Data Science Final Project Report

Team24 組員：趙仰生、郭蕙綺 貢獻程度相同

1. Coding 實作：在資料收集的部分，將資料從蝦皮與露天的平台上爬下並整理。而在預測賣出數量的部分，我們利用平台上有的features(蝦皮：標題、評價、評價數量、價格，露天：標題、評價、商家的商品數量、價格)，還有label(售出數量)來進行，其中除了標題之外，都是已經量化好的數值，而標題我們利用jieba進行斷詞與出現次數的分析，將標題轉換成一個數值，這個數值會是這個標題所有單詞在所有商品標題中出現次數的加總(只有出現頻率前一百名的單詞有分數，其餘為零，因為不希望一些沒有意義的單詞影響分數，且有過濾符號)，之後利用MinMaxScaler來標準化，然後將資料shuffle(因為我們的資料是利用平台熱銷程度來爬的，所以會有前面都是很熱賣，後面都很滯銷的資料分布，利用shuffle可以避免這個問題)，並且做outlier detection(後面會更詳細解釋)，之後將testing與training資料分開(Test = 200筆，Training = All – Test)，使用xgboost 這個模型來預測，最後得到下面所有表格：（outlier detection 以下簡稱”OD”）
2. 以下根據三個不同種類(服飾、配件、家電)進行資料基本觀察、商品預測、標題Top10單詞的分析。其中關於商品標題的分析，我們利用爬下來所有資料的標題進行單詞前10名的排序，並將其中較有意義的單詞留下，例：販售背心時，最常出現的單詞即為背心，所以“背心”可以被刪除，以及“袖”這個單詞也較無意義，也將其刪除。
   - 服飾：
     - 基本觀察：在服飾類的部分，顯而易見的是蝦皮的銷售明顯較露天好，女生服飾的部分全數均由蝦皮勝出，而男裝的部分露天勝出的是短袖、長袖、及外套這些上衣類，所以我們可以得出“服飾商品在蝦皮上有較好的表現，男裝的部分可以視情況放到露天的平台做販售”
     - 預測：
       - 蝦皮：預測算是蠻準的，可以看到容錯在10件內平均有75%，20件內平均有85%的準確率，但可以發現少數的類別準確率非常低，觀察資料後，推估是因為有的商家價格賣的超級貴或是有的商家實在太熱賣，於是進行outlier detection之後，平均準確率提昇了不少。
       - 露天：露天相對就沒有蝦皮那麼準了，推估是因為，露天拍賣少了評價數量這個feature，實際上，顧客在蝦皮上購物的時候，都會觀察評價數量來決定此商家值不值得信任(包括我們自己在內)，而露天多的商家商品數量卻相對的不是顧客觀察此商家值不值得信任的重點。
     - 前10名單詞：除了讓賣家更了解標題應該要有什麼關鍵字之外，同時也可以觀察產品的趨勢、潮流。不管是在哪一類服飾、不分男女，幾乎都可以看到“韓版”這個單詞的出現，可以看出趨勢是韓系風，賣家在販售時可以考慮關於韓系的服裝。
       - 女生服飾：“新款”這個詞是在多數女生服飾均有出現而男生卻沒有，可見女生較注重款式的新舊，推薦賣家可以在販售女生衣物時將標題加入“新款”，會較有競爭力；在長袖上衣的部分較常出現“長、寬”的詞，賣家也可以朝這方面著手販售的商品及標題；長褲、短褲的部分則露天和蝦皮有很大的不同，露天較偏重於“運動”方面，蝦皮則是日常的“牛仔、高腰”的商品。

- 男生服飾：外套及長袖部分長出現“夾克”、“襯衫”的詞，可以推薦賣家販售相關產品，長褲的部分露天較適合販售運動相關，蝦皮則是休閒相關；短褲的部分則是均有“運動、五分”的詞出現，可以看出在男生短褲的趨勢；在短袖及背心部分也是以運動為主，休閒部分則建議販售於蝦皮。
- 以上是對服飾產品標題做得一些分析，賣家可以根據得到的表格，更深入研究產品走向及標題的適合性。

