機器學習 - NTUDAC

James Yeh

James Yeh

- Yahoo 軟體工程師(Backend, ML)
- 台大電機所
- 網站開發、資料分析、機器學習
- Python 資料分析與機器學習入門
- yehjames23@gmail.com

今天我們會

環境安裝 — 機器學習初探 — 真實案例 — Shopify介紹 — 資料分析&Logistic Regression — Q&A

環境安裝

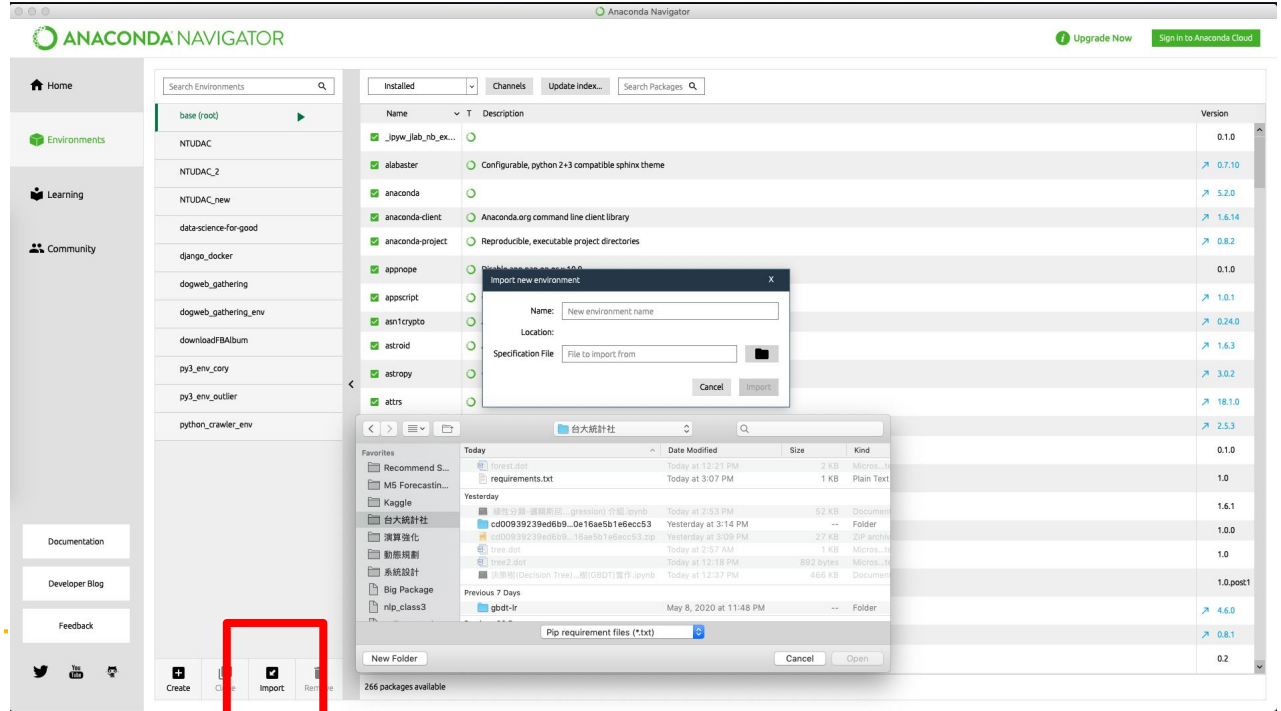
重裝一個乾淨的conda環境(避免套件衝突)

從你的終端機:

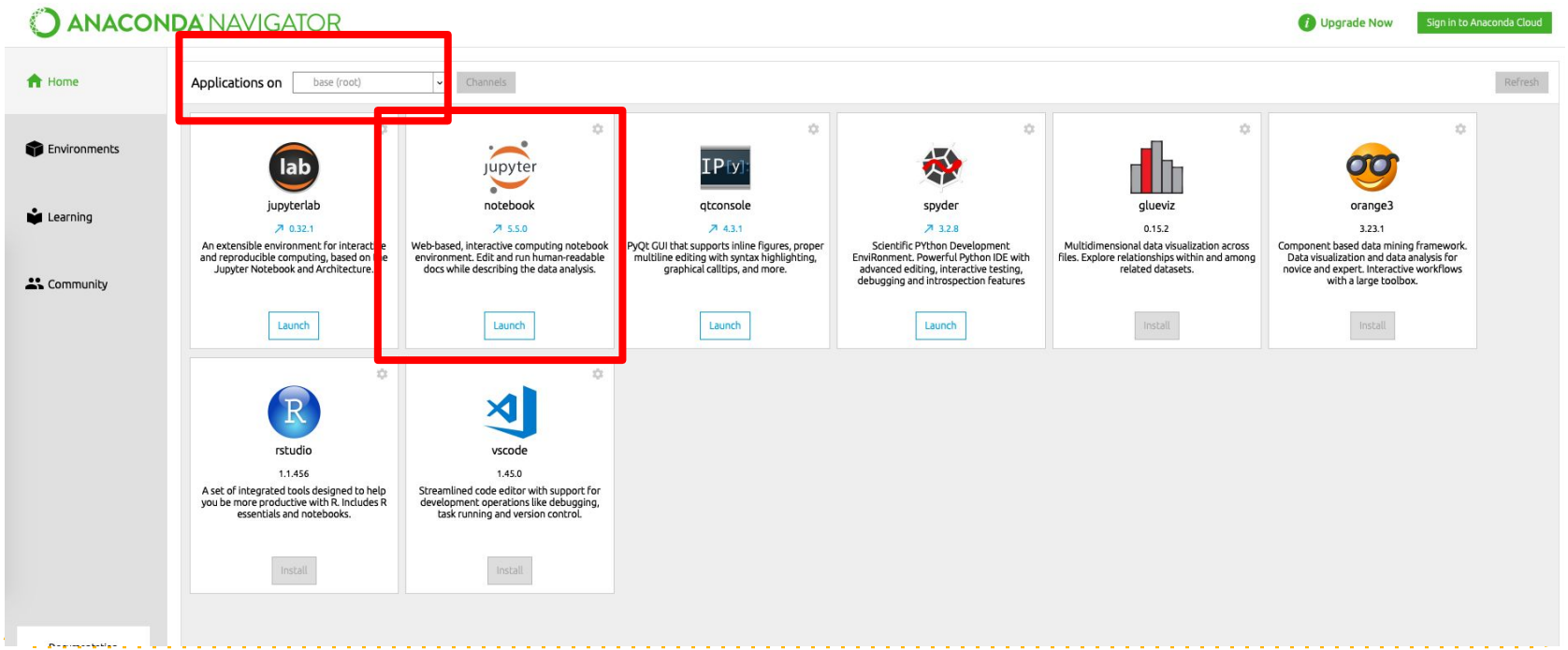
```
conda create --name NTUDAC python=3.6
conda activate NTUDAC
pip install -r requirements.txt
conda deactivate
```

