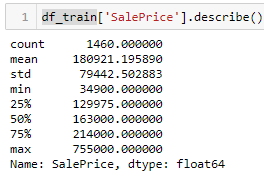
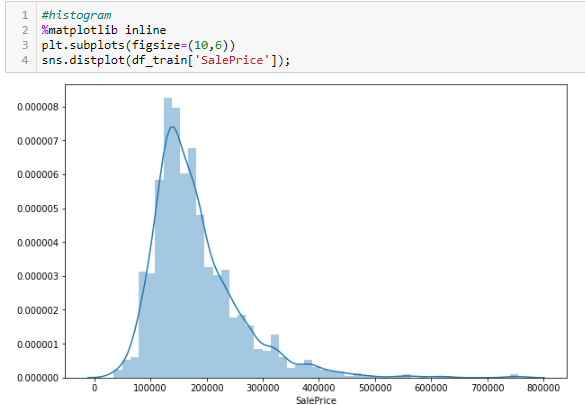
房價的分布

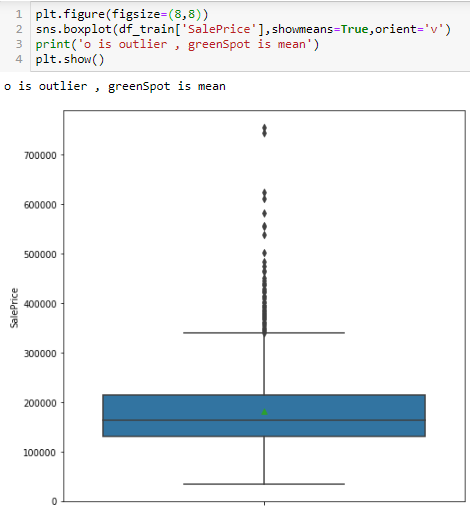
可以看到我們可以訓練的資料有1460筆，平均是180921，標準差是79442，根據經驗法則的結果，在一個常分布的結果，一個標準差、二個標準差、三個標準差分別是代表了68.27%、95.45%及99.73%的機率密度函數，也就是事件的發生機率，而75%(IQR)是214000在兩倍標準差內，所以可以知道我們的常態分佈為正常的。

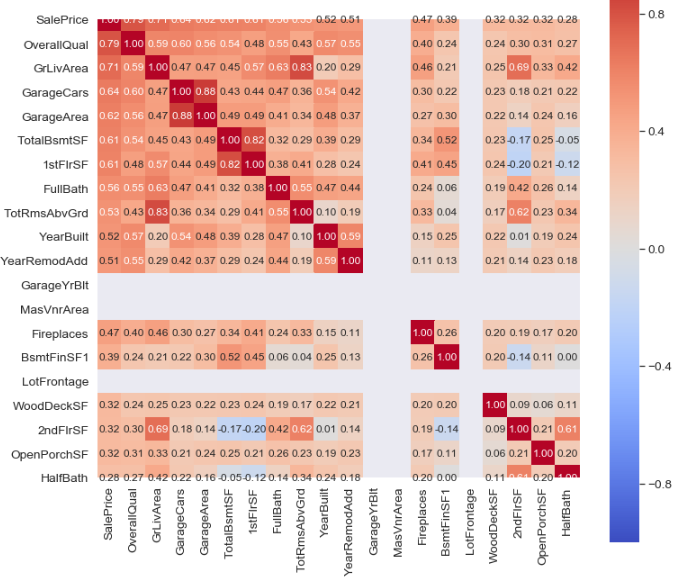
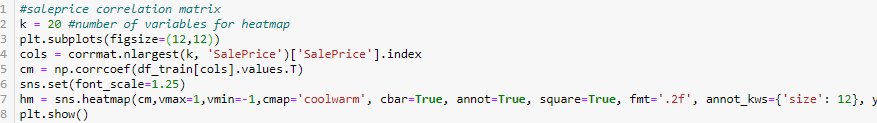
****