蝦皮-服飾(資料數/預測準度)	容錯10件以內No OD	容錯10件以內OD(price>2000/sold>5000)	容錯20件以內No OD	容錯20件以內OD(price>2000/sold>5000)	售出總數量	標題Top10單詞
外套女	2717/0.85	2689/0.825	2717/0.925	2689/0.93	186,145	新款、韓版、上衣、寬、襯衫、長、牛仔
外套男	1595/0.825	1517/0.81	1595/0.94	1517/0.88	91,856	襯衫、韓版、夾克、潮流、連帽、休閒、現貨
長袖女	3269/0.76	3259/0.79	3269/0.89	3259/0.88	327,041	襯衫、新款、外套、韓版、寬
長袖男	1392/0.135	1378/0.76	1392/0.365	1378/0.82	141,660	襯衫、外套、韓版、上衣、襯衣、潮流
長褲女	3353/0.75	3343/0.76	3353/0.845	3343/0.875	379,699	牛仔、新款、九分、寬、高腰、鬆
長褲男	3018/0.225	2985/0.85	3018/0.735	2985/0.905	335,911	牛仔、九分、韓版、寬、潮流
背心女	3067/0.56	3058/0.765	3067/0.825	3058/0.86	293,076	吊帶、打底、內衣、運動、無袖、性感、搭
背心男	1173/0.24	1162/0.705	1173/0.8	1162/0.815	198,240	運動、無袖、健身、夏季、坎肩、T恤
短袖女	3484/0.725	3578/0.73	3584/0.875	3578/0.86	323,211	T恤、新款、夏季、韓版、寬、鬆
短袖男	3320/0.805	3309/0.795	3320/0.905	3309/0.875	217,713	夏季、T恤、襯衫、潮流、寬
短褲女	3201/0.665	3186/0.79	3201/0.865	3186/0.9	339,226	牛仔、夏季、新款、寬、高腰、鬆、韓版
短褲男	2981/0.865	2970/0.855	2981/0.945	2970/0.92	148,105	五分、休閒、夏季、運動、牛仔、鬆、寬

露天-服飾(資料數/預測準度)	容錯10件以內No OD	容錯10件以內OD(price>2000/sold>5000)	容錯20件以內No OD	容錯20件以內OD(price>2000/sold>5000)	售出總數量	標題Top10單詞
外套女	3861/0.505	3423/0.71	3861/0.775	3423/0.8	112,192	連帽、風衣、新款、長、夾克、韓版
外套男	6623/0.59	5509/0.55	6623/0.8	5509/0.78	194,870	夾克、連帽、風衣、休閒、韓版、修身
長袖女	4720/0.625	4575/0.705	4720/0.74	4575/0.805	147,901	上衣、T恤、外套、襯衫、新款、衣服
長袖男	6237/0.605	6039/0.63	6237/0.74	6039/0.755	229,114	襯衫、T恤、男、運動
長褲女	1798/0.325	1734/0.545	1798/0.635	1734/0.75	112,350	運動、慢跑、愛迪達、緊身、健身、運動褲
長褲男	3960/0.21	3804/0.62	3960/0.745	3804/0.75	204,831	牛仔、運動、男女、休閒褲、男士、修身

露天-服飾(資料數/預測準度)	容錯10件以內No OD	容錯10件以內OD(price>2000/sold>5000)	容錯20件以內No OD	容錯20件以內OD(price>2000/sold>5000)	售出總數量	標題Top10單詞
背心女	1689/0.175	1621/0.445	1689/0.66	1621/0.7	63,501	內衣、運動、上衣、裙、無袖
背心男	2453/0.21	2401/0.5	2453/0.57	2401/0.625	181,138	運動、無袖、健身、外套、內衣
短袖女	6673/0.65	6583/0.72	6673/0.84	6583/0.87	249,500	T恤、新款、短、夏季
短袖男	7177/0.415	7052/0.655	7177/0.645	7052/0.72	364,413	T恤、衣服、衫、運動、夏季
短褲女	1439/0.24	1423/0.43	1439/0.6	1423/0.64	65,887	運動、牛仔、休閒、鬆、新款、慢跑
短褲男	2842/0.17	2768/0.505	2842/0.665	2768/0.69	129,209	五分、運動、休閒、健身、夏季、內褲