或從Anaconda navigator UI

1. 輸入虛擬環境名稱ex: NTUDAC
2. 選擇import pip檔案 requirements.txt

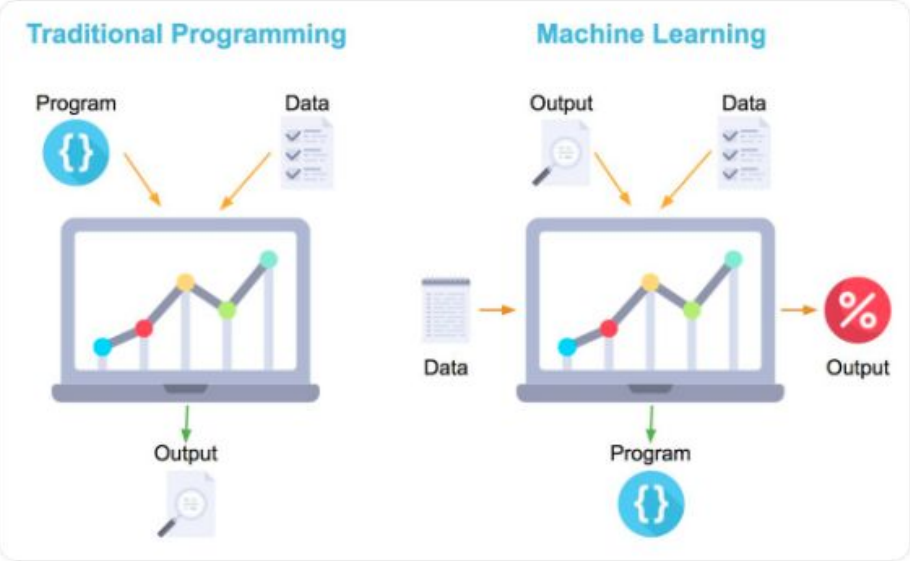


選擇剛創建的虛擬環境 開啟 Jupyter notebook



機器學習初探

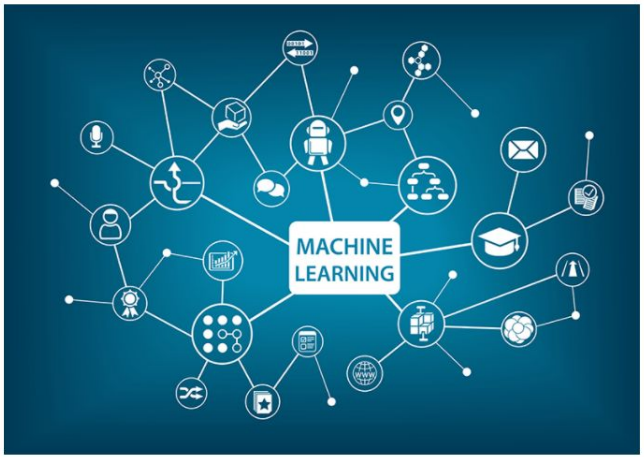
什麼是機器學習？



Machine Learning vs Traditional Programming

[資料分析&機器學習] 第3.1講：Python 機器學習以及Scikit-learn介紹

Yeh James
Oct 21, 2017 · 7 min read



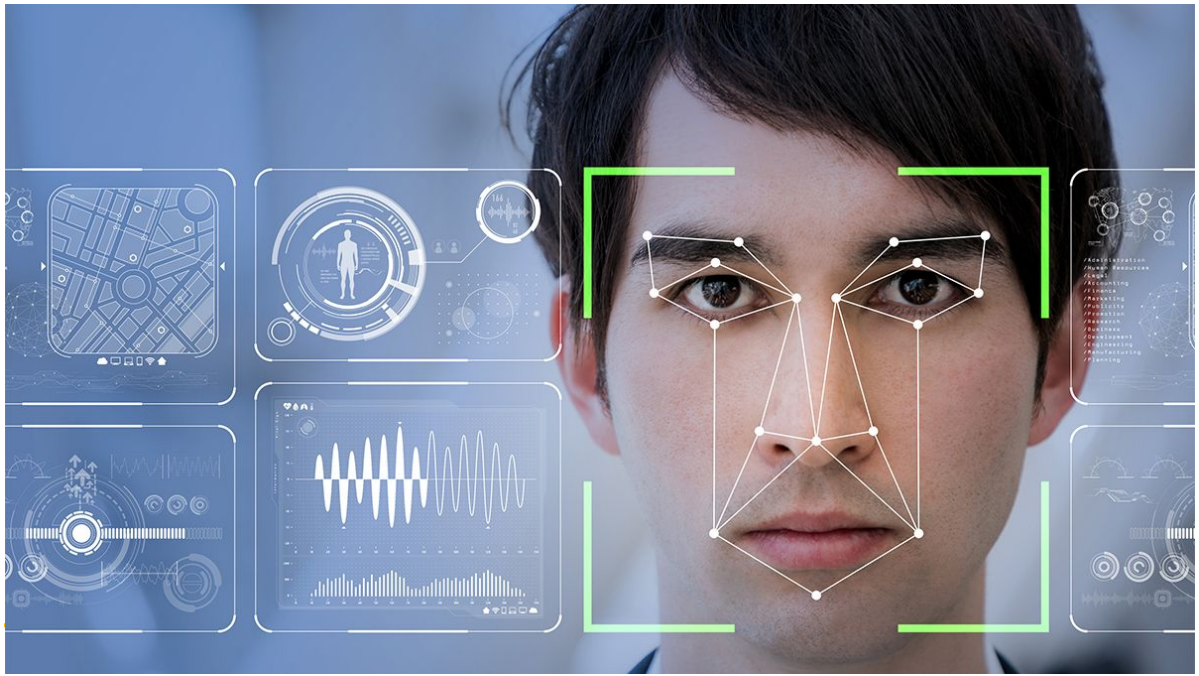
機器學習示意圖(圖片來源：madeiradata)

Python 機器學習以及 Scikit-learn介紹

機器學習可以做什麼呢?

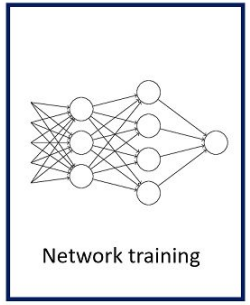
垃圾郵件分類

人臉辨識



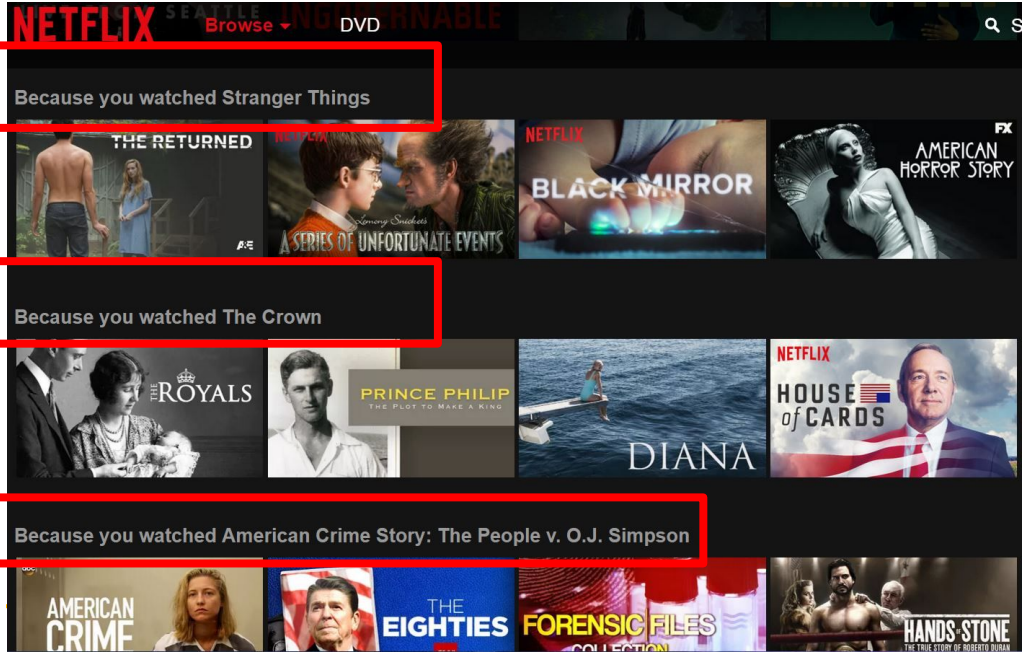
數字辨識

Data & Labels

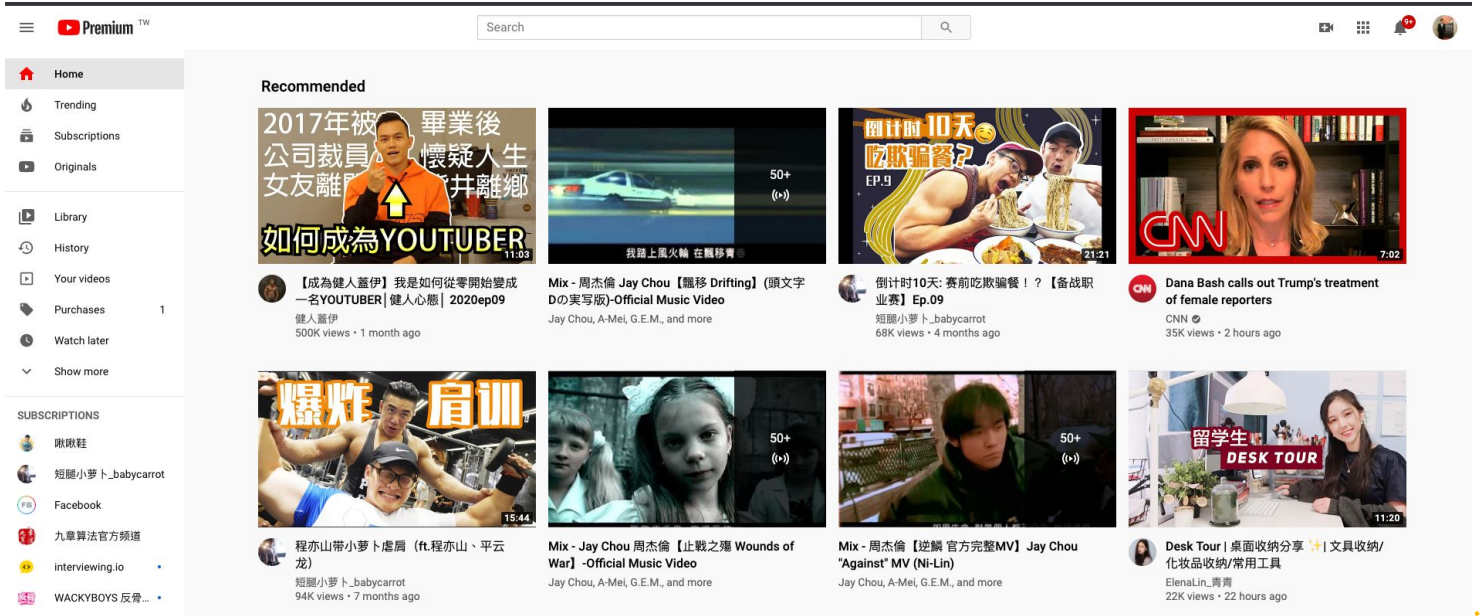


0
1
2
3
4
5
6
7
8
9

推薦系統(NETFLIX)