圖中可以看到，資料為右偏，代表離群直接在右邊的常態分佈。

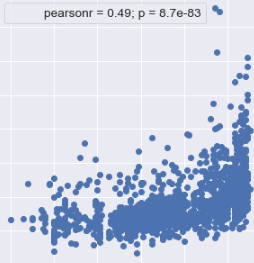
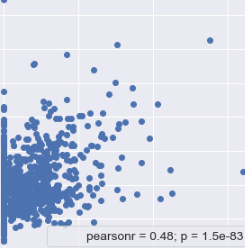
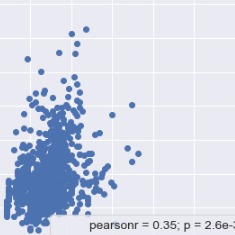
****

從我們的盒型圖也可以剛好驗證我們對於房價分布的推論，離群值都為大於mean+2倍標準差的點。

****

列出熱力圖前20項高

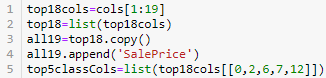
因為GarageYrBlt MasVnrArea LotFrontage資料有缺失值所以熱力圖為空白，下面計算他們與房價的相關係數，分別是0.49、0.48、0.35。 ****

**** ****   

分析數值屬性與房價的相關性

一般研究認為，相關係數0.3以下為低相關，0.3~0.7為中等相關，0.7以上為高度相關。以下的散佈圖皆採此標準為評估，並無條件捨去到小數第一位。

選取pearson’s相關係數超過30%的欄位作為之後預測房價的特徵值



類別屬性



列出類別屬性的盒型圖的程式碼(顯示結果會在下面分析時利用與分析用)

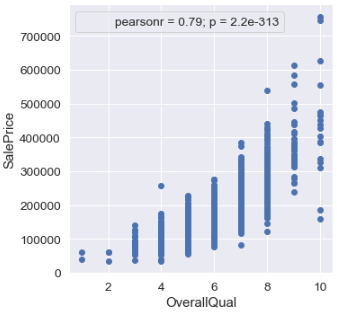
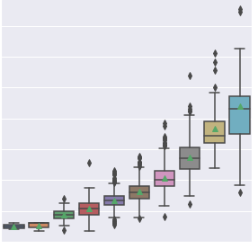


列出類別屬性的散佈圖的程式碼(顯示結果會在下面分析時利用與分析用)

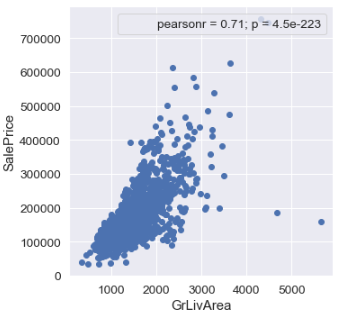


列出有超過Pearson相關係數0.7的屬性，表示與房價具有高度相關信，並作分析。

OverallQual:評估房屋的整體材料、裝飾品質的評價與房價呈現量高度的相關性，因為價值較高的商品本來就可以用架高的金錢去衡量(相對於較低)。

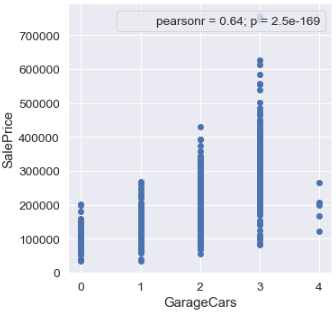
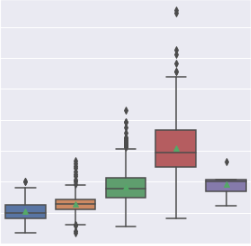
 

GrLivArea: 地上居住範圍(平方英尺)與房價呈現量高度的相關性，帶有更大“ GrLivArea”的兩個值似乎很奇怪，並且它們並未跟隨人群(大部分的樣本)。 我們可以做推測為什麼會這樣。也許他們指的是農業領域，這可以解釋價格低廉的原因。 但依舊只是推論，這兩點並不代表典型案例。 因此，我們將它們定義為離群值。



列出Pearson相關係數界在0.3~0.7的屬性，表示與房價具有中度相關信。

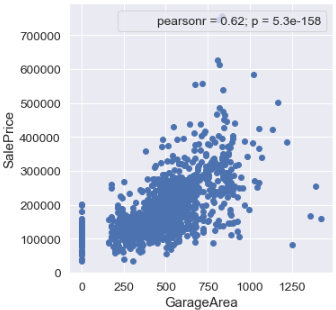
GarageCars: 車庫容量與房價呈現中度相關性。因為美國地廣，一班家庭都會有汽車當作交通工具。

