Women’s Clothing Review

104403011范哲豪、104403044吳謦廷

Abstract

本篇研究以Ecommerce Women’s Clothing Review來探討幾個問題。在Sentimental Analysis 中及Product Ratings Prediction中，我們使用Naïve Bayes、XGBoost、以及Logistic Regression三個不同的分類器，其中以Logistic Regression 的結果最好。本篇也探討關於回應中中立回應的問題，以及其後續所能夠應用的商業場域。此篇研究確立對於網路回應分析中，較爲恰當的分析工具以及其能夠使用可能的方向。

1. Overview of the Study
   1. Tasks of the study

我們在此篇研究中要做兩個task。

* Sentiment Analysis

我們要將使用者的review判斷是否為正面的(positive)或是負面的(negative)，再試著將review以星星數分高低等級。

* Product Ratings Prediction

我們會從一則review中預測該篇的rating是多少。

* 1. Why we chose to do the tasks
* 第一個task我們想要研究的是review為正面或者是負面的，從中可以推斷出此使用者是否願意推薦此產品。
* 第二個task我們想要研究的是大家在留言及評價之間的標準相近程度，從review中推估rating的高低以瞭解大家在給與評價的標準是否一致。我們認爲這個task所能帶給主管或是客服人員的資訊不高，因爲從review推出ratings的數據唯一能夠得知的是該使用者是否誤植ratings的數值，對於產品的瞭解不大。

1. Brief literature review on relevant techniques

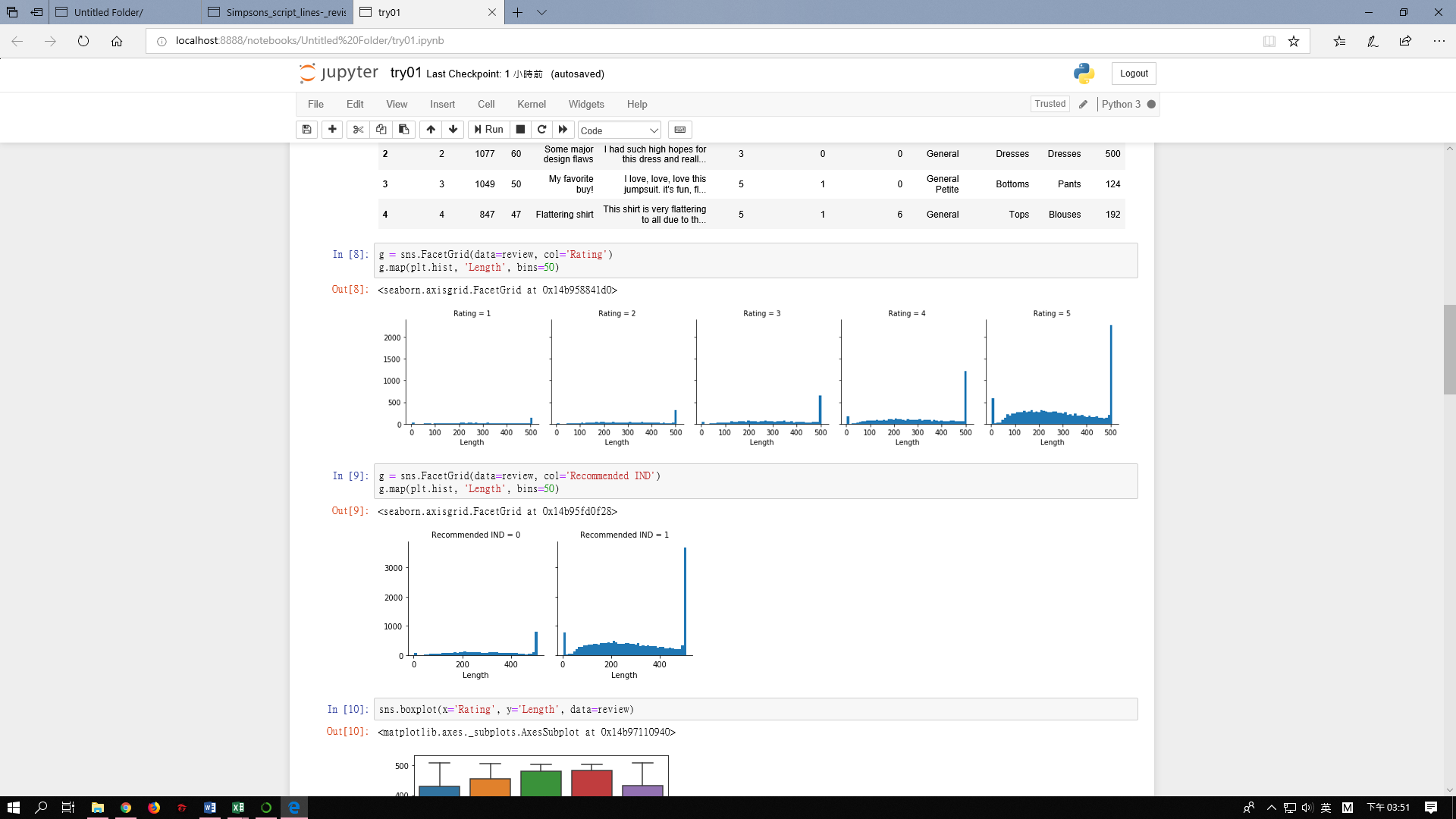
Regression methods: 這邊我們會使用到logistic regression來預測區間為1~5的ratings。Logistic regression 我們也稱作是sigmoid function，是一個S形狀的曲綫且y軸有限制介於1~0之間，而我們會得到的output是介於1~5的discrete numbers。