- 配件：
  - 基本觀察：配件類的部分則全數均是蝦皮賣的較好，而其中又為耳環、襪子更為突出，總售出數量超過千萬，露天賣的較好的則是手錶及背包，所以我們會推薦賣家要販售配件類商品時可以選擇蝦皮為他們的平台，當他們選擇露天為平台時，也可以多販售手錶及背包類產品，可能會有較好的銷售數量。
  - 預測：
    - 蝦皮與露天預測都非常不準，觀察資料後發現，配件類的outlier data超出平均非常非常多，像是耳環竟然有商家可以售出百萬筆，於是進行outlier detection之後，平均準確率提昇了不少，不過相較於服飾，配件類的預測確實遜色許多。
  - 前10名單詞：
    - 手錶：露天有“CASIO”這個詞，代表在露天上品牌銷售較多，蝦皮則有“防水、電子、運動”，產品較多元。
    - 皮帶：兩家都著重於“真皮”，可見材質在皮帶這部分的重要性。
    - 耳環、項鍊：有“鈦、鋼、銀針、皮”為關鍵字，材質是重要的一環，露天則是還有”情人、情侶“出現，可見對練方面可能較適合於露天販售。
    - 眼鏡：眼鏡的部分多數以墨鏡為主。
    - 其實在配件部分蝦皮和露天只有些許差異，較明顯的不同是蝦皮常出現”蝦皮、優選“，表示蝦皮優選的賣家較被大家關注，關於該如何下商品標題，賣家可利用此表格作參考。

蝦皮-配件(資料數/預測準度)	容錯10件以內No OD	容錯10件以內OD(price>2000/sold>5000)	容錯20件以內No OD	容錯20件以內OD(price>2000/sold>5000)	賣出總數量	標題Top10單詞
手錶	3806/0.005	3398/0.455	3806/0.005	3398/0.685	2,604,184	防水、現貨、手環、電子、石英、兒童、運動、優選
皮帶	3750/0.19	3647/0.625	3750/0.43	3647/0.695	800,315	腰帶、現貨、真皮、優選、男士
耳環	3907/0.005	3621/0.095	3907/0.005	3621/0.165	17,872,844	優選、皮、蝦、韓國、現貨、銀針、耳夾、簡約
背包	3906/0.0	3765/0.435	3906/0.0	3765/0.675	4,715,765	現貨、防水、書包、優選、雙肩包、大容量

蝦皮-配件(資料數/預測準度)	容錯10件以內No OD	容錯10件以內OD(price>2000/sold>5000)	容錯20件以內No OD	容錯20件以內OD(price>2000/sold>5000)	賣出總數量	標題Top10單詞
眼鏡	3831/0.26	3747/0.595	3831/0.545	3747/0.71	1,062,277	太陽、墨鏡、現貨、優選、偏光、皮、蝦
帽子	3912/0.12	3867/0.355	3912/0.355	3912/0.415	1,267,959	現貨、漁夫帽、遮陽帽、棒球帽、鴨舌帽、優選、老帽、韓版
項鍊	3852/0.035	3754/0.345	3852/0.045	3754/0.45	3,232,228	優選、現貨、配件、材料、皮、飾品
錢包	3794/0.0	3606/0.59	3794/0.0	3606/0.695	9,731,048	夾、長、零錢、短夾、現貨、卡位
襪子	3988/0.005	3645/0.035	3988/0.02	3645/0.06	10,536,219	襪、襪子、短襪、現貨、優選、中筒、皮、棉襪、隱形、蝦