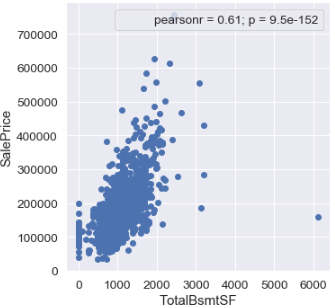
其中上下兩圖圈起來的資料價格落點是一樣的，所以可以知道他們是相同類型的資料，並且藉由這點，與車庫面積本身車庫容量就會呈現固有的相關性，車庫容量為4時或以上與房價的相關係參考性會比較低。

GarageArea: 車庫面積（平方英尺）與房價呈現中度相關性。因為車庫面積

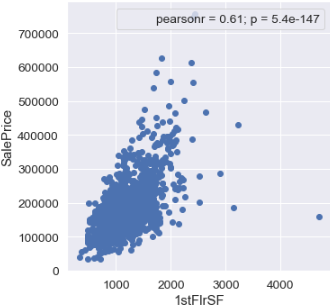
本身也與車庫容量有高度的相關性。



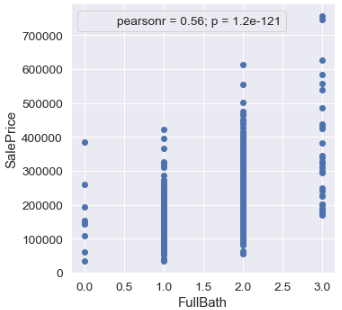
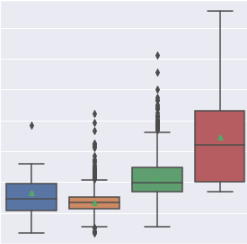
TotalBsmtSF: 地下室總面積(平方英尺)與房價呈現中度相關性(除了一個極端的離群值)。



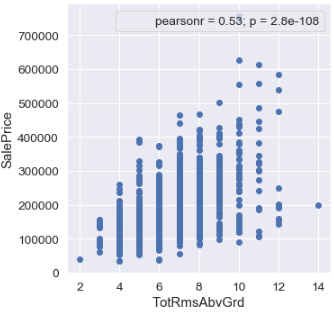
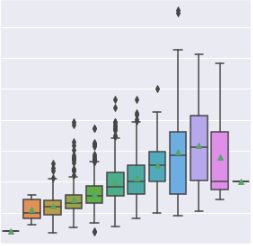
1stFlrSF: 一樓面積(平方英尺)與房價呈現中度相關性(除了一個極端的離群值)。



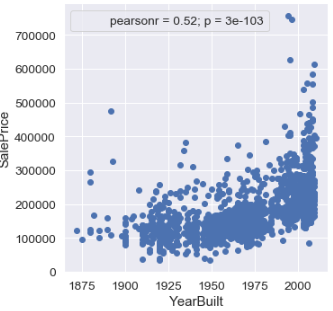
BsmtFullBath: 地下室浴室洗手間的數量與房價呈現中度相關性，

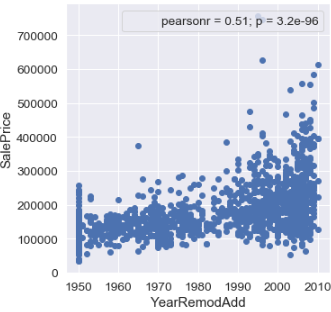
TotRmsAbvGrd: 地上房間總數量房價呈現中度相關性。推斷，因為一個家庭中的平均人數並不會太高，所以才低於地下室浴室洗手間的數量。

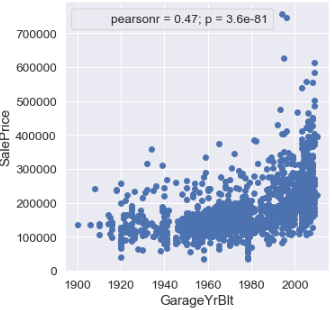
YearBuilt: 建造完工日期與房價呈現中度相關性，可以看出越老的房子與新的房子有略為呈現價格落差的趨勢，客戶喜歡新房子機率較高，可能是舊房子會有一定的程度需要翻修與在裝潢。



