



|  |
| --- |
|  |
|  |
| *Kaggle House Prices:*  *Advanced Regression Techniques* |

# 問題是什麼

Subtitle Text Here

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紀錄,透過歸納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”. 一開始穫得的數據琳瑯滿目,有需要的也有對問題毫無相關的甚至有遺失的！！所我們可以將期初的資料整理統稱為Data cleaning.

主要可以分為以三個階段

1. Raw data

2. Technically correct data

3. Consistent data

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.

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.

而這些處理數據的過程又可以分為以下幾種

1.Missing Value

2.Outliers

3.Obvious insistencies

|  |
| --- |
|  |
|  |
|  |