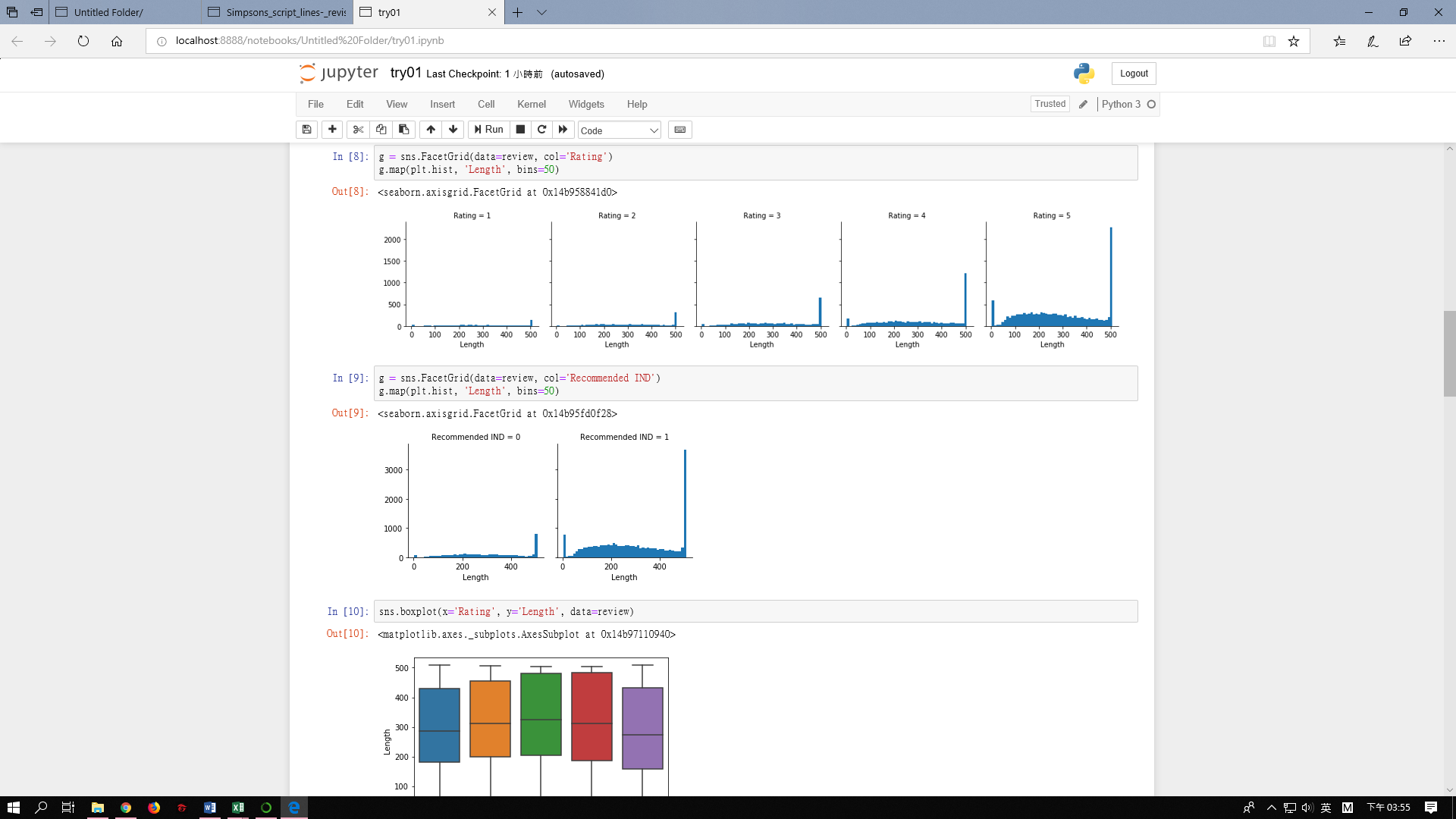
在這邊我們提出一個想法，我們在選擇使用什麽方法去推測rating的時候，認爲應該要先知道我們要拿這些rating資料的用處是什麽。如果說我們想要去除ratings低於2的產品，我們要使用的是binary logistic regression 比較恰當（要留下and丟掉）；但如果我們想將評分的使用者分成三大類以作行銷策略的不同，也就是rating=1 2分一類，rating=3 4分一類，rating=5分一類，那可能使用multi-class classification會比較適當。所以在瞭解需要拿此prediction做什麽用途後會比較清楚知道要用什麽方法。

如果使用的是multi-class classification，其所要達成的目的是要將review分成兩個以上的classes。每一則review都會分配到一個class，而在這裏我們可以分成五個class也就是rating1~5。這個classification的方法是binary classification的延伸，它能夠更清楚地瞭解各個欲分配的instance所在的class。我們可以使用naïve bayes，但是不能使用SVM這類的的classifier因爲它只能用作binary。

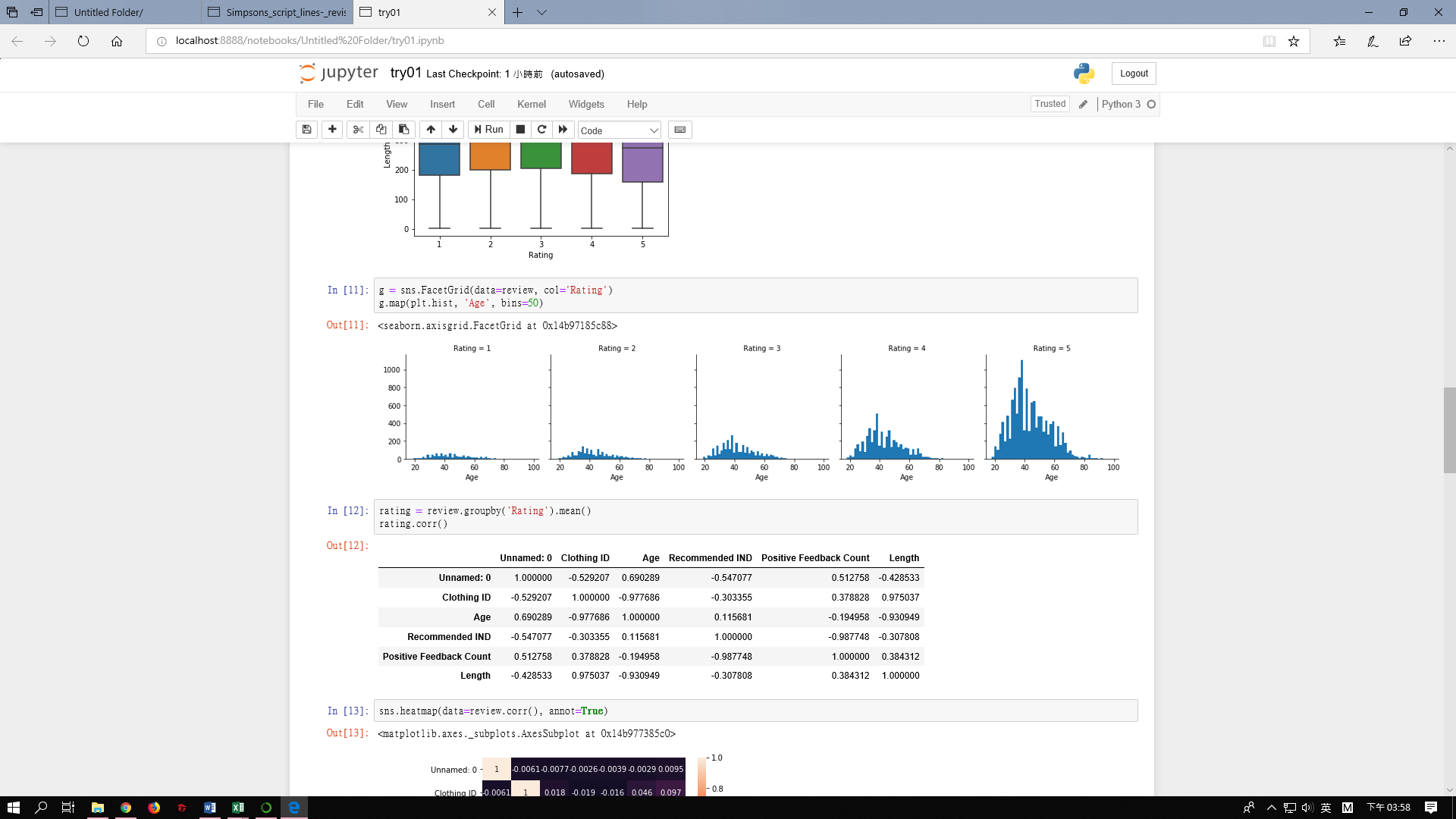
1. Methodologies
   1. Description/Stats of Dataset

我們在此使用的是kaggle上找的資料集Ecommerce clothing review。此資料集中總共有23486筆資料。





我們將每一篇review的字數長度做統計，發現review 的字數長短，對於不管是Rating的高低或是Recommended IND的不同都不會有顯著的變化，所以我們不建議用字數去作爲判斷的依據。我們也發現此資料中寫review的人為35~40歲佔大宗。如下圖。