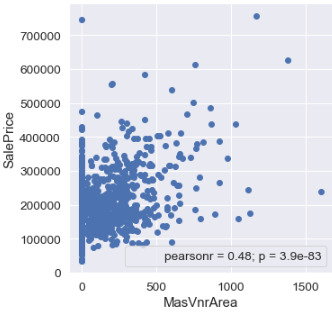
YearRemodAdd: 翻新日期與房價呈現中度相關性。翻新日期是個屋況推論的重要指標之一，所以對房價也具有一定的參考性。



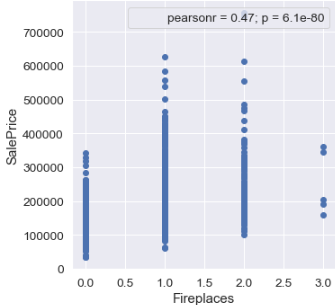
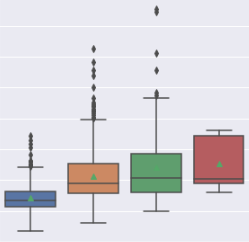
GarageYrBlt: 車庫完工時間與房價呈現中度相關性。因為通常車庫完工時間是與房屋建造完工時間是相同的(但也有可能是事後加蓋車庫的)，所以也具又一定對房價上的影響，但沒有比房屋完工時間對房價的相關係數具有較高的相關性。



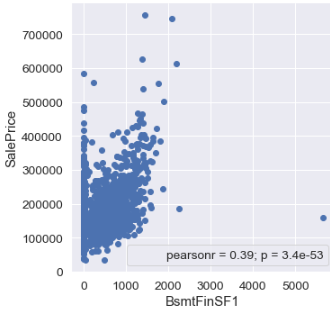
MasVnrArea: 磚瓦面積(平方英尺) 與房價呈現中度相關性。數據上有有點離散。



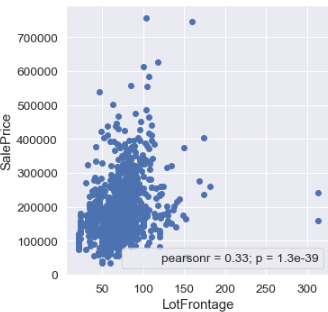
Fireplaces: 壁爐數量房價呈現中度相關性。壁爐在有些地區可以試算必要的設備，在寒冷地區也是常見的。盒型圖也可以看到趨勢往上升，但壁爐的數量使用也是固定的。所以這可能是相關係數並未高的原因。

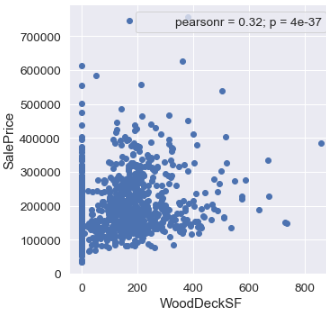
BsmtFinSF1: 第一個地下室完成的面積(平方英尺) 與房價呈現中度相關性。地下室的完成面積肯定會比地上的居住面積大，但其中有不少比事並未設置地下室的。紅色圈為離群值。



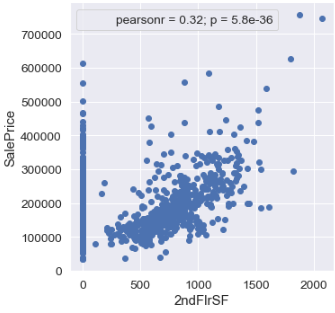
LotFrontage: 連接到物業的街道的直線(英尺) 與房價呈現中度相關性。可以看到這邊PEARSON相關係數掉落將近20%。



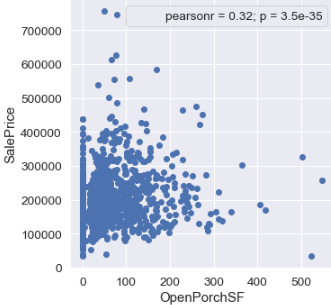
WoodDeckSF: 木製陽台面積（平方英尺）與房價呈現中度相關性。散佈圖顯示的數據顯示資料分布分散，且許多比資料未有陽台，顯示陽台雖然仍然有一定對房價影響的作用。