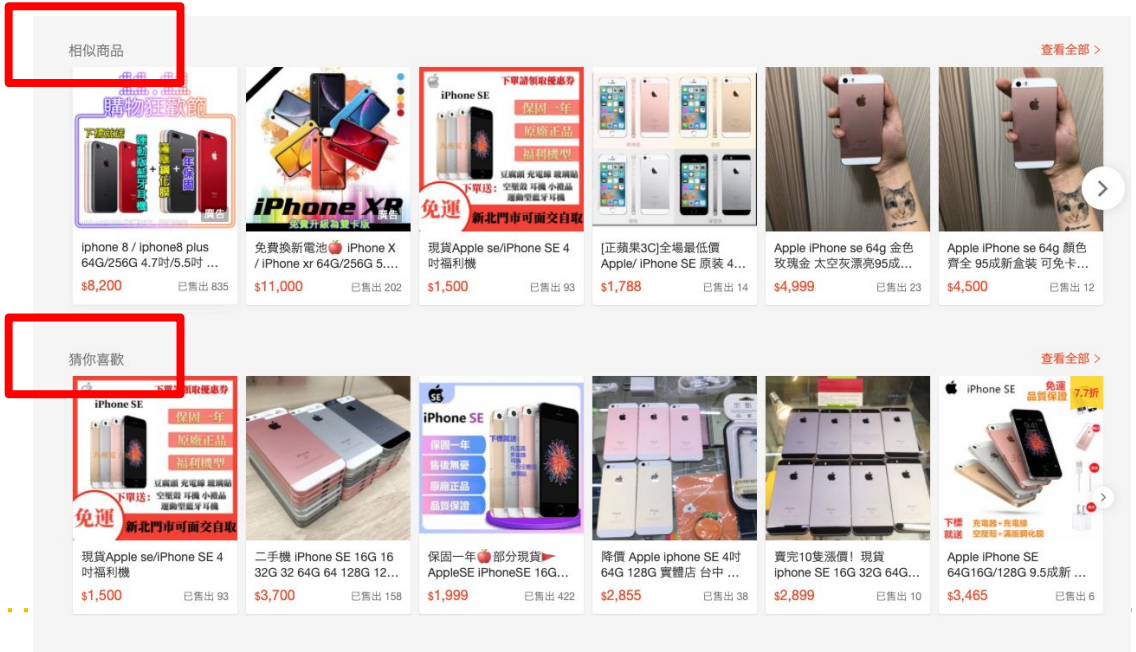
推薦系統(Youtube)



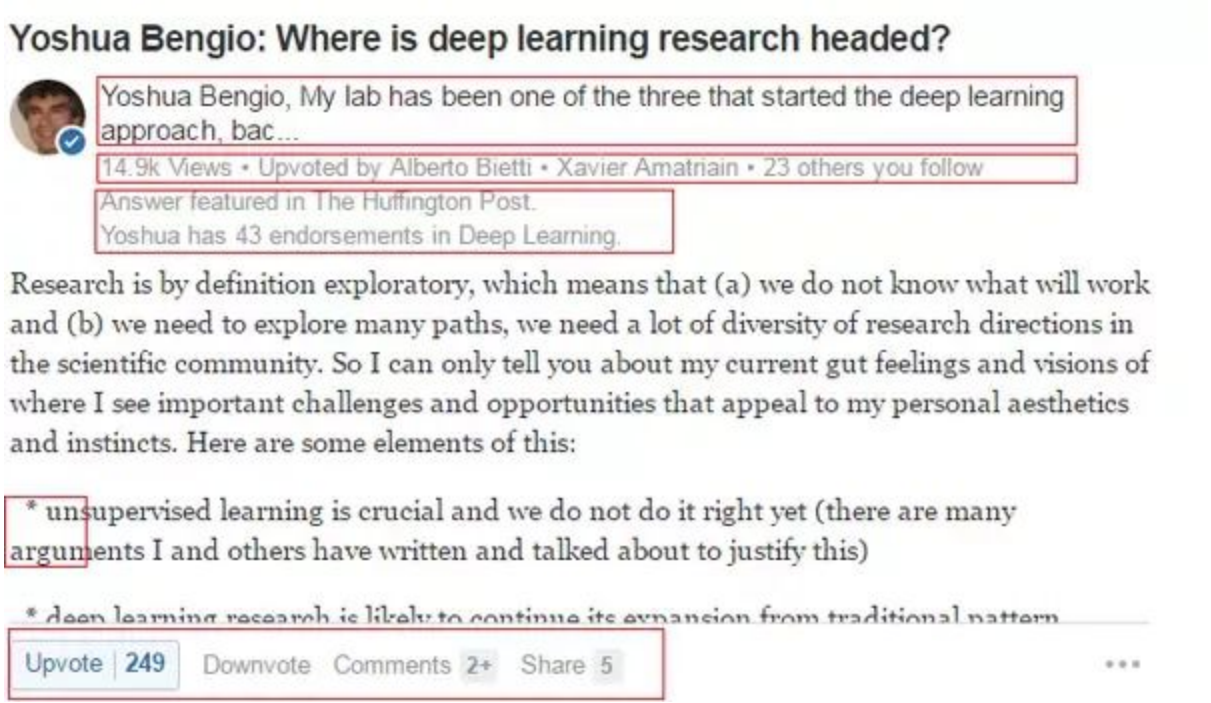
推薦系統(FB)

Search
James
Home
Create
Yeh James
News Feed
Messenger
Watch
Marketplace
Shortcuts
crypto lending bot...
2016 NTUEDS-104...
Front-End Develop... 20+
HackNTU HackCo...
NTUCC Alumni Fa...
See More...
Explore
COVID-19 Informat...
Groups
Pages 9
Events
Fundraisers
See More...
Stories
See All
Angela Yu
Chen-kuan Sun
Hong-Xuan Huang
Bill Chen
You're seeing posts from RabbieStartup 創業兔筆記 first.
RabbieStartup 創業兔筆記
14 hrs
"...Google台灣將在22日舉辦Google數位人才探索計畫活動，希望積極培養台灣的數位行銷人才投入、協助未來企業轉型。
目前，Google數位學程認證計畫在網路上提供上百堂線上課程免費學習，完成26項主題後就可以進行Google證照認證。..."
https://learndigital.withgoogle.com/digitalgarage-tw
Google 數位學程
LEARNDIGITAL.WITHGOOGLE.COM
學習線上行銷技能 - Google 提供的免費培訓課程
我們提供免費培訓課程，幫助您學習職涯和業務發展所需的數位技能...
Publish
Photo
Live
Invite
Likes
Views
Posts
2,798
2 new likes this week
Create Promotion
Listings You Recently Viewed
NT$20,000
See More
Sponsored
Create Ad
Find your blog expert on Fiverr. Today.
fiverr.com
Did you know? 81% of B2B companies use Blog as a content marketing tactic. Grow your busin...