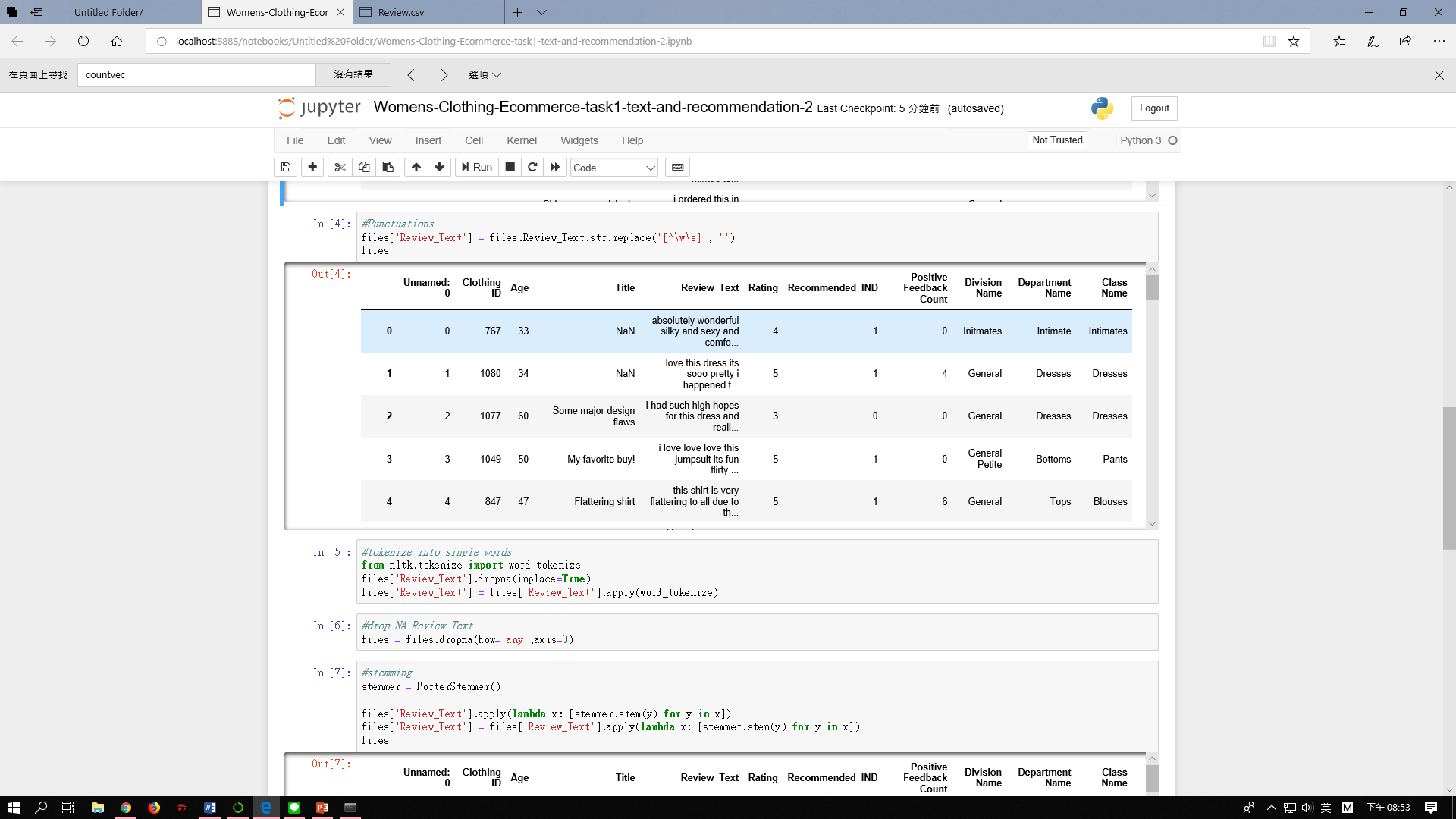
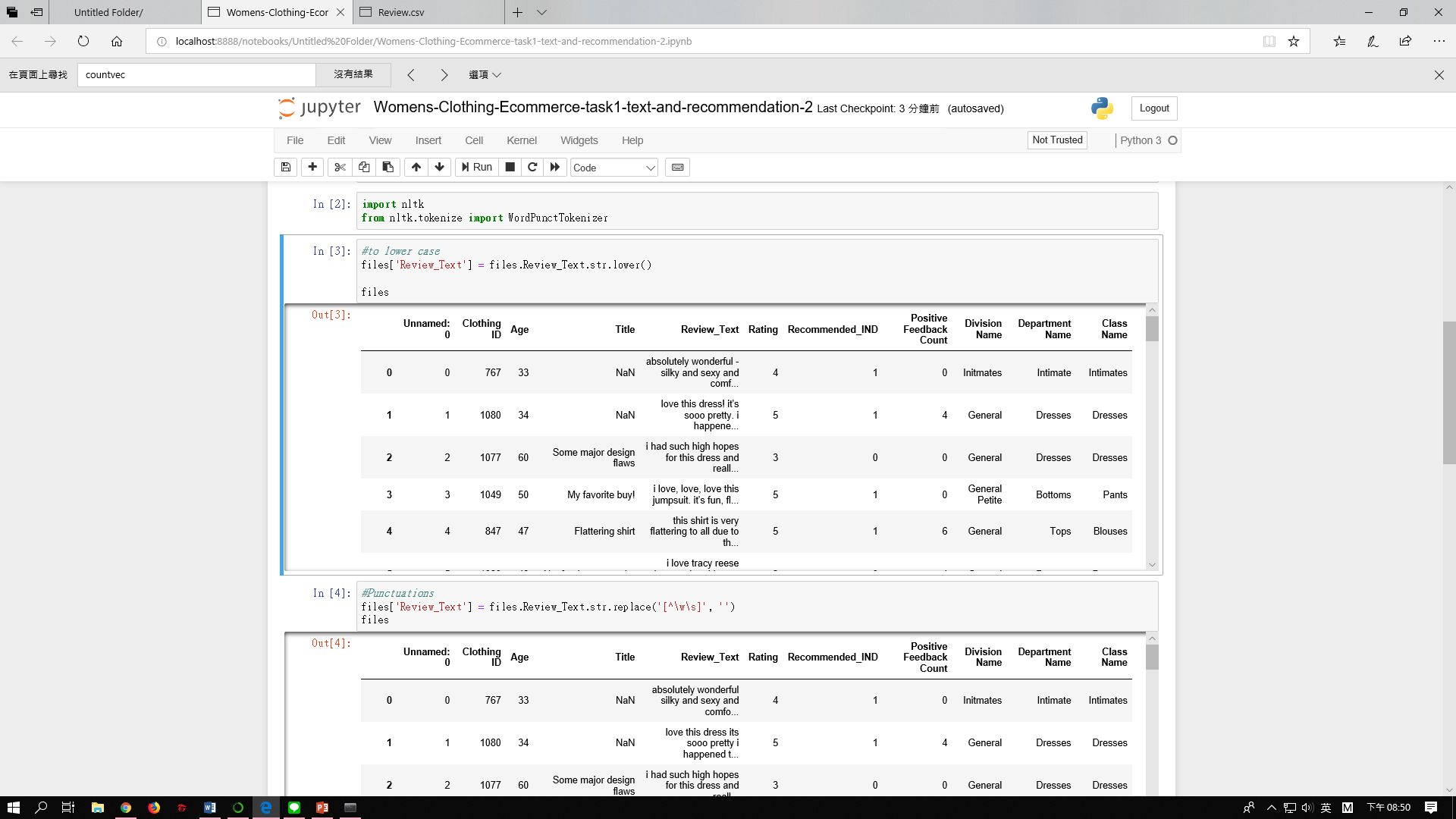
我們在此資料集中rating的這個column有1~5個類別，而我們經過數字的統計後，分析出的各類別的比例如下圖：

