**Lab 09**

|  |  |
| --- | --- |
| Name: | 易峻葦 |
| Student ID: | B08611031 |
| Total Score: |  |

1. **Multiple Choice (35 points, 5 points each question)**

|  |  |  |  |
| --- | --- | --- | --- |
| # | Answer | Explanation (Optional) | Score |
| 1 | (a) |  |  |
| 2 | (c) |  |  |
| 3 | MSE 相同 |  |  |
| 4 | (a) |  |  |
| 5 | (c) |  |  |
| 6 | (c) |  |  |
| 7 | (b) |  |  |

1. **Simple Linear Regression (30 points, 10 points each question)**

|  |  |  |
| --- | --- | --- |
| # | Description | Score |
| 1 |  |  |
| 2 | B0=0.67，B1=1，MSE=0.33，還不錯 |  |
| 3 | 斜率=1，截距=0.67    用此套件得出的回歸直線與(2)的相同。 |  |

1. **Multiple and Polynomial Regression (35 points)**

|  |  |  |
| --- | --- | --- |
| # | Description | Score |
| 1 | 根據train\_test\_split中stratify 參數的用法，我們可以填入data其中一個feature(這裡是datas[‘region’])，則分出來的訓練集以及測試集內該feature的比例就會與原先data中feature 的比例相同。然而作業中給的csv 檔有些瑕疵，像是有nan，還有不知道是不是故意的，有些球員的資料會少一個，導致後面的字料都會往前挪一格，因此region 的那格會變成國家(但最後載資料進來時發現不會影響取值，region仍為正確值)，test size 則是取0.2 的比例。  Code: |  |
| 2 | 1. 每次的準確率會依據每次將資料分成訓練集以及測試集的分布情形而有所不同，上述第一題的做法僅能夠確保region 的比例，但無法確保其他特色的比例， 進而影響到回歸的準確率。下圖為讓改成是重複跑一百次的結果，可發現即使每次訓練集跟驗證集的資料分布會不同，每次的準確率大概會在60~75%左右，還算是有一定的準確性。      1. Age- 隨著年紀的增長，一位球員的經驗以及技術必定會有所長進，因此市場價值   會跟著升高，然而當年紀再繼續增長時，球員就會面臨老化問題而失去價值，因此，一個拋物線最適合用來表達球員年紀與市場價值的關係，而每次回歸的結果也可發現平方像係數必定小於一，代表極值是最大值，符合預期。  Log2(page views): 因為觀看次數的數值通常會比較大，因此用log的方式縮小數值。  Big club:每次回歸可發現，該項係數都會是大於一的係數，由於該項不是1就0，其所代表  的意義為如果身在好的球隊必定可在市場價值上達到加分的效果，而不在也不至於會被扣分。   1. 以其中一次回歸出來的數據為例，的係數分別為[-84.64074572375665 0.08323629 5.97171287 -0.118532 1.91292639 0.73993252 8.55679854 -0.85683549]，因此要提升自己的市場價值，就必須提升係數為正的相關特徵。若想藉由page views讓自己的市場價值提高10的話，就必須提升 次pageviews。 2. 下圖為程式重複進行一百次後每次test data 與train data 的mean square error，可發現test data 的mean square error 的變動範圍比train data 大上許多，因此我認為有overfit的可能性。 |  |