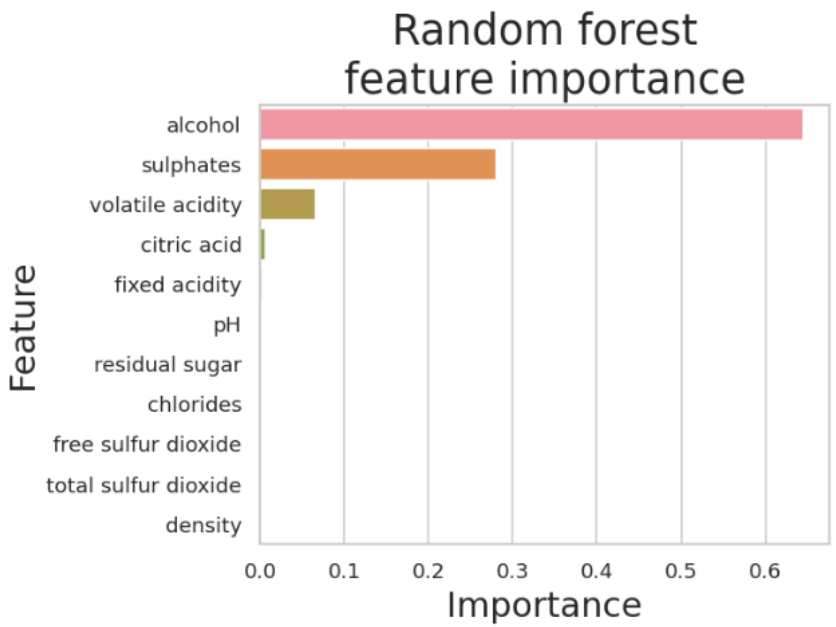
```yaml
name: your-workflow-name
on: [push]
jobs:
  run:
    runs-on: [ubuntu-latest]
    container: docker://dvcorg/cml-py3:latest
    steps:
      - uses: actions/checkout@v2
      - name: 'Train my model'
        env:
          repo_token: ${{ secrets.GITHUB_TOKEN }}
        run: |

          # Your ML workflow goes here
          pip install -r requirements.txt
          python train.py

          echo "## MODEL metrics" > report.md
          cat metrics.txt >> report.md

          echo "## Data viz" >> report.md
          cml-publish feature_importance.png --md >> report.md
          cml-publish residuals.png --md >> report.md

          cml-send-comment report.md
```

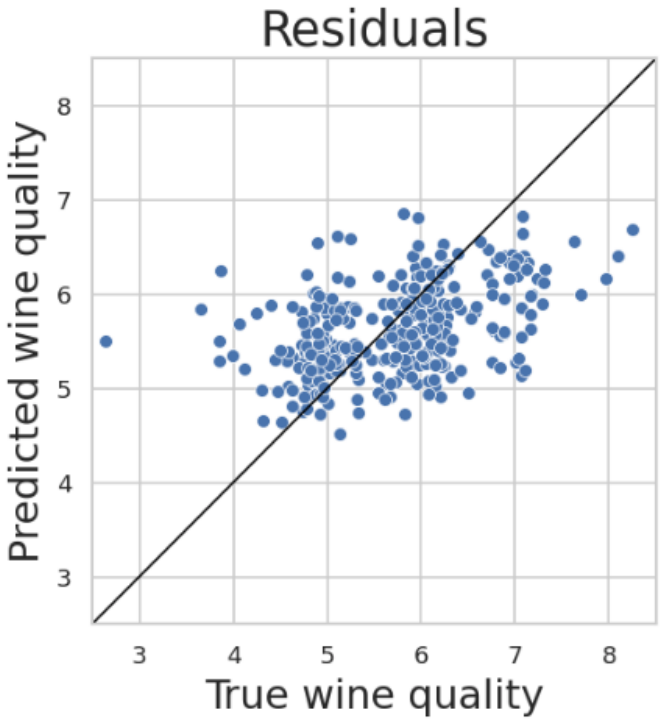
Results

MODEL metrics

Training variance explained: 33.0%
Test variance explained: 32.0%

Data viz





Results

✓ **change the depth of Random Forest 2 -> 5**

main (#2)

hao134 committed 17 minutes ago Verified

Showing **1 changed file** with **1 addition** and **1 deletion**.