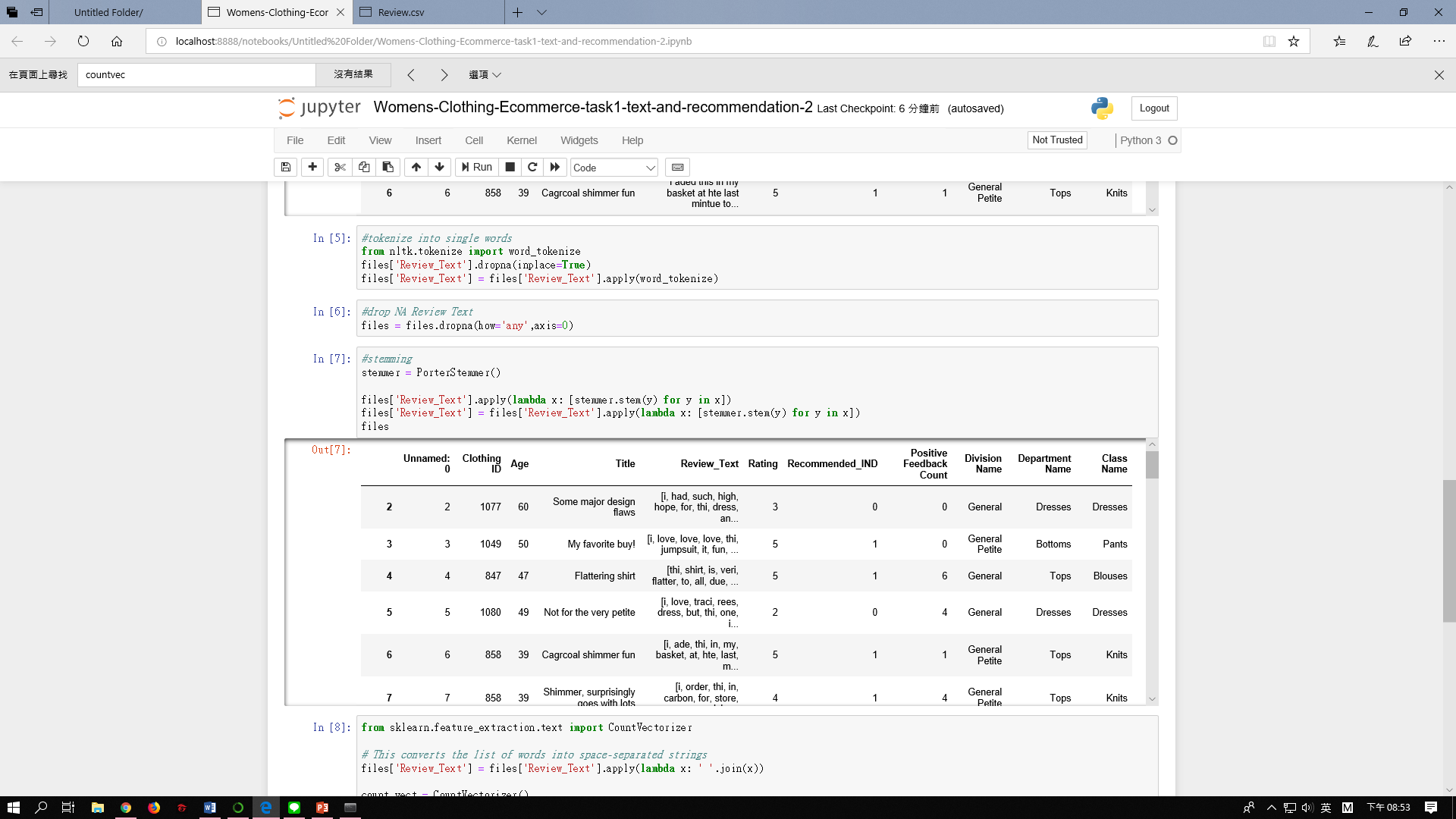
|  |  |  |
| --- | --- | --- |
| **Rating** | **Numbers of instances** | **Percentage** |
| 1 | 842 | 0.04 |
| 2 | 1565 | 0.07 |
| 3 | 2871 | 0.12 |
| 4 | 5077 | 0.21 |
| 5 | 13491 | 0.56 |

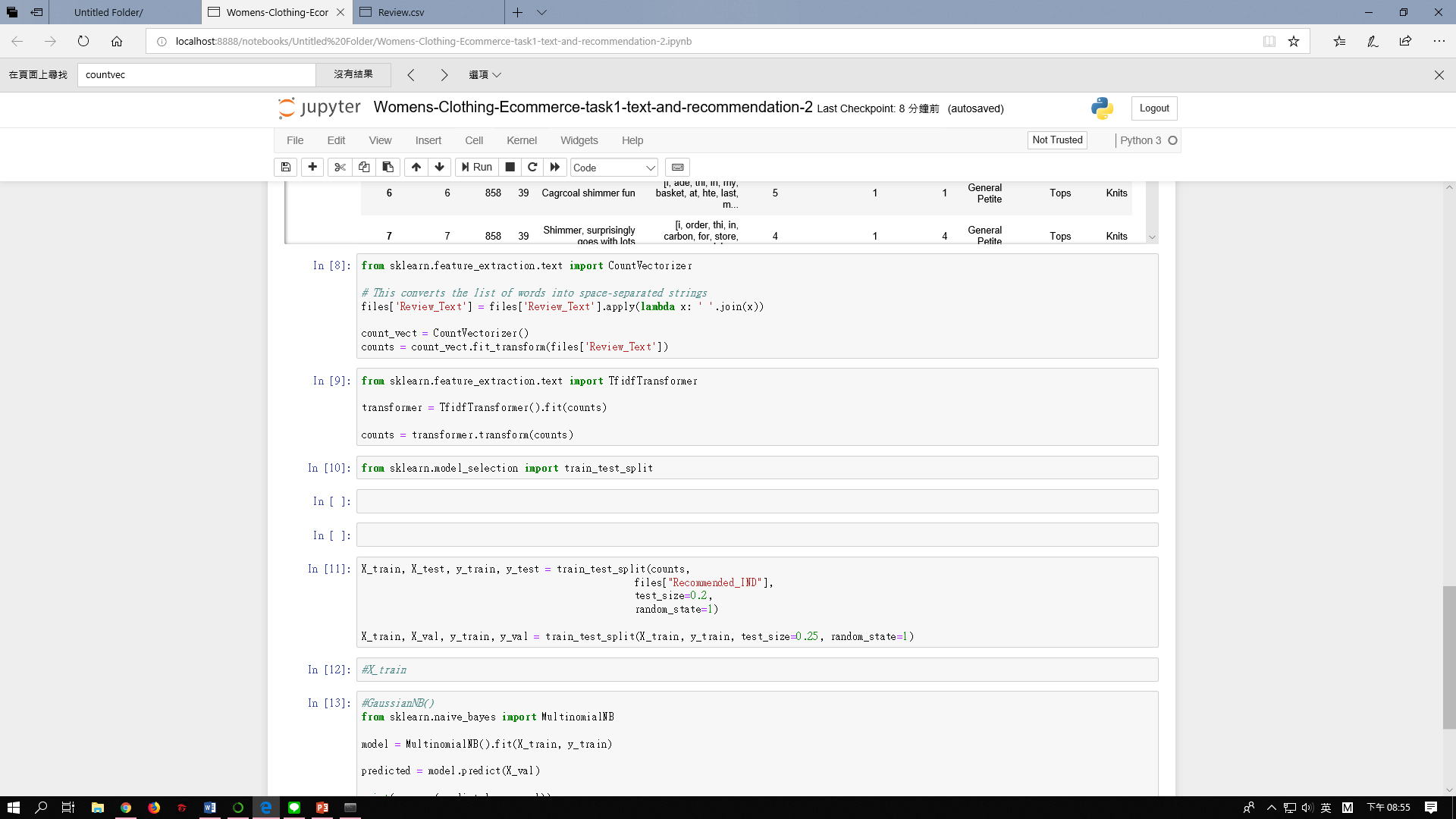
可見此資料集中類別4和5佔有超過70%，因此我們可以說這是個不平衡的資料集。針對不平衡資集我們將在會透過feature selection及sampling的方式處理。