露天-配件(資料數/預測準度)	容錯10件以內No OD	容錯10件以內OD(price>2000/sold>5000)	容錯20件以內No OD	容錯20件以內OD(price>2000/sold>5000)	賣出總數量	標題Top10單詞
手錶	7426/0.04	6720/0.38	7426/0.235	6720/0.595	1,432,500	帶、防水、妖怪、電池、手環、男、CASIO、電子
皮帶	7469/0.2	7134/0.475	7469/0.67	7134/0.75	428,186	腰、扣、腰帶、原廠、製、真皮
耳環	7285/0.24	7203/0.36	7285/0.505	7203/0.62	639,741	耳洞、韓國、鈦、鋼、款、鍊、無、鑽
背包	7519/0.08	7171/0.105	7519/0.13	7171/0.285	1,119,125	後、側、斜、防水、扣、現貨、腰包、登山
眼鏡	7493/0.035	7219/0.13	7493/0.085	7219/0.31	927,969	太陽、偏光、鏡片、運動、鼻墊、墨鏡
帽子	7330/0.46	7086/0.52	7330/0.58	7086/0.705	334,035	棒球帽、鴨舌帽、遮陽帽、保暖、現貨、男女、韓版
項鍊	7319/0.175	7186/0.275	7319/0.455	7186/0.535	561,134	鈦、鋼、情人、飾品、情侶、禮物
錢包	7344/0.17	7210/0.45	7344/0.52	7210/0.67	394,494	零錢、皮夾、夾、長、手機、皮包、鍊
襪子	7064/0.355	6948/0.555	7064/0.7	6948/0.72	341,667	襪、襪子、短襪、毛巾、雙、棉襪、1、元、製、免運

- 家電：
  - 基本觀察：在家電類的部分，則是較大型的家電，如：電風扇、烤箱、微波爐、飲水機、吸塵器，在露天上的銷售量較好，而小型家電則是在蝦皮上的表現較佳，可以推測出學生族群、小資族群可能較常使用蝦皮(因為他們較不會也較不需要購買大型電器)，所以在販售電器的方面，我們推薦賣家可以在露天販售大型電器、蝦皮則是販售小型電器。
  - 預測：
    - 蝦皮：預測不太準，觀察資料後發現，資料數實在是太少了，像是微波爐甚至連兩百筆都不到，讓我們深刻體悟到資料的重要性，沒有資料、或是資料太少的狀況下，能做的事情實在是少很多。

- 露天：相對的就在家電類比較出色，因為資料數筆蝦皮多很多，但是因為受限於沒有評價數量這個feature，準確率也沒有辦法達到像蝦皮的服飾類那樣準確。

- 前10名單詞：
  - 電風扇、電鍋：這兩樣商品在蝦皮及露天相似程度頗高，均注重於尺寸、材質、產地，所以賣家可以多著墨於這三個要素。
  - 飲水機：飲水機的部分則可以看出不同，上述有提到大型家電較適合在露天做販售，客群也較注重水質的部分，如：RO、淨水、濾芯，而蝦皮則是還有“寵物”這方面的詞出現，可以明顯看出客群及消費力的不同。
  - 微波爐、氣炸鍋：這兩樣商品在關鍵字的部分則是有許多周邊商品出現，如：手套、置物架、噴油瓶“的出現，推薦賣家在販售此類商品時可以一併販售其周邊商品。
  - 體重計：體重計的部分則是均有”小米、電子“的出現，可見小米的體重計較受歡迎，且電子體重計也取代傳統體重計。
  - 家電類的部分差異也不太大，推薦賣家可以從首先消費力著手進行平台的選擇，再根據表格選取適合標題，亦可參考販售相關周邊產品的必要性。

蝦皮-家電(資料數/預測準度)	容錯10件以內No OD	容錯10件以內OD(price>2000/sold>5000)	容錯20件以內No OD	容錯20件以內OD(price>30000/sold>5000)	賣出總數量	標題Top10單詞
電風扇	3862/0.19	3842/0.36	3862/0.415	3842/0.665	1,005,044	USB、迷你、手持、現貨、吋、電扇、優選
電鍋	746/0.43	746/0.415	746/0.78	746/0.78	79,294	人份、不、鋼、鏽、大同、10、TAC、製
烤箱	351/0.27	351/0.235	351/0.5	351/0.515	26,300	烘焙、墊、麵、現貨、優選、爐、公升、蝦皮
微波爐	資料太少	資料太少	資料太少	資料太少	7,971	NN、Panasonic、國際牌、公升、日本、微電腦、20L
氣炸鍋	415/0.03	411/0.155	415/0.055	411/0.39	78,937	、現貨、台灣、L、保固、健康、專用、科帥、免運
飲水機	383/0.12	383/0.17	383/0.295	383/0.295	54,516	寵物、優選、自動、智能、活水、現貨
吸塵器	1583/0.26	1578/0.52	1583/0.775	1578/0.76	232,344	無線、手持、現貨、車用、兩用、優選、濾網、小米
果汁機	942/0.41	941/0.45	942/0.77	941/0.835	103,623	榨汁、隨行杯、電動、迷你、現貨、優選、隨身
電磁爐	775/0.37	770/0.35	775/0.81	770/0.815	68,016	平底、盤、現貨、優選、適用、沾
體重計	563/0.09	559/0.19	563/0.2	559/0.41	141,798	電子、小米、體重機、現貨、智能