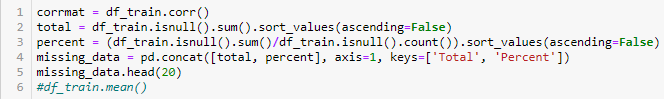
2ndFlrSF: 二樓面積(平方英尺) 與房價呈現中度相關性。，因為有些房屋未有二樓，所以通過二樓面積去評估房價的相關度會未能呈現所有樣本該有的房價資訊。

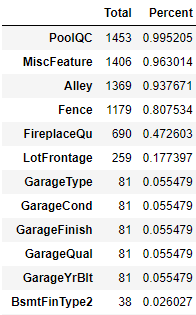
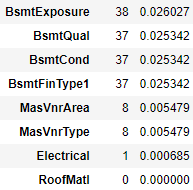


OpenPorchSF: 開放式陽檯面積（平方英尺）與房價呈現中度相關性。與木製陽台面積（平方英尺）有相同的相關係數。顯示陽台的確有一定對房價影響的作用。



觀看缺失值的狀況



**** ****

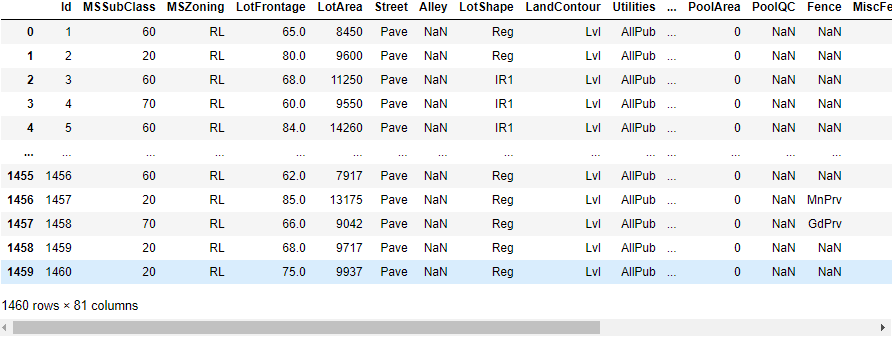
以平均數填補缺失值，並利用StandarScaler()方法標準化到Z分佈，這樣有助於運算