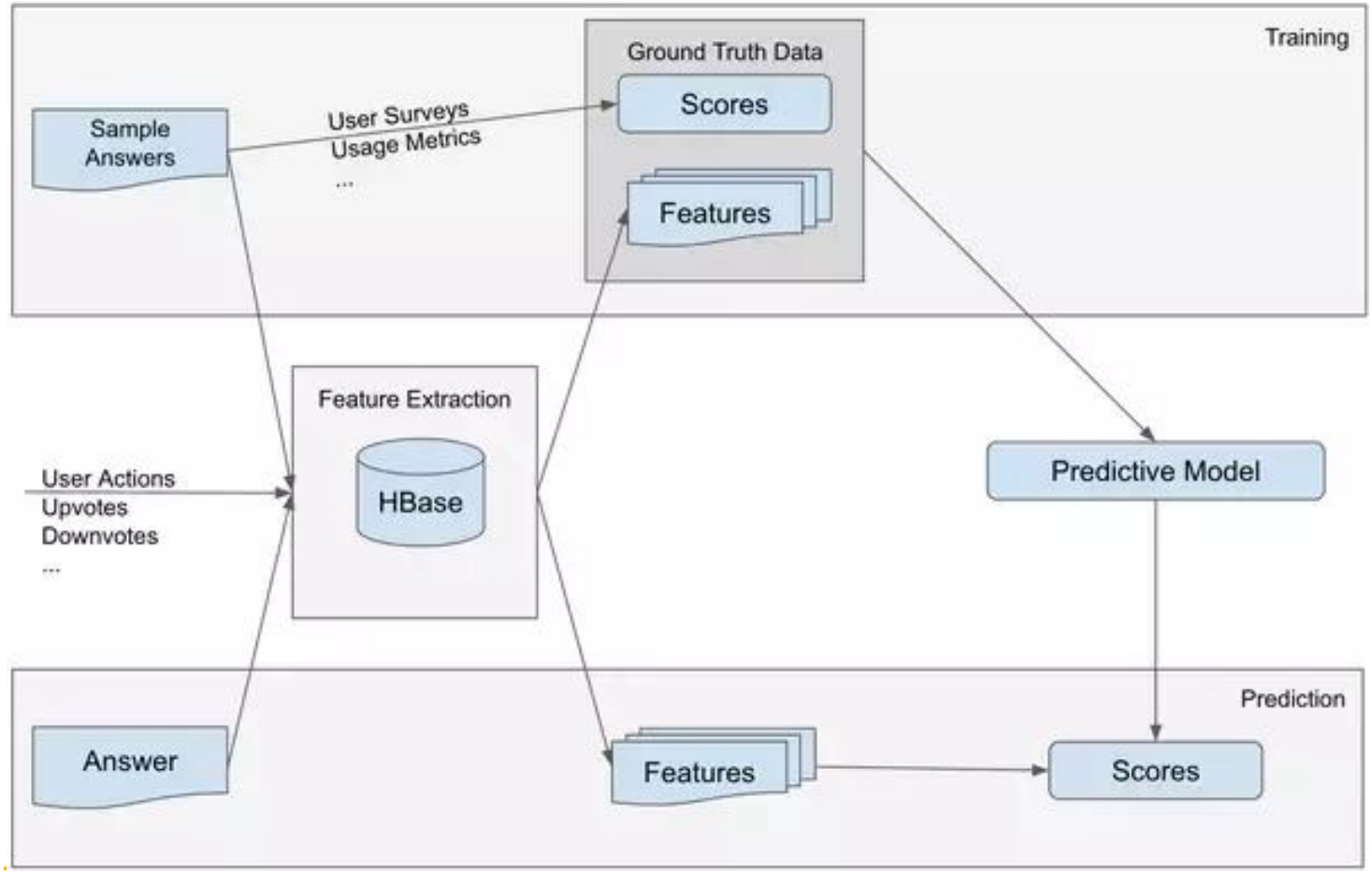
推薦系統(蝦皮)


相似商品
查看全部 >
iphone 8 / iphone8 plus 64G/256G 4.7吋/5.5吋 ...
$8,200
已售出 835
免費換新電池🍎 iPhone X / iPhone xr 64G/256G 5....
$11,000
已售出 202
現貨Apple se/iPhone SE 4吋福利機
$1,500
已售出 93
[正蘋果3C]全場最低價 Apple/ iPhone SE 原裝 4...
$1,788
已售出 14
Apple iPhone se 64g 金色 玫瑰金 太空灰漂亮95成...
$4,999
已售出 23
Apple iPhone se 64g 顏色 齊全 95成新盒裝 可免卡...
$4,500
已售出 12
猜你喜歡
查看全部 >
現貨Apple se/iPhone SE 4吋福利機
$1,500
已售出 93
二手機 iPhone SE 16G 16 32G 32 64G 64 128G 12...
$3,700
已售出 158
保固一年🍎部分現貨►AppleSE iPhoneSE 16G...
$1,999
已售出 422
降價 Apple iphone SE 4吋 64G 128G 實體店 台中 ...
$2,855
已售出 38
賣完10隻漲價！ 現貨 iphone SE 16G 32G 64G...
$2,899
已售出 10
Apple iPhone SE 64G16G/128G 9.5成新 ...
$3,465
已售出 6

真實機器學習案例(Quora)



https://www.quora.com/q/quoraengineering/A-Machine-Learning-Approach-to-Ranking-Answers-on-Quora


Training
Ground Truth Data
Scores
Features
Sample Answers
User Surveys
Usage Metrics
...
Feature Extraction
HBase
User Actions
Upvotes
Downvotes
...
Predictive Model
Prediction
Answer
Features
Scores

AI發展

AI發展已成為不可逆的趨勢

- 人臉辨識
- 推薦系統
- ....

各行各業開始使用

大數據＋模型節省人力,物力,時間

udn 聯合新聞網 (新聞發布)

自動化與無人化的軍事革新：軍用無人載具的下一階段發展

無人載具進一步快速發展的關鍵，在於結合AI人工智慧。 ... 戰機飛行，由人工智慧系統負責絕大部份的飛行工作，讓一架傳統戰機上的飛行員 ... 由此可見，軍用無人機的發展已跨入空戰的領域中，自動化飛行將取代人工遙控的

1 day ago

UDN 聯合新聞網 (新聞發布)

未來還需人類工作嗎？迎AI時代，與機器共存的預習

《人類未來方程式》中不迴避地一再提起人工智慧將取代基本勞力的主張，但同時也強調人類與機器合作能創造更高的生產力，並帶給人類更好的生活 ...

3 weeks ago

科技新報 TechNews

離全面自動化雖還早，但機器人勞力確實越來越便宜了

... 不斷發展的人工智慧所威脅甚至反噬的情節在文學作品中層出不窮，機器人取代重複性的勞動工作也是公認的自動化高度發展後可以預見的後果。

4 weeks ago

民報 (新聞發布)

【民報】【專欄】人工智慧拯救建築業

建築公司開始使用AI和機器學習來更好地規劃工作中勞力和機械的分配。機器人 ... 儘管有大量失業預測，人工智慧不太可能取代人力資源。相反，它 ...

5 days ago

何時不用AI

Allie K. Miller • 3rd+
Forbes AI Innovator of the Year | Artificial Intelligence at Amazon | LinkedIn To...
1mo •

When is machine learning NOT a good idea?

✅ No data (due to budget or access)
✅ A rules-based solution works
✅ Low ROI for your business
✅ You just want something cool
✅ No tolerance for mistakes
✅ No one to maintain it

AI學習門檻(從理論出發不適合初學者)

5. Logistic regression is an important binary classification technique in machine learning that builds off of the concepts of linear regression. Recall that in linear regression, there is a set of predictor variables (or features) $\mathbf{a}_i \in \mathbb{R}^d,\ i = 1, \ldots, n$, with corresponding outcome variables $b_i \in \mathbb{R},\ i = 1, \ldots, n$ and our aim is to fit a linear model to the data that minimizes a quadratic loss, namely

$$\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2} \sum_{i=1}^{n} \Big(\langle \mathbf{a}_i, \mathbf{x} \rangle - b_i \Big)^2 .$$

It can be easily seen that the solution to the above least-squares is the same as

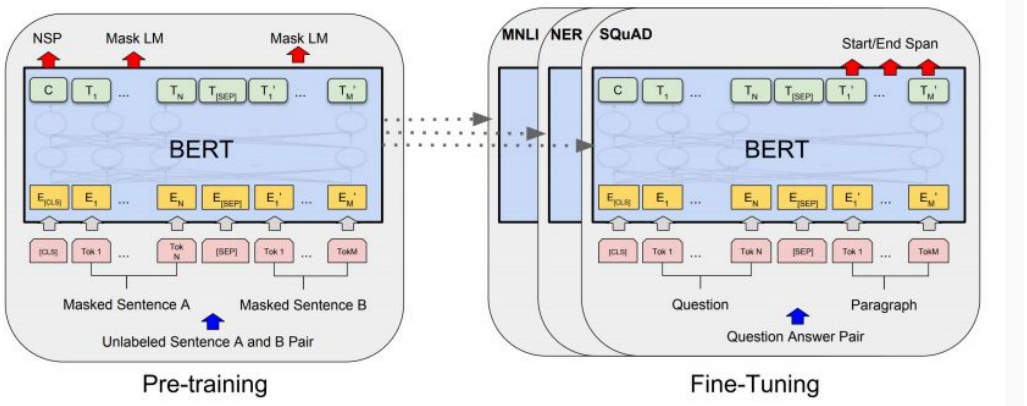
$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \triangleq \sum_{i=1}^{n} \phi(\langle \mathbf{a}_i, \mathbf{x} \rangle) - b_i \langle \mathbf{a}_i, \mathbf{x} \rangle ,$$

where $\phi(t) = t^2/2$. Logistic regression can be thought of as a modification of linear regression in two ways: first, the outcome variables are binary representing the two classes, i.e., $b_i \in \{0, 1\},\ i = 1, \ldots, n$, and second, the least-squares loss is replaced with a logistic loss, i.e., $\phi(t) = \ln(1 + e^t)$, where "ln" is natural logarithm. Logistic regression is thus the problems of finding the parameter $\mathbf{x}$ that minimizes this new loss (note that unlike linear regression, there is no closed form solution for logistic regression and one has to resort to optimization algorithms). In machine learning, this phase is often referred to as "training". After the training phase, we now have a way of classifying a new input, that is not part of your training data, i.e., predicting the label of an out-of-sample data. This phase in machine learning is often called "testing" or "generalization". More specifically, using our trained logistic model and given any new data $\mathbf{a}$, we can obtain the probability of $\mathbf{a}$ belonging to either class 0 or 1 as

$$\mathbf{Pr}(y = 0 \mid \mathbf{a}) = \frac{1}{1 + e^{\langle \mathbf{a}, \mathbf{x} \rangle}}, \quad \text{and} \quad \mathbf{Pr}(y = 1 \mid \mathbf{a}) = \frac{1}{1 + e^{-\langle \mathbf{a}, \mathbf{x} \rangle}}.$$

(a) Derive the gradient and the Hessian matrix of $f(\mathbf{x})$.