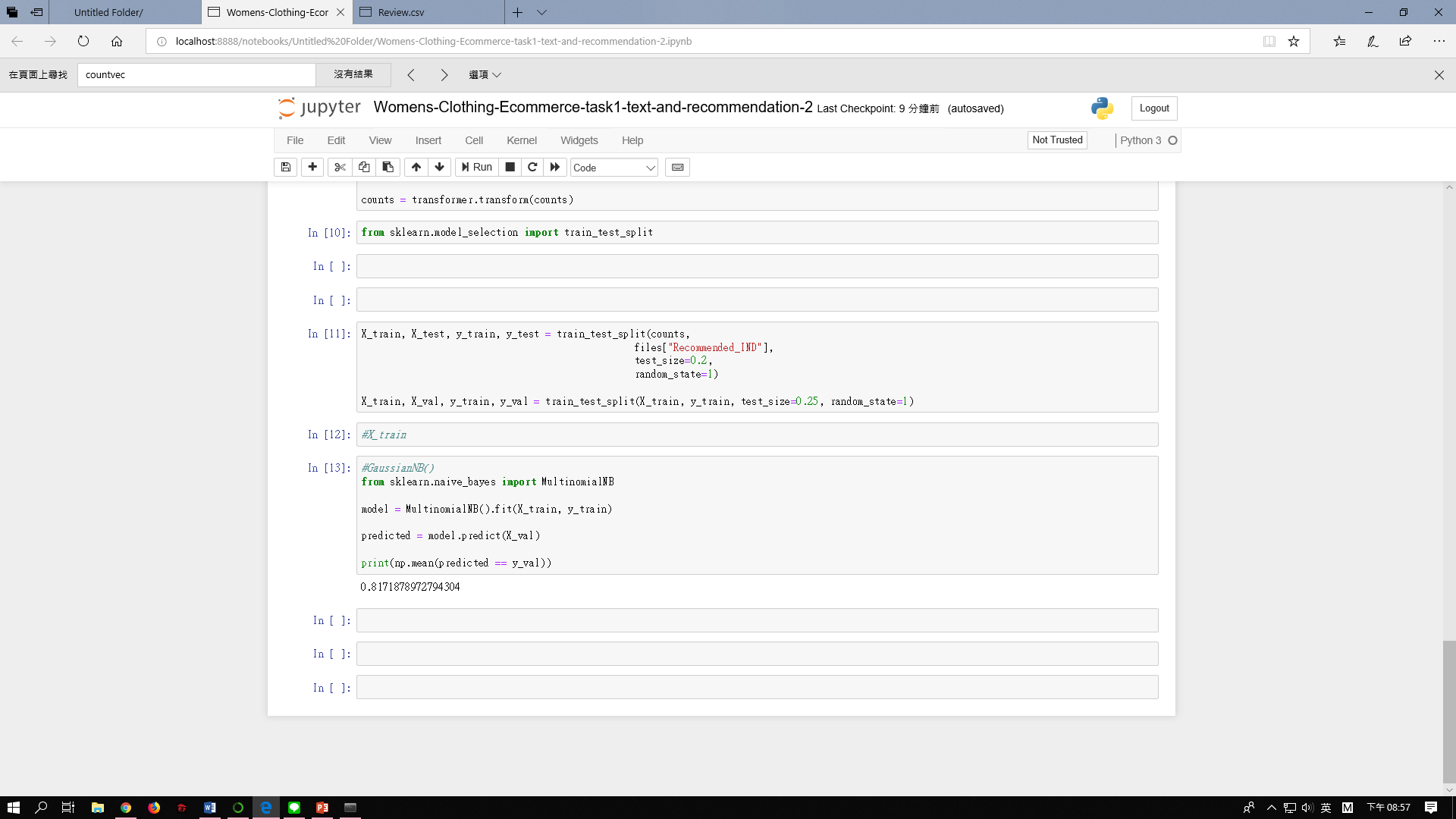
* 1. Experimental Setups: How you tried to achieve good classification performance

我們做前處理，包括：統一小寫、去除標點符號。



 Tokenize 及stemming。

Vectorising 及 TF-IDF。



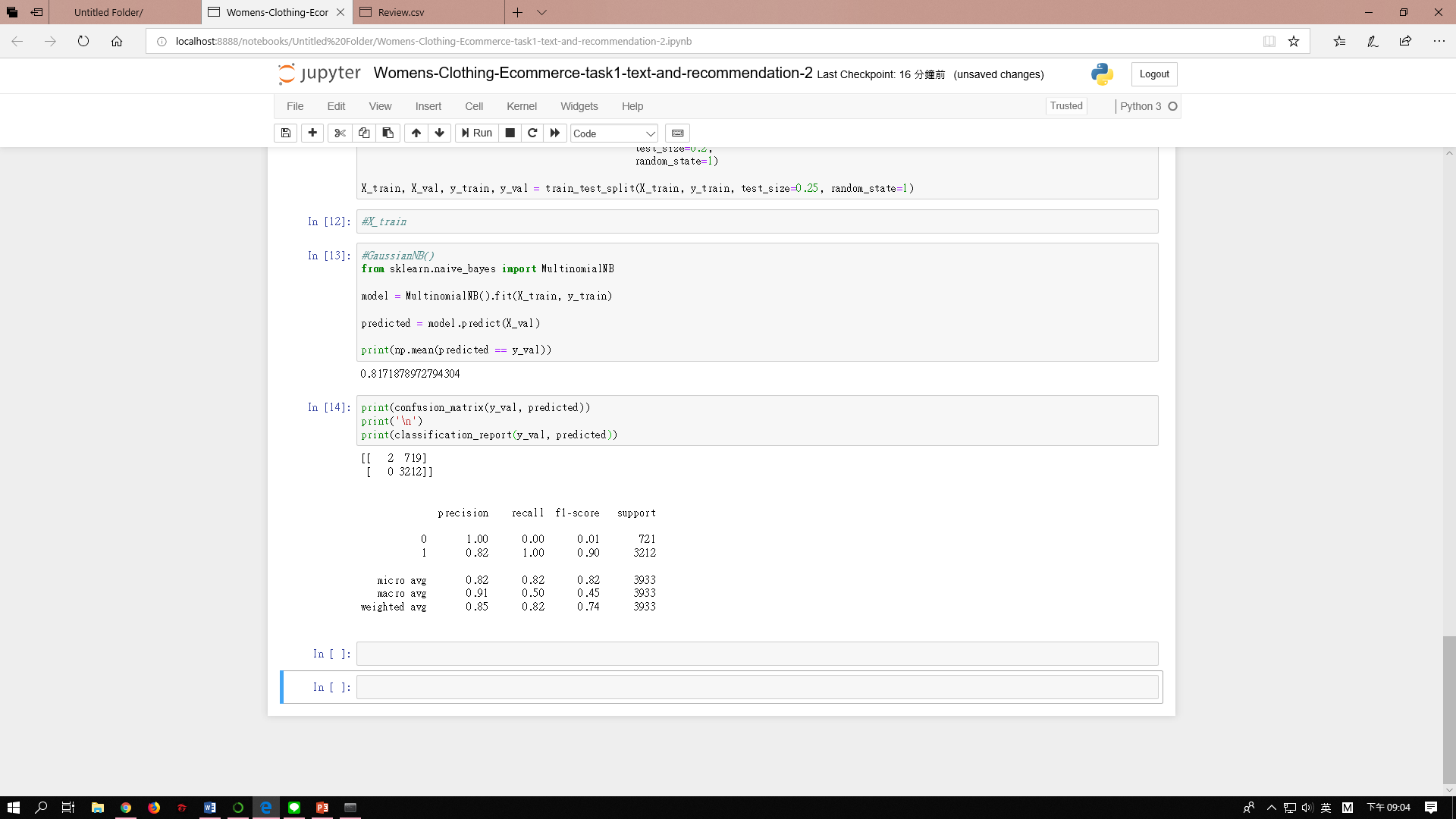
最後做train text split，先將資料分成train dataset及test dataset，在再將train dataset分成實際上train的及validation dataset。第一個task我們在此使用的是naïve bayes。