****

開始建模與訓練資料

讀取





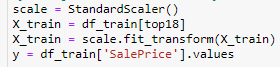
All19為相關係數與房價最高的前18欄位+房價本身



把類別資料數值化並且以該屬性的平均值填補空值



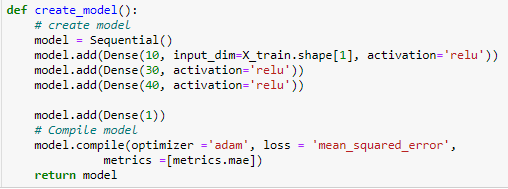
訓練資料Z標準化，並且從19欄位中選取房價當作我們的預測值y

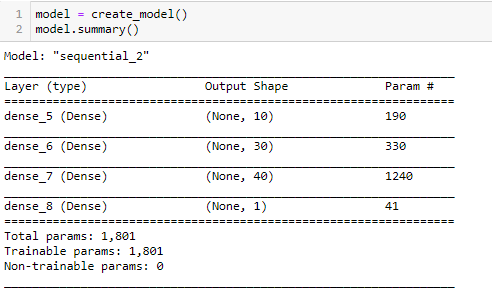


分為 67% 的訓練資料33% 驗證資料

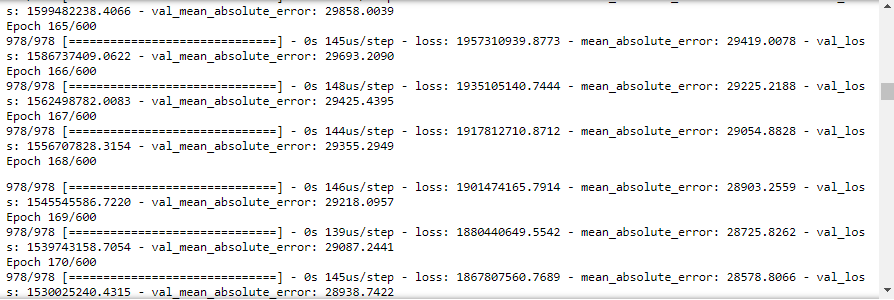


建模

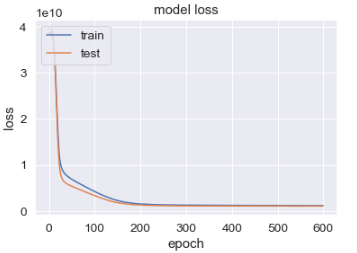
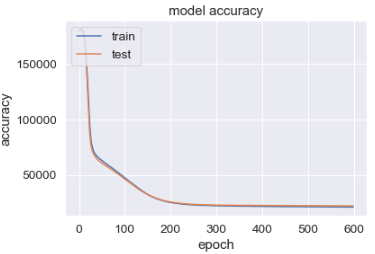
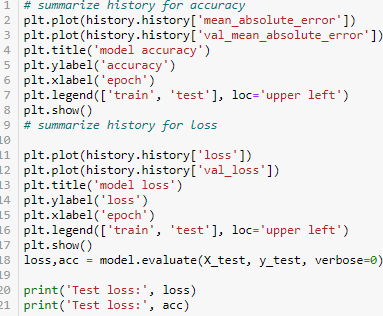


觀看模型參數與模型

訓練模型



繪圖觀看



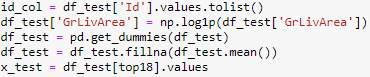
測試資料

讀取測試資料



第1行存取ID欄位、2行 log+1方法可讓標準化後的數據更好運算

3行把類別資料數值化、4行以該屬性的平均值填補空值5行相關係數最高的18個欄位當作測試資料



Z標準化



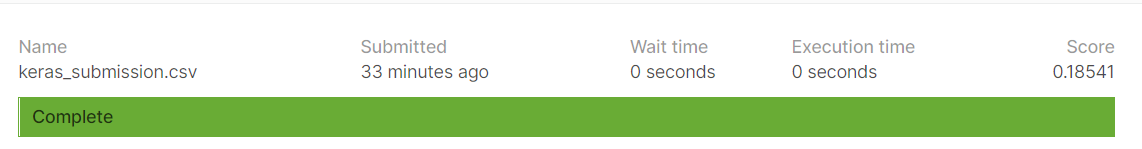
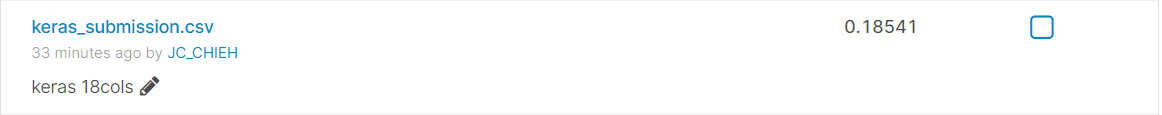
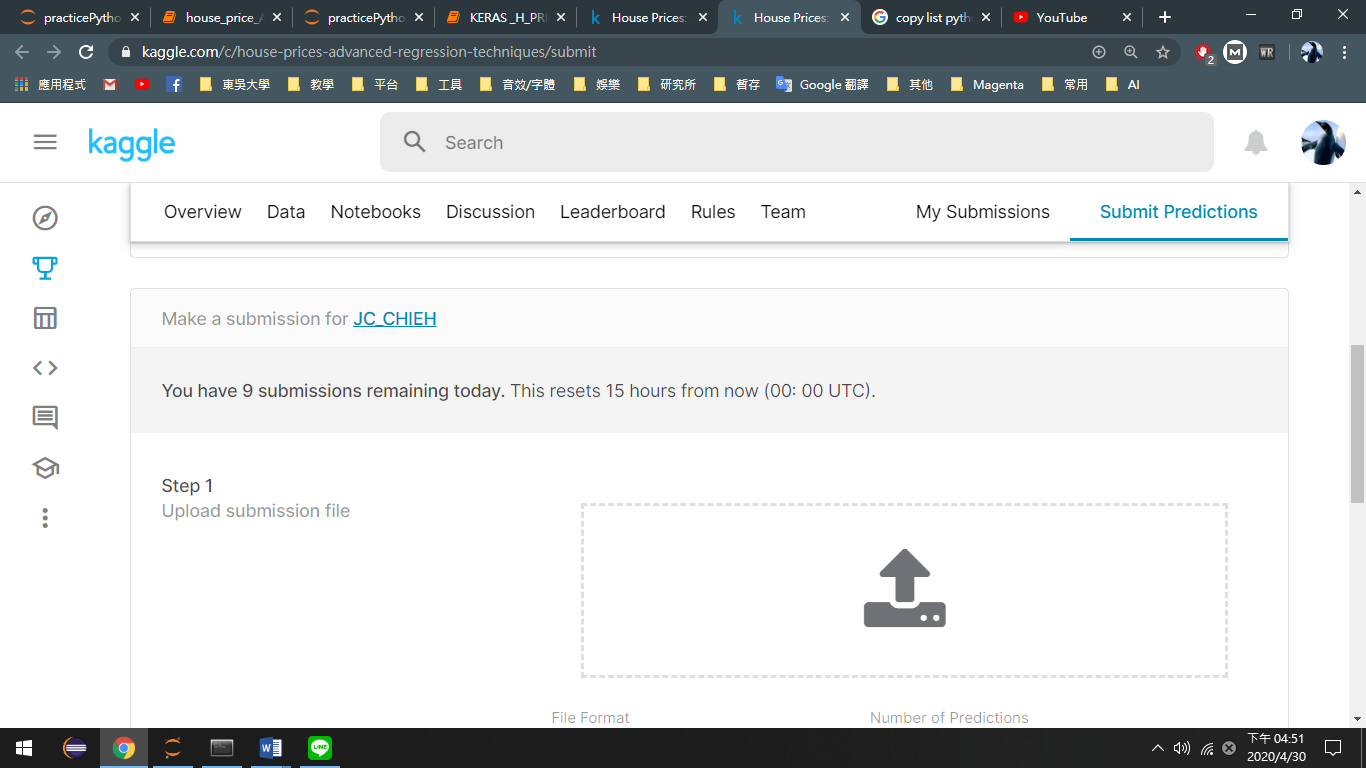
預測