AI學習門檻(許多抽象的概念跟前因後果)



實際解決問題or面試

1. 定義問題跟分析問題重要度大於ML model
2. 合理的代理問題, (你怎麼量化user的滿意度？Ads, Youtube, Alibaba)
3. model用最簡單的再慢慢提升即可(Baseline model)
4. 實際題目(第三堂課):
    a. 如何設計Facebook news feed ranking
    b. 如何幫Facebook設計一個檢測NSFW content的系統

AI相關職業

Machine learning/ AI researcher(PhD)

Data scientist(PhD ~ SQL Boy)

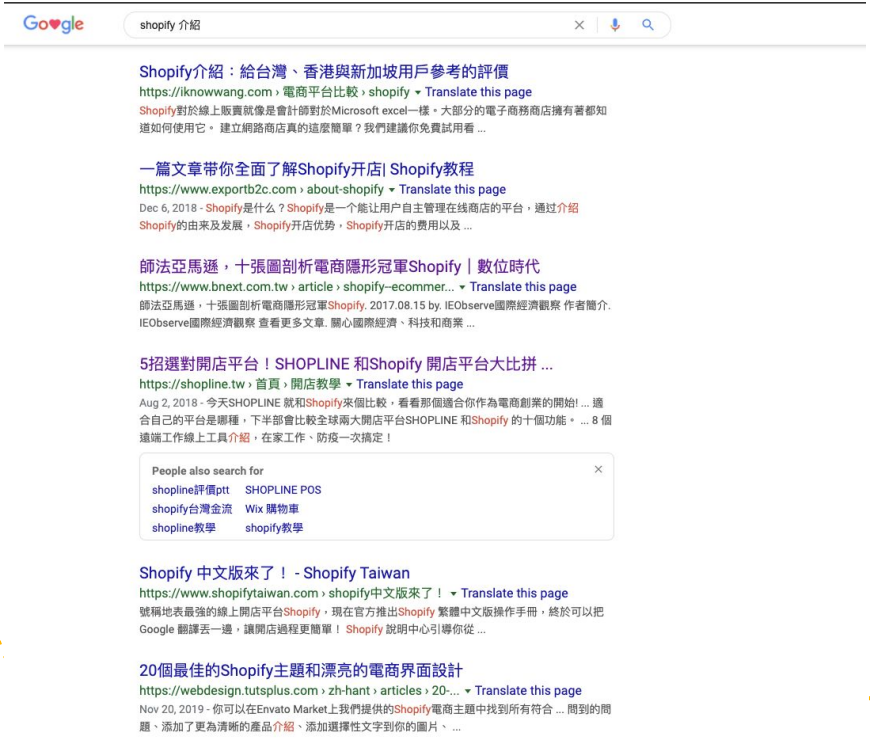
Machine learning/ AI Engineer

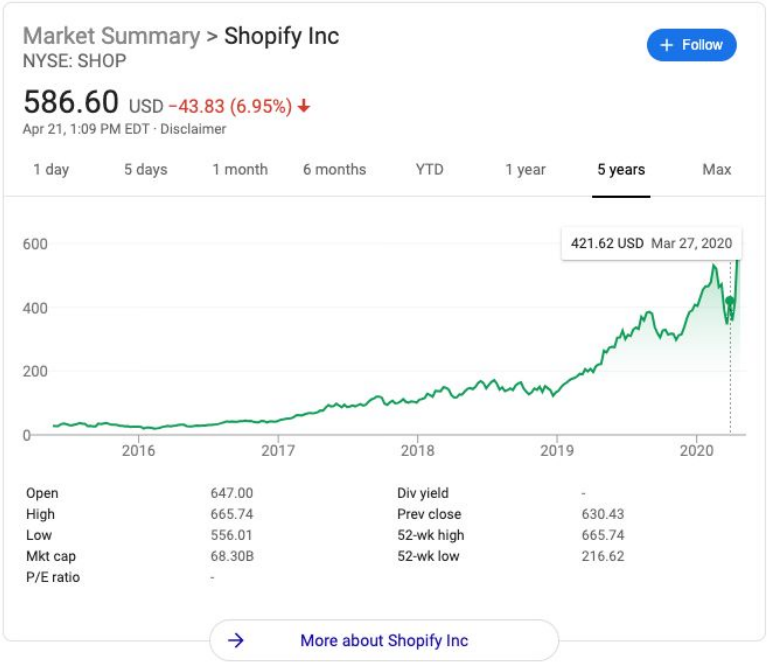
Business analyst

人類學習-觀察業界典範

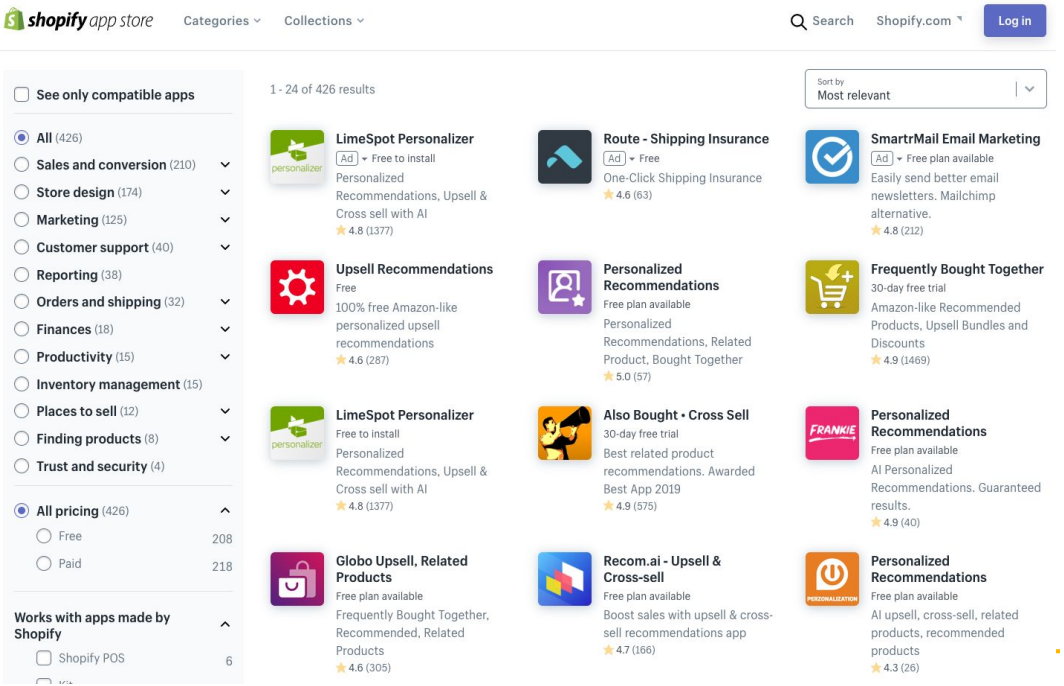
Shopify

Shopify介紹

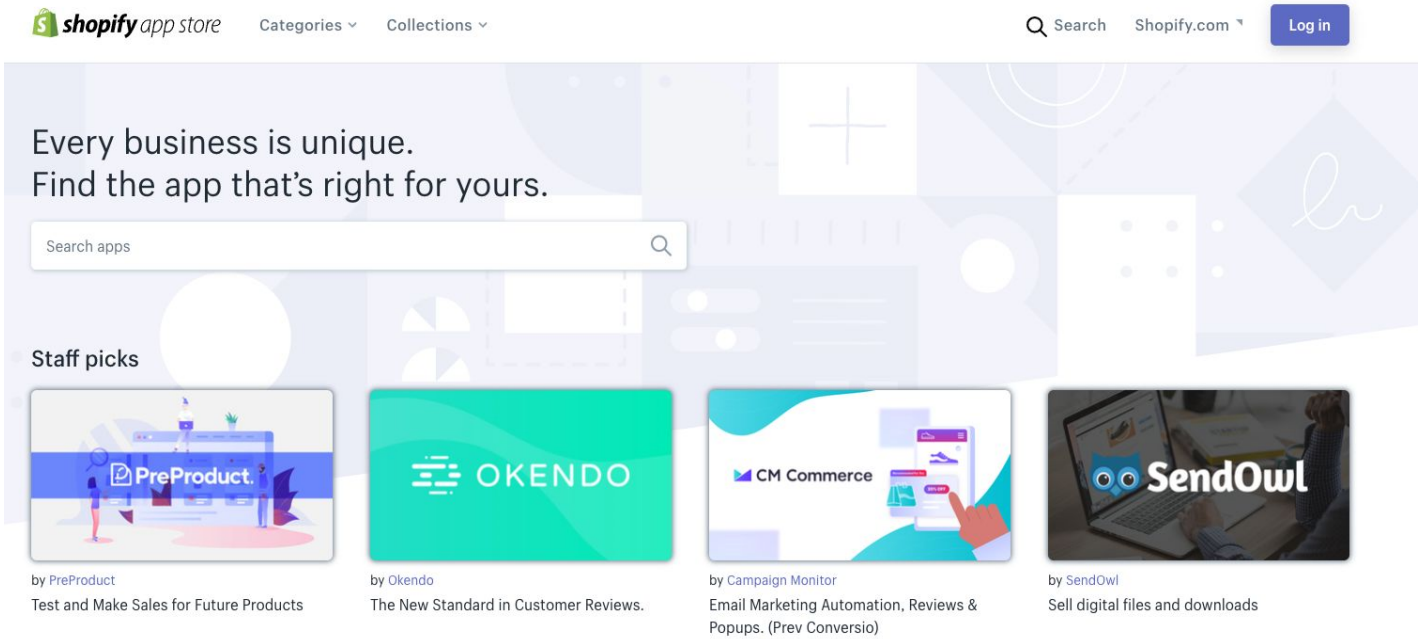

Google
shopify 介紹
Shopify介紹：給台灣、香港與新加坡用戶參考的評價
https://iknowwang.com › 電商平台比較 › shopify Translate this page
Shopify對於線上販賣就像是會計師對於Microsoft excel一樣。大部分的電子商務商店擁有著都知道如何使用它。 建立網路商店真的這麼簡單？我們建議你免費試用看 ...
一篇文章帶你全面了解Shopify开店| Shopify教程
https://www.exportb2c.com › about-shopify Translate this page
Dec 6, 2018 - Shopify是什么？Shopify是一个能让用户自主管理在线商店的平台，通过介绍Shopify的由来及发展，Shopify开店优势，Shopify开店的费用以及 ...
師法亞馬遜，十張圖剖析電商隱形冠軍Shopify｜數位時代
https://www.bnext.com.tw › article › shopify-ecommer... Translate this page
師法亞馬遜，十張圖剖析電商隱形冠軍Shopify. 2017.08.15 by. IEObserve國際經濟觀察 作者簡介. IEObserve國際經濟觀察 查看更多文章. 關心國際經濟、科技和商業 ...
5招選對開店平台！SHOPLINE 和Shopify 開店平台大比拼 ...
https://shopline.tw › 首頁 › 開店教學 Translate this page
Aug 2, 2018 - 今天SHOPLINE 就和Shopify來個比較，看看那個適合你作為電商創業的開始! ... 適合自己的平台是哪種，下半部會比較全球兩大開店平台SHOPLINE 和Shopify 的十個功能。 ... 8 個遠端工作線上工具介紹，在家工作、防疫一次搞定！