露天-家電(資料數/預測準度)	容錯10件以內No OD	容錯10件以內OD(price>2000/sold>5000)	容錯20件以內No OD	容錯20件以內OD(price>30000/sold>5000)	賣出總數量	標題Top10單詞
電風扇	5670/0.28	5545/0.35	5670/0.505	5545/0.575	460,905	吋、電扇、USB、迷你、灣、台

露天-家電(資料數/預測準度)	容錯10件以內No OD	容錯10件以內OD(price>2000/sold>5000)	容錯20件以內No OD	容錯20件以內OD(price>3000/sold>5000)	賣出總數量	標題Top10單詞
電鍋	2676/0.655	2662/0.63	2676/0.78	2662/0.84	65,865	人份、不、鋼、鏽、大同、製、10、灣
烤箱	2368/0.61	2312/0.58	2368/0.735	2312/0.715	60,567	爐、烘焙、手套、微波、廚房、蛋糕、麵
微波爐	2052/0.59	2026/0.62	2052/0.755	2026/0.76	48,186	架、手套、烤箱、廚房、爐架、置物架、收納架、電器
氣炸鍋	838/0.14	828/0.48	838/0.62	828/0.735	41,603	油瓶、配件、噴、紙、現貨、烘焙、噴霧
飲水機	3245/0.465	3233/0.59	3245/0.755	3233/0.715	132,319	淨、RO、淨水、貨號、濾心、過濾器
吸塵器	5067/0.07	4886/0.46	5067/0.38	4886/0.68	329,303	無線、配件、兩用、吸頭、手持、車用、濾網、專用、集塵袋
果汁機	1093/0.13	1080/0.165	1093/0.52	1080/0.485	74,247	榨汁、調理機、隨行杯、迷你、電動、隨身、現貨
電磁爐	1623/0.685	1617/0.715	1623/0.91	1617/0.9	23,302	平底、壺、鋼、適用、鏽、可用
體重計	558/0.005	551/0.08	558/0.02	551/0.315	102,740	電子、電子體、玻璃、電池、體重機、小米

3. 回測：我們希望回測看看好的標題是否真的可以影響賣出數量，利用蝦皮服飾這類(因為這類預測最準)，取除了標題之外所有feature的平均形成新的feature(包括評價、評價數量與價格)，並且取銷售排名前100的標題與所有商品的標題出現次數Top 10的單詞，再取交集(這樣做是因為不想要標題過於冗長，且是一個有利的標題，再者，便於使用者加入自己的店名與個人化的單詞)，形成一個新的標題，這樣我們就有一組，標題不錯而其他feature在平均值的一組data，最後將它丟進去模型裡，可以發現如下表，跟所有的商品平均賣出數量相比，這一組data確實可以賣出相對較多的商品。

	外套女	外套男	長袖女	長袖男	短袖女	短袖男
平均賣出數量	52	37	54	40	66	32
預測賣出數量	59	48	82	64	78	40

	背心女	背心男	長褲女	長褲男	短褲女	短褲男
平均賣出數量	56	80	91	52	70	29
預測賣出數量	58	107	103	68	85	39

4. 總結：在這次的project中，我們學到很多資料分析的實戰技巧，例如資料的標準化、shuffle還有outlier detection的重要性，也可以在過程中發現很多電商平台的生態以及特性，老實說，真的非常有趣！