存入預測資料



寫檔上傳



第3508名

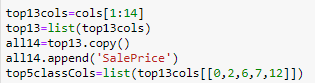
比較

原本選取pearson’s相關係數超過30%的欄位作為之後預測房價的特徵值，由於這五個欄位

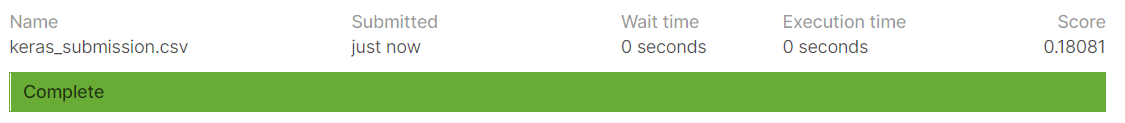
1. BsmtFinSF1: 第一個地下室完成的面積(平方英尺) pearson:0.39
2. LotFrontage: 連接到物業的街道的直線(英尺) pearson:0.33
3. WoodDeckSF: 木製陽台面積（平方英尺）pearson:0.32
4. 2ndFlrSF: 二樓面積(平方英尺) pearson:0.32
5. OpenPorchSF: 開放式陽檯面積(平方英尺) pearson:0.32

相對於前13個欄位，最少的相關係數都還有0.47的相關性，也就是幾乎一半的相關性，所以我們剔除這五個欄位，並且拿剩下13個欄位去做預測。

此圖為修改的主要地方，其餘略過。



預測結果並上傳



第3462名

Loss function為0.18081與之前的0.18541相比有略微進步，名次上升46名。

結論與評估

多層感知機為機器學習的一種，從訓練模型的觀測數據可以看到我們的loss function有良好的收斂性，並未出現over fitting的狀況，但由於多層感知機是以神經網路的方時來自及定義變數的權重，透過訓練的過程中自己更新數據，但由於目標所預測數據(房價)的範圍是廣大的數值屬性，可是提供給我們訓練的數值只有1460比資料，並且有少許資料是離群值，所以在訓練上能得到損失函數為0.18541為良好的結果。而在後面比較的結果0.18081，也可以看出運用MLP多層感知機，儘管在少掉5個欄位後，損失函數依舊影響不大，因為再訓練過程中影響不大的特徵值本身經過每一次的訓練後，他的權重都會有所改變，使得在預測房價的影響上，變得愈來愈小的影響力，直到符合他的影響程度，所以可以得知1460筆訓練資料訓練出來的結果 loss function 為0.18081 已經是 它本身的極限了，如果想要在提升，就需要再提供而外的數據。