People also search for
shopline評價ptt
SHOPLINE POS
shopify台灣金流
Wix 購物車
shopline教學
shopify教學
Shopify 中文版來了！ - Shopify Taiwan
https://www.shopifytaiwan.com › shopify中文版來了！ Translate this page
號稱地表最強的線上開店平台Shopify，現在官方推出Shopify 繁體中文版操作手冊，終於可以把Google 翻譯丟一邊，讓開店過程更簡單！ Shopify 說明中心引導你從 ...
20個最佳的Shopify主題和漂亮的電商界面設計
https://webdesign.tutsplus.com › zh-hant › articles › 20-... Translate this page
Nov 20, 2019 - 你可以在Envato Market上我們提供的Shopify電商主題中找到所有符合 ... 問到的問題、添加了更為清晰的產品介紹、添加選擇性文字到你的圖片、 ...


Market Summary > Shopify Inc
NYSE: SHOP
Follow
586.60 USD −43.83 (6.95%)
Apr 21, 1:09 PM EDT · Disclaimer
1 day
5 days
1 month
6 months
YTD
1 year
5 years
Max
421.62 USD Mar 27, 2020
600
400
200
0
2016
2017
2018
2019
2020
Open 647.00
High 665.74
Low 556.01
Mkt cap 68.30B
P/E ratio -
Div yield -
Prev close 630.43
52-wk high 665.74
52-wk low 216.62
More about Shopify Inc

Shopify App store



Shopify App store



Shopify 推薦系統介紹

Share Tweet Share Share

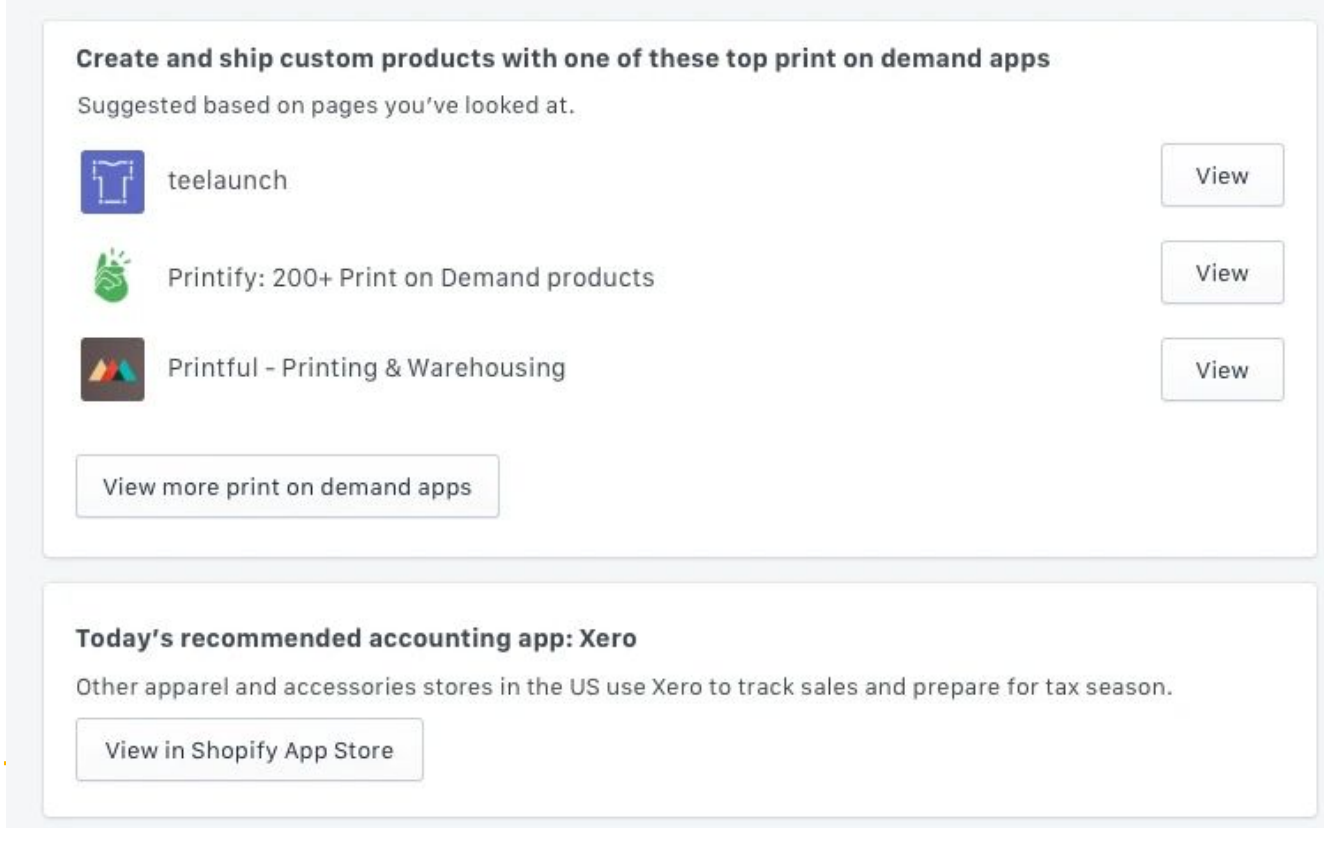
Authors: Dóra Jámbor and Chen Karako

There is a good chance you have come across a "recommended for you" statement somewhere in our data-driven world. This may be while shopping on Amazon, hunting for new tracks on Spotify, looking to decide what restaurant to go to on Yelp, or browsing through your Facebook feed — ranking and recommender systems are an extremely important feature of our day-to-day interactions.