* 1. Explanation of study limitations

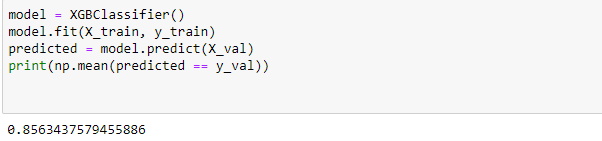
此資料集最大的困難與限制來自於我們前面提及的Data Imbalance現象，在Simpsons家族四人佔有大多數的資料的情況下，如果在訓練分類器之前，調整training資料，其實無論用哪一種分類器，其Precision幾乎不可能超過50%。

1. Results and Analysis

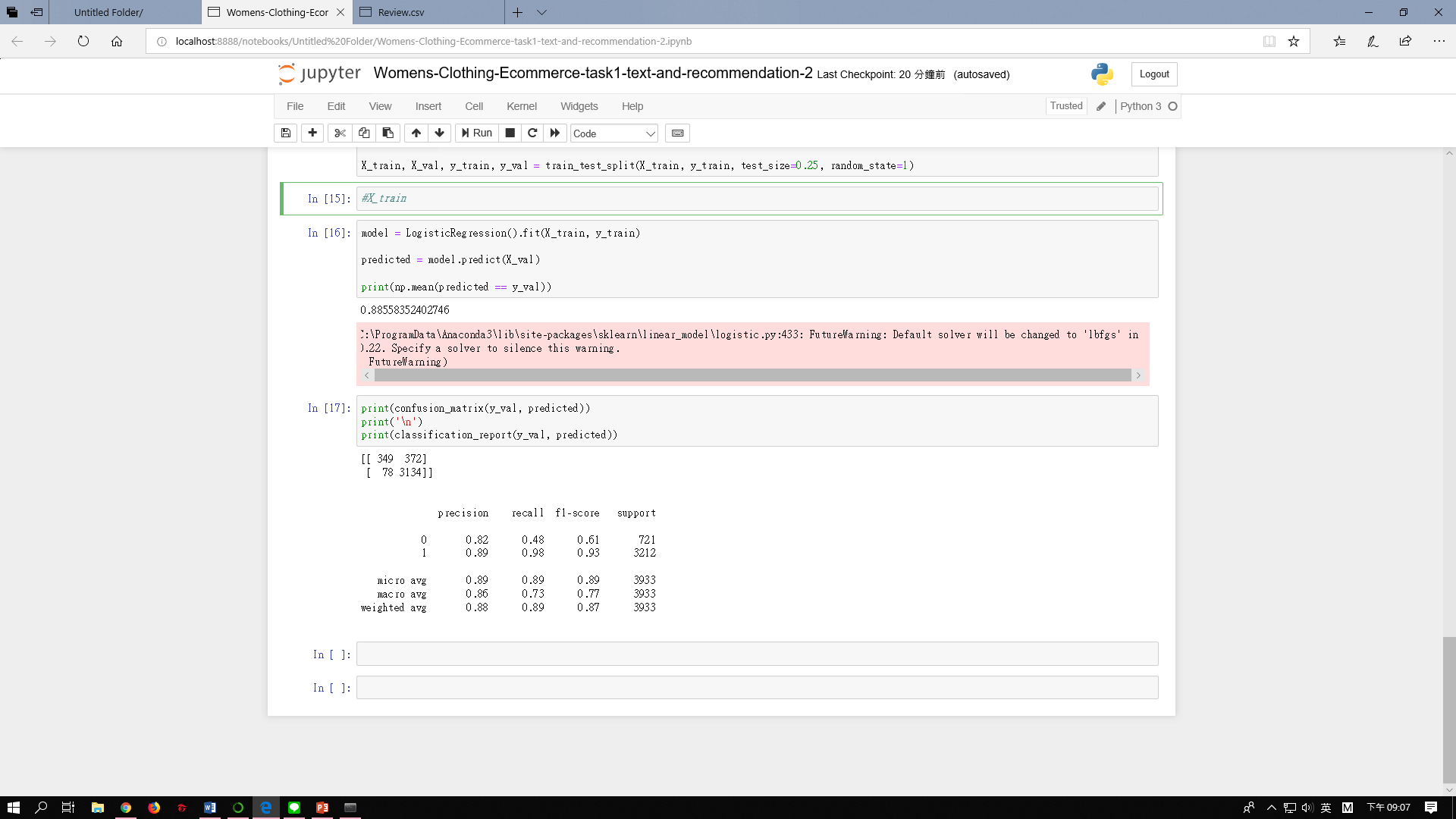