2 train.py

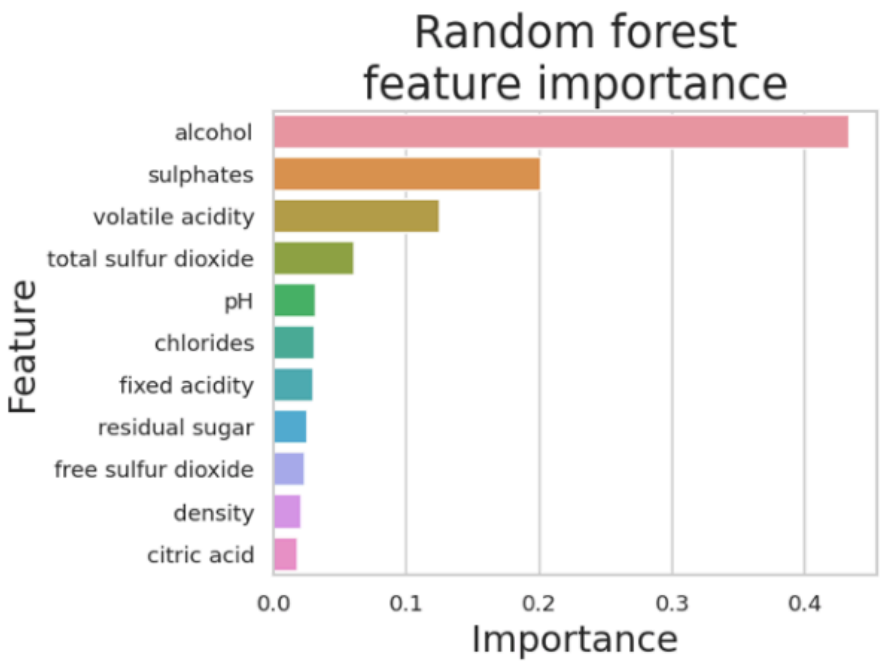
```diff
@@ -23,7 +23,7 @@
  ################################

  # Fit a model on the train section
- regr = RandomForestRegressor(max_depth=2, random_state=seed)
+ regr = RandomForestRegressor(max_depth=5, random_state=seed)
  regr.fit(X_train, y_train)

  # Report training set score
```

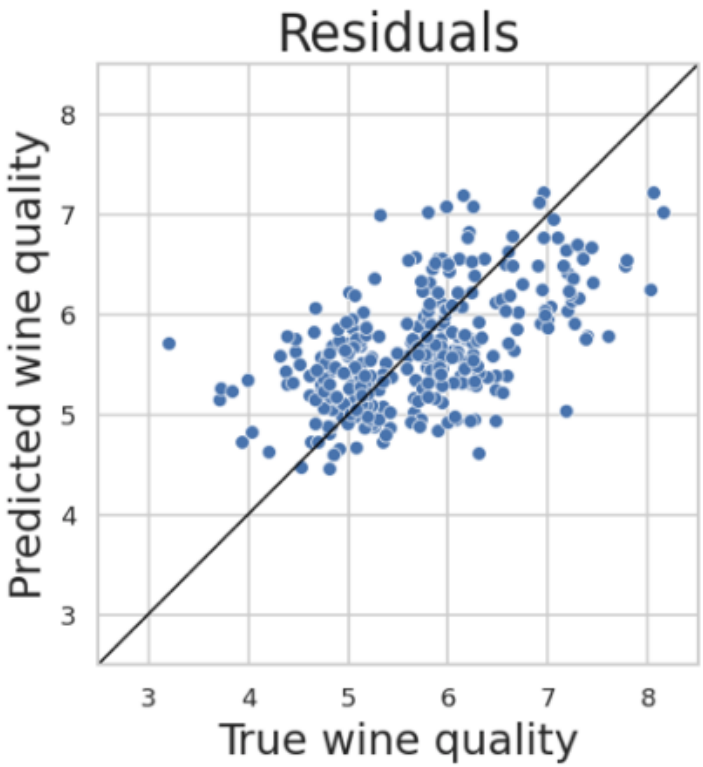
Results

MODEL metrics

Training variance explained: 54.3%
Test variance explained: 43.8%

Data viz





Summary

1. 冰水主機的資料主要是時序型資料，因此，能夠發展的模型估記有:

   ARIMA, CNN, LSTM, GRU, AutoEncoder….

2. EDA 和 Data Clean 很重要，通常時序性資料受到外在環境影響很大，哪些是偶發性哪些具有規律週期性，要透過EDA來了解清楚。

3. 為了未來的發展性著想，希望不只能訓出一個機器學習模型，還要能發展MLOps，至少要有Level 0程度的雛形，最終希望能達到 Level 1 程度。