This is no different at Shopify, a cloud-based, multi-channel commerce platform that powers over 600,000 businesses of all sizes in approximately 175 countries. Our customers are merchants that use our platform to design, set up, and manage their stores across multiple sales channels, including web, mobile, social media, marketplaces, brick-and-mortar locations, and pop-up shops.

https://engineering.shopify.com/blogs/engineering/how-shopify-uses-recommender-systems-to-empower-entrepreneurs

Shopify 推薦系統-解釋性

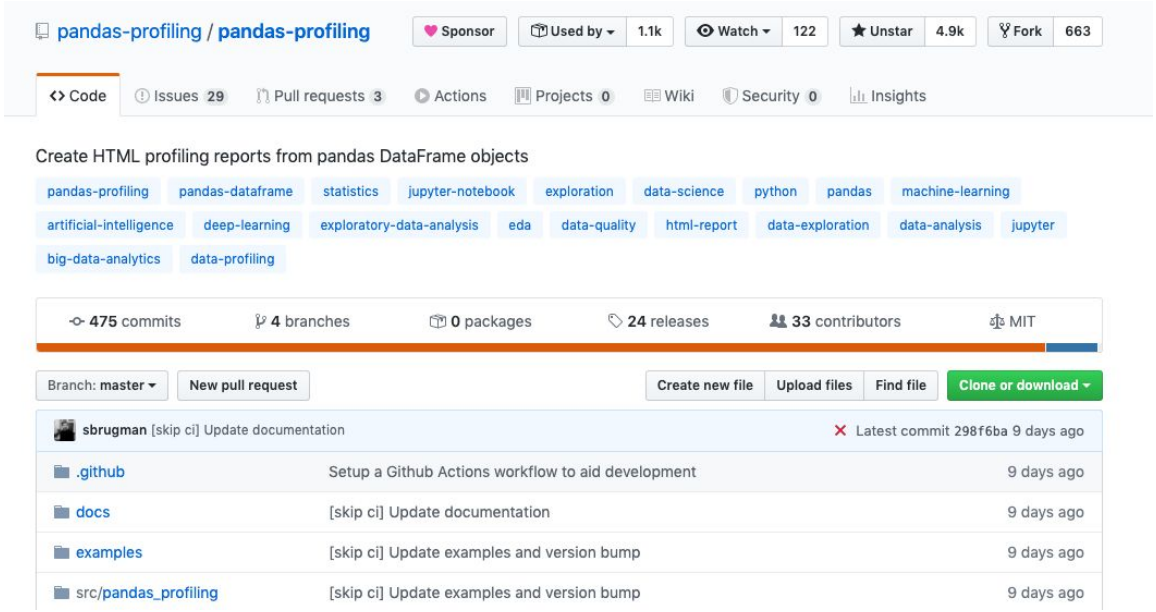


Shopify 推薦系統-結果

- Merchants receiving personalized app recommendations on the Shopify App Store had a 50% higher app install rate compared to those who didn't receive recommendations

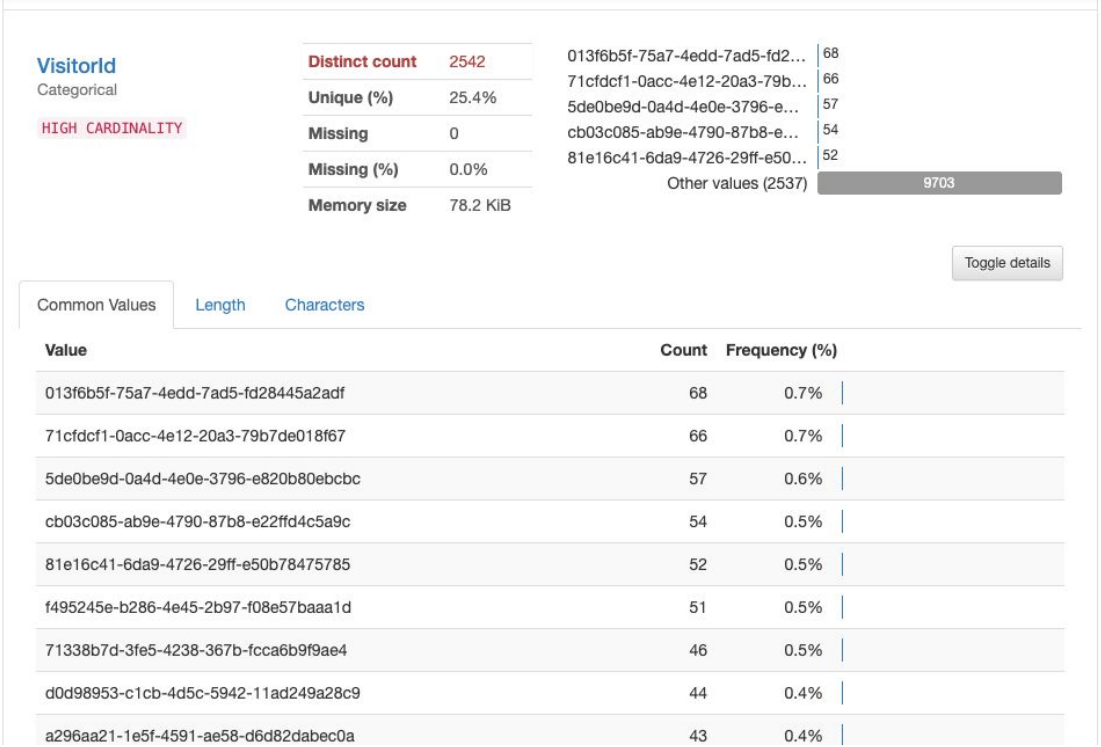
Data preprocess & analysis

資料探索神器



https://github.com/pandas-profiling/pandas-profiling#examples

客戶行為數



VisitorId
Categorical

HIGH CARDINALITY

Distinct count	2542
Unique (%)	25.4%
Missing	0
Missing (%)	0.0%
Memory size	78.2 KiB

013f6b5f-75a7-4edd-7ad5-fd2…	68
71cfdcf1-0acc-4e12-20a3-79b…	66
5de0be9d-0a4d-4e0e-3796-e…	57
cb03c085-ab9e-4790-87b8-e…	54
81e16c41-6da9-4726-29ff-e50…	52
Other values (2537)	9703

Toggle details

Common Values | Length | Characters

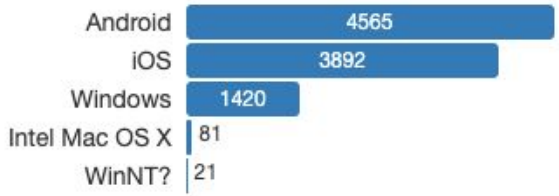
Value	Count	Frequency (%)
013f6b5f-75a7-4edd-7ad5-fd28445a2adf	68	0.7%
71cfdcf1-0acc-4e12-20a3-79b7de018f67	66	0.7%
5de0be9d-0a4d-4e0e-3796-e820b80ebcbc	57	0.6%
cb03c085-ab9e-4790-87b8-e22ffd4c5a9c	54	0.5%
81e16c41-6da9-4726-29ff-e50b78475785	52	0.5%
f495245e-b286-4e45-2b97-f08e57baaa1d	51	0.5%
71338b7d-3fe5-4238-367b-fcca6b9f9ae4	46	0.5%
d0d98953-c1cb-4d5c-5942-11ad249a28c9	44	0.4%
a296aa21-1e5f-4591-ae58-d6d82dabec0a	43	0.4%

裝置行為數

是手機App步驟比較複雜還是大家喜歡用手機App？

OperationSystem
Categorical

Distinct count	6
Unique (%)	0.1%
Missing	8
Missing (%)	0.1%
Memory size	78.2 KiB

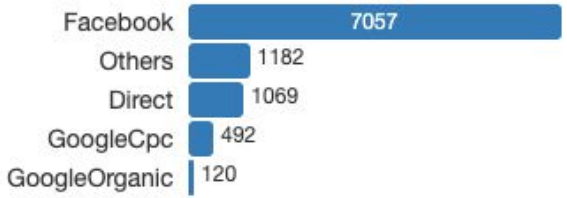


Toggle details

流量來源

TrafficSourceCateg...
Categorical

Distinct count	6
Unique (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	78.2 KiB



Toggle details

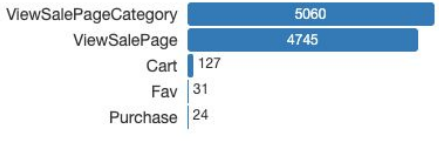
行為比例

```
profile.to_notebook_iframe()
```