第一個task，使用Naïve bayes 的結果為0.82。



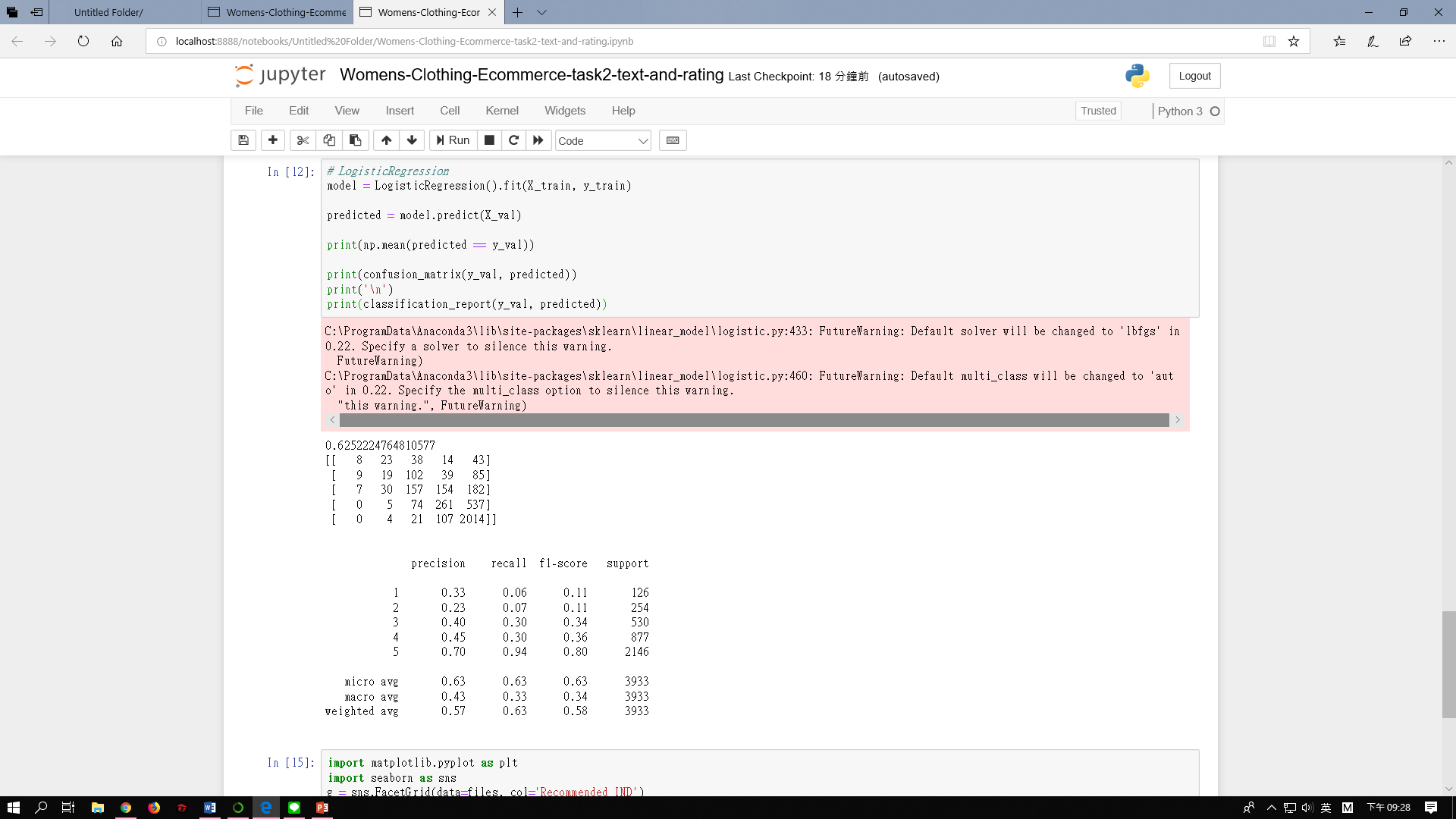
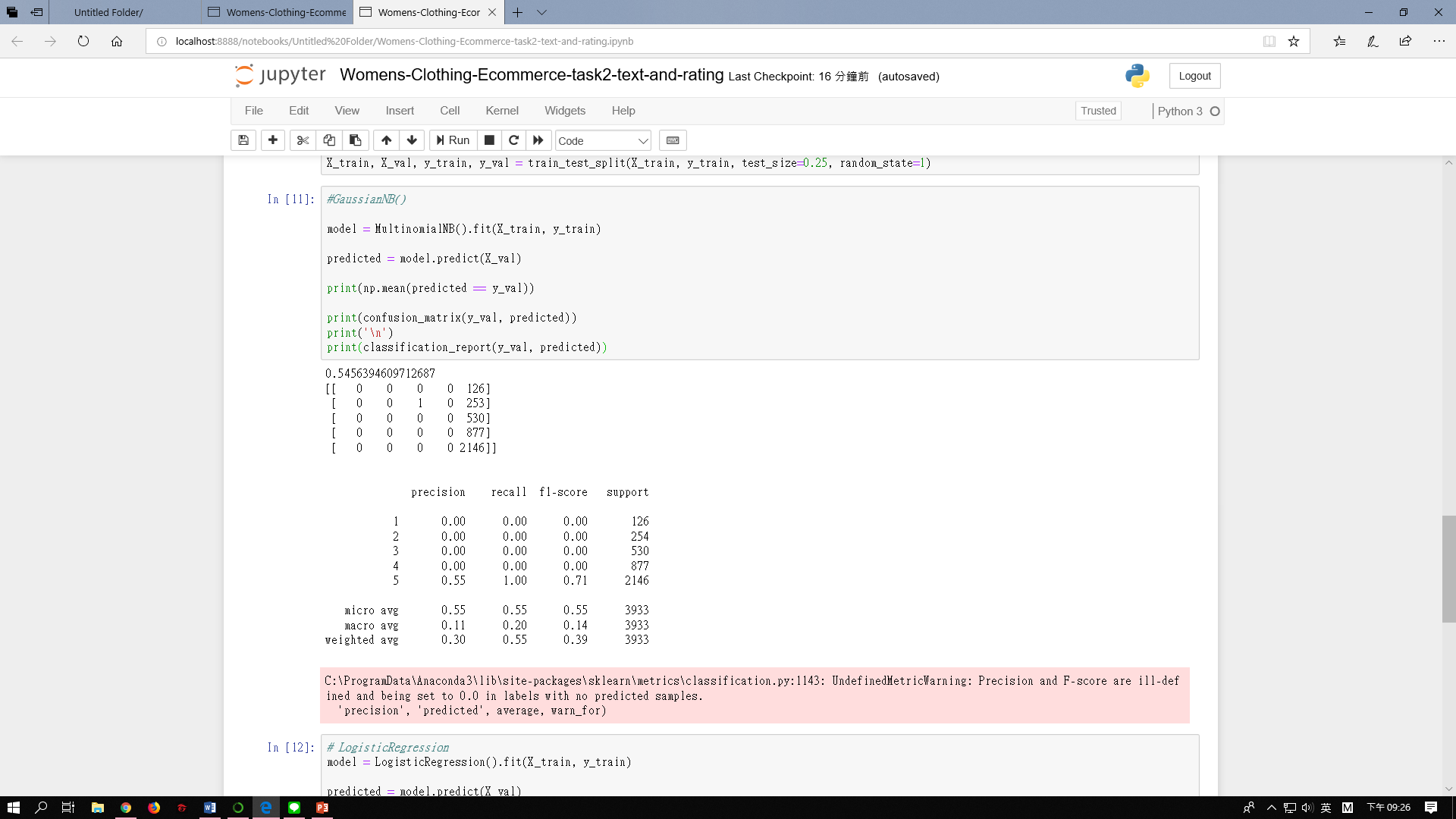
使用XGBoost 的結果為0.856。



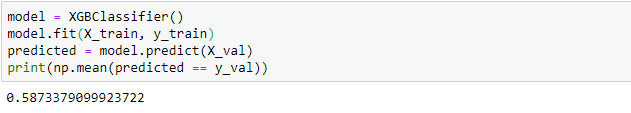
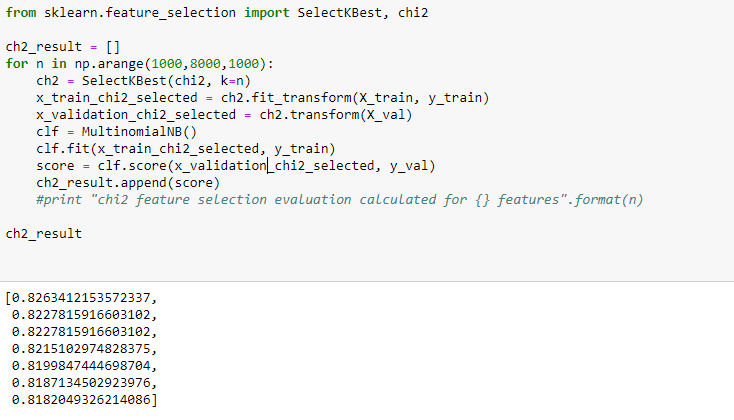
使用Logistic Regression 的結果為0.89。

