組員：108552005 高儀津(Luca\_Kao) , 108552017 申自強(Arthur.Shen)

作業名稱: Kaggle House Prices: Advanced Regression Techniques

走跳江湖常有問題憑空而出，時而看似複雜，時而又似乎有脈絡可循。其實這樣一種面對問題、解析問題、求解問題的過程，正是人類文明演進的一項催化劑!而現今人類文明的累積是經過多少時間的花費與錯誤改進而擁有的呢？這個問題還是個進行式,但可以確定的是它從未停止。面對沒有答案的問題，我們能做些什麼呢？？預測，我們可以藉由已發生的事實記錄，透過歸納與統計來得到一個近似結果的答案，協助我們朝正確的方向走。然而這些已發生的紀錄(往後我們將稱之數據)，隨著時間的累積會不斷的產生與更新，以人類對數據處理的心智能力來說，勢必無法靠一己之力來完成。這時候就需要可以快速運算的電腦來幫忙啦！！這就是我們近期常聽到的機器學習與深度學習。

本篇報告將藉由Kaggle 的初心者任務 Hourse Prices: Advanced Regresion Techniques。來說明我跟我的夥伴如何處理未知的預測問題。

凡走過必留下痕跡

這裡我們知道要透過已有的數據來產生模型，協助我們對未知的數據產生預測結果（給定一部分的特徵變數，預測實際結果)。這是機器學習中監督式學習的範疇，資料學家提供整理好的數據並提供方法讓電腦產生最佳模型。來替往後有相似特徵變數的問題，提供最佳答案。

OK 來吧”數據”!! 一開始獲得的數據琳瑯滿目，有需要的也有對問題毫無相關的甚至有遺失的！！所我們可以將初期的資料整理統稱為**Data cleaning**。

主要可以分為以三個階段:

**1. Raw data**

**2.** **Technically correct data**

**3. Consistent data**

**1.Raw Data**，我們可以稱之原始，未經整理的資料。在這個階段我們只須專心收集不必考慮資料的可靠性。這裡我們直接使用Kaggle 所提供的train\_set 及test\_set。

**2.Technically correct data**， 這裡就變得有趣些了， 因為數據需要藉由電腦來運算，而電腦運算只能對數字有效，那字串還有帶有特殊意義的字元該如何讓電腦感興趣呢？？舉例來說這裡我們將ID的欄位取消,因為ID只對個別的模型也就是一列的房屋售價有直接影響,對整體模型也就是每一列房屋售價的回歸凝合不具影響力，所以我們去掉。

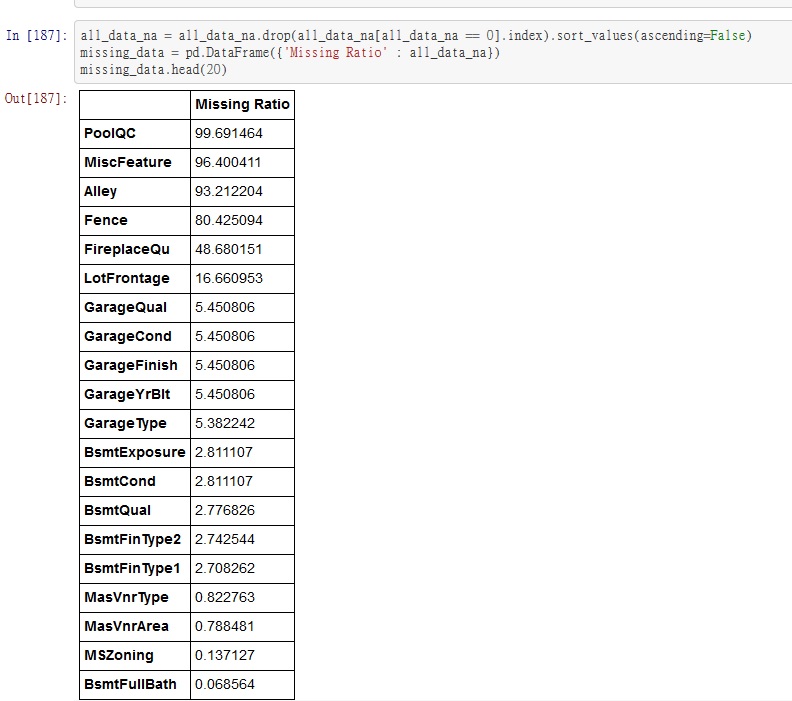
而這些使用技巧性的處理特定數據的過程又可以分為以下幾種:

**2.1.Missing Values**

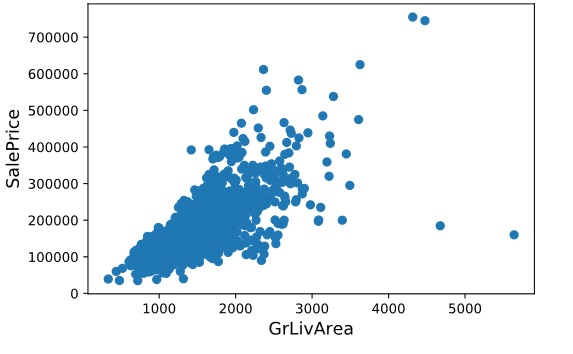
**2.2.Outliers**

**2.3.Obvious insistencies**

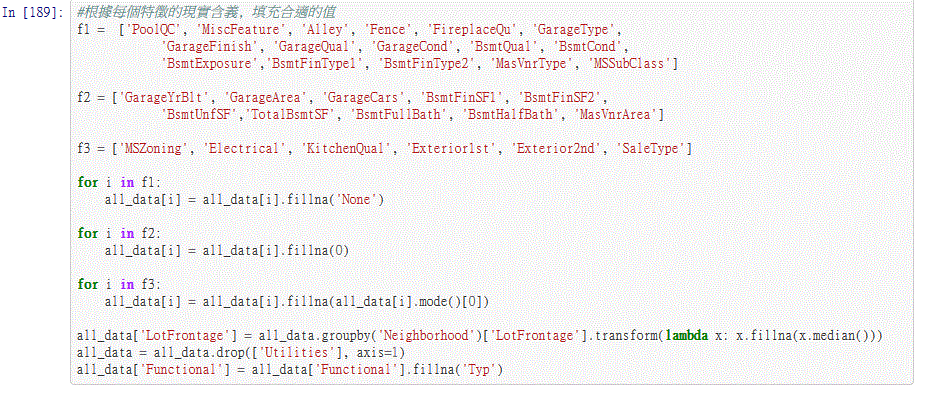
**2.1 Missing Values**, 這些特徵值的資料型別是確定的，唯獨資料值為空或不確定。這樣的值對電腦的處理上,也無法定義!!我們可以直接拋棄，亦或是將其轉換其它更具代表性的數值。這裡我們可以看Python code 180行至185行，我們使用Pandas 的資料處理功能is null,檢查每一個特徵欄位的空值狀況,並統計缺失率將缺失率大於15%的部分拋棄。



**2.2 Outliers**， 指的是異常或與均值距離相當大的實例，我們也可說是雜訊資料。舉例來說，當資料特徵屬性為年齡時，我們看到有負數的資料時，這樣的資料就算是種Outliers。這裡我們以SalePrice與 GiVivArea 來舉例，透過彼此的資料分布來看,可以確定的是彼此有正相關的影響(即存在斜率為正的方向)。但明顯看出右下角有兩點，並沒有正相關的傾向。所以我們可以判斷它為**Outliers.**它倆若加入，會導致選用如Ridge Regression 這類對雜訊相對敏感的模型過度收斂。因此將之移除。



**2.3 Obvious insistencies,** 在這個步驟原則上我們已經將顯而易見的資料確失欄位與異常資料拋棄。但對於那些與特徵屬性與**SalePrice**有相關性，但其型別為字串或為離散狀態的欄位我們尚未處理。此外，所有特徵資料欄位對於空值的描述也不一致，所以我們也必須一一將其填充。

****

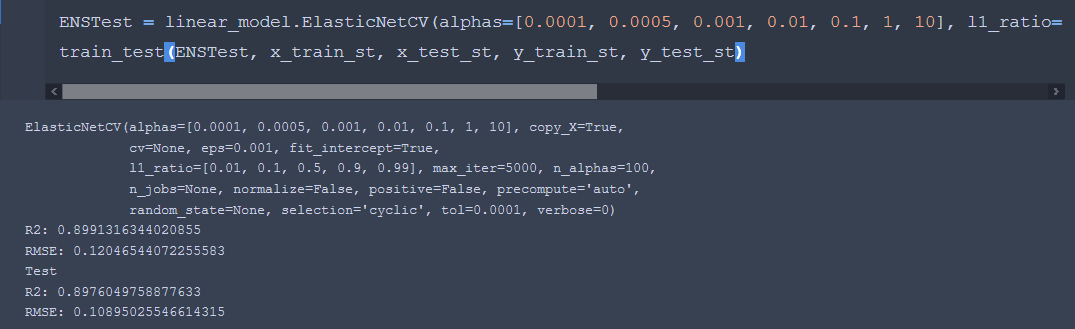
**3. Consistent data,** 資料的一致性。到目前為止我們所處理的皆為訓練資料，但

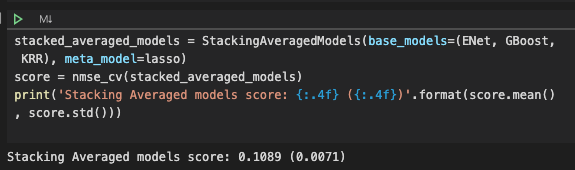
我們所關心的目標函式應該具有泛化性，也就是模型不能只有對訓練集有良好反應，但對測試及驗證集確反應不良。所以在資料預處理的步驟上也必須擴及到測試集。

模型選擇

現在我們完成資料與特徵選取，再來就是挑選適合的模型來解答我們的問題了。這裡因為我們對模型的選擇有各自的想法，而我們因為對整個機器與深度學習的領域尚處摸索階段，所以我們一位選擇了單一sklearn 的 Eleastic Net(Arthur.Shen\_lab1.ipynb),另一位選擇了用模型疊加的方式重疊model(Luca\_lab1.ipynb)，最後選擇上傳到Kaggle團隊案的Script為模型疊加的方式。

Arthur.Shen\_lab1.ipynb

****

**Luca\_kao\_laab1.ipynb**

|  |
| --- |
| **The end of report all prcocess in charge of as below**  **Data Cleaning: Both**  **Feathure selection: Both**  **Python coding: Both**  **Short report: Both**  **Kaggle Model: 108552005**  **HW-link :<https://github.com/lars10192002/Introduction_to_Deep_Learning>**  **Kaggle : <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/discussion>** |