Pandas Profiling Report — Overview | Variables | Interactions | Correlations | Missing values | Sample

BehaviorType
Categorical

Distinct count	6
Unique (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	78.2 KiB



Toggle details

Common Values | Length | Characters

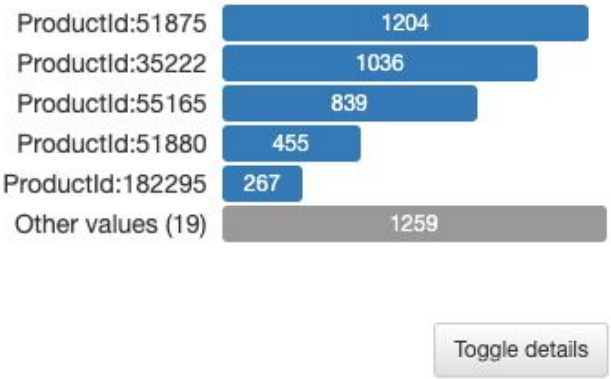
Value	Count	Frequency (%)
ViewSalePageCategory	5060	50.6%
ViewSalePage	4745	47.4%
Cart	127	1.3%
Fav	31	0.3%
Purchase	24	0.2%
Search	13	0.1%

類別比例

CategoryId
Categorical

MISSING

Distinct count	24
Unique (%)	0.5%
Missing	4940
Missing (%)	49.4%
Memory size	78.2 KiB



Hack一下
https://www.so-nice.com.tw/v2/official/SalePageCategory/51875?sortMode=Curator
https://www.so-nice.com.tw/v2/official/SalePageCategory/35222?sortMode=Curator

商品価格

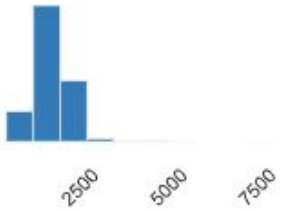
ProductPrice

Real number ($\mathbb{R}_{\geq 0}$)

`MISSING`

Distinct count	32
Unique (%)	0.7%
Missing	5104
Missing (%)	51.0%
Infinite	0
Infinite (%)	0.0%

Mean	1659.1584967320262
Minimum	390.0
Maximum	7980.0
Zeros	0
Zeros (%)	0.0%
Memory size	78.2 KiB



Toggle details

客單價

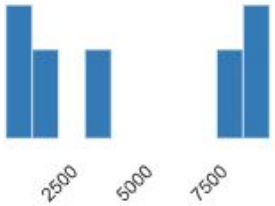
TransactionRevenue

Real number ($\mathbb{R}_{\geq 0}$)

`MISSING`

Distinct count	12
Unique (%)	50.0%
Missing	9976
Missing (%)	99.8%
Infinite	0
Infinite (%)	0.0%

Mean	4896.25
Minimum	670.0
Maximum	9480.0
Zeros	0
Zeros (%)	0.0%
Memory size	78.2 KiB

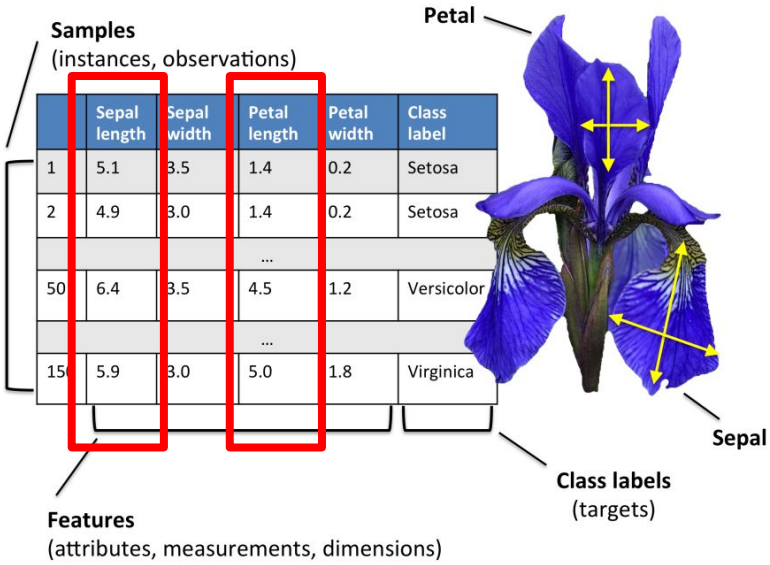


Toggle details

Live Demo 分析

Logistic Regression

資料集- Iris data set



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica

給花瓣、萼片 長跟寬

預測Iris花的品種

資料集- Iris data set

IRIS dataset

Iris Versicolor

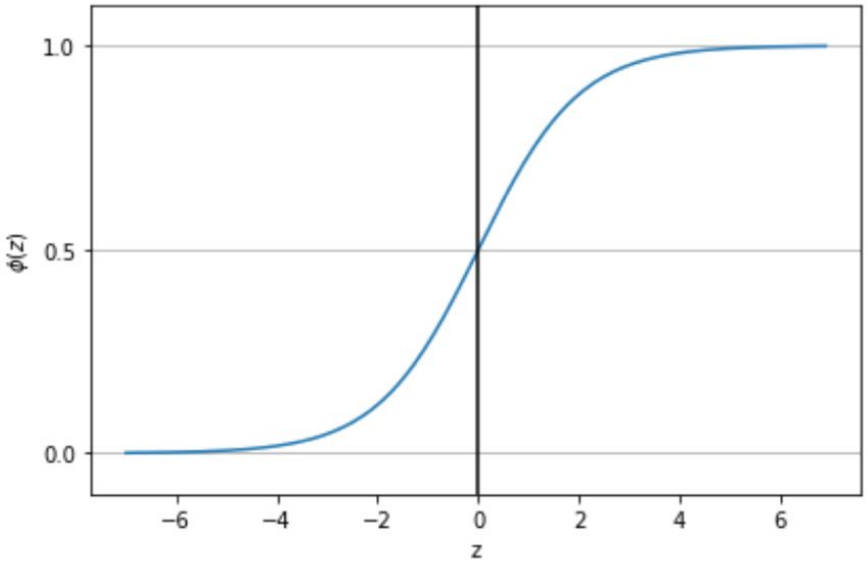
Iris Setosa

Iris Virginica

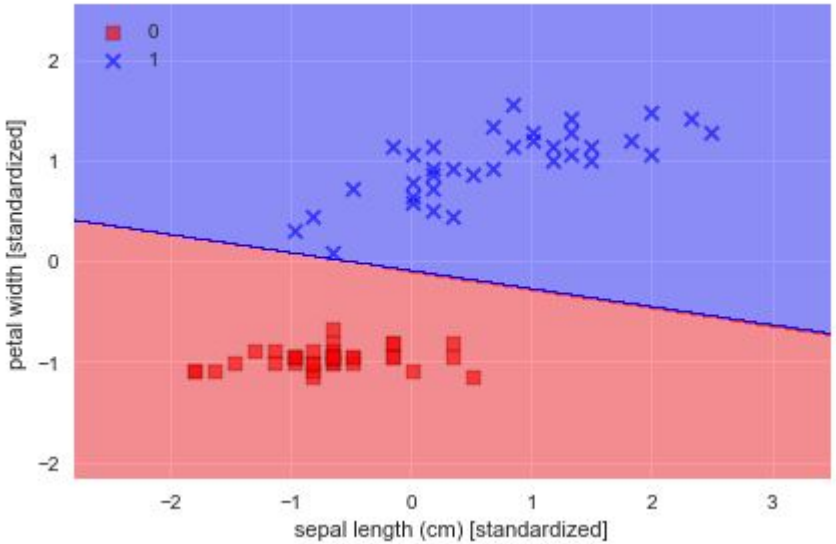
Logistic Function

$$z = w^T x$$

$$\phi(z) = \frac{1}{1 + e^{-z}}$$



Logistic Regression 能做什麼？
線性分類



aX+bY +c = 0

Live Demo

課後學習資源

學習資源

Blog

[資料分析&機器學習] 第2.3講：Pandas 基本function介紹(Series, DataFrame, Selection, Grouping)

[資料分析&機器學習] 第2.4講：資料前處理(Missing data, One-hot encoding, Feature Scaling)

[資料分析&機器學習] 第2.5講：資料視覺化(Matplotlib, Seaborn, Plotly)

[資料分析&機器學習] 第3.1講：Python 機器學習以及Scikit-learn介紹

[資料分析&機器學習] 第3.2講：線性分類-感知器(Perceptron) 介紹

[資料分析&機器學習] 第3.3講：線性分類-邏輯斯回歸(Logistic Regression) 介紹

書

Python 機器學習

學習資源

Github

https://github.com/Avik-Jain/100-Days-Of-ML-Code

Q & A

Thanks!