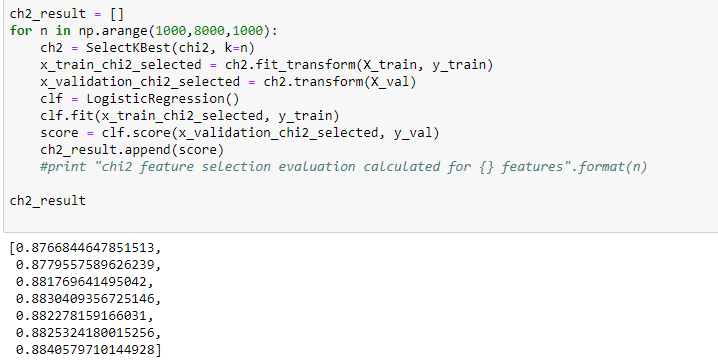


第二個task，使用使用Naïve bayes 的結果為0.55。使用Logistic Regression 的結果為0.63。

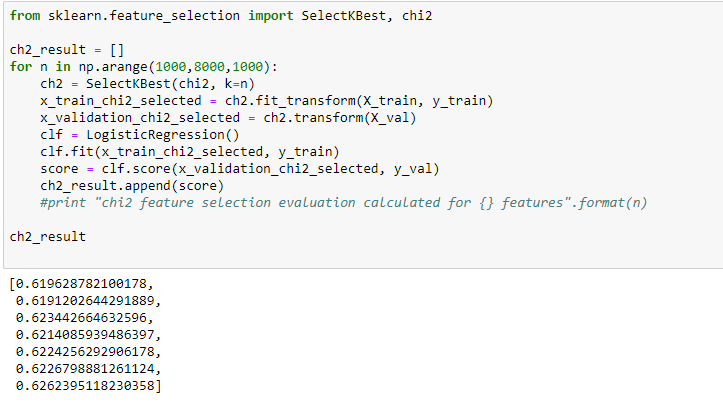
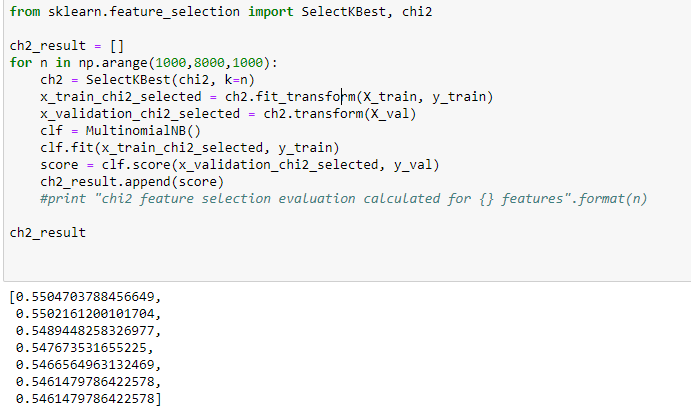


使用XGBoost的結果為0.59。





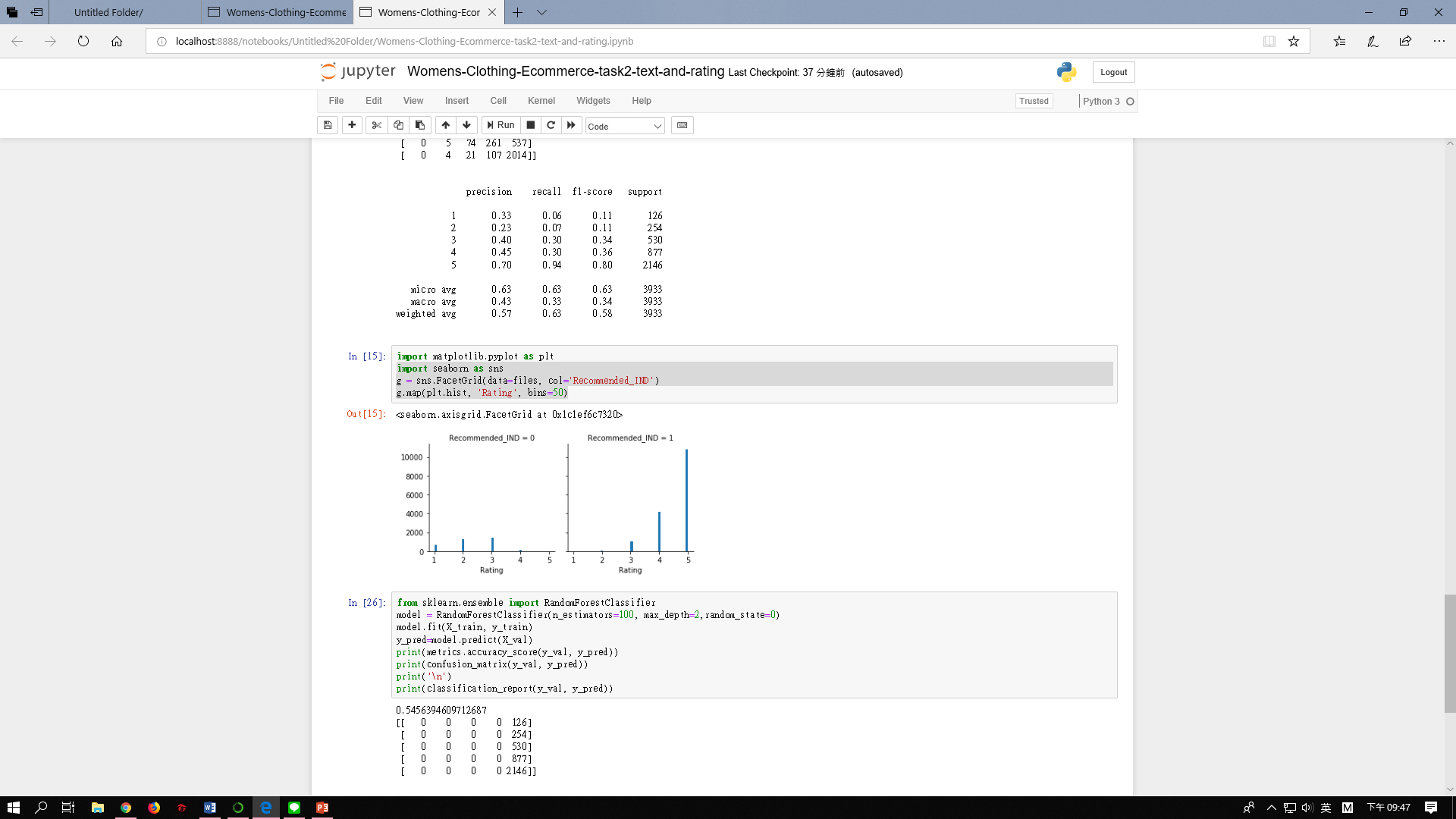
上圖第一張為task1透過chi-square作feature selection 對naïve bayes 的正確率影響，略增加了0.9個百分點，第二張是對於logistic regression下降0.9個百分點。



上圖第一張為task2透過chi-square作feature selection 對naïve bayes 的正確率影響，略增加了0.5個百分點，第二張是對於logistic regression下降0.6個百分點。

由此可見在本篇實驗中，chi-square對於naïve bayes的正確率有略微提升，但是對於logistic regression則是呈現下降的趨勢。

我們認爲在這些reviews中，neutral reviews的定義為rating=3。由下圖可知rating=3的使用者在推薦該產品的狀況是相當的。

圖中可以看到rating=1或2者不推薦該產品的可能性很高，rating=4或5者推薦該產品的可能性也很高，但是rating=3者在兩張表中呈現相當的狀況。所以我們認爲rating=3的review就是所謂的中立的留言。

如果從此實驗是以商業的角度來説的話，我們認爲可以將上述所提出的看法分成三種不同的行銷策略，對於中立review的使用者提供與其他兩個類別不一樣的服務。

此實驗中的兩個task都是從reviews中尋求對產品的評價，第一個task是分成1跟0，可以説是好或不好，第二個task是量化成五個等級。不意外地，在分成兩個類別時，分類器的表現比較好，而分成五個類別時，分類器的表現較差。

1. Study Summary

本實驗包括兩個task，第一個是sentiment analysis，其中Logistic Regression的準確率最高，達到0.89；第二個task為product ratings prediction, 其中同樣是Logistic Regression 的實驗結果最佳，達到0.63。此篇研究中，我們也討論neutral reviews 的處理方式以及後續運用，以商業角度來説能夠產生不同的商業決策。

1. Reference

Jeevanandam Jotheeswaran and Dr. S. Koteeswaran, “Sentiment Analysis: A Survey of Current Research and Techniques.”

Guo, Lu, and Wang, “Predicting Restaurants’ Rating And Popularity Based On Yelp Dataset.”

Martin, “Predicting Ratings of Amazon Reviews - Techniques for Imbalanced Datasets.”

Agarwal et al., “Sentiment Analysis of Twitter Data.”